

Understanding the Limitations of Using
Large Language Models for Text Generation

Daphne Ippolito

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Chris Callison-Burch, Professor, Computer and Information Science

Douglas Eck, Senior Research Director, Google Research

Graduate Group Chairperson

Mayur Naik, Professor of Computer and Information Science

Dissertation Committee

Lyle Ungar, Professor, Computer and Information Science

Dan Roth, Professor, Computer and Information Science

Marianna Apidianaki, Professor, Computer and Information Science

David Grangier, Research Scientist, Google Brain

Understanding the Limitations of Using
Large Language Models for Text Generation

© COPYRIGHT

2021

Daphne Ippolito

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

ACKNOWLEDGEMENT

I am deeply indebted to my advisors Chris Callison-Burch and Douglas Eck, who believed in me and my ability to conduct worthwhile research long before I believed in it myself, and who encouraged me to pursue an unconventional PhD. I am also very thankful to committee members; I especially appreciate Marianna Apidianaki for her astute feedback on paper drafts and presentations, David Grangier for our many brainstorming conversations, Dan Roth for his insights into the field of NLP, and Lyle Ungar for convincing me to try my hand at NLP research.

This dissertation would not have been possible without the support of my many friends and colleagues at Penn, including but not limited to Alyssa Hwang, Arun Kiribaran, Barry Slaff, Bryan Li, Caleb Stanford, Clark Zhang, Harry Zhang, Joao Sedoc, Krzysztof Jordan, Liam Dugan, Luke Valenta, Marcella Hastings, Maria Kustikova, Reno Kriz, Siddharth Mysore, Stephen Philips, and Veronica Qing Lyu. A special thanks to Joao, whose class on dialog systems started me on the path to this dissertation, and to Alyssa for keeping me on task while writing it. This dissertation would also not have been possible without my many collaborators, mentors, and question-answerers at Google, including but not limited to Adam Roberts, Andy Coenen, Ann Yuan, Anna Huang, Carey Radebaugh, Chiyuan Zhang, Daniel Duckworth, David Grangier, Emily Reif, Fjord Hawthorne, Hugh Zhang, Ian Simons, Katherine Lee, Matthew Jagielski, Nicholas Carlini, Noah Constant, Noah Fiedel, and Sehmon Burnam.

Finally, I would like to thank my friends and family who have supported me through the entire PhD process, and in particular, Yun William Yu, who has helped me in every facet of becoming a successful and happy researcher.

Understanding the Limitations of Using Large Language Models for Text Generation

ABSTRACT

State-of-the-art neural language models are capable of generating incredibly fluent English text. This success provides opportunities for novel forms of interaction, where human writers work collaboratively with a natural-language generation system toward a set of goals. However, it also poses several challenges. Evaluating and comparing the skill of different open-ended text generation systems is challenging, and generated text can have negative societal impact if it proliferates and people are not able to detect it. In this dissertation, I introduce a detection-based evaluation task that can be used to investigate the tradeoff between generating high-quality and generating diverse text. I also show how large neural language models' capability of memorizing large swaths of their training data complicates our ability to evaluate their skill at generating high-quality *novel* text. I also show how, despite these challenges, neural language models can be successfully employed to support creative writing tasks. In particular, I describe methods for performing style transfer into any user-provided style and for efficiently supporting fill-in-the-blank operations in addition to the more standard continuation operation. Finally, I present an interactive writing tool we built which allows creative writers to collaborate with a natural language generation system to craft stories. Users studies with both novice and professional writers provide insights into the strengths and limitations of applying natural language generation systems in real-world settings.

CONTENTS

| | | |
|-------|--|----|
| 1 | INTRODUCTION | 1 |
| 1.1 | Thesis Statement | 4 |
| 1.2 | Publications Presented | 4 |
| 2 | BACKGROUND ON TEXT GENERATION | 6 |
| 2.1 | What is a Language Model? | 6 |
| 2.2 | What is a Neural Language Model? | 7 |
| 2.3 | Encoding Text into a Vocabulary | 10 |
| 2.4 | Generating Text with a Language Model | 12 |
| 2.4.1 | Greedy Approaches | 12 |
| 2.4.2 | Search-Based Approaches | 13 |
| 2.4.3 | Generation Diversity | 17 |
| 2.4.4 | Generation Quality | 17 |
| 2.4.5 | The Diversity-Quality Tradeoff | 18 |
| 2.5 | Language Generation Tasks in This Dissertation | 20 |
| 2.6 | Controllability and Task-Specific Generation | 22 |
| 3 | DETECTING MACHINE-GENERATED TEXT | 25 |
| 3.1 | Motivation | 25 |
| 3.1.1 | Detection as a Task | 26 |
| 3.2 | Impact of Decoding Strategy on the Detectability of Machine-Generated Text | 27 |
| 3.2.1 | Dataset Construction | 29 |
| 3.2.2 | Methods for Automatic Detection | 31 |
| 3.2.3 | Method for Human Detection | 32 |
| 3.2.4 | Results | 36 |
| 3.3 | ROFT: A Largescale Study of Human Detection Ability | 42 |
| 3.3.1 | The Real or Fake Text Game | 44 |
| 3.3.2 | Experimental Design | 45 |
| 3.3.3 | Results | 50 |
| 3.3.4 | Analysis | 53 |
| 3.3.5 | Discussion | 61 |

| | | |
|-------|--|-----|
| 3.3.6 | Study Limitations | 63 |
| 3.4 | Conclusion | 64 |
| 3.5 | Summary of Contributions | 65 |
| 4 | MEMORIZATION OF TRAINING DATA | 67 |
| 4.1 | Motivation | 67 |
| 4.1.1 | Definitions of Memorization | 68 |
| 4.2 | Quantifying the Factors that Influence Memorization | 69 |
| 4.2.1 | Selection of Evaluation Data | 71 |
| 4.2.2 | Experiments | 72 |
| 4.2.3 | Replication Study–T5 | 86 |
| 4.2.4 | Replication Study–Language Models Trained on Deduplicated Data | 88 |
| 4.3 | Deduplicating Training Data Reduces Memorization | 89 |
| 4.3.1 | Large Language Model Datasets | 91 |
| 4.3.2 | Datasets Considered in this Study | 92 |
| 4.3.3 | Method for Exact Substring Duplication | 93 |
| 4.3.4 | Method for Approximate Matching with MinHash | 98 |
| 4.3.5 | Results | 101 |
| 4.3.6 | Discussion | 109 |
| 4.4 | Conclusion | 110 |
| 4.5 | Summary of Contributions | 112 |
| 5 | ENABLING APPLICATIONS IN CREATIVE WRITING | 116 |
| 5.1 | Supporting Arbitrary Style Transfer | 117 |
| 5.1.1 | Experimental Setup | 119 |
| 5.1.2 | Results | 121 |
| 5.1.3 | Limitations and Failure Modes | 129 |
| 5.2 | Models for Infilling Text | 132 |
| 5.2.1 | Supporting FITB and Continuation | 134 |
| 5.2.2 | Experimental Setup | 135 |
| 5.2.3 | Results | 137 |
| 5.2.4 | Conclusion | 144 |
| 5.3 | Wordcraft: An editor for AI-assisted writing | 144 |
| 5.3.1 | The Wordcraft Application | 145 |
| 5.3.2 | User Study with Novice Writers | 154 |
| 5.3.3 | The role of AI in co-writing | 158 |

| | | |
|-------|--------------------------------|-----|
| 5.3.4 | User Study with Expert Writers | 163 |
| 5.4 | Conclusion | 164 |
| 5.5 | Summary of Contributions | 164 |
| 6 | CONCLUSION | 165 |
| | LIST OF TABLES | 169 |
| | LIST OF ILLUSTRATIONS | 174 |
| | DISSERTATION | 179 |

1

INTRODUCTION

One of the oldest yet most elusive promises of AI is computers that can converse with humans, not just via rigidly structured templates and programming languages, but in natural language. In 1950, Alan Turing, one of the fathers of modern computing, framed this goal as an imitation game—computers ought to be able to imitate real human interaction so well that “an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.” He expected that by 1950 this game would be solved.

Natural Language Generation (NLG) is a critical component to solving the imitation game. NLG is the task of writing novel text in a human language such as English. Improvements over the last half-decade have led to natural language generation systems which are capable of producing incredibly fluent text. These systems have been applied to practical domains such as machine translation and text summarization and simplification, but they have also been applied to more fanciful ones, such as story generation, video games development, and tooling for creating writing.

AI-assisted creative writing is a particularly interesting testbed for how far we have come toward achieving Turing’s goal. Unlike in machine translation or summarization, where it is critical that the generated text is factual and stays faithful to the source material, in creative domains, the “hallucinations” and unusual word choices that are pervasive in modern NLG may be beneficial to a creative writer’s process. Indeed, ideation tools, such as decks of cards which suggest writing topics, are commonplace in creative writing circles. The creative writing domain allows us to evaluate whether NLG systems are useful for accomplishing real writing tasks and whether systems can be easily *controlled* by non-technical users. This allows us to more holistically

evaluate the strengths and limitations of machine-generated text than the narrowly targeted evaluation tasks typically employed in the academic literature.

Before discussing the applications of assisted writing, this thesis first examines two important considerations around the use of modern NLG systems. Modern NLG relies on neural language models, neural networks trained on billions of words of text in order to represent human language. Chapter 2 gives an overview of how these systems work. Understanding the limitations of these language models and the nature of the text they are capable of generating is crucial to the ultimate use of these systems in real applications.

A significant limitation for creative applications like Wordcraft is the difficulty in generating text that is diverse (containing uncommon and interesting words and phrases) and high quality (as perceived by human readers). Very often, practitioners choose text generation strategies that err on the side of human-perceived quality at the expense of lexical diversity. This decision leaves subtle signatures in the generated text which make it easy for automatic classifiers to distinguish it from genuine human-written text.

Chapter 3 focuses on this challenge and its ramifications to the detectability of generated text. The proliferation of machine-generated text, especially when it lacks attribution, is of significant concern to the public, and understanding detectability is also imperative because it gives us a proxy for how far along NLG systems are at fooling humans and whether undesired use of machine-generated text can be mitigated. In this chapter, I measure the ability of humans as well as automatic systems to detect machine-generated text, and show the connection between detectability and the diversity-quality tradeoff in generated text.

One difficulty in studying human ability to detect machine-generated text is that it can be very difficult to collect annotations. Many of the errors NLG systems make are subtle and require closely reading several sentences of text to be able to identify. Common strategies for soliciting human annotators, such as paying crowd workers a fixed dollar amount per annotation, do not tend to yield useful annotations since annotators are not incentivized to spend the extra time to do a close read. In order to be able to study human detection ability at scale, we built the Real or Fake Text game (RoFT), a website that gamifies the task of identifying machine-generated text [40]. The RoFT platform allowed us to collect over 40,000 annotations of whether players could correctly identify when a passage of text transitioned from being human-written to being

machine-generated . Chapter 3.3 presents a detailed analysis of the factors we found that most impacted detectability.

Of course, machine-generated text is most undetectable when it looks *exactly* like its training data. Large language models are worryingly capable of memorizing and regurgitating significant amounts of their training data. For example, GPT-3, a popular model that has already been incorporated into several products, when prompted with the first sentence of *Harry Potter* or *Lord of the Rings* will accurately generate the first several paragraphs of each book. This behaviour is especially problematic for the domain of AI-assisted creative writing, as writers using tools such as Wordcraft have the expectation that the generations they are being shown are unique and not plagiarized. Memorization also makes the task of studying the detectability of generated text more challenging. If an NLG system generates Chapter 1 of *Harry Potter*, should this text be labeled as human-written or machine-generated? Chapter 4 focuses on this question of memorization. First, we show how performing thorough deduplication of training data results in models that are less likely to exhibit memorization. Then we conduct experiments showing how observable memorization scales with respect to the number of times a sequence occurs in the training set, the model size, and the length of conditioning prompt [22].

Finally, Chapter 5 describes my contributions to the field of AI-assisted creative writing. It discusses the importance of introducing controllability into natural language generation systems—providing writers the ability to dictate what kind of text gets generated and decide how it interfaces with what they might have already written. In particular, we introduce methods for efficiently supporting a fill-in-the-blank paradigm, where a writer can insert text into any position of their current text [66]. We also describe a simple recipe for supporting style-transfer into any user-defined style without the need for costly training data acquisition and test-specific model training. Both these approaches are incorporated into Wordcraft, an AI-augmented text processor that provides several interfaces for writers to get feedback and suggestions from an NLG system. In user studies with both amateur and skilled writers, we found Wordcraft to be a valuable assistive tool for creative writing.

1.1 THESIS STATEMENT

In this thesis, I argue that we need to build a deeper understanding of neural network-powered language generation systems before they are safe to deploy these systems widely. To thwart NLG systems being used in ways that are detrimental to society, it is crucial to understand how machine-generated text differs from the text a human would write given the same writing task. My research focuses on two such differences: (1) how the word choices made by NLG systems cause generated text to be distinguishable from human-written text, and (2) the tendency of NLG systems to plagiarize verbatim from their training data when asked to produce novel content. Both of these issues make it difficult to evaluate machine-generated text. In the first case, the tradeoff between generating high-quality versus generating lexically diverse text makes it challenging to simultaneously optimize for text that is pleasing to human readers and text that is statistically indistinguishable from human-written text. In the second case, memorization leads to us over-representing the strength of NLG systems by attributing generalization ability to what is actually memorization ability. To this end, I argue that we should be evaluating NLG systems holistically as part of larger tools meant to *assist* human writers in tasks they wish to accomplish (in addition to evaluating on isolated individual tasks). I present a case study evaluation of human-AI writing collaboration in the domain of creative writing.

1.2 PUBLICATIONS PRESENTED

The work described in this thesis has been published in several conference papers. In all cases, the work was completed jointly with collaborators at University of Pennsylvania and/or Google Research. At the end of each chapter section, I include a summary of my primary contributions to the work.

- Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. “Comparison of Diverse Decoding Methods from Conditional Language Models.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3752–3762
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. “Automatic Detection of Generated Text is Easiest when Humans are Fooled.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 1808–1822
- Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. “RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 189–196
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. “Deduplicating Training Data Makes Language Models Better.” In: 2022
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. “Quantifying Memorization across Neural Language Models.” In: *arXiv preprint arXiv:2202.07646* (2022)
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. “A Recipe for Arbitrary Text Style Transfer with Large Language Models.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022
- Andy Coenen, Luke Davis, Daphne Ippolito, Ann Yuan, and Emily Reif. “Wordcraft: a Human-AI Collaborative Editor for Story Writing.” In: (2021)
- Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. “Wordcraft: Story Writing With Large Language Models.” In: *27th International Conference on Intelligent User Interfaces*. 2022, pp. 841–852

2

BACKGROUND ON TEXT GENERATION

Automatic text generation has been a goal of computer science researchers since the early days of computing. In recent years, template-based, grammar-based, and statistical approaches have given way to neural language models—neural networks trained to build representations of human language from millions or even billions of documents. This chapter gives a brief overview of how modern text generation systems based on neural language models generate text.

2.1 WHAT IS A LANGUAGE MODEL?

A language model is any model that assigns probabilities to sequences of words. Given a sequence of words w_1, \dots, w_n , a language model outputs the likelihood $P(w_1, \dots, w_n)$ of this sequence. An ideal language model would have high likelihood to natural-sounding text, like the sentences in this paragraph, and low likelihood to gibberish. Most language models make the assumption that the likelihood of a word is dependent only on the words that precede it. Thus, the chain rule applies:

$$P(w_1, \dots, w_n) = P(w_1) \times \dots \times P(w_i | w_1, \dots, w_{i-1}) \times \dots \times P(w_n | w_1, \dots, w_{n-1}) \quad (2.1)$$

Before the transition to neural network-based models, the most common form of language model was a statistic language model called the n -gram model. Instead of trying to estimate the probability of a word

given all preceding words, n -gram model make the Markov assumption that the probability of a word is only dependent on a fixed $n-1$ preceding words. The n in n -gram refers to the number of words used in the conditional probability distribution, and “gram” simply means “word.” For example, using a 3-gram model, we would approximate each factor in Equation 2.1 as

$$P(w_i|w_1, \dots, w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1}) \quad (2.2)$$

An n -gram models can be constructed from a corpus of text by simply counting how many times each word in the text is preceded by each possible n -gram. This is an advantage over grammar-based approaches to language modeling—such as statistical parsers—which require explicitly labeled training data, such as the Penn Tree Bank, in order to estimate probabilities.

There are several disadvantages to this n -gram based approach to language modelling. First, n -gram models tend to be sparse. If a particular $\langle n\text{-gram, word} \rangle$ pair never occurs in the corpus, then the model will assign it a probability of 0. As a result, smoothing techniques are often employed to prevent plausible but novel word sequences from being assigned a probability of zero. Second, the complexity of storing an n -gram language model grows exponentially with the choice of n . In practice, most n -gram models used n between 1 and 5, which is insufficient for modelling long-term dependencies and coherence. Third, n -gram models do not adequately represent words which are not present in their training data. Such words are typically replaced with a special out-of-vocabulary identifier. Neural language models, described in the following sections, overcome many of these limitations.

2.2 WHAT IS A NEURAL LANGUAGE MODEL?

Neural network-based language models replace the statistical models described in the previous section with a learned function (the neural network) whose output can be used to predict the likelihood of a word sequence. In

contrast to n -gram models, neural language models are capable of assigning non-zero probability to sequences never seen in their training corpora, and thus they can be used to model longer sequences. State-of-the-art neural language models can model sequences in the thousands of words.

One of the key advancements in neural language modeling was the transition from operating on sequences of discrete words to operating on sequences of continuous vector representations. The sequence of words w_1, \dots, w_n is mapped to a sequence of embedding vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$. In early work on neural language modeling, these vector representations were computed separately. Algorithms such as word2vec [102] and GloVe [114] were employed to construct embedding matrices where each row corresponded to a word in the vocabulary. In today's neural language models, the embedding matrix is typically treated as part of the neural language model, initialized randomly than optimized along with the rest of the network. Let \mathbf{E}_θ be a learned embedding matrix where each row correspond to the vector representation of one word in the vocabulary.

Typical neural language models emit $\hat{\mathbf{y}}_t$, a predicted embedding for the t th position in the sequence given the previous word embeddings in the sequence. This can be written as

$$\hat{\mathbf{y}}_t = f_\theta(\mathbf{y}_1, \dots, \mathbf{y}_{t-1}) \quad (2.3)$$

where f_θ is the neural network and $\mathbf{y}_1, \dots, \mathbf{y}_{t-1}$ are the embeddings of the previous tokens in the sequence.

To produce a probability distribution for what the next word should be given the previous words, the predicted embedding $\hat{\mathbf{y}}_t$ is multiplied by the embedding matrix \mathbf{E}_θ to produce a score for each word in the vocabulary. Then a softmax transformation is used to normalize these scores into a probability distribution. Let Y_t be a random variable representing the vocabulary item predicted for the t th position. We then have:

$$P(Y_t = i | \mathbf{y}_1, \dots, \mathbf{y}_{t-1}) = \frac{\exp(\mathbf{E}\hat{\mathbf{y}}_t[i])}{\sum_j \exp(\mathbf{E}\hat{\mathbf{y}}_t[j])} \quad (2.4)$$

where i and j are indexes into the vocabulary.

The learned weights θ are optimized using a log likelihood loss. More precisely, we can write the training loss for a sequence $\mathbf{y}_1, \dots, \mathbf{y}_n$ as:

$$\mathcal{L} = - \sum_{t=1}^n \log P(Y_t = i^* | \mathbf{y}_{1:t-1}) \quad (2.5)$$

$$= - \sum_{t=1}^n \log \frac{\exp(\mathbf{E}_\theta \hat{\mathbf{y}}_t[i^*])}{\sum_j \exp(\mathbf{E} \hat{\mathbf{y}}_t[j])} \quad (2.6)$$

$$= - \sum_{t=1}^n \mathbf{E}_\theta \hat{\mathbf{y}}_t[i^*] \quad (2.7)$$

$$= - \sum_{t=1}^n (\mathbf{E}_\theta f_\theta(\mathbf{y}_1, \dots, \mathbf{y}_{t-1})[i^*]) \quad (2.8)$$

In these equations, i^* is the index of the groundtruth word at position t in the sequence. By taking the dot product between the neural network's predicted embedding and the embedding of the true word at each position t (Eq. 2.7), we get a score for how correct the neural network's prediction for this position is. Training with an objective of maximizing the sum of these scores over every word position is equivalent to minimizing the negative log likelihood (or maximizing the likelihood) of the sequence.

In some language modelling applications, it is common to have an additional sequence which the model is conditioned on in addition to the tokens of the target sequence. This paradigm is known as an encoder-decoder or sequence-to-sequence model, and the formulation above is modified to

$$\hat{\mathbf{y}}_t = f_\theta(\mathbf{y}_1, \dots, \mathbf{y}_{t-1}; \mathbf{x}_1, \dots, \mathbf{x}_n) \quad (2.9)$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n$ is the additional input sequence. The most popular application of encoder-decoder models is machine translation, where to convert some text from French to English, the language model predicts the next word of the English sequence given the entirety of the French sequence and the preceding words of the English sequence.

Most state-of-the-art neural language models uses a variant of the Transformer architecture [154] as the neural network f_θ . Prior to the development of Transformers, recurrent neural architectures, typically based on Long Short Term Memory units [60], were most commonly employed. Transformers have several

advantages over their recurrent predecessors, most notably that operations are parallelized across all tokens in the sequence. This immensely speeds up computation time during training, and computation time is no longer dependent on the length of the sequence. Transformers also are much better than recurrent models at making connections between information that may be very far apart in the sequence. Recurrent architectures keep track of a “hidden state” which gets updated for every position in the sequence, which means it may no longer encode much information about the beginning of the sequence by the time it arrives at the end. In contrast, Transformers use an “attention mechanism” that allows any position in the input sequence to easily “attend” to any other position.

2.3 ENCODING TEXT INTO A VOCABULARY

For simplicity, the previous sections refer to the input to a language model as a sequence of words, but in practice, neural language models use a variety of different techniques to construct vocabularies of varying granularities. There is no single solution for forming the base units of language (referred to for the remainder of this chapter as “tokens”), and techniques vary significantly across languages. In English, the simplest vocabularies are character-level—each letter of the alphabet and punctuation mark becomes a token. Historically, word-level vocabularies, where each token corresponds to a word in the dictionary, were most common. Word-level vocabularies can be created by splitting a string on whitespace and punctuation. Since in most languages, the number of fully inflected words is enormous, in practice only the most common tens or hundreds of thousands of words are included in the vocabulary, and all other words are replaced with an out-of-vocabulary (OOV) token.

In recent years, subword vocabularies, which eliminate the OOV problem, have become standard in neural language modeling. Subword vocabularies are formed by choosing a budget (the desired size of the vocabulary), then running an algorithm that joins letters together into larger units, such that the most common character sequences end up as tokens in the vocabulary. While common words such as “cat” or “dog” end up as single

tokens in the vocabulary, uncommon words such as hippopotamus end up being broken into multiple tokens. Several greedy algorithms have been proposed to approximate optimally breaking up a text corpus into subwords, but byte-pair encoding (BPE) is currently the most popular [136]. Typical subword vocabulary sizes are between 32,000 and 50,000 tokens. Table 2.1 shows a sentence under a few different tokenization schemes.

Table 2.1: Examples of the string “A hippopotamus ate my homework.” tokenized using three different vocabularies. With the subword tokenizer, the rare word “hippopotamus” gets broken up into multiple tokens. For word-level tokenizers, if the word “hippopotamus” occurred very infrequently in the corpus used to build the vocabulary (or perhaps the writer of the sentence misspelled it), it would typically get replaced with an out-of-vocabulary token (row 4).

| Vocab Type | Example |
|-----------------|---|
| character-level | ['A', ' ', 'h', 'i', 'p', 'p', 'o', 'p', 'o', 't', 'a', 'm', 'u', 's', ' ', 'a', 't', 'e', ' ', 'm', 'y', ' ', 'h', 'o', 'm', 'e', 'w', 'o', 'r', 'k', '.'] |
| subword-level | ['A', 'hip', '##pop', '##ota', '##mus', 'ate', 'my', 'homework', '.'] |
| word-level | ['A', 'hippopotamus', 'ate', 'my', 'homework', '.'] |
| word-level | ['A', '[UNK]', 'ate', 'my', 'homework', '.'] |

For all of the types of vocabularies discussed, a decision must be made on whether to convert strings to lowercase before vocabulary creation. Removing case allows for a more compact vocabulary, but it also removes potentially useful information about the location of proper nouns.

Subword vocabularies were designed to be a compromise between the advantages and disadvantages of word-level and character-level vocabularies. Character-level vocabularies are usually very small, no more than a couple hundred tokens. However, the vocabulary can cover near every possible string a person could write. Word-level vocabularies cannot feasibly contain the hundreds-of-thousands of words present in English text. Realistically, only the most common words are kept, and less common ones are replaced with a special UNK token. When text is tokenized with character-level vocabularies, the resulting sequences are very long, while word-level tokenization yields shorter sequences since there is just one token per word. Lastly, word-level representations learned by a neural net tend to be more meaningful than character-level representations since

a word has semantics associated with it that are common across uses. Subword vocabularies adopt the best of both worlds, using word-level tokens for common words but falling back to subword, or in the worst case, character-level, tokenization for uncommon words. This approach eliminates the need for an out-of-vocabulary token and results in tokenized sequence lengths which are somewhere between the two strategies.

2.4 GENERATING TEXT WITH A LANGUAGE MODEL

Neural language models in themselves are capable of generating text. As described in the previous sections, most language models provide a probability distribution for what the next token in the sequence *could* be, given the previous tokens. To perform generation, an algorithm is needed that chooses which words to output given the model's predicted probability distributions. We refer to this family of algorithms as **decoding methods** because they “decode” a sequence of discrete words from the model's predictions. At each step of decoding, the decoding algorithm performs a forward pass on the neural network using the existing prompt text as input, selects a next token based on the neural network's predictions, adds this token to the prompt, and repeats until the desired number of tokens have been generated.

2.4.1 Greedy Approaches

The simplest strategies for generating text from a language model involve greedily choosing a token at each step. The easiest way to do this is to take the arg max of the distribution, repeatedly picking the token with the highest probability according to the model. This approach is simple but only allows a single generation to be produced for any given prompt. Alternatively, one can randomly sample from the vocabulary, where each vocabulary item has a chance of being picked that is proportional to the probability predicted for it by the language model. This method allows for many different sequences to be generated from the same prompt.

However, in practice, this strategy results in text that is perceived as nonsensical and otherwise low-quality. This is because the probability distributions returned by neural language models tend to be very long tailed, and the chance of sampling a rare/unusual word from this long tail is quite high. For example, if we sample from the full distribution words that could follow `The dog ate my`, with low probability we might sample `brains`, even though `homework` is much more probable.

Several strategies have been proposed to improve random sampling techniques by reducing the entropy of the distribution before sampling. Introducing a temperature parameter τ into the softmax computation allows us to smoothly shift probability mass from low-scoring items in the vocabulary to high-scoring ones.

$$P(Y_t = i | \mathbf{y}_1, \dots, \mathbf{y}_{t-1}) = \frac{\exp(\mathbf{E}\hat{\mathbf{y}}_t[i]/\tau)}{\sum_j \exp(\mathbf{E}\hat{\mathbf{y}}_t[j]/\tau)} \quad (2.10)$$

Alternatively, one can introduce sparsity into the distribution by deliberately zeroing out low-likelihood vocabulary items. Top- k random sampling accomplishes this by restricting sampling to only the k most likely tokens at each step. Nucleus sampling, also referred to as top- p random sampling, accomplishes this by restricting sampling at timestep t to the k_t most likely tokens, where k_t is selected such that these tokens cover no more than $p\%$ of the probability mass. For all three of these techniques there is a parameter (τ , k , or p) which controls the amount of randomness we want to permit in the generation. Choosing a low value for these parameters results in an increasingly peaky distribution, which, at its extreme, is the same as taking the `arg max`. Choosing a high value for these parameters results in the distribution that looks closer and closer to the original scores produced by the model.

2.4.2 Search-Based Approaches

Before the transition to Transformer-based architectures (Section 2.2), the standard convention for generation was to try to generate the most likely overall sequence from the language model. This approach made a lot of sense for the predominant use case of machine translation, where generating one correct translation was

considered more important than generating several diverse translations. Since computing the overall most likely output sequence is intractable, early work in neural machine translation found that beam search was an effective strategy to heuristically sample sufficiently likely sequences from these probabilistic models [144].

Algorithm 1 gives an overview of the beam search algorithm. “SOS” is a start-of-sequence token and “EOS” is an end-of-sequence token.

Algorithm 1 Beam Search Inference

```

1: procedure BEAM SEARCH
2:    $B \leftarrow \{SOS\}$ 
3:    $k \leftarrow \text{BeamWidth}$ 
4:    $out \leftarrow k\text{-best output list}$ 
5:   while  $|out| < k$  do
6:      $front \leftarrow$  remove all nodes from  $B$ 
7:     for  $w \in front$  do
8:        $succ \leftarrow w$ 's  $k$ -best successors
9:       for  $s \in succ$  do
10:        if  $s == EOS$  then
11:           $out \leftarrow out \cup \{s\}$ 
12:        else
13:           $B \leftarrow B \cup \{s\}$ 
14:        end if
15:      end for
16:    end for
17:    Sort  $B$ 
18:    if  $|B| > k$  then
19:      Prune  $B$  to  $k$ -best successors
20:    end if
21:  end while
22:  return  $out$ 
23: end procedure

```

As neural language models came to be applied increasingly to open-ended tasks, such as chatbot dialog or story generation, beam search was found to be ill-suited to generating a set of *diverse* candidate sequences. Since beam search only explores a limited portion of the overall search space, it tends to yield multiple variants of the same high-likelihood sequence, sequences that often only differ in punctuation or minor morphological

changes [87]. To try and solve this problem, many researchers proposed modification to beam search to encourage it to produce more diverse sets of candidate generations. We summarize several of these here:

- **Noisy Parallel Approximate Decoding.** Introduced by Cho [28], NPAD is a technique that can be applied to any decoding setting. The main idea is that diversity can be achieved more naturally by taking advantage of the continuous manifold on which neural nets embed language. Instead of encouraging diversity by manipulating the probabilities outputted from the model, diverse outputs are instead produced by adding small amounts of noise to the hidden state of the decoder at each step. The noise is randomly sampled from a normal distribution. The variance is gradually annealed from a starting σ_0 to 0 as decoding progresses (that is $\sigma_t = \frac{\sigma_0}{t}$) under the reasoning that uncertainty is greatest at the beginning of decoding. NPAD can be used in conjunction with any decoding strategy, though the paper in which it was introduced primarily showed its performance in conjunction with beam search.
- **Top-g Capping.** In beam search, it is often the case that one hypothesis h is assigned a much higher probability than all other hypotheses, causing all hypotheses in the next step to have h as their parent. Li et al. [87, 88] proposed adding an additional constraint to standard beam search to encourage the model to choose options from diverse candidates. At each step t , current hypotheses are grouped according to the parental hypothesis they come from. After grouping candidates, only the top g from each grouping are considered. The resulting $b \times g$ candidates are ranked, and the top b are selected as hypotheses for the next beam step.
- **Hamming Diversity Reward.** Vijayakumar et al. [157] proposed adding an additional diversity-promoting term, θ , to the sequence log-likelihoods before the reranking step of beam search. This term measures how different a candidate hypothesis $c_{\leq t}^{(i)}$ is from the partial hypotheses selected in the

previous step. Let $\mathcal{H}_{t-1} = \{c_{\leq t-1}^{(1)}, \dots, c_{\leq t-1}^{(b)}\}$ be these partial hypotheses. Then the beam search scoring function for the i th candidate at timestep t becomes:

$$\text{score}(c_{\leq t}^{(i)}) = \sum_{j=1}^t (\log P(c_j^{(i)} | c_{< j}^{(i)}, \mathbf{x})) \\ + \lambda \theta(c_{\leq t}^{(i)}, \mathcal{H}_{t-1})$$

where λ is a tunable hyperparameter. Vijayakumar et al. [157] try a variety of definitions for θ , including embedding diversity and n -gram diversity, but they find that Hamming distance, the number of tokens in the candidate sequence which exist in the previously selected partial hypotheses, is most effective.

- **Iterative Beam Search.** In an attempt to improve the size of the search space explored without sacrificing runtime, Kulikov et al. [81] propose an iterative beam search method. Beam search is run many times, where the states explored by subsequent beam searches are restricted based on the intermediate states explored by previous iterations. Formally, they define the set of all partial hypotheses for beam search instance i at time step t as $\mathcal{H}_t^{(i)}$. From here, the search space explored by beam search instance i can be expressed as $\mathcal{S}_i = \cup_{t=1}^T \mathcal{H}_t^{(i)}$. The i th beam search is prevented from generating any partial hypothesis that has previously been generated, that is, any hypothesis found in $\mathcal{S}_{< i} = \cup_{i'=0}^{i-1} \mathcal{S}_{i'}$.
- **Clustered Beam Search.** Tam et al. [145] proposed a clustering-based beam search method to help condense and remove meaningless responses from chatbots. Specifically, at each decoding step t , this method initially considers the top $2 * b$ candidates. From there, each candidate sequence is embedded¹, and the embeddings are clustered into c clusters using K -means. Finally, we take the top $\frac{b}{c}$ candidates from each cluster. Note that in the case any clusters have size less than $\frac{b}{c}$, we then include the highest-ranked candidates not found after clustering.

¹ We follow Tam et al. [145] and used averaged GloVe word embeddings [114].

2.4.3 Generation Diversity

For many tasks, especially open-ended ones like story generation or chitchat dialog, it is important for generated text to be “diverse.” The term “diversity” has been used in the language model literature to refer to a diverse set of properties. Some use it as a synonym for sentence interestingness or unlikeliness [58]. Others consider diversity a measure of how different two or more sentences are from each other [157, 53]. In some framings, diversity is measured across a set of generations coming from the same prompt. Given a particular prompt or input, the goal is to measure the breadth of possible generations the model will produce [100]. Diversity can also be measured as a corpus-level: given all the sentences generated by the model for all prompts, what is the overall lexical diversity?

In my research, I use three definitions of diversity. First, when performing conditional generation, I define diversity as the ability of a generative method to create a set of possible outputs that are each valid given a particular input but vary as widely as possible in terms of word choice, topic, and meaning. Second, when performing unconditioned generation using decoder-only language models, I instead consider corpus-level diversity across all the model’s generations—how much lexical variety is there over all the text the model generated? Finally, in some of my work, I ask human raters to evaluate generation interestingness, which is a measure of human-perceived diversity.

2.4.4 Generation Quality

For all generation tasks, it is important for the output text to be high quality, though this property can also be difficult to define. In some downstream applications, “quality” can be evaluated directly with human raters by asking them questions like “how good is this text?” (though definitions of “good” vary widely across the literature [82]). In others, it can be quantified as how many times a user interacts with

the generative system (for example, the number of conversation turns with a dialog agent) before losing interest.

To some extent, quality can also be measured automatically. In tasks with a clear goal, like machine translation or summarization, one can compare the generation against a gold standard. Quality is strongly associated with fluency, and it is generally true that the lower perplexity a language model assigns some text, the more fluent the text is, and thus the higher quality. However, my collaborators and I show that this relationship between quality and perplexity breaks down for extremely high-likelihood generated text [177].

In some of my research, we evaluate quality by asking humans to assess generations in terms of their fluency, adequacy, and interestingness. In Chapter 3, we propose a novel method for assessing generation quality based on the premise that humans (or a trained discriminator) ought to have a hard time distinguishing between real human-written text and model outputs when the model outputs text that is high-quality.

2.4.5 The Diversity-Quality Tradeoff

The goal of generating high-quality text is often at odds with the goal of generating diverse text. In experiments conducted with Reno Kriz [68], I found that none of the diversity-promoting search methods accomplished their stated goal of improving diversity without significant penalty to generation quality. In our experiments, we compared all of the diverse beam search methods described above with standard beam search as well as several settings of random sampling with temperature. On an open-ended dialog task, we showed that human-judged generation quality was inversely correlated with three measures of diversity (Figure 2.1).

For each utterance in the dialog task validation set, we generate 10 candidate outputs using each decoding method. To measure the diversity across the generated candidate sequences for a given input

utterance, we compute **Dist- k** , the total number of distinct k -grams divided by the total number of produced tokens in all of the candidate responses for a prompt [86]. We use $k=2$. A limitation of Dist- k is that all k -grams that appear at least once are weighted the same, ignoring the fact that infrequent k -grams contribute more to diversity than frequent ones. Therefore, we also report Zhang et al.’s [179] proposed entropy metric, **Ent- k** , defined as:

$$Ent-k = \frac{-1}{\sum_{w \in S} F(w)} \sum_{w \in S} F(w) \log \frac{F(w)}{\sum_{w' \in S} F(w')}$$

where S is the set of all k -grams that appear in candidate responses for an example, and $F(w)$ denotes the frequency of w in the candidate responses. Finally, we report **perplexity**, averaged over *all* the top 10 outputs for each example.

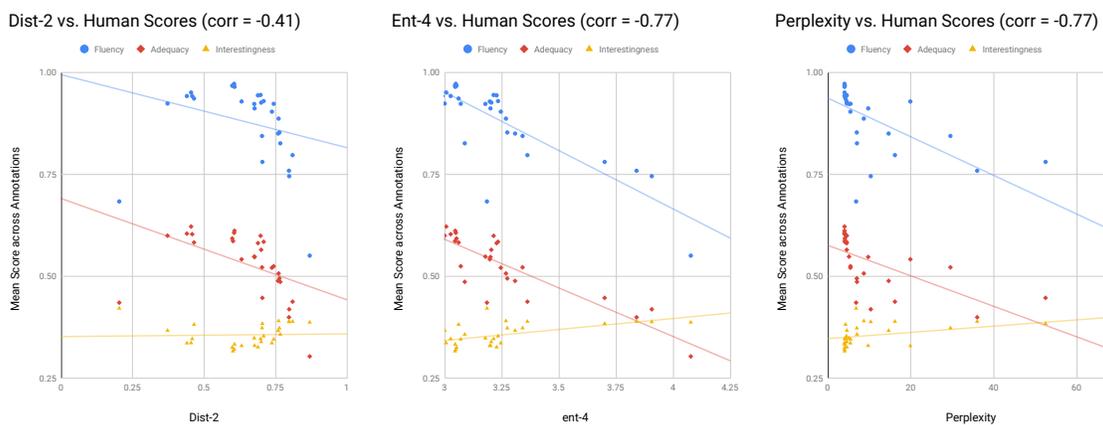


Figure 2.1: Each point corresponds to the outputs from one decoding strategy. The x-axes give the dist-2, ent-4, and perplexity scores of the generated text. The y-axes give the human-judged fluency, coherence, and interestingness of the outputs on a scale from 0 to 1. The Pearson Correlation coefficients between each statistic and the average of fluency, coherence, and interestingness are shown in parentheses.

For all three diversity measures, we see what the decoding strategies which produce the most diverse text also produce the least fluent and least adequate responses to the input utterances. For example, given the prompt “Look, nobody knows we did it.,” random sampling generates the candidate responses “We didn’t have a plan I engineered a policy.” and “Same time you pick us

up at six and get we..” These are pretty interesting but don’t make much sense. In contrast, beam search generates “I don’t know what to say,” which is neither interesting (as evaluated by human raters) nor diverse (many generated responses started with I don’t know). However, it is a reasonable response to the prompt.

One important ramification of the diversity-quality tradeoff is how detectable the generated text is to humans and automatic discriminators. The relationship between detectability and the lexical diversity of model generations is described in detail in Chapter 3.2.

2.5 LANGUAGE GENERATION TASKS IN THIS DISSERTATION

My dissertation addresses several different tasks in language generation. A brief summary of each task, as well as the means by which performance on it is evaluated, is provided here.

CONTINUATION An NLG system is asked to generate a continuation for a prompt. It is then evaluated on how close the generated continuation is to the true continuation. Automatic evaluation can either be performed using word overlap metrics such as BLEU [112], or by measuring fluency (computing the perplexity of a model on the generated continuation). Human evaluation usually involves showing a rater the prompt and generation and asking them to make some decision about it. In Chapter 3.2, the decision is to try and distinguish whether the presented text was machine-generated or not.

FILL-IN-THE-BLANK The fill-in-the-blank or infilling task is similar to continuation, except that the system also has access to the text which should occur *after* the generation. Evaluation is similar to evaluating continuation. Chapter 5.2 describes this task in detail.

CHITCHAT DIALOG Chitchat dialog is the task of predicting the next utterance in a conversation given the previous turns. As describes in the previous section, we evaluate how choice of decoding strategy impacts the ability of an NLG system to produce an utterance that is both high quality and diverse. In addition to conducting automatic evaluation with BLEU and perplexity, human evaluation can be performed by asking raters to evaluate each generated utterance in terms of fluency, adequacy, and interestingness.

TEXTUAL STYLE TRANSFER AND REWRITING Textual style transfer is the task of taking an input passage of text and a desired style and rewriting the input text to be in that style. Typical tasks include sentiment transfer (for example, rewrite a negative restaurant review to have positive sentiment) and formality transfer (rewrite informal language to be formal). In my work, I am broadly interested in the task of rewriting input text to fulfill a user-specified writing objective. These rewriting tasks are a superset of style transfer; for example a user may ask for text to be “rewritten to include the word balloon” or to “have a cliffhanger at the end.” Such rewriting could change both the content and the style. We evaluate rewriting using automatic metrics, such as measuring how often an automatic classifier identifies the rewritten task as having fulfilled the rewriting goal. We can also use metrics like BLEU score to compare against both the input sentence and a human-written groundtruth, though word-overlap metrics break down the more open-ended and transformational the rewriting task is. Finally, human raters can be used to assess quality of the rewrite.

STORY IDEATION AND BRAINSTORMING For many writers, the process of writing is collaborative. They may use an ideation tool such as a deck of trigger cards to come up with initial ideas, or they may share their in-progress draft with readers to get feedback. One of the goals of my research is to be able to use neural language models to provide an alternative collaboration source for creative writers. This includes a suite of user-defined tasks centered around allowing writers to make requests such as “what should happen next in my story” or “help me write a description of the old man introduced in the

first sentence.” Because the goals here are so broad, evaluation is best done holistically—by conducting user studies to evaluate whether NLG outputs are useful to human writers in their writing goals.

2.6 CONTROLLABILITY AND TASK-SPECIFIC GENERATION

In the early days of neural language modeling, it was common to train a separate neural language model for each NLG task of interest. For example, if one wanted a system capable of producing chatbot dialog, one would train their neural language model on a dialog dataset (or close approximate) such as OpenSubtitles [158]. If one wanted a system able to perform text summarization, one would likewise train a model from scratch on a dataset such as the CNN/Daily Mail corpus [107, 134]. At the time, the neural networks being used for these sorts of tasks were relatively small, and training and maintaining one model per task, was mostly feasible.

In 2018, Howard and Ruder [62] and Radford et al. [117] concurrently proposed the idea of pre-training a single universal task-agnostic language model. To accomplish any specific language task of interest, that model could subsequently be trained for extra steps on the training data of the desired task, a process known as finetuning. The idea of finetuning a more general model for a specific task had already taken off in computer vision, where researchers had shown a convolutional neural network pre-trained on the ImageNet task of classifying the contents of images could be finetuned for tasks ranging from image segmentation to cancer detection.

General-purpose language models intended for generation tasks tend to be pre-trained on massive datasets scraped from the internet (Table 2.2). It is common to use both decoder-only models trained only to predict the next token given the previous ones [118], as well as encoder-decoder architectures trained with a de-noising loss, where the input is a corrupted version of the text, and the task is to recover the uncorrupted text [119, 85]. Table 2.3 gives examples of several pre-training objectives that

Table 2.2: A survey of datasets which have been used to train large general-purpose neural language models.

| Dataset | Size | Public? | Language | Models trained on it |
|---------------------------|----------------|---------|----------------|--------------------------|
| C4 [119] | 365M examples, | Yes | Most English | T5 |
| mC4 [170] | 8.5B examples | Yes | 101 languages | mT5 |
| The Pile [50] | 825 GiB | Yes | Mostly English | GPT-Neo, Megatron-Turing |
| RealNews | | Yes | | |
| PanGu- α train set | 1.1TB | No | Chinese | PanGu- α |
| WebText | 40 GiB | No | Mostly English | GPT-2 |
| GPT-3 train set | 500B tokens | No | Mostly English | GPT-3 |

have been employed by popularly-used pre-trained models. Each of these models has been finetuned for a large diversity of downstream tasks.

Finetuning such models has yielded immense success in tasks across the field of natural language processing. Chapter 5.2 focuses on the feasibility of finetuning for the fill-in-the-blank task. There are however several limitations to the paradigm of pre-training followed by finetuning. As state-of-the-art neural language models increase in number of parameters, the computational expense of finetuning is becoming increasingly prohibitive. Furthermore, the need to store (potentially in GPU memory) one set of model weights per task makes it very difficult to build downstream applications which need to perform several different tasks. In addition, finetuning only works where there is enough data to fine-tune on. Overfitting is a significant challenge when training or finetuning in low-resource settings, where there may only be a handful of training examples.

For these reasons, various approach have been proposed for replacing the finetuning step with methods which require either no or minimal weight training. Brown et al. [20] introduce the technique of few-shot prompting. By constructing a textual prompt which contains several exemplars of the goal task, a general-purpose language model can be made to perform the task. Lester et al. [84] introduce *prompt tuning* as an improvement over few-shot prompting that trains a small neural network to produce an optimal prompt in embedding-space for the goal task.

Table 2.3: Examples of pre-training objectives used in popular general-purpose models. In these examples, the original training sequence is “The hippopotamus ate my homework. It made me very mad.”

| Model | Input | Objective |
|---------------|--|--|
| BERT [36] | [cls] The hippopotamus [mask] my homework. [SEP] It made me very [mask] . [sep] | Predict tokens for [mask] positions and predict whether the two sentences are in the correct order. |
| T5 [119] | The hippopotamus [x] made me [y] mad. | Predict missing sequences. |
| BART | It _ me very mad. The hippopotamus my _. | Predict the original uncorrupted sequence from a version that has been noised (token masking/deletion, text infilling, document rotation, sentence shuffling). |
| GPT [118, 20] | The hippopotamus ate | Predict the next token given the previous tokens. |

In this dissertation, I explore both finetuning and few-shot prompting. Chapter 5.1 uses prompting techniques for the task of textual style transfer, while Chapter 5.3 shows how they can be used for a variety of story editing operations. Chapter 5.2 focuses on fill in the blank, a task where finetuning outperforms other more training-efficient methods.

3

DETECTING MACHINE-GENERATED TEXT

3.1 MOTIVATION

State-of-the-art generative language models are now capable of producing multi-paragraph excerpts that at a surface level are virtually indistinguishable from human-written content [174, 118, 2]. Often, only subtle logical fallacies or idiosyncrasies of language give away the text as machine-generated. These errors can be difficult for humans to detect because they require a close reading and/or domain knowledge.

Deceptive text, whether human- or machine-generated, has entered the sphere of public concern [32]. It propogates quickly [159], sets political agendas [152], influences elections [5], and undermines user trust [161, 141]. Recently, Adelani et al. [2] have shown that automatically generated reviews are perceived to be as fluent as human-written ones. As generative technology matures, authors, well-meaning or otherwise, will increasingly employ it to augment and accelerate their own writing. In the initial year following the release of GPT-2, a model whose authors claimed was so impressive that “malicious applications of the technology” were a significant concern [118], there was relatively little inquiry into the textual properties that cause humans to give generated text high human-like ratings compared to those that cause automatic systems to rate it highly.

This task of trying to guess whether text is coming from a robot or a fellow human was made famous by the Turing Test [151]. It has been a standard challenge for evaluating chatbot systems [96, 35] but has not been adopted more generally for evaluating natural language generation systems. The related (but not identical) task of asking human raters to judge the quality of machine-generated excerpts remains the gold-standard for evaluating open-domain generation systems [82]. However, using detection as a method for evaluating NLG

systems has several advantages over asking for quality ratings. First, text quality—i.e., the properties that make one passage of text “better” than another can be inherently subjective to define. The detection task offers a mostly objective definition; we can precisely measure what fraction of examples are correctly identified as machine-generated or human-written. Second, detection can be framed both as an automatic evaluation task and a human evaluation one. Valuable insights can be garnered by looking at text sequences that automatic detection systems label incorrectly but humans label correctly, and vice versa. Third, studying detection allows us to evaluate the real risk of NLG systems to the public. Thus, in the research presented in this chapter, I show how the detection task is an effective means of evaluating progress in NLG and determining the relative strengths of different NLG systems.

In Section 3.2, I explore the relationship between choice of decoding strategy and the detectability of generated text. When using a decoding strategy that samples from a language model’s full predicted distribution, systems end up frequently choosing words from the long tail of the distribution. Choosing one of these words can lead to odd or contradictory phrases and semantic errors. Humans readers are quick to notice these types of errors. If we instead use a decoding strategy that reduces the entropy of the distribution before sampling, the result is improved generation quality at the cost of diversity. We show that humans have a hard time identifying that text is machine-generated when sampling is heavily restricted to only high-likelihood words, but automatic detection systems easily pick up on the resulting lexical anomalies. In other words, humans are fooled where text classifiers are not, and vice versa.

In Section 3.3, I conduct a large-scale study of the detectability of generated text by human annotators, expanding upon the pilot human evaluation experiments described in Section 3.2. I show how factors such as model size and text genre impact detectability, and I introduce ways to gamify the detection task.

3.1.1 Detection as a Task

There are many ways one could frame the detection task. The simplest is to treat it as binary classification: given an excerpt of text, label it as either human-written or machine-generated. The excerpt is either entirely

human-written or contains some amount of machine-generated text. This is the definition I use in Section 3.2 where the positive examples of machine-generated text contain at most one initial word of human text.

One limitation with this definition is that it is not realistic. Most users of NLG systems prompt the language model with much more than one word when they ask it to generate text. Therefore, in Section 3.3, we instead formulate the task as one of boundary detection. Given a passage that starts off with a variable number of human-written sentences which are followed by a generated continuation, the goal is to identify the point of transition between human and generated.

Others have studied different variations of the detection task. GROVER was designed to not only generate convincing news excerpts but to also identify them using a fine-tuned version of the generative model itself [174]. GLTR, expecting attackers to use sampling methods that favor high-likelihood tokens, aims to make machine-generated text detectable by computing histograms over per-token log likelihoods [52]. Bakhtin et al. [8] frame human-text detection as a ranking task and evaluate their models' cross-domain and cross-model generalization, finding significant loss in quality when training on one domain and evaluating on another. Schuster et al. [133] argue that the language distributional features implicitly or explicitly employed by these detectors are insufficient; instead, one should look to explicit fact-verification models. Finally, discriminators for whether text is machine-generated are a promising research direction in adversarial training [92, 89] and in automatic evaluation of generative model quality [111, 74, 96].

3.2 IMPACT OF DECODING STRATEGY ON THE DETECTABILITY OF MACHINE-GENERATED TEXT

Even when the language model used for generation is fixed, choice of decoding strategy has a huge impact on the detectability of generated text. Using top- k random sampling, a decoding method where only the selection of high-likelihood words is permitted, means the system is less likely to make a poor choice and create the type

of mistakes that are easy for humans to detect. Since humans are not proficient at identifying when a model subtly favors some utterances more often than a human author would, they don't notice the over-representation of high-likelihood words in the generated text. In contrast, automatic detection systems excel at identifying statistical anomalies and struggle to build deeper semantic understanding. Top- k in particular creates text that is easy for machines to detect but very hard for humans. Thus, we observe the general trend: as the number of unlikely words available to be chosen is increased, humans get *better* at detecting fakes while automatic systems get *worse*.

In this section, I present a study of three popular random decoding strategies—top- k , nucleus, and full random sampling—applied to GPT-2 [118]. As described in Chapter 2, a challenge with random sampling is that predicted probability distributions often contain a long tail of vocabulary items that are individually low-probability but cumulatively comprise a substantial amount of probability mass. Holtzman et al. [61] observe that choosing tokens from this tail often leads to incoherent generations. Top- k sampling, nucleus sampling, and (in the extreme) beam search have all been proposed to heuristically promote samples with higher per-token likelihoods. Top- k and nucleus sampling both do so by setting the likelihood of tokens in the tail of the distribution to zero. Top- k restricts the distribution to all but the k most likely tokens, where k is a constant [42]. Nucleus sampling, also called top- p , truncates the distribution at each decoding step t to the k_t -most-likely next tokens such that the cumulative likelihood of these tokens is no greater than a constant p [61].

I draw a large number of excerpts generated by each strategy and train a family of BERT-based [36] binary classifiers to label text excerpts as human-written or machine-generated. I find large differences in human rater and classifier accuracy depending on the decoding strategy employed and length of the generated sequences. Regardless of strategy, we find human raters achieve significantly lower accuracy than the automatic discriminators. I also show that when a decoding strategy severely modifies the unigram token distribution, as top- k does, humans have trouble detecting the resultant generated text, but automatic classifiers find it the easiest to discriminate. Worryingly, I further find that classifiers are brittle; they generalize poorly when trained to discriminate samples from one strategy and then evaluated on samples from another.

3.2.1 Dataset Construction

To study detectability, I formed datasets that probe how excerpt length and decoding strategy impact detection performance. Each dataset is approximately balanced between positive examples of machine-generated text and negative examples of human-written text. While they all share the same human-written examples, each dataset contains a different set of machine-generated examples sampled using one particular decoding strategy. I also built additional datasets by truncating all of the examples to a particular sequence length, testing out sequence lengths between 2 and 192 tokens long.

By training a separate classifier on each dataset, we are able to answer questions about which decoding strategy results in text that is the easiest to automatically disambiguate from human-written text. We are also able to answer questions about how the length of the examples in the training set impacts our ability to automatically classify excerpts of that same length as either human-written or machine-generated.

All of our generated text samples are drawn from GPT-2, a state-of-the-art Transformer-based generative language model that was trained on text from popular web pages [118]. While we use the GPT-2 LARGE model with 774M parameters, we found that similar trends to those reported here hold in experiments with smaller language models. GPT-2 was state-of-the-art at the time these experiments were conducted; it has subsequently been surpassed by other larger LMs.

I consider three different decoding strategy settings:

- Sample from the untruncated distribution
- Top- k , choosing $k=40$ [118].
- Nucleus sampling (aka top- p), choosing $p=0.96$ [174].

In addition, I form “negative” examples of human-written text by taking excerpts of web text that come from the same distribution as GPT-2’s training data. By picking text that resembles GPT-2’s training data, we ensure that our classifiers can’t simply take advantage of stylistic differences between the human-written text corpus and the kind of text GPT-2 was trained to generate.

Table 3.1: The number of excerpts used for training, validation, and testing. Three decoding strategies—top- k with $k=40$, nucleus sampling with $p=0.96$, and full random sampling ($p=1.0$)—were employed. The language model was either prompted with a single word (1wordcond) or used without any prompt sequence (nocond).

| Model | Decoding Method | Prompt | # train | # valid | # test |
|---------------|-----------------|-----------|---------|---------|--------|
| large-744M | k40 | 1wordcond | 211148 | 4226 | 4191 |
| large-744M | k40 | nocond | 218825 | 4362 | 4360 |
| large-744M | p0.96 | 1wordcond | 210587 | 4248 | 4208 |
| large-744M | p0.96 | nocond | 209390 | 4174 | 4185 |
| large-744M | p1.0 | 1wordcond | 209334 | 4169 | 4173 |
| large-744M | p1.0 | nocond | 208219 | 4187 | 4168 |
| human-written | n/a | n/a | 201344 | 4031 | 4030 |

For each decoding method, I construct a training dataset by pairing 250,000 generated samples with 250,000 excerpts of web text. 5,000 additional paired samples are kept aside for validation and test datasets. Lastly, I filter out excerpts with fewer than 192 WordPiece tokens [167] (excerpts might be quite short if the model produces an end-of-text token early on). The final dataset sizes are shown in Table 3.1.

A crucial question when generating text with a language model is whether or not to provide a priming sequence which the language model should continue. Unconditioned samples, where no prompt text is provided, in conjunction with top- k sampling, lead to pathological behavior for discriminators as the first token of the generated text will always be one of k possible options. On the other hand, if long sequences of human text are used as a prompt, the space of possible generated sequences is larger, but the detection problem shifts from one of “how human-like is the generated text?” to “how well does the generated text follow the prompt sequence?”.

Since in this study I am interested in the former simpler question, I create two datasets, one with no priming, and one with the minimum amount of prompting possible: a single token of web text. This means that for every excerpt of web text in the training set, there is an excerpt of machine-generated text that starts with the same token. I find that the ability of automatic detectors can be strongly impacted by even this short prompt.

Table 3.2: Performance (accuracy and AUC) of the fine-tuned BERT classifier and several simple baselines on detecting length-192 sequences generated with one word of priming (1wordcond). Note that p1.0 refers to untruncated random sampling, where we sample from 100% of the probability mass. The last column shows human performance on the same task where accuracy with a 50% baseline is computed by randomly pairing samples from each decoding strategy with a human-written sample.

| Method | BERT | | BagOfWords | | HistGLTR | | Hist50Buckets | | TotalProb | Human |
|-----------------|------|------|------------|------|----------|------|---------------|------|-----------|-------|
| | acc | AUC | acc | AUC | acc | AUC | acc | AUC | acc | acc |
| k40-1wordcond | 0.88 | 0.99 | 0.79 | 0.87 | 0.52 | 0.52 | 0.69 | 0.76 | 0.61 | 0.64 |
| p0.96-1wordcond | 0.81 | 0.89 | 0.60 | 0.65 | 0.53 | 0.56 | 0.54 | 0.56 | 0.63 | 0.77 |
| p1.0-1wordcond | 0.79 | 0.92 | 0.59 | 0.62 | 0.53 | 0.55 | 0.54 | 0.55 | 0.65 | 0.71 |

To study the effect of generation length, I construct variations of the above datasets by truncating all excerpts to ten possible lengths ranging from 2 to 192 WordPiece tokens [167]. In total, we obtain sixty dataset variations: one per sampling method, truncation length, and choice of priming or no priming.

3.2.2 Methods for Automatic Detection

The primary discriminator we employ is a fine-tuned BERT classifier [36]. We fine-tune one instance of BERT per dataset variation described above. For the longest sequence length, $n=192$, we compare BERT’s performance with several simple baselines that have been proposed in other work.

FINE-TUNED BERT We fine-tune BERT-LARGE (cased) on the task of labeling a sentence as human- or machine- generated. The models are trained for 15 epochs, with checkpoints saved every 1000 steps, and a batch size of 256. All results are reported on the test set using the checkpoint for which validation accuracy was highest.

BAG-OF-WORDS For each sequence, we compute a bag-of-words embedding where each dimension corresponds to a token in GPT-2’s 50,000 token BPE vocabulary [136], and we count how many times that token appears in the text sequence. We then train a logistic regression binary classifier to predict human- or

machine-written given this 50,000-dimensional embedding. We experimented with truncating embedding size by removing entries for infrequent vocabulary words, but this did not improve performance.

HISTOGRAM-OF-LIKELIHOOD RANKS Following GLTR [52], we compute the probability distribution of the next word given the previous words in a text sequence according to a trained language model (in our case the same GPT-2 model that was used for generation). At each sequence position, we rerank the vocabulary words by likelihood, and record the rank of the ground-truth next word within this list. These ranks are then binned. GLTR uses four bins, counting (1) the number of times the top 1 word is seen, (2) the number of times words ranked 2 through 5 are seen, (3) words ranked 6-100, and (4) words ranked >100. However, we observe higher accuracy when 50 bins are spread uniformly over the possible rankings. This means that since there are 50,000 vocabulary words, the first bin counts the number of times the actual next word was within the 1,000 mostly likely next words, the second bin counts the 1,001-2,000th, and so on. We then train logistic regression binary classifiers to predict human- or machine-written given either the 4-dimensional histograms or 50-dimensional histograms as input.

TOTAL PROBABILITY Solaiman et al. [140] propose a very simple baseline consisting of a threshold on the total probability of the text sequence. An excerpt is predicted as machine-generated if its likelihood according to GPT-2 is closer to the mean likelihood over all machine-generated sequences than to the mean of human-written ones.

3.2.3 Method for Human Detection

The human evaluation task is framed similarly to the automatic one. We ask the raters to decide whether a passage of text was written by a human or by a computer algorithm. Figure 3.1 shows screenshots of the instructions and user interface for the annotation task. Raters are allowed to choose between four options: “definitely” or “possibly” machine-generated and “definitely” or “possibly” human-written. They are first

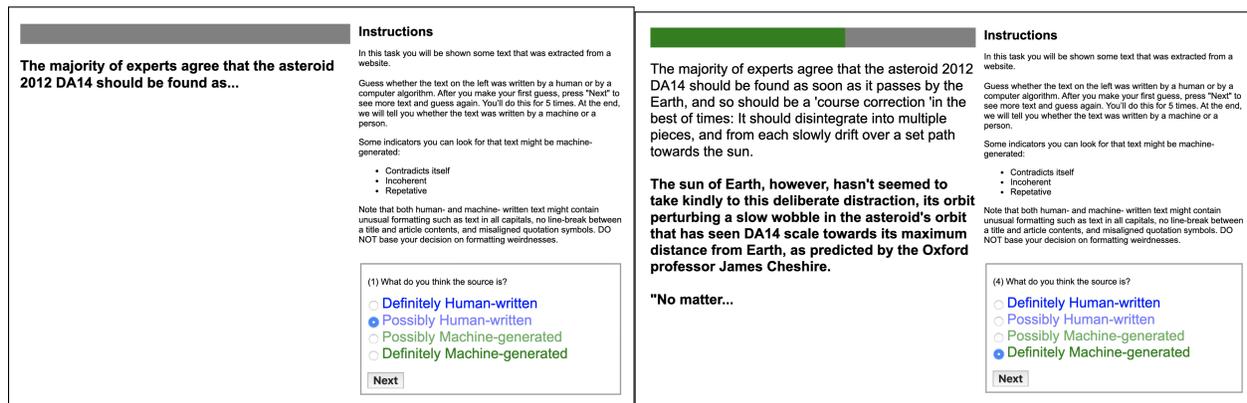


Figure 3.1: The interface of the task used for human evaluation. Each time the user presses next, the passage’s length is doubled. On the left, we show the first step of evaluation, on the right, the second to last.

shown an excerpt of length 16 WordPiece tokens. After they make a guess, the length of the excerpt is doubled, and they are asked the same question again. This continues until the entire passage of length 192 tokens is shown. Passages are equally likely to be human-written or machine-generated, with the machine-generated excerpts being evenly split between the three sampling strategies considered in this paper.

Initially, Amazon Mechanical Turk (AMT) raters were employed for this task, but rater accuracy was poor with over 70% of the “definitely” votes cast for “human” despite the classes being balanced. Accuracy, even for the longest sequences, hovered around 50%, indicating random guessing. The same study was then performed with university students who were first walked through ten examples (Table 3.4) as a group. Afterward, they were asked to complete the same tasks that had been sent to the AMT workers. No additional guidance or direction was given to them after the initial walk-through. We will refer to this group as the “expert raters.” Among them, 52.1% of “definitely” votes were cast for human, and accuracy on the longest excerpt length was over 70%.

The human evaluation dataset consisted of 150 excerpts of web text and 50 excerpts each from the three decoding strategies. Each question was shown to at most three raters, leading to 900 total annotations from the untrained workers and 475 from the expert raters. A more detailed breakdown can be found in Table 3.3.

Instructions

In this task you will be shown some text that was extracted from a website.

Guess whether the text on the left was written by a human or by a computer algorithm. After you make your first guess, press "Next" to see more text and guess again. You'll do this for 5 times. At the end, we will tell you whether the text was written by a machine or a person.

Some indicators you can look for that text might be machine-generated:

- Contradicts itself
- Incoherent
- Repetitive

Note that both human- and machine- written text might contain unusual formatting such as text in all capitals, no line-break between a title and article contents, and misaligned quotation symbols. DO NOT base your decision on formatting weirdnesses.

This excerpt was actually a test that you're reading!

Is this surprising?

Yes

No

Figure 3.2: For some of the questions, the text "Dear AMT Worker: to show you're reading, please select definitely [X] for this one." was inserted into the last text segment, and "Did you read carefully?" was appended to the end.

Table 3.3: The number of human annotations collected. In total, there were 50 examples from each sampling strategy and 150 examples of web text. Each example was shown to at most three raters.

| # Annotations | Expert Raters | AMT Workers |
|-----------------|---------------|-------------|
| webtext | 239 | 450 |
| k0-1wordcond | 87 | 150 |
| k40-1wordcond | 75 | 150 |
| p0.96-1wordcond | 74 | 150 |
| total machine | 236 | 450 |

| | |
|---------|---|
| Human | <p>I recently got the chance to try the new Oil Essentials line. With six potent blends to choose from—at \$13 each—these cute little bottles offer a great, affordable way to partake in the skin and hair care oil craze.</p> <p>I tested each product in the line, massaging them onto my face every night before bed and running any leftover oil through my hair to tame frizziness. You could also add a few drops to your bath, favorite moisturizer, or even your shampoo and conditioner.</p> <p>Here's a quick rundown of each oil.</p> <p>Revitalize: Omega 3, 6, 9 & Evening Primrose</p> <p>This was the first one I tried (I went in ROYGBIV order to keep things straight) and my first impression was that it smells lovely but a little strong. The fragrance smells genuinely like flowers.</p> |
| Machine | <p>Red Lanterns, the lead exposure to a movie starring the Batman solo movie alum Margot Robbie taken under Wonder Woman's wing have reignited that rivalry with their whispery premiere. They played it as much as they possibly could, even though people who didn't ever watch Justice League or might have missed it waiting in line for the theater were still talking about as I spilled coffee.</p> <p>The gist? An overextended (OK, a sore) Adam West films set up a Legion of Super-Heroes situation. How aggro? Super laws and paramilitary groups watch over the world's superheroes, which is a mix of that schtick ending, Planet Of The Apes II bit, and the Batman/Venom bit of last appeared in The Seventh Seal when Chris O'Donnell infiltrated one of the teams at some point, also wearing Staff.</p> |
| Machine | <p>He is considered to be the most terrifying man on the planet and people stay away from him. A guy asks him to do something and he says, "My girlfriend's so important to me... I don't need to fight her any more." And then, boom, there's some in a corner crying inappropriately.</p> <p>Men: It's gone in five minutes. Why do I have to be so sad? It's cute," says female member, who asks to remain anonymous. "It's what grew up to drive me crazy when I was a kid, seeing these women become the nurturing, wealthy things they are in this professional world I truly love."</p> <p>And it's nothing to do with her success. These men still actively fear being around the idea of a woman who might win Oscars, make movies or be audacious drivers.</p> |
| Human | <p>Dropbox and Google Drive are very different services that appeal to different users. While Drive is connected to the entire Google Apps (now known as G Suite) ecosystem, Dropbox is a lightweight, simple alternative for file storage. While both are useful, users need to look beyond features, and make sure the service they choose can adequately protect their data. Here's how Dropbox encryption and Google Drive encryption stack up.</p> <p>Dropbox and Google Drive Encryption</p> <p>To their credit, both Dropbox and Google Drive protect user files with encryption. Both also allow users to enable two-step verification, which requires an extra code texted to the user's phone to access the account, making it harder for hackers to access a user's data.</p> |
| Human | <p>EVE Isk Per Hour(Eveiph) is hands down the best tool I've ever used to make isk in New Eden. It is a market helper program that is able to do a great deal of the work that is typically done by a traders spreadsheet. I've used it to go from a 200m/month trading income to 3b/month on my main trading character.</p> <p>Above you can see the blueprint manufacturing page which is located on the first tab of Eveiph. Here you can see the components required to make an item, the settings for the blueprint, and a brief market analysis of what you can expect to make manufacturing the item and selling it at the market you've selected. You can enter the amount of runs you want to make, the ME and PE of your blueprint and click add to shopping list, and it will be added to a list of items to purchase when you are next at a trade hub.</p> |
| Machine | <p>So, not only was the speech a thoroughly mediocre diatribe about what he now thinks we should do for the next 45 minutes, but also how much credit we should give to Mumford and Sons for bringing Obama to the campaign trail. Behold:</p> <p>At the DNC, we drew strength from something even more powerful than the power of words. We drew strength from the power of families in this country. We drew strength from the power of family values. We drew strength from the power of a common purpose--We drew strength from our shared commitment to fighting against everything that undermines our potential in this country and our freedom. It is with that same conviction that we launch this campaign today and we urge every American in America to join us tonight.</p> <p>To allow the same attempt to succeed in this election.</p> |
| Machine | <p>The year is twenty-eight, and the boy is Harry, the sixth year at Hogwarts School of Witchcraft and Wizardry. He can't walk without spells covering his feet (or in his case, his feet are so badly burned that he, for practical purposes, can't even walk for that long without them) and he's just starting to feel more secure about things. This is a pretty dull aspect of the book, I'd say. They probably spent way too much time on the fact that he can't use the stick of silver from his wand, despite his friends bewitching all the knives they had.</p> <p>Harry had been having some difficulty getting to sleep until Hermione pulled him out of his state of near-death-conversation. Thanks to Hermione's meddling, he's gotten some sleep for the past two days. They also learnt a fair amount about getting used to his new surroundings.</p> |
| Machine | <p>Coincidentally, just a few days after the first tweet came out, a fellow named Kevin McReynolds sent out an interview with GQ to promote their upcoming issue. McReynolds describes himself as "a conservative Catholic" who "cannot fathom this guy being a real person and should be ashamed that he was able to be elected president."</p> <p>It's true. If you believe Hillary Clinton gave away 20 percent of the American Uranium to Russia, then you should be ashamed that you voted for Trump. No one should be able to give or receive anything that's not supposed to, so long as they have a warrant. If you've been in a relationship for more than six months with a person who's also convicted of being a felon (or convicted of stealing), that's just stupid, especially as a married man. If you're married to someone convicted of a crime, and they go on their honeymoon with you, that's a felony, not a honeymoon.</p> |
| Human | <p>CHIP DESIGNER Texas Instruments unveiled a family of system on chip (SoC) processors aimed at automakers today, which are designed for use in self-driving cars. Named the TDA2x, the SoC family integrates safety features, such as aiding auto designers to create advanced driver assistance systems (ADAS), which in turn help "reduce the number of collisions on the road and enable autonomous driving experiences".</p> <p>"TDA2x device family combines an optimal mix of high performance, vision analytics, video, graphics and general purpose processing cores in a low power envelope, enabling a broad range of ADAS applications including front camera, surround view and sensor fusion," Texas Instruments said in its release.</p> |
| Machine | <p>Description</p> <p>This classic blend of coffee, cream, and sugar is the perfect drink! It is a smooth and creamy coffee with hints of cream and sweet sugar that can be enjoyed even after a full day of work or playing! The sugar provides a wonderful texture to the coffee beans, so that it can be scooped out into a cup.</p> <p>Available in four flavours: vanilla cream, caramel cream, coffee creme, and chocolate cream.</p> <p>Note: Coffee can be prepared in less than 120 minutes. Note: Serves one.</p> |

Table 3.4: The 10 examples that “expert” raters were guided through before they were asked to perform the detection task. These are hand-selected to showcase the spectrum of generated text and human-written text.

3.2.4 Results

SIMPLE BASELINES Table 3.2 shows the performance of the baseline discriminators on length-192 sequences, as compared with fine-tuned BERT. Reassuringly, BERT far surpasses all simple baselines, indicating that it is not fully possible to solve the detection problem without complex sequence-based understanding. The simplest baseline, TotalProb, which makes a decision based on the likelihood of the sequence, performs surprisingly well (over 60% accuracy for all sampling methods) relative to the methods which involve training logistic regression models.

Logistic regression on bag-of-words is the best of the baselines, beating out the histogram-based methods. While Gehrmann et al. [52] report an AUC of 0.87 on classifying text as real or generated using logistic regression on the four buckets of the GLTR system, we report AUC between 0.52 and 0.56 for this task. The discrepancy is likely due to the fact that the human-written text in our discriminator training set comes from the same distribution as the text used to train the language model, while in GLTR the human text comes from children’s books, scientific abstracts, and newspaper articles. The selection of training data for learned detection systems is crucial. In real-world applications, the choice ought to reflect the genres that builders of text-generation systems are trying to impersonate.

FINE-TUNED BERT In Figure 3.5a, we see that discriminator accuracy as a function of excerpt length and sampling method. As can be intuitively expected, as sequence length increases, so too does accuracy. For unconditioned text decoded with nucleus (p0.96) and untruncated (p1.0) random sampling, we find discriminator accuracy increases from 55%, near random, to about 81% for the longest sequences tested. In contrast, discriminators trained and evaluated on top- k achieve over 80% accuracy even on short 16-token excerpts.

Why are top- k ’s samples so easy to detect? It is because there are only a small number of word sequences that can start a generation when we limit to only ever choosing the 40 most likely tokens at each generation step. In Figure 3.3b, we see the percentage of probability mass concentrated in the k most common token types for

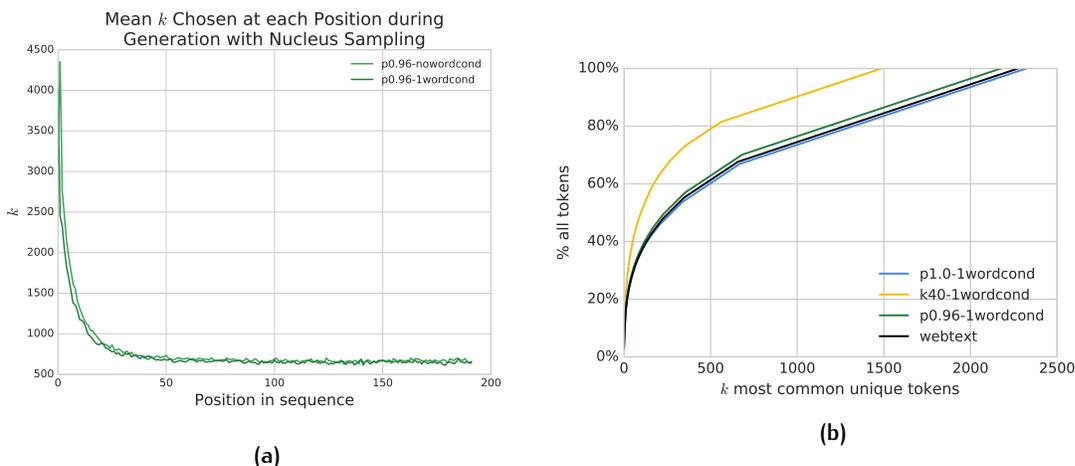


Figure 3.3: In (a), the average (over sequences in the test set) k chosen at each step during generating with nucleus sampling is plotted. Adding a single word of priming strongly impacts the k s chosen for the first few positions, but this difference quickly dissipates. In (b), we consider the first token generated in each sequence by top- k , and plot what fraction of these are captured by the k most common unique tokens from the vocabulary. Overall, at its first step, top- k concentrates 80% of its probability mass in the 500 most common tokens from the vocabulary.

each sampling method. While random sampling and nucleus sampling are very similar to human-written texts, we see top- k concentrating up to 80% of its mass in the first 500 most common tokens. The other sampling methods as well as human-written texts require at least 1,100 token types for the same. It is clear that top- k 's distribution over unigrams strongly diverges from human-written texts—an easy feature for discriminators to exploit. In fact, See et al. [135] note that it takes setting k to 1000 to achieve about the same amount of rare word usage and fraction of non-stopword text as human writing.² This makes it very easy for the model to pick out machine-generated text based on these distributional differences.

Instead of unconditioned generation, which in actuality means conditioning always on the same thing (an empty sequence), we can instead prompt with human-written text that the NLG system then extends. Doing so causes more rare words to be incorporated into the top- k of the unigram distribution. Adding even a single human word of priming significantly reduces the performance of detectors trained with top- k random sampling. Without priming, a discriminator trained on sequences of length 2 can classify with $\sim 90\%$ accuracy

² when decoding from the GPT-2 small model with 117M parameters.

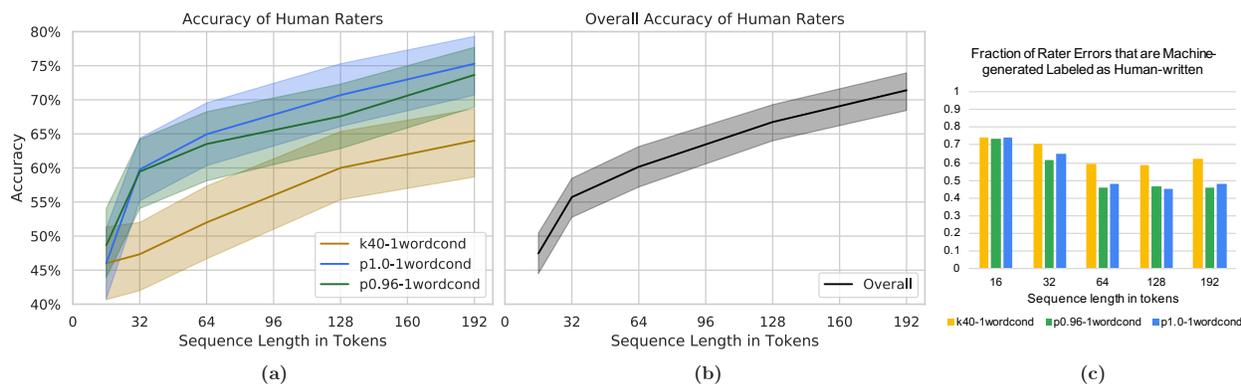


Figure 3.4: (a) and (b) show human rater accuracy of correctly identifying an excerpt as human-written or machine-written, shown with 80% confidence internals, in (a), broken up by decoding strategy and in (b), overall. Accuracy increases as raters observe more tokens. (c) shows that for short excerpts, most rater mistakes are them incorrectly thinking machine-generated text is human written. The two errors types become more balanced at longer lengths.

the provenance of the text (Figure 3.5a). By adding even just a single token prompt, accuracy drops to $\sim 65\%$. Even on the longest 192-length sequences, top- k discriminator accuracy is 6% lower on the primed dataset than the unprimed one.

When generating with nucleus or untruncated random sampling, adding a priming token is not as impactful, as these methods are already sampling from a large fraction (or all) of the probability distribution. This is seen in Figure 3.3a where at the very first step of unprimed generation, nucleus sampling selects from 3075 possible vocabulary words, and at later positions selects from on average more than 500. Untruncated random sampling always selects from the entire 50,000 word vocabulary, whereas top- k only selects from k .

TRANSFERABILITY In Table 3.5, we show how discriminators trained with samples from one decoding strategy can transfer at test time to detecting samples generated using a different decoding strategy. Unsurprisingly a discriminator trained on top- k generalizes poorly to other sampling methods: accuracy drops to as low as 42.5%, *worse than chance*. Conversely, training the discriminator with sequences sampled from the untruncated distribution leads to little transferability to detecting top- k samples. Only the discriminator trained with nucleus sampling (a compromise between unmodified sampling and top- k) was able to detect sequences

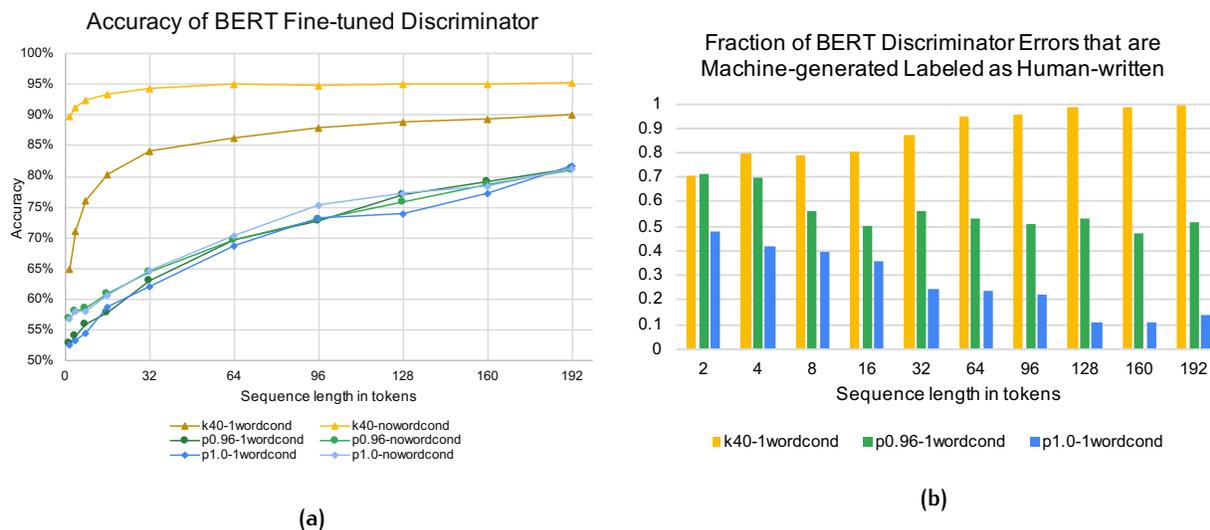


Figure 3.5: In (a), accuracy increases as the length of the sequences used to train the discriminator is increased. In (b), we see that the BERT fine-tuned discriminator predicts about the same number of false-positives as false-negatives when trained with samples generated using top- p sampling. However, for top- k , it more often mistakes machine-generated text to be human-written, while for untruncated random sampling the opposite is the case.

from the other sampling strategies without too much of a hit to accuracy. As expected, a discriminator trained on an equal portion of data from each decoding method does reasonably at detecting all three.

Perhaps this lack of transferability is related to each discriminator’s calibration. Indeed, the degree to which a discriminator’s average prediction deviates from 50% is a direct indicator of its accuracy. In Table 3.6, we observe that of the three BERT discriminators, only that trained on top- p samples predicts ‘machine-generated’ on approximately 50% of in-domain examples as expected. This same discriminator’s behavior holds on datasets generated by other sampling strategies as well. In contrast, we observe that discriminators trained on top- k and untruncated random samples severely underestimate the percentage of machine-generated excerpts in out-of-domain datasets. Even within domain (Figure 3.5b), we find both discriminators heavily favor a single class, increasingly so as the number of tokens increases.

Table 3.5: Accuracy of BERT fine-tuned discriminator when trained on samples from one strategy (rows) and evaluated on another (columns). Trained on samples with 192 tokens. The ‘mixed’ dataset is one containing an equal portion of samples from each strategy.

| | | Eval | | |
|-------|---------|---------------|-------------|-------------|
| | | top- <i>k</i> | nucleus | random |
| Train | top-k | 90.1 | 57.1 | 43.8 |
| | nucleus | 79.1 | 81.3 | 78.4 |
| | random | 47.8 | 63.7 | 81.7 |
| | mixed | 88.7 | 74.2 | 72.2 |

Table 3.6: Average probability of ‘machine-generated’ according to each length-192 discriminator. The expected in-domain probability is 0.5. One token of conditioning.

| | | Eval | | |
|-------|---------|---------------|---------|--------|
| | | top- <i>k</i> | nucleus | random |
| Train | top-k | 60.9 | 27.9 | 14.5 |
| | nucleus | 49.2 | 51.7 | 48.9 |
| | random | 7.3 | 22.6 | 38.3 |

HUMAN ACCURACY Overall human performance across all sampling methods is shown in Figure 3.4b. Even with the multi-paragraph 192-length excerpts, human performance is only at 71.4%, indicating that even trained humans struggle to correctly identify machine-generated text over a quarter a time. However, it is worth noting that our best raters achieved accuracy of 85% or higher, suggesting that it is possible for humans to do very well at this task. Further investigation is needed into how educational background, comfort with English, participation in more extensive training, and other factors can impact rater performance.

To break up the accuracies by sampling method in a way that is comparable to the results shown for the automatic discriminators, we pair each machine-generated example with a randomly selected one of webtext to create a balanced dataset for each sampling strategy. Performance is shown in Figure 3.4a. Top-*k* produces the text that is hardest for raters to correctly distinguish, but as shown in Section 3.2.4, it is the easiest for our automatic detection systems. Samples from untruncated random sampling and nucleus sampling with $p=0.96$ are equivalently difficult for raters to classify as machine-generated. Our human evaluation results suggest

| Truth | Raters | p1.0 | k40 | p0.96 | Truth | Raters | p1.0 | k40 | p0.96 |
|--|--------|------|-----|-------|--|--------|------|-----|-------|
| H | M | H | H | M | H | H | M | M | M |
| <p>EDIT:OKAY!, I guess that'll work for now. >_ http://www.teamfortress.com/ and then go buy the game and experience some of the best online gaming I have ever played. ^_^Both girls had a really fun time and I had a GREAT time making both of these costumes. Everything was altered even a little bit(dying the pants a darker grey and painting the boots and shirts) But my piece de resistance would have to be my eldest's Medi-Gun.If you have any questions about the costumes, I would be happy to assist you!Oh and here's a video of my daughter before the costume was completed.Thanks!</p> | | | | | <p>Image copyright Getty Images Image caption Women mourn over the coffin of one of the victim's of Sunday's bombing in Ankara ¶Who'd be in Turkey's shoes right now? ¶Since July last year, hundreds of soldiers and civilians have been killed in terrorist attacks. Suicide bombs have torn into crowds of demonstrators and tourists. Military convoys have been targeted in the heart of the capital. ¶A long-running Kurdish insurgency, once thought to be close to resolution after years of painstaking efforts to build bridges, has erupted once more. ¶The country is awash with Syrian and other refugees. The government has been under pressure to stop them moving on into Europe and prevent would-be jihadis travelling the other way. ¶How dangerous is Turkey's unrest? ¶Tears and destruction amid PKK crackdown ¶Turkey v Islamic State v the Kurds</p> | | | | |
| Truth | Raters | p1.0 | k40 | p0.96 | Truth | Raters | p1.0 | k40 | p0.96 |
| M | M | H | - | - | M | M | - | - | H |
| <p>First off, this thread has done a pretty good job of describing in detail yet another broken touchscreen. That's the difference between a smartphone and a PC with no prying eyes having to snap shots for the police to find. ¶What I would like to address is the mindset that generally surrounds Chrome OS users. To me this is analogous to saying that Apple does "hate their Windows", or that HP does "hate their Macs" as if http://twitter.com/ (and that quote is from two years ago), that anyone who covers smartphones and tablets from a "PC" perspective is just jealous. ¶Chrome OS is for browsing the web, PC processors can do stronger things in that regard, Windows is a juggernaut on those fronts. This is how I see it. Yes, it can be slow. And yes, you need a fast CPU</p> | | | | | <p>FOR ALABAMA, GOOD WEEKS ¶AND A TOUR OF CAIRO ¶THE ALABAMA COMMITTEE ON THE STUDY OF THE AMERICAN SECURITY AGENDA. ¶America's future has been mapped out in carved stone. Metro Atlanta's last US congressman, Bill Posey, was a inextricable integral element of the Citadel project as it became another metaphor for Atlanta's transformation from an industry backwater into the finance and information hub of the nation's capital. Meanwhile, Cobb County – Atlanta's geode of change – is home to some of the largest industrial parks in the South, a regional cultural center, a 100-year-old manufacturing town and a potent symbol of the former city's cherished Georgian past. The gentry still live there, the defunct industrial landscapes carry the names of</p> | | | | |
| Truth | Raters | p1.0 | k40 | p0.96 | Truth | Raters | p1.0 | k40 | p0.96 |
| M | H | - | - | M | M | H | - | M | - |
| <p>Exidentia at Eurnari, is an upcoming Cryptopia event which is currently still in development. Be a part of the first live stream of this year's event on 15-16 January 2016! ¶Since the release of v1.22, Exidentia has received a fair amount of user feedback. This event takes place in the underwater Cryptopia they have built. During this event, you will learn about the ocean and areas around it, and be reached by a treasure hunter that helps you explore the different areas. ¶There will be six different levels in this event that you will become acquainted with: thought Polar Lava, Ocean Seared Cones and Celestine Floors, Sea Damaged Aerie Bricks, coast Puddle (congipit stopping at red water), Shaikh Swamp and Bugmite. At rotating points, you will learn how to access various types of creatures</p> | | | | | <p>Ever since the opening of the North American College of Art Education in 1990, the demand for art education in America has grown steadily, and in recent years we have seen the rise of students that pursue art education not in the classroom but at art academies. This year saw another 50 percent increase in the number of art academies in the United States offering courses – with an additional 10 percent of students in 2017 taking art. ¶Some major changes have occurred in recent years with regard to the art curriculum and the way students learn, and we will explore each of these in coming months as we look at the various forms of art education. There is no one-size-fits-all approach for this or any other field of study, and students who begin a course in art education may change their plans based on what they see that course, including what lessons they have completed and the resources available, to create meaningful experiences of artistic creation. ¶One important area</p> | | | | |

Table 3.7: Some 192-token examples where at least two expert raters agreed with each other, but were not in agreement with the automatic discriminators. The first row shows examples where the ground-truth was human-written, the second shows machine-generated examples where the corresponding discriminator guessed incorrectly, and the third shows machine-generated examples where the discriminator was correct, but raters got it wrong.

that much lower p -values than the 0.92 to 0.98 range proposed in Zellers et al. [174] might be necessary in order to generate text that is considered significantly more human-like to human raters than the text produced by using the untruncated distribution.

Table 3.7 gives several examples where human raters and our BERT-based discriminators disagreed. When raters incorrectly labeled human-written text as machine-generated, often the excerpts contained formatting failures introduced when the HTML was stripped out. In the middle two examples, topic drift and falsehoods such as Atlanta being the “information hub of the nation’s capital” allowed humans to correctly detect the generated content. However, in the bottom two examples, the high level of fluency left human raters fooled.

Overall we find that human raters—even “expert” trained ones—have consistently worse accuracy than automatic discriminators for all decoding methods and excerpt lengths. In our experiments, randomly-selected pairs of raters agree with each other on a mere 59% of excerpts on average. (In comparison, raters and discriminators agree on 61% to 70% of excerpts depending on the discriminator considered). We surmise that the gap between human and machine performance will only grow as researchers inevitably train bigger, better detection models on larger amounts of training data. While improved detection models are inevitable, it is unclear how to go about improving human performance. GLTR proposes providing visual aids to humans to improve their performance at detecting generated-text, but it is unlikely that their histogram-based color-coding will continue to be effective as generative methods get better at producing high-quality text that lacks statistical anomalies.

3.3 ROFT: A LARGESCALE STUDY OF HUMAN DETECTION ABILITY

Our pilot study in Section 3.2 showed how choice of decoding strategy impacts human ability to detect machine-generated text. However, there are many other factors which influence detectability that we were not able to include in this study, including the domain of the text being used for evaluation and the architecture and manner in which the underlying language model was trained. In addition, we were interested in studying

the annoators themselves—how do annotator background as well as the incentive structure set up for soliciting annotations impact performance on the detection task? We therefore saw the necessity of designing a platform for conducting large-scale studies of the detection task.

Previous studies, including our pilot, focused on the binary classification task—given a text example that is either entirely human-written or entirely machine-generated (aside from an initial prompt), annotators must predict whether it is human-written or machine-generated. For example, Clark et al. [29] demonstrated that annotators are able to distinguish GPT-2 XL generations with at best 62% accuracy, but they perform no better than random chance on GPT-3 outputs [20]. Even after training evaluators to improve their detection abilities, detection accuracy on GPT-3 was only able to converge to around 55%. An older study by Brown et al. [20] reported similarly low performance (52%) on the detection of machine-generated news articles.

For our large-scale study, we instead framed detection as a boundary-detection task: given a document that starts off as human-written and at some point transitions to machine-generated, can annotators detect the transition point? The boundary detection setting is more informative than the classification setting because it better aligns with how LMs are used to generate text in practice; in typical usage, a generative system is provided with a prompt and asked to produce a continuation. By measuring human skill at the boundary detection task, we were able to evaluate the relative performance of different generative systems, build a better understanding of how incentive structure influences the quality of the annotations acquired, and make progress toward quantifying the risks associated with large language model goals.. Furthermore, because the annotation platform we built was public, we could achieve these research goals while simultaneously educating the public about how to spot generated text.

In total, we collected over 20,000 annotations with the goal of answering the following research questions:

- How do model size, decoding strategy, and prompt genre impact dectability?
- What kinds of errors and textual properties do humans associate with machine-generated text?
- Do annotators who take longer per annotation or spend more time on the task do better?
- Are their external factors (such as knowledge of NLG) which make some annotators better at the task?

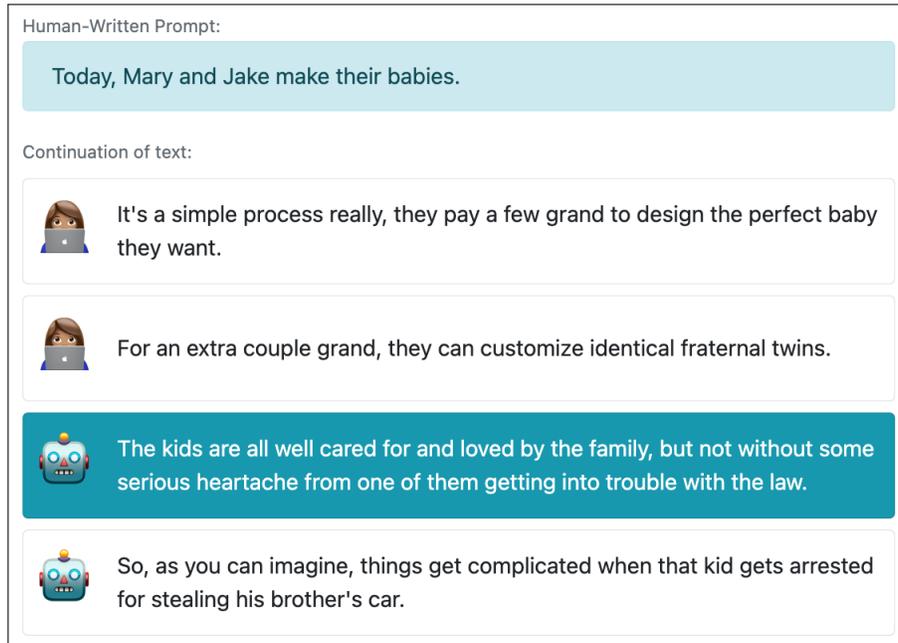


Figure 3.6: In the boundary detection task, players see one sentence at a time and try to guess when they transition from human-written to machine-generated.

3.3.1 The Real or Fake Text Game

Our study uses data collected through the “Real or Fake Text” (ROFT) annotation platform [40]. ROFT is a turn-based game where a player first selects a domain of text (news articles, recipes, short stories, or speeches). The player then plays a series of game rounds. Figure 3.6 shows a screenshot from a game round. In each round, the player is shown a starting sentence which they are told comes from a real human-written document. They are then shown subsequent sentences, one at a time. Each subsequent sentence may be the true continuation of the document, or it may be text generated by a language model. Once the sentences transition to being machine-generated, they will stay so for the rest of the 10-sentence passage.

After being shown each sentence, the player must guess whether that sentence was machine-generated or human-written. If the user selects “human-written,” another sentence is displayed. If the player deems the current sentence to be written by a machine, the game round ends and the true author (machine or human)

for each sentence is revealed, potentially allowing the player to improve their intuitions over time. Before submitting their selection, the player is able to select a reason to explain their choice of sentence. They may select from a pre-defined set of reasons (Table 3.14) or else write a custom reason. Thus, the player’s goal in ROFT is to correctly identify the sentence at which a passage transitions from being human written to being generated by a language model. This setting is considerably more realistic than prior work, since in the real world, generating with a prompt is the standard way to achieve controllability, and malicious actors will not reveal what portion of a generation is the human-written prompt.

In total we collected over 20,000 annotations. We found that players vary substantially in their detection ability, and that factors such as the amount of time taken to complete a game round and total number of game rounds played sometimes correlate with success. Furthermore, we examine some of the trends and errors which distinguish real from generated text and look at whether annotators could pick up on these trends. Finally, we discuss the difficulty in incentivizing players to improve in their ability over time.

3.3.2 Experimental Design

Datasets

In order to answer questions of how textual genre and writing style affect detectability of machine-generated text, we selected four diverse categories of prompts. For each category, documents were sentence-segmented, and only documents with 11 or more sentences were retained. For each document, the first h sentences were used as the prompt, where h is a uniform random number between 1 and 10 (inclusive). The remaining $10 - h$ sentences of each 10-sentence game round were a machine-generated continuation. Our four genres of prompts are as follows:

NEWS ARTICLES. Documents were drawn from the New York Times Annotated Corpus [130], which contains 1.8 million articles published by the Times between 1987 and 2007. Our hypothesis was that this

domain would be challenging for models since news requires factual accuracy, which state of the art models have been shown to struggle with [106, 93].

PRESIDENTIAL SPEECHES. Documents were drawn from the presidential speech corpus [16], which contains 963 speeches given by presidents of the United States, with dates ranging from 1789 to 2015. Our hypothesis was that the sort of first-person rhetoric found in these speeches would be easy for models to impersonate since political speech and first-person speech are plentiful in web-based training data.

STORIES. Fictional stories were selected from the Reddit Writing Prompts dataset [42], a corpus of amateur short stories scraped from the r/WritingPrompts sub-Reddit. We hypothesized that this domain would be challenging for players since the writing quality of the stories is not especially high (which lowers the bar for the model generation quality), and factuality is not as important in a fictional domain.

RECIPES. Recipes were extracted from the Recipe1M+ dataset [99]. Recipes were parsed slightly differently than the other domains. We set the “first sentence” of each document as the name of the recipe and the ingredient list, and each subsequent “sentence” was a step in the recipe. Some recipe steps were more than one sentence. We hypothesized that this dataset would be difficult for models due to the closed-ended nature of the task and the reliance on common sense.

Awarding Points

In each game round, the player is awarded points based on how close their selection was to the true boundary³. Players were awarded 5 points for correctly choosing the boundary sentence and $\max(5 - n, 0)$ points for a guess n sentences after the boundary. Players were not awarded points for guessing a sentence before the boundary. Players were able to see how many points they earned in each category on their profile page and compare their performance with fellow players on the leaderboard page. In the Findings section (Section

³ For our purposes, the “boundary” sentence is considered to be the first machine-generated sentence in the passage

3.3.3), we report mean score earned as the predominant evaluation metric. Table 3.10 shows the correlation between mean score and other sensible metrics. We see that mean score is strongly positively correlated with both perfect guess accuracy and correct side of boundary. Mean score is only weakly correlated with distance after boundary due to the harsh scaling of points; only guesses within five sentences to the right of the boundary receive any points. While imperfect, this harsh scaling is by design, as without it later sentences will give significantly more points in expectation.

Player Recruitment and Annotation Filtering

Players were recruited from two sections of an Artificial Intelligence course for Master’s students and senior undergraduates at the University of Pennsylvania. We only analyze fully anonymized data from students who consented to having their annotations used for research purposes.

The first section (Group A) was asked to play 30 minutes of the ROFT annotation game for a fixed amount of points of class credit. Students in this section were not given any instructions beyond how to create an account. The second section (Group B) was explicitly told they would be awarded to 2 points of extra credit toward their final grade. The amount awarded was $\min(2p/250, 2)$ where p was the number of points the student earned on the ROFT leaderboard. Students in Group B were given detailed instructions and examples of signs to look out for that text was machine-generated. Table 3.9 gives statistics on the annotations collected from each class.

We note that university students taking an advanced artificial intelligence course are not reflective of the global population of English speakers, and the results presented in this paper may not reflect the general population’s ability to detect machine-generated text.

In total, we collected 42,165 annotations over 7,895 different game rounds. The annotations were then filtered in the following ways. If a player guessed the same boundary position for a series of 5 or more rounds in a row, we removed all the annotations in the series because the player was likely no longer actually playing the game as designed. We also removed annotations from the two players cheated by exploiting Javascript vulnerabilities. Finally, for the recipes genre, a bug during dataset curation resulted in an over-representation

Table 3.8: Statistics on the annotation tasks (game rounds) available in our system. The second column shows the number of game rounds available for each system. The discrepancies in number of annotations per dataset is partially due to the fact that players were able to choose which domain they performed annotations in.

| Genre | # Rounds | # Annotations | | Avg Ann/Gen | Systems | Decoding Strategies |
|----------|----------|---------------|-------|-------------|---------|---------------------|
| | | Raw | Final | | | |
| News | 1,838 | 7,806 | 4,488 | 2.97 | | |
| Stories | 9,864 | 8,007 | 4,614 | 2.53 | | |
| Recipes | 7,258 | 17,978 | 7,709 | 2.13 | | |
| Speeches | 297 | 8,374 | 4,835 | 16.28 | | |

Table 3.9: Statistics on the students who were invited to complete annotations on ROFT. “Avg Ann / Part” is the average number of annotations per participating student, while “Avg Score / Part” is the average score. “Avg Time” is the average time it took a participant to read one sentence. Standard error is shown.

| Class | # | | Avg Annotations / Participant | Avg Score / Participant | Avg Time / Annotation (s) |
|---------|----------------|---------------|-------------------------------|-------------------------|---------------------------|
| | # Participants | # Annotations | | | |
| Group A | 141 | 6,527 | 46 | 1.966 | 5.651 |
| Group B | 102 | 15,119 | 148 | 2.134 | 6.443 |
| Overall | 241 | 21,646 | 90 | 2.083 | 6.338 |

of “all-human” game rounds played; for better balance during analysis, we randomly removed a portion of these annotations. Our final filtered dataset consisted of 21,646 annotations over 7,257 game rounds. For News, Stories, and Recipes, we had on average over 2 annotations per game round, while for Speeches, a smaller dataset, we had on average 16. Table 3.8 gives a detailed breakdown of the dataset across genres and generation systems.

Continuation Sources

In order to answer questions related to how model attributes affect generated text we employed different methods of text generation for each category. For Recipes, New York Times, and Stories, we generated continuations with GPT-2 XL using nucleus sampling [holtzmanetal2020] with $p = 0.4$ and a repetition

| Metric | ρ |
|------------------------------|--------|
| (a) Correct side of boundary | 0.74 |
| (b) Perfect guess | 0.88 |
| (c) Distance after boundary | 0.31 |

Table 3.10: Average points earned is the main metric reported in the Results section. This table shows the Spearman’s rank correlation between average points per user and several other possible metrics: (a) the fraction of times the user correctly guessed on or after the boundary; (b) the fraction of times the user guessed exactly on the boundary; and (c) the average number of sentences after the boundary of the user’s guess (giving new score for guesses before the boundary).

penalty of 1.2 [77]. For Recipes, we additionally generated continuations with a GPT-2 XL model finetuned on recipes.

For New York Times and Stories, we experimented with varying the p used for decoding, testing out $p = 0.0$ (argmax) and $p = 1.0$ (sampling directly from the model’s predicted distribution). As an additional sanity check on annotator skill, we also included 100 game rounds in the News domain where instead of transitioning to an LM-generated continuation, the passage transitioned to a completely different news article selected at random. We expected these game rounds to be trivial for players.

For Stories, we experimented with different model sizes, generating continuations with both GPT-2 Small (117M Parameters) and GPT-2 XL (1.5B Parameters). Lastly, for Presidential Speeches, we generated continuations using the CTRL model [77] rather than GPT-2. CTRL has the option to specify a control code indicating what domain to generate text in. For half of the generations, we used the “[Politics]” control code while for the other half we randomly selected a control code each time. We expected use of the politics control code to result in generations which more on topic.

Table 3.8 gives the statistics of the game rounds included in ROFT. Overall, {**TODO:**}% of game rounds were exclusively human-written.

3.3.3 Results

The collected annotations allow us to investigate several questions. Error bars on all figures and tables are 95% confidence intervals.

Can humans detect generated text?

Players correctly guessed exactly on the boundary sentence 23.4% of the time. For game rounds which contained at least one generated sentence, players were able to eventually guess machine-generated 72.3% of the time, even if they missed the exact boundary. Players incorrectly identified 61.3% of all-human game rounds as containing machine-generated text.

The average number of points (§3.3.2) received per round by our players was 2.08, well above random chance. For comparison, if a player uniform randomly guessed every round, their expected per-round score would be 1.31, and if they always guessed the last sentence, their expected per-round score would be 1.5.⁴ For the remaining analyses, we will use average points earned as the primary measure of detection ability. This measure correlates with other possible metrics (Table 3.10).

Out of the 214 annotations we collected on the “sanity check” game rounds, the mean score was 2.75, significantly higher than any of the true LM-backed systems. Also, for these annotations, the error type “irrelevant” was selected about twice as often as all other error types combined, validating that players were paying attention to the task at hand.

How much does player ability vary?

There was a large variance in the skill of individual players. Out of the 116 players who completed 50 game rounds, 19 earned a total score of 70 or fewer points (one std below the mean score) in their first 50 rounds,

⁴ These expectations assume that the true boundary position is equally likely to be at any position. Figure 3.7 shows the true distribution of boundaries, which was not quite uniform.

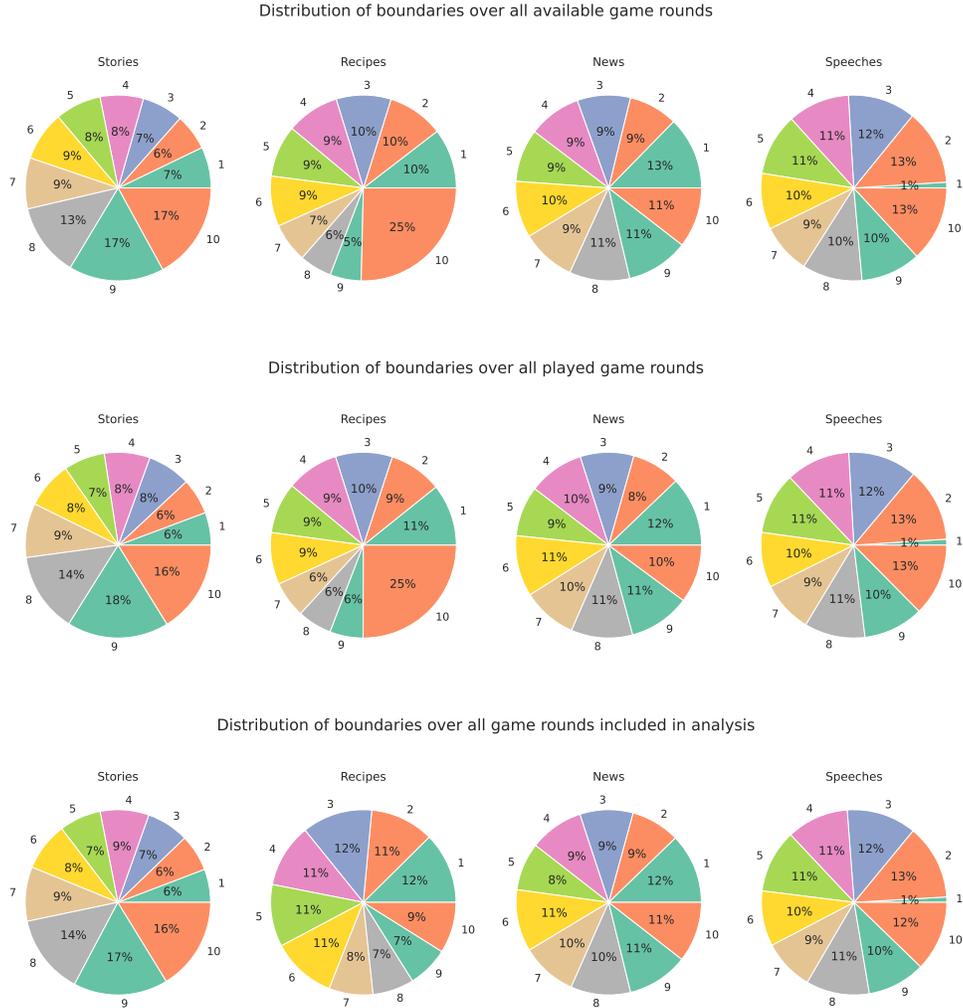


Figure 3.7: The distribution of boundary sentence positions over all game rounds available on RoFt (top), all game rounds that received annotations (middle), and all game rounds included in this paper’s analysis after filtering out problematic annotations (bottom).

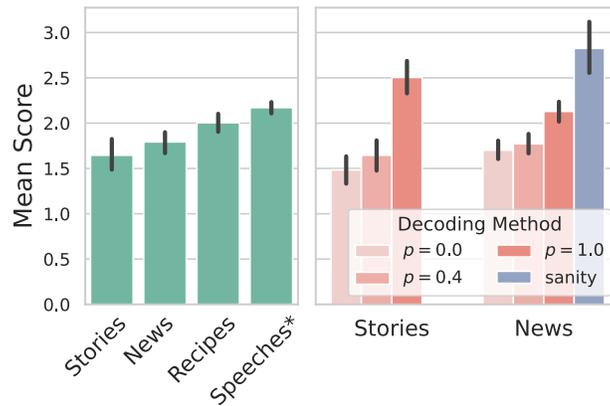


Figure 3.8: (left) Comparison of mean player score across different genres with GPT-2 XL $p=0.4$ (and CTRL $p=0.4$ for speeches). (right) Comparison of mean player score across different values of p for nucleus sampling (GPT-2 XL), as well as a “sanity-check” baseline.

while 15 earned a total score 127 or greater points (one standard deviation above the mean score). Four of these raters scored two standard deviations above the mean score.

We also found that under the right conditions, players can improve over time. There was no correlation between number of rounds played and player score for Group A. However, Group B, who were given extra credit proportional to their game score, did show slight improvement (Table 3.12). There was also lower variance in points earned among students in Group B (Figure 3.10), possibly because they were more incentivized to do well at the task.

We can also measure inter-annotator agreement with the Krippendorff’s alpha co-efficient. This statistic measures how much disagreement there is between players compared to the amount of disagreement one would expect by chance. Two players are considered to have agreed if they both guessed “machine-generated” on any sentence after the true boundary or if they both guessed the entire passage was human-written. Over all annotations, we found $\alpha=-0.25$, indicating there was less agreement than could be expected from random guessing, suggesting different annotators were better at identifying different kinds of problems with LM-generated text. However, among our top 25% of players (measured by mean score), there was high inter-annotator agreement, with $\alpha=0.44$, suggesting that good annotators made similar errors.

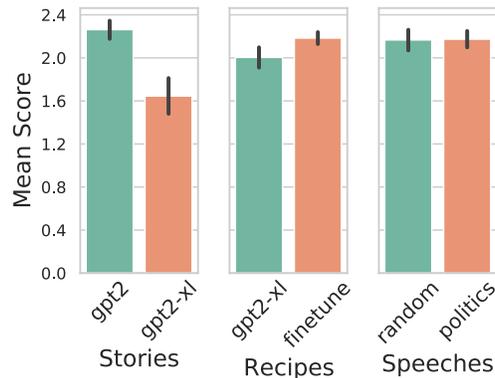


Figure 3.9: **(left)** For Stories, as model size increases (using $p=0.4$), detection becomes harder. **(middle)** For Recipes, extra finetuning does not significantly impact detectability. **(right)** For Speeches, using a “[Politics]” control code (with the CTRL model) has no impact on detectability compared to using a random control code.

3.3.4 Analysis

Are some genres easier to detect?

We found that generated text was easier to identify in the recipes and speeches genres than in the stories and news genres. Figure 3.8 (left) shows the average points received on each genre for game rounds that used comparable LMs, while Table 3.11 gives a more detailed breakdown across models.

For recipes, we expect that the task was made easier by the fact that the first human-written “sentence” in each game round was a semi-structured ingredients list, making it easy for players to check for contradictions—a step saying to mix in cream is probably generated if there is no cream ingredient. In addition, recipes often assume implicit unwritten knowledge, which language models struggle to get right—a step saying to crack eggs into a bowl must precede a step saying to whisk the eggs. Indeed, if we look at the reasons given by our players for saying “machine-generated,” recipes had a much larger percentage of “common_sense” errors (26%) than did either News (10%) or Stories (10%). It is worth noting that this result slightly contradicts the one reported by Clark et al. [29] who reported that generated recipes were more difficult to detect than news or stories; more targeted research is necessary to fully understand the relationship between domain and generation performance.

| Dataset | p | n | Mean Score |
|-----------|-----|-------|-------------|
| News | 0.4 | 1,197 | 1.793±0.109 |
| Stories | 0.4 | 468 | 1.645±0.168 |
| Speeches* | 0.4 | 4,252 | 2.171±0.062 |
| Recipes | 0.4 | 1,811 | 2.004±0.098 |

Table 3.11: The mean scores for each domain on annotations involving XL-sized models for $p=0.4$. Asterisk denotes generation by CTRL. Interval is $\alpha = 0.95$ confidence.

We believe the speech genre was easier for players not because speeches are intrinsically more difficult to generate but because we struggled to get the CTRL model to produce high-quality, non-repetitive generations, even though it is about the same size model as GPT-2 XL. It was necessary to incorporate repetition penalties during generation with CTRL, which helped but did not solve the quality issues.

Does model size make a difference?

Previous work has shown that language model performance scales with number of parameters [75], so we expected players to be worse at detecting generations from larger models. Indeed, we found that players scored significantly higher when generations came from GPT-2 small (117M parameters) than when they came from GPT-2 XL (1.5B parameters) (Figure 3.9.).

Are diverse generations easier to detect?

Choice of decoding strategy is known to have significant impact on text quality [177] and detectability [65]. Choosing a lower value of p when generating with a nucleus sampling [61] decoding strategy produces less diverse but also less noisy text than choosing a higher value of p . In our experiments, we did not find statistically significant differences in player skill between $p=0.0$ (greedy) and $p=0.4$ sampling (Figure 3.8). However, players were significantly better at $p=1$ (pure random sampling) than the lower values, validating claims from earlier papers that LMs struggle to generate high-quality text with similar diversity to human-written text. Interestingly, generations from GPT-XL using $p=1.0$ were easier for players to detect than generations from

GPT-2 small using $p=0.4$. This highlights the importance of decoding, as improper selection of decoding strategy may cause a language model to perform worse than one that is one tenth its size.

Do control codes affect detectability?

CTRL is a 1.6B parameter LM trained with controllability in mind. At inference time, one can pass in a control code, such as “[Politics]” or “[Horror]” to include the style of the generated text. We investigated the efficacy of these control codes on the genre of presidential speeches by using “[Politics]” for half the generations and randomly selecting control codes for the remaining half. We found that use of the politics control code did not significantly affect players’ ability to distinguish real from fake text. This is not to say that control codes do not affect generation; however, it does suggest that the cues used by players to detect generations may not be related to genre-specific details, as least not within the genre of political speeches. Further work is needed to investigate whether control codes could have influenced detectability in other genres.

Does finetuning affect detectability?

We had expected that finetuning on in-domain text would result in a model that was better able to fool humans. Counter to expectations, there was a small increase in player detection ability when generations came from GPT-2 finetuned on recipes compared with generations from pre-trained GPT-2. This is despite the fact that the finetuned model had close to half the perplexity of the pre-trained model on a held out test set of 50,000 recipes (4.781 vs. 8.979). While we can only speculate as to the amount of recipe knowledge present in the pre-trained model (GPT-2’s training data is not publicly available), it is possible the pre-trained model already contained enough understanding of recipe-like text that it was not critical to do the extra-finetuning. Perhaps finetuning would have had more impact in a specialized or jargon-laden domain (e.g. legal, medical).

How much time did game rounds take?

To understand how much time game rounds took, we logged how many seconds players spent on each sentence decision. We controlled for instances of players leaving a game open mid-annotation by applying $\min(120, t)$

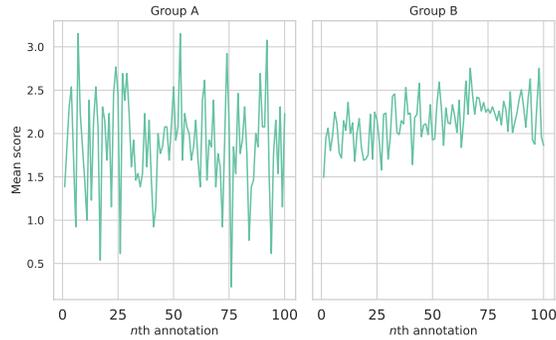


Figure 3.10: Performance over time for the two player groups (§3.3.2). Players in Group B, who were given extra instruction and incentives, improved over time while those in Group A did not.

| Group | k | n | Spearman ρ |
|-------|-----|-----|-----------------|
| A | 50 | 22 | -0.03 |
| A | 100 | 13 | -0.06 |
| B | 50 | 88 | 0.29 |
| B | 100 | 81 | 0.42 |

Table 3.12: The Spearman’s rank correlation coefficient between the number of annotations performed before the current annotation and the score on the current annotation, for all n players who have performed k or more annotations. Players in Group B, who were given extra instruction and incentives, improved over time while those in Group A did not.

to all recorded times t . We computed total time per annotation by summing the times for each sentence-level decision. We found players took longer on annotations where they ended up receiving more points, and players gradually got faster over time (Figure 3.13). While one might expect longer sentences to take more time to read and make decisions on, we found no correlation between time taken and length of sentence ($\rho=-0.10$), indicating that players take time to think about the task beyond just reading the sentence.

What sentence-level features could be used to detect generated text?

It has been well-studied how generated text differs in basic, measurable ways from human-written text, often due to the choice of decoding strategy. In particular, we measured how sentence length, part-of-speech

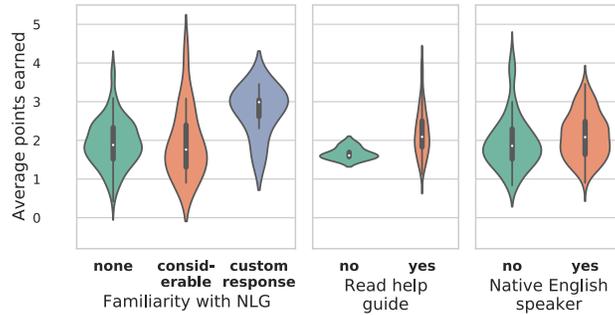


Figure 3.11: Violin plots showing results of our mandatory exit survey. A violin plot is a box plot that also provides a density estimation. Results shown are filtered to only include players who did at least 20 rounds. We see that reading the help guide, being a native English speaker, and providing a custom response for your familiarity with NLG all contribute to a higher mean score while high domain expertise does not seem to have an effect (except in the case of short stories, where variance is lower for domain experts).

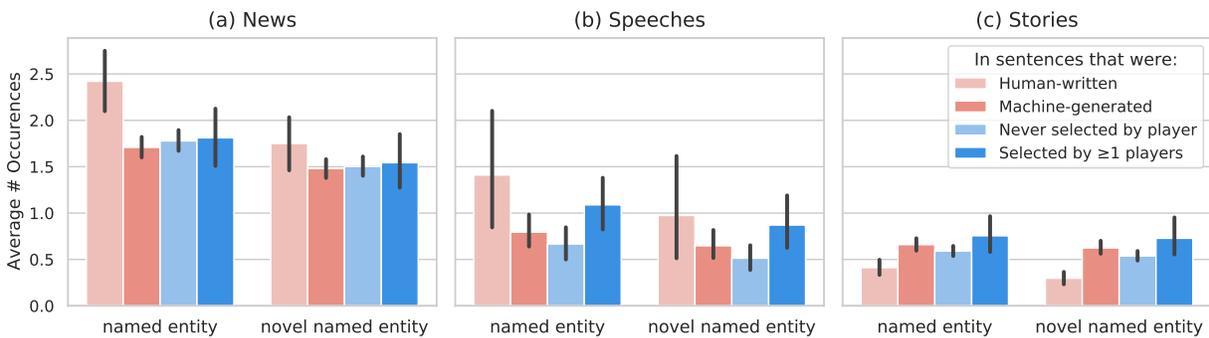


Figure 3.12: We see that human sentences tended to have a different number of named entities than generated sentences. Players picked up on the correct trend in Stories, but not in News or Speeches.

distribution, and presence of named entities and novel words differed between the generated and human-written sentences in our dataset, and whether players were able to pick up on these differences. Figure 3.12 shows the results for named entities, where novel named entities are ones which occurred in the current sentence but not in any previous sentences. We found surprisingly different trends across different genres. On News and Recipes, the generated sentences tended to have fewer named entities than in human-written sentences. Annotators did not pick up on these trends, though they may have picked up on the fact that for Stories, the generated sentences tended to have slightly more named entities.

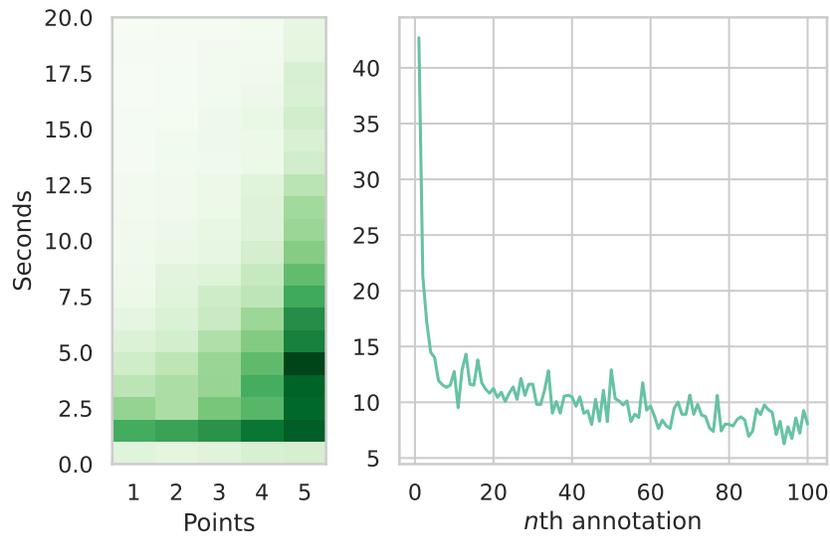


Figure 3.13: **(left)** Histogram showing the relationship between points earned and the number of seconds an annotation took. Annotators tended to earn more points on annotations they spent longer on. **(right)** Among players who completed at least 100 annotations, average annotation speed decreased with increased experience at the task.

In News and Speeches, machine-generated sentences tended to be shorter than human-written ones, a trend players did not pick up on. However, for Stories, the generated sentences were on average longer than the human ones, and annotators tended to select longer sentences as the boundary. Additionally in Stories, generated sentences has on average a greater proportion of adjectives and adverbs, but annotators did not pick up on this trend.

Does familiarity affect detectability?

All participating players filled out an exit survey after completing their annotations. The questions on this survey are in Table 3.13. Figure 3.11 shows some of the results. First, there was not much difference in performance between participants who reported they had never heard of GPT-2/3 and those who reported having considerable familiarity with them. Interestingly, participants who answered “other” and wrote custom responses did end up being better at the task. (For example, we released the extra credit assignment a week

after a prominent NLG researcher gave a colloquium talk, and a couple responses we received references hearing about them in her talk.) Second, participants who admitted that they did not read the help guide tended to perform poorly; all the best players did read the guide. Third, there was not much difference in ability between native and non-native English speakers. The very strongest players were not native English speakers. Finally, we did not observe any correlation between self-reported familiarity with a given genre and detection skill on that genre.

What are the most reliable errors to look for when detecting generated text?

Each time a player specified a sentence was machine-generated, they had the option to specify why they made this decision, selecting from a set of pre-defined options (Table 3.14) or else writing down a custom reason. Table 3.15 shows for each reason, the average number of points earned when that reason was specified. Like Clark et al. [29], we see that conditioning on bad grammar is by far the least reliable way to detect generated text. In addition, we see that over 30% of all reasons given for thinking generated text was generated was because the text was “irrelevant or unrelated to the previous sentences.” This result stayed consistent across all models and domains. We note that the three most reliable reasons given (“common_sense,” “irrelevant,” and “contradicts_sentence”) were also the three most common, indicating that improving these attributes will lead to the biggest improvements in generation performance.

Figure 3.14 shows the full text of the reasons players could choose between for their boundary decisions, as well as some “other” responses we received. Figure 3.14 shows a more detailed breakdown of the percentage of errors made by different models. We see that using $p = 1.0$ results in a higher percentage of “irrelevant” errors (36%) than $p = 0.0$ (31%) and $p = 0.4$ (28%) while models decoded using $p = 0.0$ in turn have a higher percentage of “generic” errors. We also see that smaller models tend to make more “irrelevant” errors than larger models (39% vs. 28%). More research is necessary to understand not only the distribution of the types of errors made by certain generative models but also the ways in which that distribution changes given factors such as domain, model size, and decoding strategy.

| Question | Response Type |
|---|---|
| What did you (or what are you planning to) major/minor in? | Free Text |
| Are you a native English speaker? | Yes/No |
| How often do you consult a recipe when preparing food? | Daily (5) Once to a few times per week (4) Once to a few times per month (3) Once to a few times per year (2) Never (1) |
| How often do you read news from credible news publishers (Wall Street Journal, New York Times, etc.)? | Daily (5) Once to a few times per week (4) Once to a few times per month (3) Once to a few times per year (2) Never (1) |
| How often do you read fiction on the internet (fan fiction, creative writing sub-reddits, ebooks, etc.)? | Daily (5) Once to a few times per week (4) Once to a few times per month (3) Once to a few times per year (2) Never (1) |
| What is your familiarity with GPT-2 and GPT-3? | I've used them before (OpenAI API, HuggingFace, etc.) (4) I've been excitedly following them. (3) I've read about them in the news or a blog post. (2) I've never heard of them. (1) |
| Did you read the RoFT Guide before you tried the game? | Yes/No |
| Do you agree for the data being collected on this form along with any annotations you make to be used in an anonymized, aggregated way for research on students' ability to detect machine-generated text? Your answer on this question will not affect your grade. | Yes/No |

Table 3.13: The text of the exit survey questions given to players after completing their annotations

| Reason | Description |
|-----------------------|---|
| grammar | is not grammatical |
| repetition | substantially repeats previous text or itself |
| irrelevant | is irrelevant or unrelated to the previous sentences |
| contradicts_sentence | contradicts the previous sentences |
| contradicts_knowledge | contradicts your understanding of the people, events, or concepts involved |
| common_sense | contains common-sense or basic logical errors |
| coreference | mixes up characters' names or other attributes |
| generic | contains language that is generic or uninteresting |
| other | <ul style="list-style-type: none"> ▷ Bacon is not sauted ▷ Mr. vs President Clinton ▷ navel and sternum seem like very unusual word choices ▷ It's unlikely that President Nixon will be quoting a one-month old report when he talks about progress made to date ▷ lemon, zest of some things dont sound right? 34 cups of splenda and 14 cups of vinegar? ▷ doesn't rhyme like rest ▷ Grammar substantially improves from the previous sentences |

Table 3.14: (top) The possible reasons players could select for why text was machine generated, and (bottom) several examples of custom reasons players wrote.

3.3.5 Discussion

In the ROFT user study, we demonstrated the viability of the boundary detection task as a framework for soliciting human evaluation of natural-language generation systems. We conducted the largest study of generated text detectability to date and, in the process, replicated many previous major results in the field, such as the improved performance of bigger models [75] and the difficulty in incentivizing annotators to improve over time [29]. We confirmed the result from Section 3.2 showing that less random decoding strategy settings result in generated text that is harder for humans to detect. While this trend was true in both domains we tested it in, the impact of decoding strategy was more stark in the Stories domain than the News domain. In addition, we have provided new insights into the ways in which humans interact with partially-generated text.

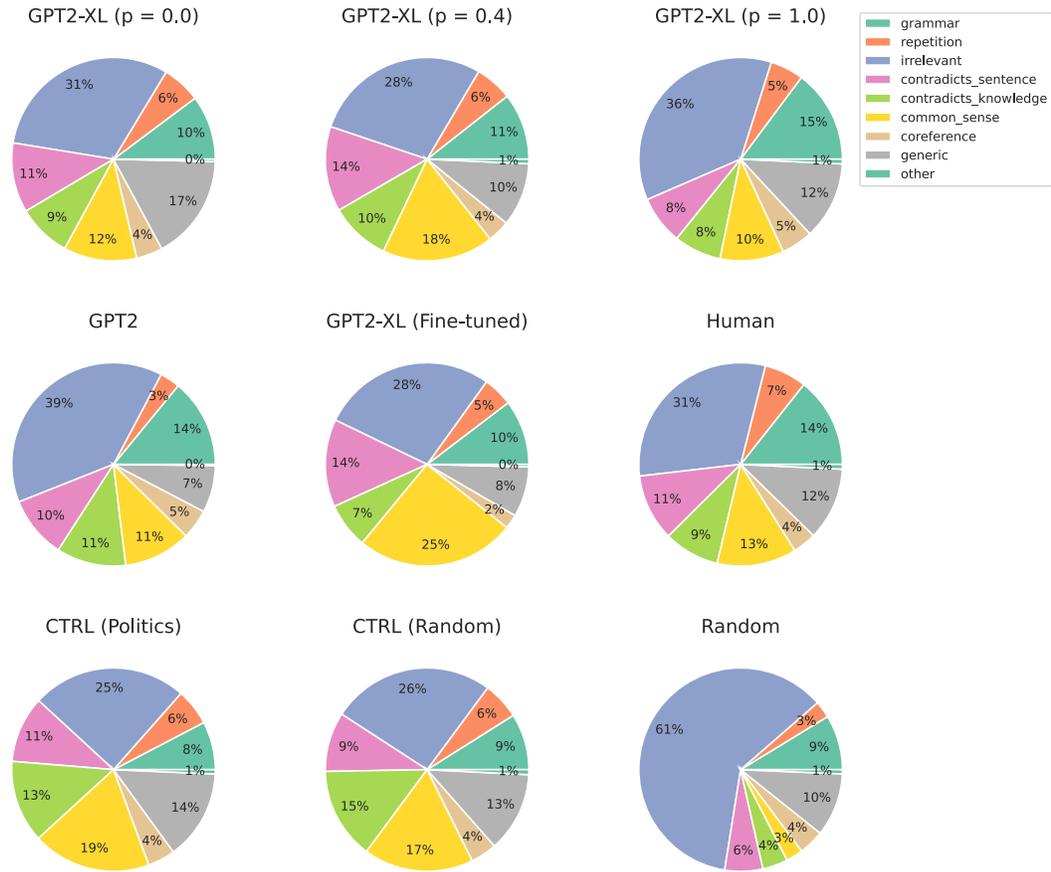


Figure 3.14: The reasons provided by players as to why a given example was generated broken up per model that generated the text

Future work could build off out study by testing a larger set of models, genres, and other experimental conditions ((finetuning, topic control, decoding strategy, etc.). In addition, our study was limited in that we assumed continuations always happened on the sentence boundary, but that is not always necessarily the case. Future work could look at continuations that do not happen exactly on the boundary between sentences. We also believe that more investigation is needed into exactly what annotators are thinking when they make their decisions and how we can give annotators the right tools to explain their thought processes.

| Reason | <i>n</i> | Mean Score |
|-----------------------|----------|---------------|
| common_sense | 2,432 | 2.566 ± 0.086 |
| irrelevant | 4,259 | 2.530 ± 0.064 |
| contradicts_sentence | 1,606 | 2.527 ± 0.105 |
| contradicts_knowledge | 1,411 | 2.262 ± 0.111 |
| coreference | 542 | 2.249 ± 0.176 |
| repetition | 728 | 2.128 ± 0.154 |
| other | 75 | 2.040 ± 0.483 |
| generic | 1,546 | 1.920 ± 0.101 |
| grammar | 1,539 | 1.780 ± 0.105 |

Table 3.15: The number of times each reason was given for text being machine-generated, and the mean score over those annotations. We see that when players select reasons like “grammar” or “generic,” they are much less likely to be correct than when selecting “common_sense” or “irrelevant.”

Finally, we expect our released dataset of generations and annotations to be of broad use to those studying detection. It would be worthwhile to study how well automatic systems perform at the detection task, and whether we can predict when generated text will be especially difficult for human annotators to recognize.

3.3.6 Study Limitations

Work on the detectability of machine-generated text sits at an interesting balancing point. On one hand, gamifying and publicizing the detection task may help to raise the public’s awareness of their susceptibility to machine-generated text, and work such as ours paves the way for future research on techniques for helping the public to improve at detection. On the other hand, we show that the detection task is a viable method for evaluating generation systems. For researchers aiming to build better generative language models, decreasing human detection ability might a very reasonable goal to optimize for. As much as our project seeks to better understand and improve human detection, our results can just as easily be used to make generative models even less detectable than they already are. Despite this drawback, we nonetheless believe it is important to

study detection as a means of assessing the risks that language models pose and protecting against future harm.

One significant limitation in our work is in our choice of participants. We acknowledge that university students (many of whom have studied computer science) may not be representative of the larger population. It will be important for future work to take on a broader user study conducted with a more diverse population with the goal of understanding how the unique backgrounds of different annotators contribute to their ability to detect generated text.

Another limitation in our study was in the incentives given to participants to perform well. Many of the students given extra credit proportional to the amount of points they scored learned they could exploit the point system by always picking one of the later sentences as the boundary. They found that rapidly guessing Sentence 9 as the boundary on every game round was a more effective strategy for maximizing earned points per time spent than taking the time to carefully read the text in each round. One alternative system which could reduce this bias would be to show all ten sentences in the passage at once rather than show them one at a time. The player would get a certain number of tries to guess the index of the boundary sentence and would be scored based on the number of tries this takes. This would resolve the bug in our current system that some sentence positions have a high point value in expectation.

3.4 CONCLUSION

In this chapter, I introduce the task of detecting when text is machine-generated. This task can either be a human evaluation task (asking annotators to identify when text is generated) or an automatic one (building discriminators which can identify it). In Section 3.2 I conduct a preliminary study of how humans and finetuned BERT classifiers perform on this task. I show that decoding strategies which result in generated text that tends to fool humans lead to text that is likely to fool automatic detection systems. However, the results in this section are limited in that they don't break down detection by text genre, don't investigate passages

which are partially human-written and partially generated, and assume the automatic classifiers have access to training data from the same generative model we are aiming to detect samples from.

Section 3.3 aims to fill in several of these gaps by conducting a large-scale user study of human detection ability. In the ROFT user study, I demonstrate the viability of a boundary detection task as a framework for soliciting human evaluation of natural-language generation systems. I conduct the largest study of generated text detectability to date and, in the process, replicate many previous major results in the field, such as the improved performance of bigger models [75] and the difficulty in incentivizing annotators to improve over time [29]. We confirmed the result from Section 3.2 showing that less random decoding strategy settings result in generated text that is harder for humans to detect. While this trend was true in both domains we tested it in, the impact of decoding strategy was more stark in the Stories domain than the News domain. In addition, we have provided new insights into the ways in which humans interact with partially-generated text.

Future work could build off our study by testing a larger set of models, genres, and other experimental conditions ((finetuning, topic control, decoding strategy, etc.). In addition, our study was limited in that we assumed continuations always happened on the sentence boundary, but that is not always necessarily the case. Future work could look at continuations that do not happen exactly on the boundary between sentences. We also believe that more investigation is needed into exactly what annotators are thinking when they make their decisions and how we can give annotators the right tools to explain their thought processes.

3.5 SUMMARY OF CONTRIBUTIONS

My initial research on the detection task was published as “Automatic Detection of Generated Text is Easiest when Humans are Fooled” in the 2020 Proceedings of the Association of Computational Linguistics [65]. The work was performed in conjunction with Daniel Duckworth, with the mentorship of Douglas Eck and Chris Callison-Burch. I proposed the initial idea of studying detection, and Daniel and I both worked on designing and implementing the experiments and analyzing the results.

The Real or Fake Text annotation platform was introduced as a system demonstration at the 2020 Conference on Empirical Methods in Natural Language Processing [40]. The RoFT website was implemented by Arun Kirubarajan, Liam Dugan, and myself, with the assistance of Run Shi, and the mentorship of Chris Callison-Burch. The user study using annotations from RoFT was designed and run, and its results analyzed, by Liam and myself.

4

MEMORIZATION OF TRAINING DATA

4.1 MOTIVATION

Machine-generated text is most undetectable when it looks exactly like its training data. In fact, the log-likelihood loss used during training explicitly encourages models to be able to exactly reproduce their training data. The result is models that are capable of exactly reproducing multi-paragraph sequences verbatim from their training data. As models have grown from millions to trillions of parameters [43], with their training sets similarly growing from millions to trillions of tokens, they are at increased risk of memorizing their training data. The problem is made worse by the fact that these enormous datasets are only minimally curated. For example, Carlini et al. [24] found that while most instances of memorization are innocuous, such as news articles or religious text, models are also capable of memorizing things like contact information or the names of real individuals (referenced outside of news contexts). This sort of memorization is harmful if it breaches expectations of privacy or content ownership from those whose data is included in the train set. Memorization also reduces generalizability if models are biased toward examples that are not representative of the underlying distribution of natural language.

In this chapter, we quantify the properties which raise the risk of memorization, notably the size of the model and the number of times a document occurred during training. We also show how passing in a long prompt to the model increases the chance we will extract memorized content. Finally, we describe one actionable step—thorough dataset deduplication—which can be employed before training to diminish the risk of memorization.

4.1.1 Definitions of Memorization

Memorization is an intuitive concept. When presented with two passages of text, humans are able to make judgement calls over whether they are similar to each other. In order to systematically study memorization, it is necessary to develop definitions which attempt, imperfectly, to capture this intuition. In this chapter, we employ several different imperfect measurements of memorization, each capturing a slightly different notion of what it means for a language model to memorize.

There is extensive prior work that qualitatively studies memorization in neural language models. Much of this work has defined memorized as *extraction*, the ability to extract training data examples from a trained model. These works have demonstrated that it possible to recover various forms of memorized data including URLs, phone numbers, or other forms of personal information [24, 184], or in other work, synthetically injected “canaries” [23, 59, 147, 148]. However, extraction attacks typically aim to demonstrate the existence of extractable data, rather than precisely quantifying how much models memorize. For example, the unprompted memorization evaluation of Carlini et al. [24] found just 600 examples of memorization in GPT-2.

In contrast, in Section 4.2, our goal is to establish tight approximations to the fraction of a dataset that can be adversarially extracted from a language model. We use the following definition for memorization:

Definition 1. A string s is *extractable with k tokens of context* from a model f if there exists a (length- k) string p , such that the concatenation $[p || s]$ is contained in the training data for f , and f produces s when prompted with p using greedy decoding.

For example, if a model’s training dataset contains the sequence “*My phone number is 555-6789*”, and given the length $k = 4$ prefix “*My phone number is*”, the most likely output is “*555-6789*”, then we call this sequence extractable (with 4 words of context). We choose greedy decoding because it is a cheap to run (compared to beam search), and we found that memorization is correlated with generated sequences being assigned very high likelihood.

In Section 4.3, we use a slightly different definition of “extractable”. Instead of generating with greedy decoding, we use top- $k=50$ random sampling, and instead of checking if the generation verbatim matches the

true continuation, we check if the edit similarity between the generation and the true continuation is above a chosen threshold. Section 4.3 also measures memorization in unconditioned generation. We generate 100,000 sequences each of 512 tokens, and then counted the fraction of generated tokens which are part of 50-token long substrings that occur in the train set.

All of these definitions are inspired by the broad literature on privacy attacks in machine learning. For example, membership inference attacks [139, 172] allow an adversary to detect the presence of a given example in a model’s training dataset, and other forms of data leakage permit an adversary to learn dataset properties [48, 45]. We focus on extraction attacks due to their relevance for language modeling—extraction demonstrates significant leakage from a model, and grows with data duplication [83], a common feature of large-scale text datasets.

Various formulations of memorization in deep neural networks have been studied in previous papers [23, 24, 44, 176, 109]. One leading general memorization definition is differential privacy [41], which is formulated around the idea that removing any user’s data from the training set should not change the trained model significantly. However, while differential privacy protects a single user’s private information, it is ineffective for memorization of duplicated data and does not capture the complexity of linguistic data [17]. Also, differentially private deep learning algorithms [1] generally suffer from expensive computation, slow convergence, and poor model utility, despite recent advances [6].

4.2 QUANTIFYING THE FACTORS THAT INFLUENCE MEMORIZATION

It is important to quantify factors that lead to increased memorization of a model’s training set. Indeed, recent work has shown that *training data extraction attacks* are a practical threat for current language models [24]. In a training data extraction attack, an adversary “attacks” a trained neural network by interacting with it with the goal of establishing whether an item was in the training data. While current attacks are effective, they only represent a lower bound on how much memorization occurs in existing models. For example, by querying the

GPT-2 language model, Carlini et al. [24] (manually) identified just 600 memorized training examples out of a 40GB training dataset. This attack establishes a (loose) lower bound that at least 0.00000015% of the dataset is memorized. In contrast, we are able to show that the 6 billion parameter GPT-J model [12, 160] memorized at least 1% of its training dataset, The Pile ([50]) (see Section 4.2.2).

In addition to these loose estimates of models' memorization capabilities, there is a limited understanding of how memorization varies across different neural language models and datasets of different scales. Prior studies of memorization in language models either focus on models or datasets of a fixed size [23, 176, 147] or identify a narrow memorization-versus-scale relationship [24, 83]. McCoy et al. [101] broadly study the extent to which language models memorize, but their focus is on how to avoid the problem and ensure novelty of model outputs, rather than on studying model risk through identifying maximum memorization.

The research presented in this section addresses both of the above open questions by comprehensively quantifying memorization across three families of neural language models and their associated datasets. We leverage access to each model's original training set to provide order-of-magnitude more precise bounds on the amount of extractable data than in prior works.

To construct a set of prompts from the model's training set, we feed varying-length prefixes of the training data back into the trained model, and verify whether the model has the ability to complete the rest of the example verbatim.

This allows us to measure memorization across models, datasets, and prompts of varying sizes. We identify three properties that significantly impact memorization:

1. **Model scale:** Within a model family, larger models memorize 2-5× more data than smaller models.
2. **Data duplication:** Examples repeated more often in a model's training set are more likely to be extractable.
3. **Context:** It is orders of magnitude easier to extract sequences when given a longer prompt.

Our analysis suggests that future research on neural language modeling will need to take steps to prevent future (larger) models from memorizing their training datasets.

4.2.1 Selection of Evaluation Data

Having chosen a definition, we next describe our evaluation procedure. Ideally, we would consider every sequence $x = [p \parallel s]$ contained in the model’s training dataset (where x has been split into a length- k prefix p and a suffix s). For each sequence, we would report if the model exactly reproduces s when prompted with p , following Definition 1. Unfortunately, performing this test on every sequence in the training data would be prohibitively expensive. For example, the largest 6 billion parameter GPT-Neo model has a throughput of roughly one 100-token generation per second on a V100 GPU. Extrapolating to the 800GB training dataset, this would require over 30 GPU-years of compute.

Instead, we query on a small subset of the training data. This subset should be small enough that it is feasible to test for extraction, but also large enough that it gives statistical confidence. In our experiments, we choose subsets of roughly 50,000 sequences. The primary criteria when choosing a subset of the training data is to obtain a representative sample that allows us to draw meaningful conclusions from the data. Yet, naively sampling from the data independently at random to construct a representative subset of the data distribution is not the best approach. Indeed, prior work has identified that one of the most important factors that contributes to training data memorization is how often that data has been *duplicated* (i.e., how often the same sequence is repeated either exactly or approximately-exactly). Because the frequency of training data duplication follows an exponential distribution [83], a fully random sample of only 50,000 sequences (accounting for $\leq 0.02\%$ of the dataset) is unlikely to contain *any* signal that would allow us to accurately measure the tail of this distribution.

Instead, we construct a duplication-normalized subset. For each sequence length $\ell \in \{50, 100, 150, \dots, 500\}$, and integer n , we select 1,000 sequences of length ℓ that are contained in the training dataset between $2^{n/4}$ and $2^{(n+1)/4}$ times. We do this until we reach an n for which 1,000 sequences are not available. This gives us 1000 sequences that repeat between 6 and 8 times ($\approx 2^{11/4}$ and $\approx 2^{12/4}$) and also 1000 sequences that repeat between 724 and 861 times ($\approx 2^{38/4}$ and $\approx 2^{39/4}$). This biased sampling allows us to more accurately measure memorization as a function of a sample’s duplication factor, without querying the entire dataset.

Note that constructing this duplicate-normalized data subset requires some work, as efficiently identifying duplicate substrings in an 800GB training dataset is computationally challenging. We make use of the suffix array construction from Lee et al. [83] to identify sequences which are present in the training set a specific number of times.

For each sequence length between 50 and 500 tokens, this collection process gives us roughly 50,000 examples duplicated varying numbers of times, totaling roughly 500,000 sequences. For each length ℓ sequence, we prompt the model with the first $\ell - 50$ tokens and report the sequence as “extractable” if the next 50 tokens emitted by the model exactly match the 50 token suffix of this sequence. Fifty tokens corresponds to an average of 127 characters or 25 words⁵, well over the length of a typical English sentence. Finally, we compute the average probability that a sequence is extractable by averaging over all lengths ℓ .

4.2.2 Experiments

Model and Dataset

We primarily study the GPT-Neo model family [12, 160] trained on the Pile dataset [50]. The GPT-Neo models are decoder-only language models trained with the objective of predicting the next token in a sequence given the previous ones. They come in four sizes: 125 million, 1.3 billion, 2.7 billion and 6 billion parameters. The Pile is a dataset containing text collected from various sources (e.g., books, Web crawls, open source code) that totals 825GB. At the time this research was completed, the largest GPT-Neo model was the largest language model available for public download, and The Pile is the largest public text dataset available.

Bigger Models Memorize More

We begin by considering the impact of model size on memorization, expanding on prior studies which qualitatively established a relationship between the size of GPT-2 models and their ability to memorize <30

⁵ As measured by `spaCy` on the GPT-Neo training set.

URLs [24]. In contrast, we study *a million* model generations in order to describe how model scale relates to memorization.

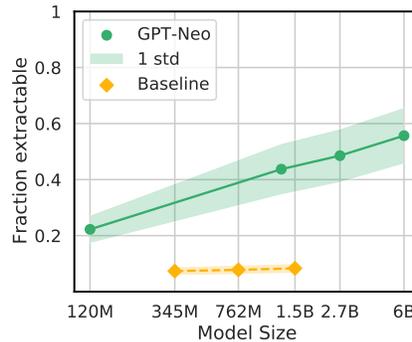


Figure 4.1: We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. Larger models memorize a larger fraction of their training dataset, following a log-linear relationship. This is not just a result of better generalization, as shown by the lack of growth for the GPT-2 baseline models. The dark shaded region is one std away from the mean, and the lighter shaded region represents the min and max over all document lengths.

RESULTS. The results of this experiment are given in Figure 4.1. The y-axis reports the fraction of generations which exactly reproduce the true suffix for their prompt, averaged over all prompt and sequence lengths we experimented on. We find that larger models memorize significantly more than smaller models do, with a *near-perfect log-linear fit* (R^2 of 99.8%): a ten fold increase in model size corresponds to an increase in memorization of 19 percentage points.⁶

To confirm that larger models are indeed *memorizing* more data, and not simply *generalizing* better, we also perform the same analysis with the GPT-2 model family as a baseline. The GPT-2 family of models are similarly sized, and also trained on Internet-scraped data. If our result that larger models memorize more was due to the general predictive strength of larger models, and not the memorization of specific training data, we would expect a similar relationship between comparably sized GPT-2 models trained on similar data. Put differently, this baseline allows to establish what fraction of the training data is sufficiently “easy” that any

⁶ This trend cannot continue indefinitely; the maximum percentage is 100%. We do not address these complications as our results max out at ~60%, but future work may need to handle these additional difficulties when extrapolating to even larger models.

language model could correctly predict the 50-token suffix, even if the example had never been seen before during training. For example, a language model that has seen multiple examples of number sequences during training could learn to correctly complete other number sequences that were not seen in training.

We find that approximately 6% of the examples in our evaluation dataset can be correctly completed by GPT-2, compared to 40% for a similarly sized 1.3B parameter GPT-Neo model. A qualitative analysis (see examples in Appendix Figure 4.12) suggests that examples “memorized” by GPT-2 are largely uninteresting sequences (e.g., number sequences, repetitions of the same few tokens, or common phrases). Therefore, we conclude that when larger models have a higher fraction of extractable training data, it is because they have memorized the data; it is not simply because the larger models are generally more accurate.

Repeated Strings are Memorized More Often

Prior work has provided preliminary evidence that memorization in language models increases with the number of times sequences are repeated in the training set [24, 83]. We expand on this observation and systematically measure the effect number of repetitions has on memorization. Using our experimental methodology, we measure the fraction of sequences which are extractable, for sequences in each bucket of duplicate counts, varied between 2 duplicates and 900 duplicates. Each bucket consists of 1,000 distinct sentences, and we compute the average amount of memorization for each bucket.

RESULTS. Figure 4.2 shows an analysis of our results, aggregated over all sequence lengths. We find a clear log-linear trend in memorization. While the model struggles to regurgitate strings which are repeated just a handful of times, this probability increases dramatically as strings have more repetitions. These small memorization values at low numbers of repetitions corroborate the impact of training dataset deduplication on memorization observed by Lee et al. [83]. However, we find that memorization does still happen, even with just a few duplicates—thus, deduplication will not perfectly prevent leakage. While this relationship is perhaps obvious, and has been corroborated for specific training examples in prior work [23, 24], our results show that it holds *across the entire training set*.

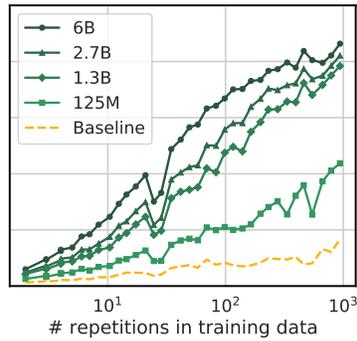


Figure 4.2: We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. Examples that are repeated more often in the training set are more likely to be extractable, again following a log-linear trend (baseline is GPT-2 XL).

Longer Context Discovers More Memorization

The previous two questions evaluated how data collection and model training decisions impact the leakage of a model’s training data when it is provided a fixed number of tokens from a sequence as context. As a result, those experiments suggest particular actions that could be taken to mitigate memorization (by reducing model size, or limiting the number of duplicate examples).

However, even when the model is fixed, it is possible to vary the amount of extractable training data by controlling the length of the prefix passed to the model. By studying how the number of tokens of context impacts extractability, we demonstrate the difficulty of *discovering* memorization—language models may only exhibit their memorization under favorable conditions.

RESULTS. In Figure 4.3, we observe that the fraction of extractable sequences increases log-linearly with the number of tokens of context. For example, 33% of training sequences are extractable from the 6B model at 50 tokens of context, compared to 65% with 450 tokens of context.

We call this the **discoverability phenomenon**: some memorization only becomes apparent under certain conditions, such as when the model is prompted with a sufficiently long context. This makes “discovering” memorization difficult.

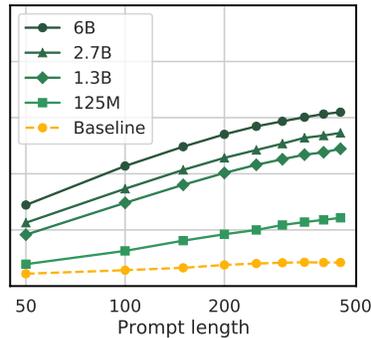


Figure 4.3: We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. As the number of tokens of context available increases, so does our ability to extract memorized text.

The discoverability phenomenon may seem natural: conditioning a model on 100 tokens of context is more specific than conditioning the model on 50 tokens of context, and it is natural that the model would estimate the probability of the training data as higher in this situation. However, the result is that some strings are “hidden” in the model and require more knowledge than others to be extractable.

From one point of view, it is good that some memorization is difficult to discover. This makes it harder for attackers to perform training data extraction attacks [24], or otherwise exploit memorization. Indeed, if an exact 100 token prompt is required to make the model output a given string, then, in practice, an adversary will likely be unable to perform the attack. The difficulty in discovering memorization also reduces the likelihood of *non-adversarial* training data regurgitation. For example, the GitHub CoPilot model [26] reportedly rarely emits memorized code in benign situations, and most memorization occurs only when the model has been prompted with long code excerpts that are very similar to the training data [184]. Practitioners building language generation APIs could (until stronger attacks are developed) significantly reduce extraction risk by restricting the maximum prompt length available to users.

Viewed differently, however, the difficulty of discovering memorization can also harm our ability to audit privacy in machine learning models. Because existing approaches for provably-correct privacy-preserving training of machine learning are applied only rarely in practice [1, 147, 120], it is common to attempt post-hoc

privacy auditing [71, 70, 108]. Our results suggest that correctly auditing large language models will likely require prompting the model with training data, as there are currently no known techniques to identify the tail of memorized data without conditioning the model with a large context. Improving upon this limitation is an interesting problem for future work.

Alternate Experimental Setting: Random Dataset Sampling

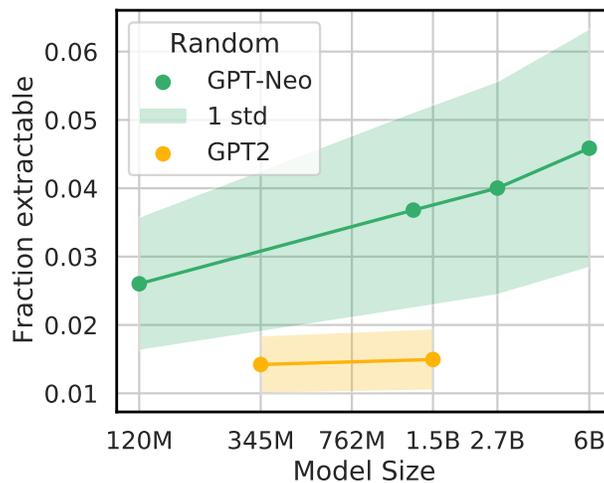


Figure 4.4: Percentage of sequences extracted as a function of model scale where we sample randomly from the training set.

The experiments presented thus far use subsets of the training data that were explicitly sampled according to how many duplicates of the example exist in the data. We now explore what would happen if we instead choose a truly random subset of the training data, where each sequence is sampled uniformly. Specifically, we randomly sample 100,000 sequences from The Pile dataset of length 100, 200, 500, and 1000; prompt the model with the first $N - 50$ tokens; and then test for memorization by verifying if the model can emit the remaining 50 tokens perfectly. We explore the result of this analysis in Figure 4.4 and Figure 4.5. We again vary the size of the models we train and the context length we provide to understand how this impacts memorization—but this time through prompting the models with randomly sampled training sequences. As

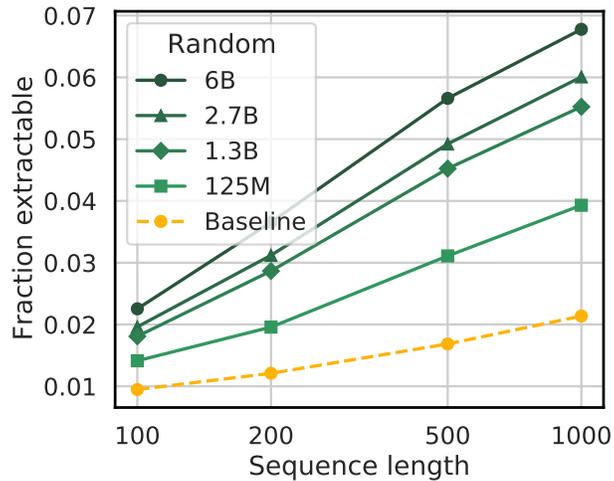


Figure 4.5: Percentage of sequences extracted as we vary the length of the prompt. For each sequence length n , $n-50$ tokens are used as the prefix, and we check for extraction of the remaining 50 tokens.

expected, the absolute probability of memorization is much lower than in Figure 1 where we prompted models with training data from the sampled duplication-normalized subset.

As before, we observe similar trends with model scale and context length. Larger models memorize more training examples than smaller models—and much more than the baseline GPT-2 model that was not trained on The Pile. Similarly, providing more context to a model increases the likelihood we can discover memorization. In Figure 4.5, we prompt models with: prompt length = sequence length – 50. We see that the longer prompts are easier to predict correctly than shorter prompts. The baseline GPT-2 model is nearly twice as accurate on sequences of length 1000 (prompt length = 950) compared to sequences of length 100 (prompt length = 50).

We can extract the last 50 tokens of a length-1000 sequence with nearly 7% probability for the largest GPT-J 6B model compared to 4% probability for the smallest 125M GPT-Neo model. (And both of these are much larger than the 2% probability of extraction for the 1.5B parameter GPT2-XL model.) These results taken together allow us to establish an estimated lower bound that there is 1% of The Pile dataset that is extractable by the 6B GPT-J model, but not by GPT-2 XL.

Alternate Experimental Setting: Beam Search Decoding

We previously defined memorization as the ability of a model to generate the exact true continuation when the *most likely* token is chosen at every step of decoding. However, using greedy decoding strategy does not produce the overall most likely sequence. Many language model applications use other decoding strategies, such as beam search (an algorithm for efficiently searching over the exponential space of sequences that could possibly be generated) to find the one with highest possible likelihood. To understand how our choice of decoding strategy affects the amount of memorization we measure, we compare greedy decoding with beam search in Figure 4.7(c).

We find that using beam search (with $b = 100$) results in only slightly more extracted memorization. The average difference in fraction of extractable memorization is just under 2 percentage points on average, with a maximum of 5.6. Interestingly, beam search and greedy decoding generated the same output 45% of the time.

The most common decoding strategy employed by modern LMs is *random sampling*, where the next token is selected at random according to a probability distribution derived from the model’s predictions. McCoy et al. [101] found that random sampling resulted in generated text with a greater number of novel n -grams. Since the goal of our study is to maximize discoverability—an antithetical goal to maximizing linguistic novelty—we do not present experiments that use random sampling.

Alternate Definition of Extractability

Our main experiments report a sequence as “extractable” if the model’s generated continuation is identical to the true suffix within that training example. This method is a loose lower bound on memorization. Consider two sequences x_1, x_2 both contained in the training dataset. Suppose these two sequences share the same prefix, and differ only in the final suffix; that is, $x_1 = [p||s_1]$ and $x_2 = [p||s_2]$. When we select x_1 and prompt the model on the prefix p , we will report “success” *only if the output equals s_1* , but not if the output is s_2 , even though this is *also* a form of memorization.

| Prompt | Continuation (== 6B) | 2.7B | 1.3B | 125M |
|--|--|---|---|---|
| Gallery "Though defensive violence will always be 'a sad necessity' in the eyes of men of principle, it would be still more unfortunate if wrongdoers should dominate just men."- St. Augustine "A new idea is first | condemned as ridiculous, and then dismissed as trivial, until finally it becomes what everybody knows." - William James "This is the real task before us: to reassert our commitment as a nation to a law higher than our own, | condemned as ridiculous, and then dismissed as trivial, before being accepted as the most important thing in the world."- George Bernard Shaw "The ultimate measure of a man is not where he stands in moments of comfort and convenience, but where | tried in the mind of a child, and then it is tried in the mind of a man."- Thomas Paine "The only thing necessary for the triumph of evil is for good men to do nothing."- Edmund Burke " | and foremost a man's duty to his fellow man. It is a duty to be a man of principle, and to be a |
| _GPL(crypto_unregister_alg); int crypto_register_template(struct crypto_template *tmpl) { struct crypto_template *q; int err = -EEXIST; | down_write(&crypto_alg_sem); list_for_each_entry(q, &crypto_template_list, list) { if (q == tmpl) | list_for_each_entry(q, &crypto_alg_list, list) { if (tmpl->name && tmpl->name != q->alg_cra_name) | q = kzalloc(sizeof(*q), GFP_KERNEL); if (!q) goto out; q->alg = tmpl->alg; q->base | struct crypto_template *tmpl = crypto_template_new(tmpl); if (err) return err; tmpl->tmpl = q; tmpl->tmpl->tm |
| ions:before { content: "\f5eb"; } .fa-discord:before { content: "\f392"; } .fa-discourse:before { content: "\f393 | "; } .fa-divide:before { content: "\f529"; } .fa-dizzy:before { content: "\f567"; } .fa-dna:before | "; } .fa-digg:before { content: "\f391"; } .fa-dochub:before { content: "\f394"; } .fa-docker:before { | "; } .fa-digg:before { content: "\f96c"; } .fa-dollar-sign:before { content: "\f155"; } .fa-digniter | "; } .fa-discus:before { content: "\f394"; } .fa-drapal:before { content: "\f395"; } .fa-drapal-discord |
| new users as an exploration tour and getting started guide, with exercises at the end of each chapter. For more advanced trainees it can be a desktop reference, and a collection of the base knowledge needed to proceed with system and network administration. This | book contains many real life examples derived from the author's experience as a Linux system and network administrator, trainer and consultant. They hope these examples will help you to get a better understanding of the Linux system and that you feel encouraged to try out things on | book is designed to give the reader a firm understanding of the technologies needed to install and manage Linux systems, using the various available tools and techniques for the task. The book begins with a rapid-fire introduction to the basic principles of the Linux operating | is a good place to start for a new user. A: I would recommend the book "Linux Networking" by David S. It is a very good book for beginners. A: I would recommend | is a great way to get started with a new project. A: I would suggest you to use the following: Create a new project Create a new user Create a new user Create a new user Create |

Figure 4.6: Text examples that are memorized by the 6B model, but not by smaller models. Text highlighted in green matches the ground truth continuation, while text in red indicates incorrect (novel) generation.

We now consider how our results would change if we instead checked that the generation $[p|f(p)]$ from a prompt p was contained *anywhere* in the training dataset. This gives a strictly larger measurement of memorization. By comparing these two methods (checking for memorization within the ground truth continuation, and within the entire dataset), we can understand how the choice of measurement affects the results in our experiments.

Searching within the entire dataset finds more memorized content than comparing with the ground truth (Figure 4.7). For examples at 100 repetitions 32.6% of outputs are contained somewhere in the dataset but just 15.8% match the ground truth continuation. This difference becomes more pronounced as the number of repetitions increases. The maximum difference between these approaches is 28.4%, at 2,200 repetitions.

We refrain from using this approach for our main experiments, because this definition requires substantially larger computation resources; it requires querying whether hundreds of thousands of sequences are contained in an 800GB training dataset. Therefore, to promote reproducibility, the remainder of this section continues with testing the generated suffix against the single expected training suffix.

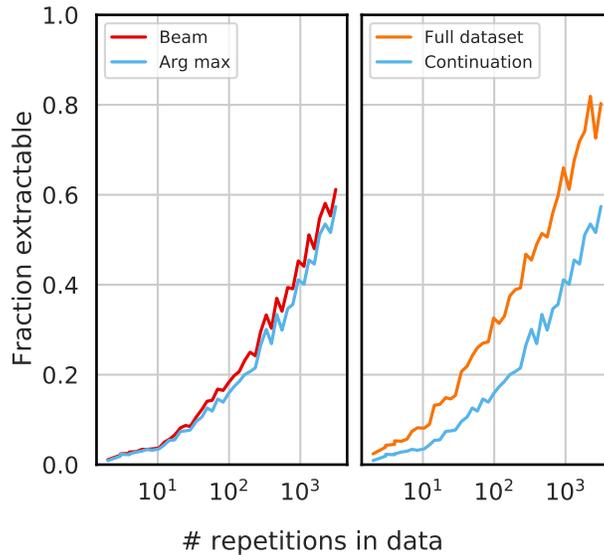


Figure 4.7: (left) Using beam search with $b=100$ slightly increases the data extracted. (right) We observe considerably more memorization when checking whether the generated sequence occurs anywhere in the entire training set (Section 4.2.2). However, this approach is very computationally expensive so we do not use it for experiments.

Qualitative Examples of Memorization

We now turn to inspect the training sequences memorized by the models.⁷ Figure 4.8 in the appendix shows examples of sequences that are memorized by *all* the models. We found most of these universally-memorized sequences to be “unconventional” texts such as code snippets or highly duplicated texts such as open source licenses.

More interestingly, Table 4.1 summarizes the total number of sequences that are memorized by one model but not another. Increasing model size leads to large numbers of nonoverlapping memorized sequences, although every model has some amount of memorization not shared by each other model. (Even the 125M model memorizes a few sequences the 6B model does not.)

In Figure 4.6, we present qualitative examples that are only memorized by the largest (6B) model. In these examples, the 50-token generations of the 6B model match the groundtruth continuations exactly, but the

⁷ For these results we sample 50-token prompts, 50-token continuations, and randomly sample across duplication counts.

| Model | Memorized | Not Memorized By | | | |
|-------|-----------|------------------|-------|-------|-------|
| | | 125M | 1.3B | 2.7B | 6B |
| 125M | 4,812 | - | 328 | 295 | 293 |
| 1.3B | 10,391 | 5,907 | - | 1,205 | 1,001 |
| 2.7B | 12,148 | 7,631 | 2,962 | - | 1,426 |
| 6B | 14,792 | 10,273 | 5,402 | 4,070 | - |

Table 4.1: The number of sequences memorized by one model, and not memorized by another. Not all sequences memorized by a small model are also memorized by a larger model. As a model gets larger, it memorizes more unique sequences.

generations from the smaller models match *neither* the groundtruth continuations of the prompted examples *nor* any other training examples with the same prompts. We highlight some interesting patterns in these sequences: while the generations from the smaller models do not match the training data, they are generally thematically-relevant and locally consistent. However, a closer inspection reveals that those generations are syntactically sound but semantically incorrect. Figure 4.8 lists examples that are memorized by models of *all* sizes, in the sense that the 50-token generations match the groundtruth continuations of the prompts.

In Figure 4.9 we show examples that are only memorized by the smallest model, using similar criterion as when we filter examples that are only memorized by the largest model. There are significantly fewer examples that are only memorized by the smallest model (35) than only memorized by the largest model (2860). The first row of Figure 4.9 is particularly interesting: the groundtruth continuation contains a typo due to formatting cutoff. While the smallest model memorized the typo, larger models try to fix the typo.

In Figure 4.10 and Figure 4.11 we show examples that are memorized but not heavily duplicated in the training set, and examples that are heavily duplicated but not memorized, respectively. Finally, we show examples that are memorized by GPT2-XL in Figure 4.12.

| Prompt | Continuation (== GPT2-XL) |
|---|--|
| 165.00 \$175.00 \$180.00 \$185.00 \$190.00 \$195.00 \$200.00 \$210.00 \$220.00 \$230.00 \$240.00 \$250.00 \$260. | 00 \$270.00 \$280.00 \$290.00 \$300.00 \$310.00 \$320.00 \$330.00 \$340.00 \$350.00 \$360.00 \$370.00 \$380.00 \$ |
| ISRC3DEC2, ARIZONA_ISRC3DEC2MIX_INPUT_1_SOURCE); ARIZONA_MUX_ENUMS(ISRC3DEC3, ARIZONA_ | ISRC3DEC3MIX_INPUT_1_SOURCE); ARIZONA_MUX_ENUMS(ISRC3DEC4, ARIZONA_ISRC3DEC4MIX_INPUT_ |
| the councils of government, we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the militaryindustrial complex. The potential for the disastrous rise of misplaced power exists and will persist. We must never let the weight | of this combination endanger our liberties or democratic processes. We should take nothing for granted. Only an alert and knowledgeable citizenry can compel the proper meshing of the huge industrial and military machinery of defense with our peaceful methods and goals, so that security and |
| full spreadsheet NEW! Buy our book • Facts are Sacred: the power of data (on Kindle) More open data Data journalism and data visualisations from the Guardian World government data • Search the | world's government data with our gateway Development and aid data • Search the world's global development data with our gateway Can you do something with this data? • Flickr Please post your visualisations and mash-ups on |
| Original press release Get ahead of the crowd by signing up for 420 Investor, the largest & most comprehensive premium subscription service for cannabis traders and investors since 2013. Published by NCV Newswire The NCV Newswire | by New Cannabis Ventures aims to curate high quality content and information about leading cannabis companies to help our readers filter out the noise and to stay on top of the most important cannabis business news. The NCV Newswire is hand-curated by |
| of sole responsibility of the author(s). The Centre for Research on Globalization will not be responsible for any inaccurate or incorrect statement in this article. The Centre of Research on Globalization grants permission to cross-post Global Research articles on community internet sites | as long the source and copyright are acknowledged together with a hyperlink to the original Global Research article. For publication of Global Research articles in print or other forms including commercial internet sites, contact: www.globalresearch.ca |

Figure 4.12: Text examples that are from The Pile and memorized by GPT2-XL. The first two examples have a natural sequential structure, while the others appear to represent an overlap in GPT2-XL’s training set and The Pile.

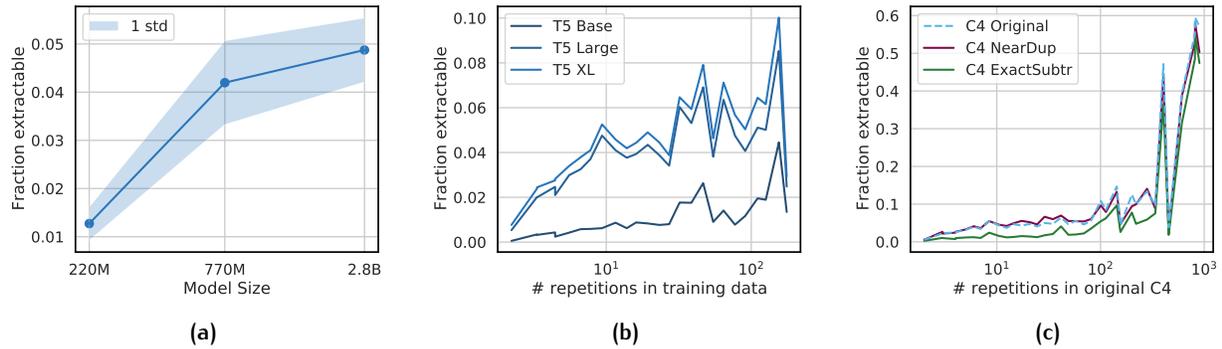


Figure 4.13: (a) Masked language model objective: Larger models have a higher fraction of sequences extractable on T5; with one standard deviation of variance shaded in dark and the minimum and maximum shaded light. (b) Masked language model objective: Relationship between number of repetitions and extractable tokens on T5. (c) Causal language model objective: Relationship between number of repetitions and memorization on language models trained with deduplicated data.

4.2.3 Replication Study–T5

The above analysis presents convincing evidence that memorization scales in a log-linear relationships with model size, data duplicates, and context length. We now replicate this analysis for different language model families trained on different datasets and with different training objectives, and performed the same memorization analysis on

We expected our results to cleanly generalize across settings—and this was indeed the case for model scale. Yet, we found the situation to be more complicated when considering training data duplication.

The T5 v1.1 models are masked encoder-decoder models trained to reproduce spans that were randomly deleted from an input sequence. The models vary in size from between 77M and 11B billion parameters. These models were trained on C4, a cleaned and filtered version of the English web pages from the Common Crawl, which totals 806 GB in size. At 11 billion parameters, the largest T5 model is the largest publicly available masked language model, making these T5 models a good candidate for studying how memorization scales with model size.

We must first define what is meant by “extractable data” for the masked language modeling task. T5 models are trained by removing a random 15% of tokens from each training sequence (i.i.d), and the model must

then “fill in the blanks” to restore the tokens that were dropped from the input. As a result of this different training objective, Definition 1 is not directly applicable because the model does not operate on a *prefix* and output a *suffix*. We instead define a sequence as memorized if the model *perfectly* solves the masked language modeling task on that sequence. For example, we call a 200-token sequence memorized if the model can use the 170 ($= 200 \cdot 0.85$) tokens of context to perfectly predict the remaining 30 tokens ($= 200 \cdot 0.15$). Because this token-dropping procedure is stochastic, it is possible that one set of dropped tokens might yield an output of “memorized” and another might not. For simplicity, we inspect only one set of masked tokens per sequence; because we are already averaging over 50,000 sequences this additional randomness does not harm the results of our analysis.

We are able to reproduce the model scaling effect shown in Figure 4.1 for the T5 model family; Figure 4.13a presents these results. Increasing the number of parameters in the model similarly increases the ability of the model to perfectly solve the masked prediction task.

Surprisingly, while the overall scaling trend holds true here, we discover an order of magnitude less data in masked models than in a comparably sized decoder-only language model. For example, the 3B parameter T5-XL model memorizes just 3.5% of sequences repeated 100 times, compared to the 53.6% of sequences repeated 100 times memorized by GPT-Neo 2.7B (with a context length of 150). We believe (without evidence) that this difference can be explained because the choice of tokens to mask at training-time varies for each duplicate of a training example, while decoder-only language models are always provided the same prediction task each time an example and its duplicates are seen during training.

Next, we turn to reproducing the analysis on the effect of duplicate examples in the models’ training data on memorization. The situation here becomes significantly less clear. As we can see in Figure 4.13b, while sequences that have been duplicated more often are easier to memorize, there is not an obvious log-linear scaling relationship to be found. In particular, compared to the smooth curves we observe in the case of GPT-Neo evaluated on The Pile, there is significant variance in the results for T5 models trained on C4. Even more surprising is that this variance is *statistically significant*: sequences repeated between 159 and 196 times are memorized with probability no more than 5.1% with 99.7% confidence (three standard deviations

of variance), however sequences repeated between 138 and 158 (that is, *less often*) are memorized with probability at least 6.2% with 99.7% confidence. That is, for some reason, sequences that occur ~ 140 times are *more likely to be memorized, despite occurring less often*, even if we assume a three-sigma error in both measurements simultaneously.

In order to explain this counter-intuitive phenomenon, we qualitatively study each of these two buckets of examples to understand why there is a pronounced difference. Surprisingly, we find that most of the duplicate examples contained in the 138-158 repeat bucket are mostly whitespace tokens, making these sequences much easier to predict correctly than sequences found at other repeat counts. This effect, to a lesser extent, can be found in other buckets which contain many approximately near duplicates.

4.2.4 Replication Study—Language Models Trained on Deduplicated Data

We further attempt replication with a family of 1.5B parameter decoder-only language models trained on deduplicated versions of C4. Described in detail in Section 4.3, this model family consists of one model trained on C4 (the same dataset as T5), one model trained on a version of C4 that was deduplicated by removing all documents which were near-duplicates of other documents, and one model trained on a version of C4 that was deduplicated by deleting any string of length-50 tokens that occurred more than once. We were most interested in whether models trained on deduplicated data would still exhibit increased memorization of examples which were repeated frequently in the original, non-deduplicated C4 dataset. Figure 4.13c plots this fraction of sequences memorized by each of the three models. We draw two interesting conclusions from this data.

First, we find that models trained on deduplicated datasets memorize less data than models trained without deduplicated datasets. For example, for sequences repeated below 35 times, the exact deduplicated model memorizes an average of 1.2% of sequences, compared to 3.6% without deduplication, a statistically significant ($p < 10^{-15}$) increase by a factor of 3. Second, while deduplication does help for sequences repeated up to ~ 100 times, it does not help for sequences repeated more than this. We observe a spike beginning at 408

repeats: the extractability of the smallest spike is larger than any value before the spike (largest is at 265 repeats, $p < 10^{-20}$). We hypothesize that this is due to the fact that any deduplication strategy is necessarily imperfect in order to efficiently scale to hundreds of gigabytes of training data. Thus, while it may be possible to remove *most* instances of duplicate data, different and valid definitions of duplicates can mean deduplication is not exhaustive.

4.3 DEDUPLICATING TRAINING DATA REDUCES MEMORIZATION

As described in the previous section, the presence of duplicate strings in training data is a significant source of memorization in neural language models. For the most part, duplicates exist not because we want them there, but because it is tricky to remove them, especially as training datasets have grown from single gigabytes to as much as a terabyte over the past few years [25, 170, 54, 20]. In this section we document the pervasiveness of duplicate content in the training data for large language models; we find that 10% of the sequences in several common NLP datasets are repeated multiple times. While naive deduplication is straightforward (and the datasets we consider already perform some naive form of deduplication), performing thorough deduplication at scale is both computationally challenging and requires sophisticated techniques.

The simplest technique to find duplicate examples would be to perform exact string matching between all example pairs, but we show this is insufficient since the web contains many documents which are near-duplicates of each other. Thus, we introduce two complementary, scalable methods for performing deduplication on documents which have substantial overlap but may not be identical.

- *Exact* substring matching identifies strings that are repeated verbatim in the train set multiple times. This allows us to identify cases where only part of a training example is duplicated (§4.3.3). Using a suffix array [98], we are able to remove duplicate substrings from the dataset if they occur verbatim in more than one example.

- *Approximate* full document matching uses MinHash [15], an efficient algorithm for estimating the n -gram similarity between all pairs of examples in a corpus, to remove entire examples from the dataset if they have high n -gram overlap with any other example (§4.3.4).

We identify four distinct advantages to training on datasets that have been thoroughly deduplicated.

1. Over 1% of tokens emitted unprompted from a model trained on standard datasets (e.g., C4) are part of a memorized sequence (See §4.3.5)—even though the 1.5 billion parameter model is much smaller than the 350GB dataset it was trained on. By deduplicating the training dataset we reduce the rate of emitting memorized training data by a factor of 10×.
2. Train-test overlap is common in non-deduplicated datasets. For example, we find a *61-word sequence*⁸ in C4 [119] that is repeated 61,036 times verbatim in the training dataset and 61 times in the validation set (0.02% of the samples in each dataset). This train-test set overlap not only causes researchers to over-estimate model accuracy, but also biases model selection towards models and hyperparameters that intentionally overfit their training datasets.
3. Training models on deduplicated datasets is more efficient. Processing a dataset with our framework requires a CPU-only linear-time algorithm. And so because these datasets are up to 19% smaller, even including the deduplication runtime itself, training on deduplicated datasets directly reduces the training cost in terms of time, dollar, and the environment [11, 142, 113].
4. Deduplicating training data does not hurt perplexity: models trained on deduplicated datasets have no worse perplexity compared to baseline models trained on the original datasets. In some cases deduplication reduces perplexity by up to 10%. Further, because recent LMs are typically limited to training for just a few epochs [118, 119], by training on higher quality data the models can reach higher accuracy faster.

⁸ “by combining fantastic ideas, interesting arrangements, and follow the current trends in the field of that make you more inspired and give artistic touches. We’d be honored if you can apply some or all of these design in your wedding. believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!”

To summarize, data duplication offers significant advantages and no observed disadvantages. In the remainder of this section, we present our text deduplication framework and study the extent of duplicate content in common NLP datasets (e.g., C4, Wiki-40B, and LM1B). We then examine the impact of deduplication on test perplexity and on the frequency of emitting memorized content. Finally, we analyze to what extent perplexity on existing, released models are skewed as a result of overlap between the train and test/validation splits.

4.3.1 Large Language Model Datasets

While we believe our results are independent of model architecture, we perform our analysis on Transformer-based decoder-only language models [154] trained for open-ended text generation. These current state-of-the-art models are trained on internet text. For example, the GPT-2 family of models Radford et al. [118] is trained on WebText, a dataset of web documents highly ranked on Reddit—however this dataset was not made available publicly. A common dataset starting point is CommonCrawl, an index of public webpages. Among the models trained on CommonCrawl include GPT-3 [20] with the addition of book datasets, GROVER [174] on a restricted subset filtered to news domains called RealNews, and T5 [119] on a cleaned version of common crawl called C4. Other models are trained on more curated Internet sources—for example Guo et al. [56] used high quality processed Wikipedia text from 40 different languages to train monolingual 141.4M parameter language models. Non-English models necessarily use different datasets; Zeng et al. [175] for instance introduced PANGU- α , a family of models with up to 200B parameters that were trained on a non-public corpus of cleaned and filtered Chinese-language documents from CommonCrawl and other sources. Since many of these datasets are not public, we deduplicate three that are: Wiki-40B, C4, and RealNews—as well as the One Billion Word Language Model Benchmark [25], a smaller dataset commonly used for evaluation.

Others have observed that popular datasets contain problematic duplicate content. Bandy et al. [9] observe that the Book Corpus [182], which was used to train popular models such as BERT, has a substantial amount

of exact-duplicate documents according to. Allamanis et al. Allamanis [4] show that duplicate examples in code datasets cause worsened performance on code understanding tasks.

When models are trained on datasets constructed by crawling the Internet, it is possible the model will train on the test set of downstream target tasks. For example, Radford et al. [118, §4] performed a post-hoc analysis to identify 8-gram overlaps between GPT-2’s training set and datasets used for evaluation, and Dodge et al. [37] analyzed C4 and found that up to 14.4% of test examples for various standard tasks were found verbatim (normalizing for capitalization and punctuation) in the dataset. A more proactive approach removes contaminated data. Trinh and Le [150, Appendix B] removed documents from their CommonCrawl-based train set that overlapped substantially with the commonsense reasoning used for evaluation. And GPT-3 [20, §5] did the reverse and removed downstream evaluation examples from their training data by conservatively filtering out any train set examples with a 13-gram overlap with any evaluation example. Up to 90% of tasks were flagged as potentially contaminated.

In our research, we do not focus on the impact of duplicate text in pretrained models on downstream benchmark tasks; instead we address how duplicate text in the LM training and validation sets impacts model perplexity and the extent to which generated text included memorized content.

4.3.2 Datasets Considered in this Study

We analyze the presence of duplicate text in four datasets of varying sizes that have been used for training natural language generation systems, producing general-purpose pre-trained models, and for language model benchmarking. While the analysis we perform is restricted to English datasets, we expect that non-English datasets suffer from similar issues and could likewise benefit from de-duplication.

- **Wikipedia (Wiki-40B)** consists of multi-lingual cleaned Wikipedia text [56]. We take the English portion, which contains 2.9M Wikipedia pages with an average length of 768 BPE tokens. The dataset

creators do not indicate any deduplication was performed aside from removing redirect-pages (e.g., “sunflower” to “Helianthus”).

- **One-Billion Word benchmark (LM1B)** contains 30M sentences of news commentary [25]. Unlike the other datasets we analyze, LM1B’s examples are one sentence long rather than multi-sentence documents. The average example length is 32 BPE tokens. While this dataset is extremely standard for benchmarking language models, Radford et al. [118, Sec 4] note it has 13.2% overlap of the test set with the train set.
- **Colossal Cleaned Common Crawl (C4)** is made up of 360M web documents, with an average length of 486 BPE tokens [119]. C4 was introduced as a pre-training dataset for T5, a set of encoder-decoder models which have been widely used in fine-tuned downstream tasks. The dataset was previously deduplicated in a more sophisticated process than the prior two datasets. Each paragraph was hashed and paragraphs resulting in hash collisions were removed. This was followed by a pass that removed placeholder text, code, and prohibited words. See Dodge et al. [38] for a detailed breakdown of the source text in C4.
- **RealNews** is a subset of the Common Crawl consisting of articles from news domains [174]. It contains 31M documents with average length 793 BPE tokens. RealNews was deduplicated by inserting a hash of the first 100 characters of each document into a bloom filter [13] and then excluding any document which resulted in a hash collision. Like C4, examples with duplicate URLs were excluded.

4.3.3 Method for Exact Substring Duplication

OVERVIEW We consider a dataset $D = \{x_i\}_{i=1}^N$ as a collection of *examples* x_i . Each of these examples is itself a sequence of *tokens*: $x_i = [x_i^1, x_i^2, \dots, x_i^{s_i}]$.

Due to the diversity of possibilities in human language, it is rare for the same idea to be expressed identically in multiple documents unless one expression is derived from the other, or both are quoting from a shared source.

This observation motivates deduplicating exact substrings. We call our approach EXACTSUBSTR. When two examples x_i and x_j share a sufficiently long substring (that is, a substring for which $x_i^{a..a+k} = x_j^{b..b+k}$), that substring is removed from one of them.

SUFFIX ARRAYS This exact-substring-matching criterion, while conceptually simple, is computationally prohibitive with naive (quadratic) all-pair matching. To improve the efficiency, we concatenate all the examples of the entire dataset D into a giant sequence S , and construct a Suffix Array \mathcal{A} of S . A suffix array [98] is a representation of a suffix tree [163] that can be constructed in linear time in $\|S\|$ [76] and enables efficient computation of many substring queries; in particular, they allow us to identify duplicated training examples in linear time. Suffix arrays have the advantage over suffix trees in that they are 10–100× more memory efficient [98], requiring just 8 bytes per input token, though they are asymptotically less efficient for some query types. They have been used widely in NLP, such as for efficient TF-IDF computation [171] and document clustering [27].

The suffix array \mathcal{A} for a sequence S is a lexicographically-ordered list of all suffixes contained in the sequence. Formally,

$$\mathcal{A}(S) = \arg \text{sort all_suffixes}(S)$$

For example, the suffixes of the sequence “banana” are (“banana”, “anana”, “nana”, “ana”, “na”, “a”) and so the suffix array is the sequence (6 4 2 1 5 3). In practice, we construct S from the BPE tokenization of the text (§4.3.5).

SUBSTRING MATCHING After constructing \mathcal{A} , it is straightforward to identify duplicated training examples. Suppose that the sequence s was repeated exactly twice in the training dataset S at positions i and j , that is, $S_{i..i+|s|} = S_{j..j+|s|}$. Then the indices i, j will occur adjacent to each other in the suffix array \mathcal{A} .

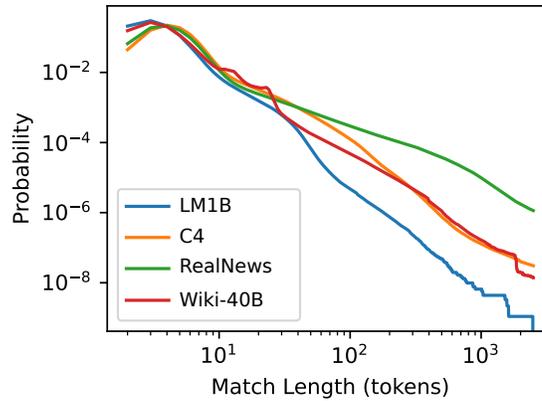


Figure 4.14: For each substring of length k , we plot the probability that there exists a second identical length- k substring in the same train set. Matches with length under 10 subword tokens are common, and account for 90% of tokens. We choose a threshold of 50 for experiments.

Finding all repeated sequences is thus a matter of linearly scanning the suffix array from beginning to end and looking for sequences $\mathcal{A}_i, \mathcal{A}_{i+1}$ that share a common prefix of at least some threshold length. Any satisfying sequences are recorded.

SETTING A THRESHOLD OF DUPLICATES One important question is how long a substring match must be before we ought to count it as a duplicate. In Figure 4.14, we plot the frequency of substring matches within the four datasets we will consider. For each substring of length k , we compute the probability that there exists another sequence of length k identical to this one; formally:

$$m(k) = \Pr_{i \in [N]} [\exists j \neq i : S_{i..i+k} = S_{j..j+k}].$$

We choose 50 tokens as the threshold to be conservative: the “bend in the knee” occurs at 10 tokens, and manual inspection of length-25 matches found no false positives. We then doubled this value to have an exceptionally large margin for error.

PARALLEL LINEAR TIME CONSTRUCTION. We build a parallelized linear time suffix array algorithm. As a building block, we make black-box use of the SA-IS algorithm for constructing a suffix array in linear time Nong et al. [110] and Ko and Aluru [78]. Unfortunately, this algorithm is not easily parallelized directly, so we introduce a simple divide and conquer approach to parallelizing the array construction.

We build our implementation in Rust and extend an existing suffix array library⁹ with three modifications. The first two are straightforward implementation differences: we modify the code to allow datasets larger than 4GB, and we remove the requirement that strings parse as valid UTF-8 sequences in favor of raw byte sequences. Our third change is more significant: we re-implement the algorithm so that we can stream the suffix array itself off disk.

PARALLEL PARTIAL SUFFIX ARRAY CONSTRUCTION. Our divide and conquer suffix array construction algorithm starts by partitioning the dataset into K different “splits” with SA-IS run over independently on each split in parallel. This algorithm still requires $O(N)$ work but runs in $O(N/K)$ wall-clock time. This gives us N separate suffix arrays \mathcal{A}^i .

Given two suffix arrays A_1 and A_2 for two sequences S_1 and S_2 it’s not completely trivial to construct a single suffix array A for $S = S_1 \parallel S_2$ because of the boundary conditions. Instead, we don’t build the data $S = S_1 \parallel S_2$ but rather let $S'_1 = S_1 \parallel S_2[\text{upto}K]$ for some K greater than the longest substring match. Then we build the arrays on S'_1 and S_2 . To merge the arrays together we can remove the items from the first array after index $|S_1|$ and merge-sort insert them into the second.

PARALLEL MERGE OF PARTIAL SUFFIX ARRAYS. We now merge these separate arrays together into a single suffix array \mathcal{A} . Consider the simpler case of two partial suffix arrays B and C that we would like to merge together. We can achieve this by letting $i = 0$ index B and $j = 0$ index C . Each iteration of the algorithm then pushes B_i into \mathcal{A} if $S_{B_i} < S_{C_j}$ and C_j otherwise, repeating until $i = |B| - 1$ and $j = |C| - 1$.

⁹ <https://github.com/BurntSushi/suffix>

To generalize to K splits, we need only replace the single comparison above with a min-heap requiring $O(\log K) \ll 10$ work on each iteration.

Observe that in the general case this algorithm is $O(Nm \log(K))$ where N is the length of the dataset, m is the average length of a prefix match, and K is the number of splits. It is therefore incorrect to call this algorithm linear time in the general case, for ours it is. Because the length of the longest match is bounded above by the length of the longest sequence, as long as the size of the dataset is independent of the length of the longest sequence in the dataset, this algorithm remains efficient.

Again, we can parallelize this operation among L simultaneous jobs (in practice we set $K = L$ as the number of threads on our machine). In the $K = 2$ case, job l processes $i \in [jN/L, (j+1)N/L]$, choosing the bounds of j by binary searching into C so that $S_{B_i} < S_{C_j} < S_{B_{j+1}}$. The case where $K > 2$ is identical except that we repeat this over all K partial suffix arrays.

COMPUTATIONAL ANALYSIS We run our algorithm on a single VM on the cloud with 96 cores and 768GB of memory. Our algorithm is efficient, for example processing the Wiki-40B training set (3 million examples containing 4GB of text) in 2.3 minutes wall-clock time (2.1 CPU-hours of work). The 350GB C4 dataset takes under 12 hours (wall-clock) to build a suffix array; although we are still memory constrained and so this corresponds to ~ 1000 CPU-hours. Once the suffix array has been constructed, it takes under an hour to deduplicate the C4 dataset.

Note that this algorithm still requires that the dataset itself fits in memory (so that we can efficiently index in arbitrary positions), but we do not need to fit the entire suffix array into memory. This is fortunate since our suffix array requires an $8\times$ space overhead. For example, the suffix array for the 350GB C4 is 1.5TB.

Compared to the cost of training a language model on this dataset, the additional work required to deduplicate the training dataset is negligible.

Table 4.2: Qualitative examples of near-duplicates identified by NEARDUP from each dataset. The similarity between documents is highlighted. Note the small interspersed differences that make exact duplicate matching less effective. Examples ending with “[...]” have been truncated for brevity. More data available in Appendix.

| Dataset | Example | Near-Duplicate Example |
|----------|---|---|
| Wiki-40B | \n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...] | \n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...] |
| LM1B | I left for California in 1979 and tracked Cleveland’s changes on trips back to visit my sisters . | I left for California in 1979 , and tracked Cleveland’s changes on trips back to visit my sisters . |
| C4 | Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination! | Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination! |

4.3.4 Method for Approximate Matching with MinHash

OVERVIEW We also perform *approximate* deduplication based on matching entire examples. This method, which we call NEARDUP, is a good complement to the *exact* substring matching, especially for web crawl text, as it handles the very common case of documents being identical except for interspersed templated fields (such as the last row of Table 4.2).

MinHash [15] is an approximate matching algorithm widely used in large-scale deduplication tasks [156, 47, 57], including to deduplicate the training set for a large Chinese-language LM [175]. Given two documents x_i and x_j , the main idea is to represent each document by its respective set of n -grams d_i and d_j . We can then use hash functions to approximate the *Jaccard Index* [69]:

$$\text{Jaccard}(d_i, d_j) = |d_i \cap d_j| / |d_i \cup d_j| \quad (4.1)$$

If the Jaccard Index between d_i and d_j is sufficiently high, it is likely that documents are approximate matches of each other. To efficiently approximate the Jaccard index, MinHash constructs document signatures by

sorting each of the n -grams via a hash function, and then keeping only the k smallest hashed n -grams. There are multiple ways to construct estimators of the Jaccard index from these kinds of signatures [31].

In our implementation, we use 5-grams and a signature of size 9,000. The probability that two documents are considered a potential match is

$$\Pr(d_i, d_j | \text{Jaccard}(d_i, d_j) = s_{i,j}) = 1 - (1 - s_{i,j}^b)^r \quad (4.2)$$

where $b = 20$ and $r = 450$ are user-settable parameters to control the strength of the filter.

For each pair of documents identified as a potential match, more computationally expensive similarity metrics can be employed as a subsequent filtering step. In particular, we identify two documents as duplicates if they are matched by the MinHash algorithm and their *edit similarity* is greater than 0.8. The edit similarity between token sequences x_i and x_j is defined as:

$$\text{EditSim}(x_i, x_j) = 1 - \frac{\text{EditDistance}(x_i, x_j)}{\max(|x_i|, |x_j|)} \quad (4.3)$$

To build clusters of similar documents, we construct a graph that has an edge between two documents if they are considered a match. Then, we use the method introduced in Łacki et al. [185] to identify connected components.

IMPLEMENTATION DETAILS For our MinHash based deduplication method, documents are first space tokenized, then each consecutive 5-gram is hashed using tabulation hashing. The set of these hashes is the signature for the document. For each element in a document’s signature, the element is hashed using k other hash functions. The minimum hashed element for each of the k hash functions is stored. These minimum hashes are then partitioned into r buckets, with b hashes per bucket. These b hashes are augmented into a

single value, then if two documents have the same value in at least one bucket, they'll be marked as a potential match. The probability that two documents are considered a potential match is equal to

$$\Pr(d_i, d_j | \text{Jaccard}(d_i, d_j) = s_{i,j}) = 1 - (1 - s_{i,j}^b)^r \quad (4.4)$$

where $s_{i,j}$ is the Jaccard index between the two documents. For document pairs that were identified as potential matches, we computed their actual Jaccard index, and if that was above 0.8, we computed their edit similarity. Document pairs with edit similarity higher than 0.8 were identified as duplicates. After some experimentation, we chose to use $b = 20$, and $r = 450$, so $k = 9,000$, so as to make sure a collision at the desired Jaccard index threshold of 0.8 had a high probability of occurring

COMPUTATIONAL ANALYSIS Let N be the number of documents and T be the maximal number of tokens in a document. Edit similarity has a worst case complexity of T^2 , so the worst case complexity is

$$O(N + bk^2T^2N) = O(N) \quad (4.5)$$

since b , k , and T are all $\ll N$. The left term is the complexity of grouping by the signatures, and the right represents the pathological worst case of all documents falling into the same B buckets.

The highly distributed NEARDUP implementation we employed is one used for large-scale production tasks at Google. On the English C4 dataset, the algorithm consumed approximately 41.5 kWh of energy. Note that our choices of k and b were designed to produce very high recall, and with different parameters, the algorithm could be made much more energy efficient while producing similar results.

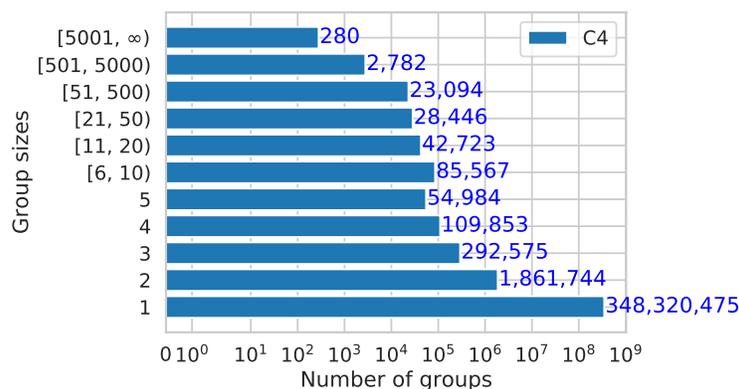


Figure 4.15: The distribution of near-duplicate cluster sizes from running NEARDUP on C4.

4.3.5 Results

We deduplicate each of the four datasets with both of our two techniques. When text was duplicated across multiple data splits, we prioritized keeping a copy in the test or validation set and removing it from the train set.

Amount of Text Removed

With NEARDUP, we found that the web-scrape datasets contain between 3.04% (on C4) to 13.63% (on RealNews) near duplicates (Table 4.3). Near-duplicate text is much less common in Wiki-40B, forming only 0.39% of the train set.¹⁰ In C4, the majority (1.8M) of near-duplicate clusters consisted of just a single pair of examples that matched against each other, but there were 280 clusters with over 5,000 examples in them (Figure 4.15), including one cluster of size 250,933.

On average with EXACTSUBSTR, we remove more total content than with NEARDUP (despite EXACTSUBSTR not removing any examples outright)—for example removing 7.18% of the tokens in C4. The exception is LM1B, where EXACTSUBSTR removes 8× less data than NEARDUP. On investigation, we find this is due to the fact that LM1B documents are significantly shorter: 90% of all documents are under 50 tokens, and

¹⁰ Most duplicates we saw were automatically generated pages, such as the outcomes of sports games. This shows the strength of manual curation for creating high-quality datasets.

Table 4.3: The fraction of examples identified by NEARDUP as near-duplicates.

| | % train examples with dup in train | | % valid with dup in train |
|----------|---------------------------------------|-------|------------------------------|
| C4 | 3.04% | 1.59% | 4.60% |
| RealNews | 13.63% | 1.25% | 14.35% |
| LM1B | 4.86% | 0.07% | 4.92% |
| Wiki40B | 0.39% | 0.26% | 0.72% |

Table 4.4: The fraction of tokens (note Table 4.3 reports the fraction of *examples*) identified by EXACTSUBSTR as part of an exact duplicate 50-token substring.

| | % train tokens with dup in train | | % valid with dup in train |
|----------|-------------------------------------|--------|------------------------------|
| C4 | 7.18% | 0.75% | 1.38% |
| RealNews | 19.4% | 2.61% | 3.37% |
| LM1B | 0.76% | 0.016% | 0.019% |
| Wiki40B | 2.76% | 0.52% | 0.67% |

so are not even candidates for potential matches even if the entire sequence matched verbatim. We find that both NEARDUP and EXACTSUBSTR remove similar content—77% of the training examples that NEARDUP removes from C4 have at least one verbatim length-50 match found by EXACTSUBSTR.

Properties of Duplicated Text

While the authors of both RealNews and C4 explicitly attempted deduplication during dataset construction, the methods were insufficient to capture the more subtle types of duplicate text commonly found on the internet. In C4 and Wiki-40B, we qualitatively observe that much of the text identified as near-duplicated is computer-generated. The text is identical except for the names of places, businesses, products, dates, and so on. Because these examples frequently differ by just a few words at a time, deduplication strategies relying on exact string matching would fail to identify a match. Example duplicate pairs from each dataset can be found in Table 4.2. Table 4.5 shows the URLs had the largest proportion of examples identified by NEARDUP

Table 4.5: On the left, we show the URLs that had the greatest proportion of examples marked as near-duplicates by NEARDUP (filtered to URLs which occurred at least 10 times). On the right, we show the 20 most frequent URLs in C4 for which all examples were marked as near-duplicates by NEARDUP.

| RealNews Url | # Total | Frac Dups | C4 Url | # Total | Frac Dups |
|----------------------------------|---------|-----------|--------------------------------|---------|-----------|
| medicalnewstoday.com. | 12 | 1.00 | hairtechkearney.com | 4883 | 1 |
| dodbuzz.com | 301 | 0.99 | keywordsking.com | 1786 | 1 |
| undertheradar.military.com | 187 | 0.97 | sydneyitalianfruitshops.online | 1178 | 1 |
| q.usatoday.com | 33 | 0.94 | moewiki.usamimi.info | 1001 | 1 |
| ad-test.thirdage.com | 354 | 0.94 | swarovskijewelryoutlet.org | 984 | 1 |
| amp.nymag.com | 15 | 0.93 | forzadurto.org | 980 | 1 |
| citizenwire.com | 1022 | 0.93 | producerati.com | 971 | 1 |
| paycheck-chronicles.military.com | 363 | 0.92 | sourceryforge.org | 908 | 1 |
| product-reviews.net | 73403 | 0.92 | heavenz-kitchen.com | 876 | 1 |
| kitup.military.com | 196 | 0.92 | little-eclipse.com | 822 | 1 |
| gcaptain.com | 33903 | 0.92 | walops.com | 819 | 1 |
| dev.screenrant.com | 70 | 0.91 | 16thstlaunderland.com | 713 | 1 |
| live.swissinfo.ch | 66 | 0.91 | theroyalstarinfo.com | 696 | 1 |
| news.theepochtimes.com | 82 | 0.87 | code4kt.com | 684 | 1 |
| opinion.toledoblade.com | 986 | 0.87 | nffalconsjerseys.us | 682 | 1 |
| cdn.moneytalksnews.com | 121 | 0.86 | quiltingbeeshop.com | 676 | 1 |
| amp.fox23.com | 14 | 0.86 | ulifeinsurancemiami.com | 675 | 1 |
| sales.rollingstone.com | 20 | 0.85 | wowkeyword.com | 673 | 1 |
| ftp.screenrant.com | 20 | 0.85 | taspetro.com | 671 | 1 |

as near-duplicates. For C4, these tend to be websites that sell many similar products and thus have a large amount of templated text. For RealNews, content aggregators seem especially common.

For RealNews and LM1B, derived from news sites, we observe that many near-duplicates occur because the same news article appears on multiple news sites with slightly different formatting. For example, in LM1B, there is one example that starts “*MINEOLA , N.Y. - New York officials say [...]*” and another that starts “(*AP*) - *New York officials say [...]*”. The two examples are otherwise identical.

Train / Test Set Leakage

Both deduplication methods identify overlap between the train set and the validation set (Table 4.3). For example, 4.6% of the C4 validation set and 14.4% of the RealNews validation set examples had an approximate duplicate in their respective training sets. Such duplication is problematic since it could cause evaluation metrics to be unfairly inflated for models that are better at memorizing their train sets. We evaluate the effect of this leakage on publicly released models in Section 4.3.5.

Impact of Training on Deduplicated Data

We trained 1.5B parameter “XL”, decoder-only, Transformer-based language models similar to GPT-2, on C4-ORIGINAL, C4-NEARDUP, and C4-EXACTSUBSTR, respectively. We use the T5 codebase and model architecture from Raffel et al. [119], and each model was trained for about two epochs on its respective dataset. To better understand the amount of variance in the perplexities of trained models, we also trained three different random seeds of the 110M parameter “base” model for each of the above three datasets—for a total of nine base-sized models.

For all experiments, we used a Byte Pair Encoding (BPE) vocabulary trained on C4-NEARDUP with a budget of 50K tokens, which resulted in a vocabulary the same size as GPT-2’s. We trained with a maximum sequence length of 512 tokens (for longer documents, we randomly extracted subsequences of this length.) Each model was trained for about two epochs. Since both C4-ORIGINAL and C4-EXACTSUBSTR contain approximately 365M examples, we performed 152K steps with a batch size of 4800 (or approximately 2

epochs). C4-NEARDUP contains approximately 350M examples, we performed 146K steps (or approximately 2 epochs). On a 128-core TPU v3 pod slice, XL models trained on C4-ORIGINAL and C4-EXACTSUBSTR took approximately 131 hours (5.5 days) to train, while the XL model trained on C4-NEARDUP took approximately 126 hours to train. Like T5, models were trained with the Adafactor optimizer [137]. A constant learning rate of 0.01 was used for the base models and 0.001 for the XL models.

The 1.5B parameter XL models had 24 layers, each with 32 attention heads. The model embedding size was 2,048, the feed forward layers had a hidden size of 5,120, and the key/value dimension size for the attention heads 64. The 110M parameter base models had 12 layers, each with 12 attention heads. The model embedding size was 768, the feed forward layers had a hidden size of 2,048, and the key/value dimension size for the attention heads 64.

VALIDATION SET PERPLEXITY We computed the perplexity of our trained models on the validation sets of LM1B and Wiki-40B, and on subsets of the C4 validation set (Figure 4.16). For the base size, we observe that all models have similar perplexity on the original C4 validation set and on validation set examples that were identified as unique (no near-duplicate in either train or validation). However, both models trained on deduplicated data have significantly higher perplexity on validation set examples that have duplicates in the training set than the model trained on the original C4. EXACTSUBSTR-deduplicated results in higher perplexity than NEARDUP-deduplicated. These trends holds true for the XL sized model as well. While this may suggest EXACTSUBSTR duplication results in models least overfit on the train set, note that both of these techniques have used separate duplicate thresholds and a different choice of thresholds could change the results.

When evaluating on the validation sets of LM1B and Wiki-40B, we found that models trained on NEARDUP-deduplicated C4 consistently achieved lowest perplexity. EXACTSUBSTR deduplication decreases perplexity of the XL model by almost 3 points perplexity on Wiki-40B which is much larger than the variation of about 1 point perplexity we observed in the base models. This is despite seeing fewer tokens of training data overall.

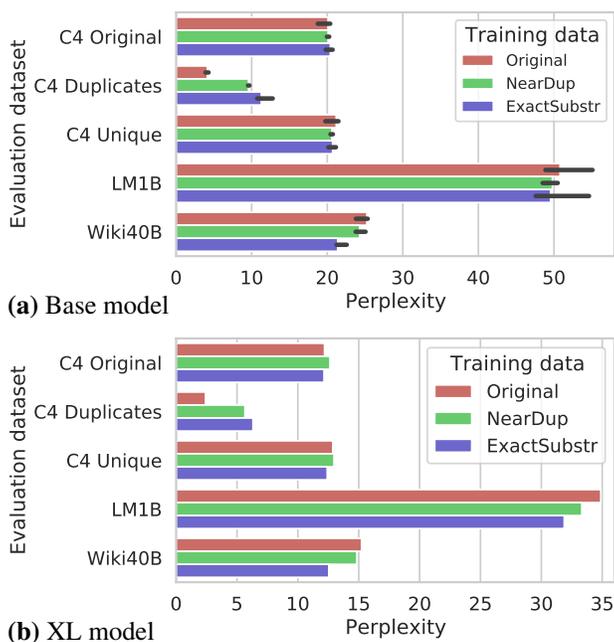


Figure 4.16: Impact of deduplicating the training set on validation perplexity. In (a), we plot the results from T5 base (110M parameters) across three training runs with different random initializations. The black bar represent the lowest perplexity to the highest perplexity, and the colored bar the median perplexity. In (b), we plot the results from T5 XL (1.5B parameters). For C4, we evaluate on *C4 Original*, the original validation set; *C4 Unique*, a subset of the validation set identified by NEARDUP as having zero matches across C4; and *C4 Duplicates*, a subset of the validation set identified by NEARDUP as having a match in the C4 train set.

Lastly, we note all our XL models achieved <35 perplexity on LM1B, which is less than the 42.16 perplexity reported for the 1.5B GPT-2 using a vocabulary the same size as ours.

MEMORIZATION EXHIBITED IN GENERATED TEXT Data duplication has the effect of biasing the trained LM towards particular types of examples. This can contribute to a lower diversity of generations, and increased likelihood that the generated content is copied from the training data [24]. For our generation experiments, we use top- k random sampling with $k = 50$ and experiment with prompted and unprompted generation.

We first evaluate memorization tendencies in the case where the model is asked to generate text without any prompt sequence. We generate 100,000 samples, each up to 512 tokens in length. For each generated token,

Table 4.6: When generating 100k sequences with no prompting, over 1% of the tokens emitted from a model trained on the original dataset are part of a 50-token long sequence copied directly from the training dataset. This drops to 0.1% for the deduplicated datasets.

| Model | 1 Epoch | 2 Epochs |
|----------------|---------|----------|
| XL-ORIGINAL | 1.926% | 1.571% |
| XL-NEARDUP | 0.189% | 0.264% |
| XL-EXACTSUBSTR | 0.138% | 0.168% |

we say the token is memorized if it is part of a 50-token substring that is exactly contained in the training data. On XL-ORIGINAL, over 1% of the generated tokens belong to memorized sub-sequences (see Table 4.6). This is $\sim 10\times$ more memorization than XL-EXACTSUBSTR or XL-NEARDUP. Some example subsequences that were copied verbatim from the train set can be found in Table 4.7.

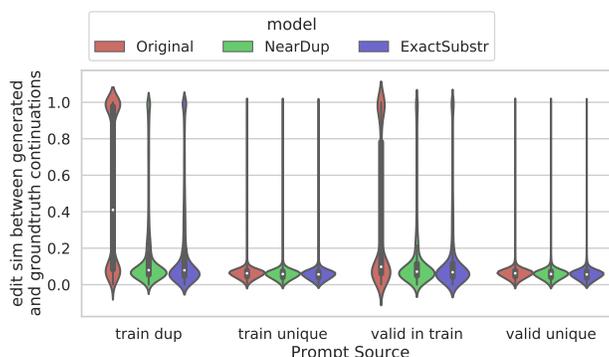


Figure 4.17: Memorized continuations distribution

In most real use cases, language model generation is controlled by providing a prompt for the model to continue. We experiment with four possible prompt sources: training examples identified by EXACTSUBSTR as having near-duplicates in the train set (train dup), training examples identified as unique (train unique), validation set examples with a near-duplicate in the train set (valid in train), and validation set examples identified as unique across all splits (valid unique). We select the first 32 tokens of each example as the prompt, which means we can evaluate the fraction of generations which are near-duplicates with the ground-truth continuation for the prompt. Figure 4.18 shows the proportion of generations which meet this requirement, while

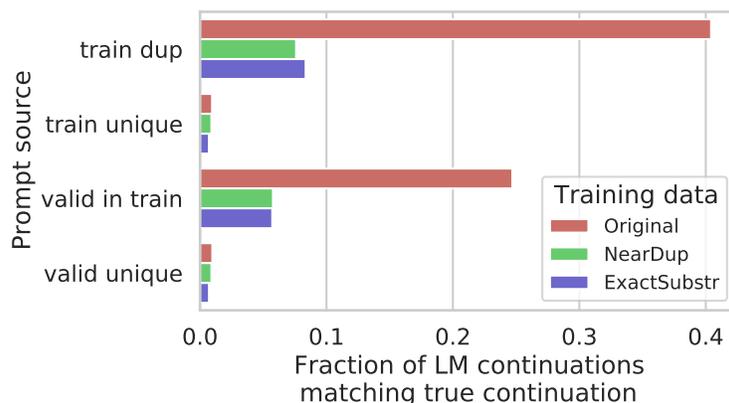


Figure 4.18: The proportion of generations which have edit similarity above 0.8 with the groundtruth continuation when using the LM to generate continuations for 32-token prompts identified by NEARDUP as either duplicated or unique.

Figure 4.17 shows the distribution in edit similarities between the generations and ground-truth continuations. When the prompt comes from duplicate examples in the train set, XL-ORIGINAL reproduces the groundtruth continuation over 40% of the time. XL-EXACTSUBSTR and XL-NEARDUP still copy the groundtruth more often when the prompt comes from a duplicate example than when the prompt comes from a unique example, suggesting that more stringent deduplication may be necessary to remove memorization tendencies entirely.

Impact on Deduplicating Validation Sets

Train-test leakage does not just impact models trained on C4. Table 4.9 shows that the presence of near-duplicates of the evaluation set in the train set has a significant impact on model perplexity for two standard models: Transformer-XL [34], which was trained on LM1B, and GROVER [174], which was trained on RealNews. For Transformer XL, the perplexity halves on examples identified as near-duplicates. For GROVER, the difference, though not quite as stark, is present in both model sizes considered.

Existing models also suffer from the problem of generating text from their train sets. We find that 1.38% of the tokens in the official release of 25k GROVER-Mega outputs are part of verbatim matches in RealNews of at least length 50. Likewise, more than 5% of the tokens in ~200k sequences outputted by GPT-Neo 1.3B [12] are part of a 50 token matches of its training data, the Pile [49].

4.3.6 Discussion

The results in this section demonstrate the crucial nature of the text used to train and evaluate language models. While recent work has focused the potential harms that could arise from problematic datasets [10, 51], less work has been done to quantitatively analyze properties of real language model train datasets, like Dodge et al. [38] has done for C4. The analysis in this section addresses just one particular axis, that of data duplication.

Our experiments measured what could be quantified: the amount of duplicate content in common datasets, the effect of deduplication on trained model perplexity, and the reduction of memorized content in trained models through deduplication. We do not focus on the nature of the data being removed by deduplication or memorized by LMs.

Privacy is an important subject for future work, as memorized training data has significant privacy consequences. By this, we mean the standard privacy definition that a model should not reveal anything particular to the specific dataset it was trained on, as opposed to another training dataset from a similar distribution [139].¹¹ Training on standard datasets that have not yet been deduplicated results in models that are particularly sensitive to examples that happened to be repeated multiple times, and this has negative privacy implications. For instance, it could violate a person’s expectations of privacy if their publicly available personal data appeared in a different, surprising context. Downstream applications of LMs, such as the game AI Dungeon¹², should also not output memorized content like adverts for real products.

We stress that in our experiments, we do not distinguish between undesired memorized text (such as phone numbers), innocuous memorized text (common phrases), and text we may want to be memorized (such as a quote by a public figure), and instead treat all instances of the LM generating text that closely matches the training set as problematic. While we qualitatively observed that much of the identified memorized content was relatively innocuous, a more systematic study of the risks associated with the detected memorization was beyond the scope of this work.

¹¹ Another interpretation of privacy focuses on the sensitivity of the data involved, when a model is trained on and able to reproduce personal identifiers or other forms of “private data.” Our definition is more expansive.

¹² <https://play.aidungeon.io/>

We also do not investigate the negative consequences of deduplication. Some language tasks explicitly require memorization, like document retrieval or closed-book question answering. Also, text that gives attribution is often duplicated across documents, so removing duplicate substrings could correspond to removing *just* the attribution, which could result in models that learn the content without its attached attribution. Deduplication is also not sufficient to remove privacy-sensitive data like bank passwords and medical records which should never be used in training data.

Lastly, the exact technique used to perform deduplication seems less important than performing stringent deduplication in the first place. On the whole, deduplication does not harm, and sometimes improves, model perplexity, despite the fact that the deduplicated datasets are smaller and faster to train on. It is especially important that there are no duplicates between the training and testing sets, because overlap here explicitly encourages selecting models that memorize the training data. Lastly, deduplication helps to reduce some of the privacy concerns around LMs memorizing their training data.

4.4 CONCLUSION

In this chapter, I present several factors which impact one's ability to extract memorized content from language models. First, I demonstrate the relationship between the length of prompt passed to the language model and the chance it will verbatim generate the true continuation to the prompt. Second, I show how modeling and training dataset choices have significant impact on memorization capacity. Larger language models memorize more than smaller ones, and the presence of duplicate examples in train sets significantly increases the chance those examples will get memorized.

In Section 4.3, I present a comprehensive quantitative analysis of the effect of mitigating the duplication problem. I discuss how thorough deduplication is difficult due to the presence of many near-duplicate examples, and I present two algorithms for handling this. By training language models on data before and after deduplication, I make the case that while current LMs do accurately model the distribution of their training

data, this does not necessarily imply they will model the desired *underlying* data distribution. In particular, when the training data distribution is skewed (e.g., by containing many duplicates of some sequences) larger models with more capacity are likely to learn these unintended dataset peculiarities. It therefore becomes even more important to carefully analyze the datasets used to train ever larger models, as future (larger) models are likely to remember even more details than current (smaller) models.

The findings in this chapter indicate that current large language models likely memorize a significant fraction of their training datasets. Memorization scales log-linear with model size—by doubling the number of parameters in a model we can extract a significantly larger fraction of the dataset. Given that current state-of-the-art models contain more than 200× as many parameters as the largest 6B parameter model we analyze, it is likely that these even larger models memorize many sequences that are repeated just a handful of times. At the same time, I have shown that this memorization is often hard to discover, and for an attack to actually extract this data it will be necessary to develop qualitatively new attack strategies. Fortunately, it appears that (for the comparatively small models we study) training data inserted just once is rarely memorized, and so thoroughly deduplicating training datasets is a practical technique to mitigate many the harms of memorization.

Ultimately, whether memorization is a desired property of a language model, or else risky and unwanted, depends both on the nature of the text that has been memorized and on the downstream applications of the trained model. However, since the trend has been towards creating datasets and models that are application-agnostic, researchers should think carefully about the limitations of the data they have collected and the how the model’s intended usage constrains what should be part of the training set. Developing techniques to memorize or forget specific sequences depending on the end application is a promising research direction.

4.5 SUMMARY OF CONTRIBUTIONS

This chapter describes work that from two papers: “Deduplicating Training Data Makes Language Models Better” published in ACL 2022 and “Quantifying Memorization Across Neural Language Models” which at the time of writing this dissertation is under submission. Both projects were conducted with my collaborators Nicholas Carlini, Matthew Jagielski, Katherine Lee, Andrew Nystrom, Florian Tramér, and Chiyuan Zhang.

For Section 4.2, I helped come up with the idea for the paper; ran inference on the deduplicated language models, wrote code to check for memorization, prepared data for the T5 results, created plots, and contributed to the framing and experimental design.

For Section 4.3, I co-led the project with Katherine Lee; ran the approximate matching data deduplication pipelines; extracted prompts and evaluation datasets; ran eval pipelines; and contributed significantly to project planning and data analysis.

Table 4.7: A selection of substrings identified by EXACTSUBSTR as being in C4 multiple times. The number of times this exact substring occurs in C4 is also given.

| Text | Freq in C4 |
|--|------------|
| HD wallpaper. This wallpaper was upload at April 19, 2019 upload by admin in.You can download it in your computer by clicking resolution image in Download by size:. Don't forget to rate and comment if you interest with this wallpaper. | 40,340 |
| to the address posted below. Include our failure information form,a packing slip with your Company name, contact person, and Email address or phone number. Upon receipt of your repair, we'll inspect it and then contact you with a quote or evaluation notice. Normal turn aro und for repair is 5 to 7 business days, with "Rush Repair" available. | 5,900 |
| is a great place to begin your search. Whether you are a first-time home buyer or you are already familiar with the home buying process, you can be assured that you have the best tools and the perfect agent available to help with your | 5,358 |
| pics at these awesome group starting P letter. Desktop wallpapers were first introduced way back in the 1980s and have gained immense popularity since then. It is possible to come across more than 80 million sites on the web offering some sort of wallpaper. | 848 |
| flowers will let them know you're thinking of them and wishing them well. Cheerful yellow flowers bring their own sunshine and will get right to work on lifting spirits, and a colorful vase will bring loads of smiles to friends and visitors! Get Well flower arrangements from | 479 |
| our premier 24 hour emergency* plumbing and heating solutions. We realise that when your heating fails or pipes and drains leak it can cause havoc with your routine and even cause damage to your property. When a plumbing problem occurs that requires an immediate response we provide qualified local plumbers throughout | 56 |
| is to remove all images that violate copyrights. Please contact us to request that images be removed or to assign proper credit. The images displayed on this site may be used for Free or educational purposes only. If you would like to use any of the images displayed on this site for any other purpose, please obtain permission from the owner. www. | 48 |
| list of fishing locations, providing interactive maps that show each location's GPS coordinates, nearby facilities (like restaurants, gas stations, marinas and fishing shops), their current and forecasted weather and, if available, their water conditions.\nFind any of the 8 | 5 |
| . Dyer, Ph.D., is an internationally renowned author and speaker in the field of self-development. He's the author of 30 books, has created many audio programs and videos, and has appeared on thousands of television and radio shows. | 5 |

Table 4.8: A selection of substrings generated by XL-ORIGINAL with no prompting (and top- k with $k=50$) that were identified by EXACTSUBSTR as being in C4 multiple times. The number of times each substring was found in C4 is given. We observe that most memorized generations tend to be from advertisements.

| Generated Text | Freq in C4 |
|---|------------|
| , you'll need to be knowledgeable to make the very best decisions. We will make sure you know what can be expected. We take the surprises from the picture by giving accurate and thorough information. You can start by talking about your task with our client service staff when you dial 888-353-1299. We'll address all of your questions and arrange the initial meeting. We work closely with you through the whole project, and our team can show up promptly and prepared. | 5,497 |
| then Waterside Lodge are well equipped for the task. Our fully equipped family sized lodges offer a comfortable luxurious stay for a fantastic price, giving you beautiful views of the lakes and the surrounding countryside. Offering luxurious self-catering holidays in our fully featured Scandinavian holiday lodges. Perfectly located to explore the beaches, coastline. All of our lodges are sized for 6 people and are furnished to the highest standards to ensure you have a stay like no other. At Waterside Lodge the stay itself is only half of the package, Waterside lodge is situated closely to the Heritage Coast which makes our lodges the perfect stay for anyone wanting to get away and have a relaxing countryside break from the city. Whilst you stay with us be sure to take advantage of all the activities Waterside Lodge has to offer. Such as the use of our on-site fishing lakes for the keen fisherman, free internet access, outside relaxation areas, comfortable lounges and much more. | 571 |
| you are only looking to find rent to own homes in your city or are open to exploring all kinds of rent to own home listings, our database does it all. One of the best aspects of iRentToOwn.com is that, besides options to rent to buy a house, it has numerous other categories of home sale options. These include bank foreclosure homes, pre-foreclosure homes, short sales, HUD/government foreclosures, auction homes and owner-financing/FSBO (For Sale By Owner) homes. With help from the convenient search features offered by our site, shoppers are able to find their ideal lease to own home, real estate company, and more in South | 51 |
| , IL employs journeyman as licensed to work by themselves, without direct supervision, installing wiring, outlets and fixtures. Our journeyman also does service work, troubleshooting when a breaker fails or a light stops working. Our journeyman does not offer permits that must be issued by our master. Our journeyman follows our master's plans and directions. Our journeyman's responsibilities will vary based on the work that needs to be done. Our journeymen are skilled with residential, commercial and industrial installations and repairs.ust work from six years as an apprentice, under direct supervision of our master, and pass a journeyman test. This person also must have some classroom education on the National Electrical Code and fundamental electricity in a technical school a program affiliated with the National Joint Apprenticeship Training Council. Journeyman training combines hands-on work with education on basic electricity. | 6 |
| combustion process of a petrol engine is never perfect. Dangerous gases, such as nitrogen oxide, carbon monoxide and hydrocarbons will arise and it is the job of the catalytic converter to reduce these to safer emissions. These cat converters can fail by becoming clogged, or if the engine has bad exhaust valves or the plugs fail, causing unburned fuel to overheat the converter. Mettam's Mufflers can resolve these issues with your Karr | 5 |
| .,ANDREW Find the ancestral town: Many a researcher is stuck behind records that say, BIRTH-PLACE: IRELAND without saying where in Ireland, or whatever other country. Remember that your immigrant ancestor's siblings probably were born in the same ancestral town, so check all of their records, too. Around 1900, the Roman Catholic churches reported marriages to the churches where the persons were baptised, and before the wedding, they would require a baptismal certificate from that church, without marriage notations, to make sure that the persons were not already married, ordained, or whatever, and were free to marry. Do check the Catholic records especially for ex loco and the home town. If your ancestor's sister had a daughter who generated a marriage or death record saying, MOTHER'S BIRTHPLACE: and the exact town, then y | 2 |

Table 4.9: For each model, the perplexity of the official validation set (*Orig*), valid set examples which were identified by NEARDUP as matches of train set examples (*Dups*), and valid set examples identified by NEARDUP as unique (*Unique*). Due to the size of the RealNews validation set, we evaluated on only the first 25k examples meeting each condition.

| Model | Dataset | Orig | Dups | Unique |
|----------------|----------|-------|-------|--------|
| Transformer-XL | LM1B | 21.77 | 10.11 | 23.58 |
| GROVER-Base | RealNews | 15.44 | 13.77 | 15.73 |
| GROVER-XL | RealNews | 9.15 | 7.68 | 9.45 |

5

ENABLING APPLICATIONS IN CREATIVE WRITING

One application where NLG has considerable potential is in the development of tools for creative writing. AI-assisted creative writing is an attractive testbed for NLG systems because ideation tools are already part of writers' arsenal, and mistakes like hallucinating false facts are less problematic in fiction than in domains like automatic news summarization, where faithfulness to the real world is crucial. In addition, science fiction writers have been grappling with the concept of human-like machines and artificial intelligence for at least as long as computer scientists have, so it stands to reason they might be early adopters of AI-powered creative writing tools.

In this chapter, I describe work I have done toward bridging the gap between what most language models do by default (predict a continuation for a prompt) and the operations writers actually would want. First, I will show how existing neural networks can be modified to support fill-in-the-blank style tasks in addition to the more common paradigm of continuation. Filling in the blank is a common functionality requested by writers. Second, I will present a recipe for performing sentence style transfer into an arbitrary range of styles—such as rewriting text to be more Shakespearean, metaphorical, or melodramatic—without any exemplars of the task or task-specific model training. Third, to test out how these and other NLG-based tools can be used in practice, I present Wordcraft, a word processor augmented with a variety of “smart” writing controls and suggestion tools. I describe how a single large language model is used to support all of the “smart” functions in Wordcraft, and I examine Wordcraft's strengths and limitations through user studies conducted with both amateur and professional writers.

5.1 SUPPORTING ARBITRARY STYLE TRANSFER

Text style transfer is the task of rewriting text to incorporate additional or alternative stylistic elements while preserving the overall semantics and structure. Early approaches to style transfer required *parallel* text data [183, 122], where every input in the source style has a corresponding output in the target style. Because the availability of such data is limited, however, there has been a shift toward approaches which instead rely on *non-parallel* monostyle data [90, 73, 94, 79]. Most recently, *label-free* methods have taken advantage of the natural manifold of language (i.e., that sentences that are nearby to each other in a text passage are also similar stylistically) to train style transfer models that require only a few exemplars in the target style for inference [169, 125]. This is true even for approaches which claim to be label-free [169, 125]. Hence, there is a clear need for new methods that both reduce the training data requirements and expand the scope of styles supported [72, 63].

In this section, we present *augmented zero-shot learning*, a prompting method that allows large language models to perform text style transfer to arbitrary styles, without any exemplars in the target style. Our method builds on prior work showing that sufficiently large LMs such as GPT-3 can perform various tasks ranging from classification to translation, simply by choosing a clever prompt to prepend to the input text for which the model is asked to continue [19, 14]. Although large LMs are trained only for continuation, recent work has shown that they can perform a variety of NLP tasks by expressing the task as a prompt that encourages the model to output the desired answer as the continuation [95, 116, 164, 19, 132]. The simplest approach, **zero-shot prompting**, directly uses natural language to ask the large LM to perform a task, as shown in Figure 5.1a. Zero-shot learning, however, can be prone to failure modes such as not returning well-formatted or logical outputs (see §5.1.3). However, zero-shot prompts are prone to failure modes such as not returning a well formatted or logical answer. This problem can often be overcome by prepending exemplars to the prompt that demonstrate what successful completions may look like. This approach, called **few-shot prompting**, has been shown to achieve higher performance, but requires exemplars for the exact task that we want the model to perform (Figure 5.1b).

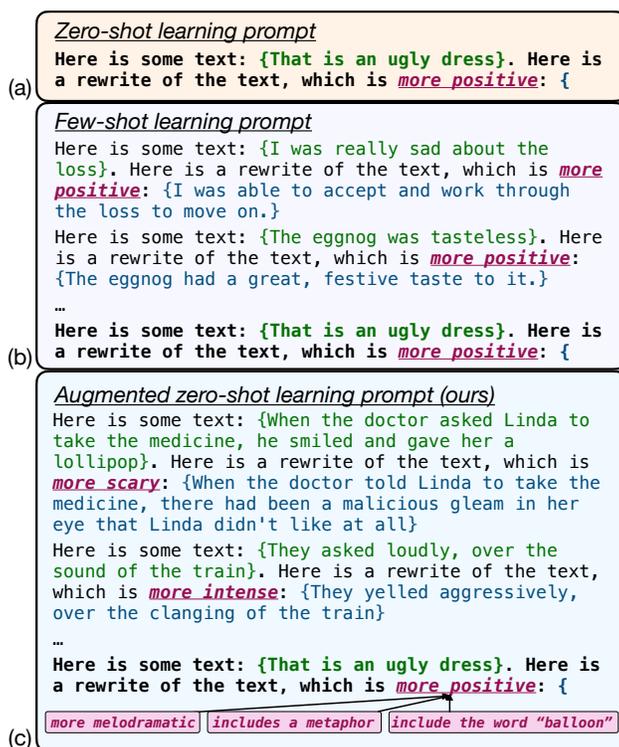


Figure 5.1: Zero-shot, few-shot, and augmented zero-shot prompts for style transfer. The boldface text is the zero-shot prompt, and the plain text is the additional priming sequence. The full prompts used in this paper are shown in Table 5.8. We encourage readers to examine the outputs of our model at <https://bit.ly/3fLDuci>.

To remove the need for these labeled exemplars for each style transfer task, we propose *augmented zero-shot learning*, a method for performing multi-task style transfer using a single set of exemplars. Instead of prompting the model with exemplars specific to the exact style transfer task we wish to perform, we prompt the model with examples of a variety of sentence rewriting operations, as shown in Figure 5.1c. This intuition is inspired by Reynolds and McDonell [124]’s observation that successful prompts constrain the behavior of the large LM away from failure modes—in our case, we aim to preserve the flexibility of a zero shot prompt while encouraging the model to produce outputs of a specific template. We keep the the format of the exemplars constant and insert the desired sentence transformation into the same format. In this way, the augmented zero-shot formulation supports arbitrary sentence rewriting tasks without the need to write any task-specific exemplars. Thus, it works for a wide range of styles, including modifying the text to be “*more melodramatic*,” “*insert a metaphor*,” or “*include the word balloon*.”

Augmented zero-shot learning is simple and facilitates the application of style transfer to a wider range of styles than existing work. We also foresee that our strategy of prompting an LLM with non-task specific examples will prove useful for other tasks besides style transfer.

5.1.1 Experimental Setup

Style Transfer Tasks

We consider six style transfer tasks that we deem non-standard, listed in Table 5.1. These styles were chosen to be representative of most frequent style adjustments made by users of an AI-assisted text editor that employs our method (discussed further in §??). As source sentences, we use 50 sentences randomly drawn from the Reddit Writing Prompts validation set [42], excluding those that already clearly exhibited one of the styles or were ungrammatical/incoherent. We use human evaluation for these styles, since not all styles have readily available classifiers.

Table 5.1: Example style transfer outputs from augmented zero-shot learning for non-standard styles.

| | |
|-------------------------------------|---|
| Original | There, in the middle of the street, stood an old man in a weatherbeaten brown coat. |
| More descriptive | The man’s brown trenchcoat was coated with snow, and as he leaned against the telephone booth, he cast a long shadow across the road in the late afternoon sun. |
| Include a metaphor | There, in the middle of the street, stood an old man whose face resembled the weatherworn bark of a gnarled tree. |
| Include the word “ <i>park</i> ” | There, in the middle of Central Park, stood an old man in a weatherbeaten brown coat. |
| Include the word “ <i>balloon</i> ” | There, in the middle of the street, stood an old man with several colourful balloons tied to the straps of his coat. |
| More comic | There, in the middle of the street, stood an old man in a raggedy brown coat that was much too big for him. |
| More melodramatic | There, in the middle of the road, stood a grizzled old man, the light of life faded from his sunken eyes. |

We also evaluate our method on two standard style transfer tasks: sentiment and formality. We use the Yelp polarity dataset [178] for sentiment and Grammarly’s Yahoo Answers Formality Corpus (GYAFC) dataset for formality [122].¹³ These datasets allow us to evaluate performance of augmented zero-shot learning in the context of prior supervised methods which have been used on these tasks.

Model

Augmented zero-shot learning requires a large language model. We primarily use LaMDA, a left-to-right decoder-only transformer language model [153] with a non-embedding parameter count of 137B [149]. The pre-trained LaMDA model, which we refer to as *LLM*, was trained on a corpus comprising 1.95B public web documents, including forum and dialog data and Wikipedia. The dataset was tokenized into 2.49T BPE tokens with a SentencePiece vocabulary size of 32K [80]. We also use *LLM-Dialog*, the final LaMDA model which was finetuned on a curated, high-quality subset of data identified to be in a conversational format. Decoding was done with top- $k=40$. To show that the success of augmented zero-shot learning is not restricted to these

¹³ Hosted by Luo et al. [97].

two large LMs, we also perform experiments with GPT-3 (Table 5.3). For GPT-3, decoding was done with nucleus sampling using $p=0.6$ [61].

The full prompts used for *LLM* and GPT-3 are shown in Figure 5.8. For *LLM-Dialog*, the prompt was instead formulated as a conversation between one agent who is requesting rewrites and another who is performing the rewrites.

5.1.2 Results

NON-STANDARD STYLES For our six non-standard styles, we asked six professional raters to assess <input sentence, target style, output sentence> tuples. These raters are fluent in English, live in India, and work full time labeling and evaluating data. To decrease inter-rater discrepancy and ensure that our instructions were clear, we had an initial calibration session where they test-rated a small portion of the data (around 10 datapoints which were then omitted from the results) and asked us any clarifying questions. For each style, we compare outputs from our method plus the three baselines for 50 sentences.

Each tuple was scored by three raters (3,600 ratings total) on the following three axes which are standard to textual style transfer [103]:

1. **transfer strength** the amount that the output actually matches the target style
2. **semantic preservation** whether the underlying meaning of the output text, aside from style, matches that of the input and
3. **fluency** whether the text is coherent and could have been written by a proficient English speaker

Following Sakaguchi and Van Durme [129], transfer strength and semantic preservation were rated on a scale from 1–100. A screenshot of the evaluation UI is shown in Figure 5.2. We use *dialog-LLM*, and compare it with three other methods:

1. **zero-shot** a baseline where no exemplars are provided

Instructions: In this task, your goal is to identify whether a desired transformation has been successfully applied to a sentence, without changing the overall meaning of the sentence. Each question contains a sentence marked "original sentence," a desired transformation, and an output sentence where the transformation has been applied.

Each of these questions relates to the same original text and desired transform, but each has a different output transformed sentence. Please rate each transformed sentence along the following three axes:

1) Transferred Style Strength: Does the transformed text has the applied style/transform compared to the original text? For example, if the original text is "I went to the store" and the style is "more angry":

| example | score | reasoning |
|--|-------|---|
| "The store is where I went" | 0 | The transformed text is no more angry than the original text. |
| "I went to the stupid store" | 50 | The transformed text somewhat relates to the style. |
| "When I went to the store, I couldn't believe how rude the storekeeper was to me!" | 100 | The text is clearly more angry. |

2) Meaning: Does the transformed sentence still have the same overall meaning as the original? It is OK if extra information is added, as long as it doesn't change the underlying people, events, and objects described in the sentence. You should also not penalize for meaning transformations which are necessary for the specified transformation. For example, if the original text is "I love this store" and the style is "more angry":

| example | score | reasoning |
|---|-------|---|
| "It is raining today" | 0 | the transformed text is about something totally different. It would be hard to tell that the texts are related at all. |
| "they were out of chicken at the store" | 50 | The transformed text is mostly related to original- some modifications of the meaning have been made but they are not egregious |
| "I adore the store." or "The store was really horrible; it took forever to do my shopping." | 100 | The text talks about the same concepts as the original, just with different or more words |

3) Fluency: Is this sentence fluent english and does it make sense?

| example | score | reasoning |
|--|-------|------------------------|
| "who said that? I thought we were going to go together!" | Yes | This text makes sense |
| "who, she said it up to me and to me together!" | No | The text is incoherent |

Original text: "Everyone in my world had different eye colours."

Desired transformation: more melodramatic

Transformed text: "Everyone in my world had the most intensely colorful eyes, and no one in this world can possibly understand how beautiful they were."

1) Transferred Style Strength: The transformed text has the applied style/transform.

50

_____●_____

2) Meaning: The meaning is preserved between the original and transformed texts (ignoring the ways that the style/transform would change the meaning)

50

_____●_____

3) Fluency: the transformed text is fluent English and it makes sense.

Yes

No

Figure 5.2: The rating UI used for human evaluation. The user may be shown a number of blue squares at once with the same original text and different outputs.

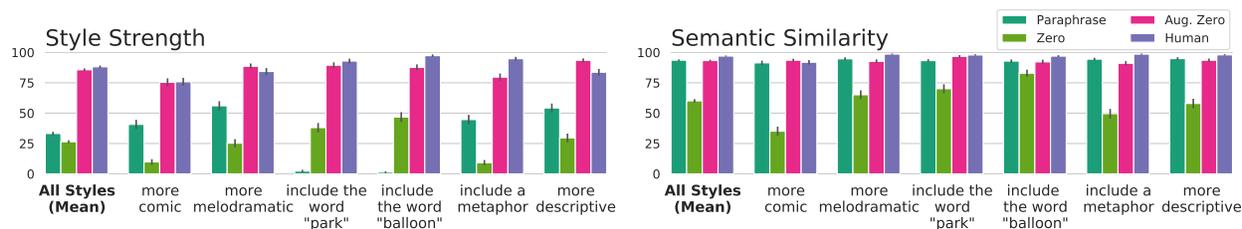


Figure 5.3: Human evaluation of style transfer for six atypical styles. Our method is rated comparably to the human-written ground truth. Error bars are mean standard error.

2. **paraphrase** our normal augmented zero shot prompt, but with the target style of “*paraphrased*”, as a control
3. **human** ground-truth transformations written by the authors

Figure 5.3 shows these results. We found that the outputs of our method were rated almost as highly as the human-written ground truth for all three evaluations. The zero-shot baseline performed the worst in all categories: 25.4% of the time, it did not return a valid response at all (see §5.1.3), compared with 0.6% for augmented zero shot. The strong performance of the paraphrase baseline at fluency and semantic similarity shows that large LMs are capable of generating high quality text that remains true to the input sentence’s meaning. Overall, the average length of the input sentences was 66 characters, whereas the average length of augmented zero-shot outputs was 107 characters. For context, human paraphrase outputs were 82 characters.

For a subset of the tasks, some automatic evaluation was also possible. We found that the “*balloon*” and “*park*” transformations successfully inserted the target word 85% of the time. For “*more descriptive*” and “*include a metaphor*” the transformed text was, as expected, longer than the original (by 252% and 146% respectively, compared with 165% and 146% for human baselines).

STANDARD STYLES To better contextualize the performance of our method with prior methods, we also generated outputs for two standard style transfer tasks: sentiment and formality. Figure 5.4 shows human evaluations (same setup as before) for our outputs as well as the outputs from two popular prior style transfer methods, Unsup MT [115] and Dual RL [97]. The outputs from our method were rated comparably to both

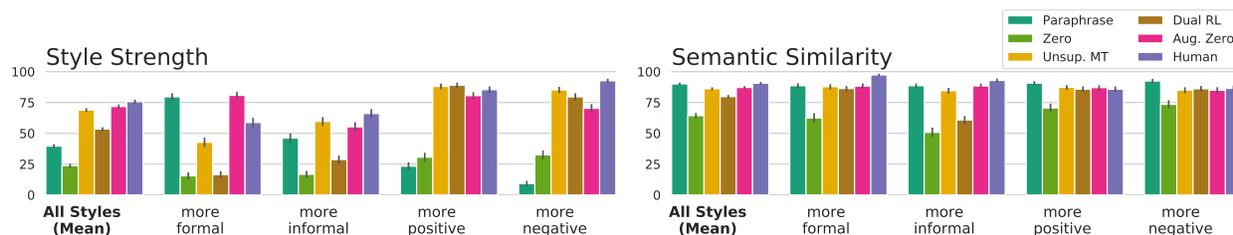


Figure 5.4: Human evaluation of sentiment and formality transfer. Our method is rated comparably to human-written ground truth as well as prior methods. Error bars show Standard Error of the Mean. Unsup. MT is Prabhume et al. [115]; Dual RL is Luo et al. [97].

human generated responses and the two prior methods, using the same rating setup as the non-standard styles, with six outputs and baselines for four styles across 50 sentences, rated independently by three raters, totalling 3,000 total ratings.

Furthermore, following Li et al. [90] and Sudhakar et al. [143], we perform automatic evaluation for sentiment style transfer since there are classifiers available for these styles. We note that although automatic evaluations can diverge from human ratings, they can still be a good proxy as we could not perform human evaluation against every prior method due to time and resource constraints. We automatically evaluate

1. **transfer strength** using a sentiment classifier from HuggingFace Transformers [165]
2. **semantic similarity** to human examples provided by Luo et al. [97] via BLEU score
3. **fluency** measured via perplexity, as predicted by GPT-2 (117M).

Table 5.2 shows these automatic evaluations, with four main takeaways. First, augmented zero-shot prompting achieves high accuracy and low perplexity compared with baselines. The BLEU scores, however, the outputs of our model had low BLEU scores with respect to human generated outputs 5.2. Based on qualitative examination of outputs, we believe that this is because our model outputs often used different language from human annotations, despite having high semantic similarity with the source sentence. For instance, for transferring the sentiment of “*ever since joes has changed hands it’s just gotten worse and worse*” to positive sentiment, our augmented zero-shot learning model outputted “the establishment has continued to provide excellent service, improving steadily since its change of ownership.” This will have low BLEU with

Table 5.2: Comparing augmented zero-shot prompting with supervised style transfer methods on the Yelp sentiment style transfer dataset using automatic evaluation. Acc: accuracy; PPL: perplexity. The inference-only table shows our method applied to 3 different sizes of GPT-3, plus our own LLM.

| | Acc | BLEU | PPL |
|-------------------------------|------|------|-------|
| <u>SUPERVISED METHODS</u> | | | |
| Cross-alignment [138] | 73.4 | 17.6 | 812 |
| Backtrans [115] | 90.5 | 5.1 | 424 |
| Multidecoder [46] | 50.3 | 27.7 | 1,703 |
| Delete-only [90] | 81.4 | 28.6 | 606 |
| Delete-retrieve [90] | 86.2 | 31.1 | 948 |
| Unpaired RL [168] | 52.2 | 37.2 | 2,750 |
| Dual RL [97] | 85.9 | 55.1 | 982 |
| Style transformer [33] | 82.1 | 55.2 | 935 |
| <u>INFERENCE-ONLY METHODS</u> | | | |
| GPT-3 ada, aug zero-shot | 31.5 | 39.0 | 283 |
| GPT-3 curie, aug zero-shot | 53.0 | 48.3 | 207 |
| GPT-3 da vinci, aug zero-shot | 74.1 | 43.8 | 231 |
| LLM: zero-shot | 69.7 | 28.6 | 397 |
| five-shot | 83.2 | 19.8 | 240 |
| aug zero-shot | 79.6 | 16.1 | 173 |
| LLM-dialog: zero-shot | 59.1 | 17.6 | 138 |
| five-shot | 94.3 | 13.6 | 126 |
| aug zero-shot | 90.6 | 10.4 | 79 |

the ground truth with respect to human references, which is simply “*ever since joes has changed hands it’s just gotten better and better.*” Though we do not see this as an inherent problem, increasing the BLEU for the purposes of comparison can be done in an easy way via candidate selection, as our model returns sixteen possible continuations. In applications for which we prefer model outputs to have high lexical similarity to the source sentence, we could select the candidate of the sixteen with the highest BLEU score compared with the original source sentence. We find that this candidate selection step can substantially improve the BLEU score with the ground truth target sentences, as we show in Table 5.3.

Table 5.3: Sentiment style transfer results with candidate selection (cand. select.). Candidate selection means that of the sixteen examples returned by our model, we choose the one with the highest BLEU with the source sentence.

| | Acc | BLEU | PPL |
|------------------------|------|------|-----|
| <u>LLM-128B</u> | | | |
| Zero-shot | 69.7 | 28.6 | 397 |
| + cand. select. | 31.4 | 61.5 | 354 |
| Five-shot | 83.2 | 19.8 | 240 |
| + cand. select. | 61.5 | 55.6 | 306 |
| Augmented zero-shot | 79.6 | 16.1 | 173 |
| + cand. select. | 65.0 | 49.3 | 292 |
| <u>LLM-128B-dialog</u> | | | |
| Zero-shot | 59.1 | 17.6 | 138 |
| + cand. select. | 46.8 | 24.2 | 166 |
| Five-shot | 94.3 | 13.6 | 126 |
| + cand. select. | 81.3 | 47.6 | 345 |
| Augmented zero-shot | 90.6 | 10.4 | 79 |
| + cand. select. | 73.7 | 40.6 | 184 |

Table 5.4: For sentiment style transfer, we show accuracy, BLEU, perplexity (PPL), and sequence length (l) reported as the average of positive \rightarrow negative and negative \rightarrow positive. In (a), augmented zero-shot prompting of GLM and GPT-3 are compared with prior supervised style transfer methods. In (b), we compare zero-shot, five-shot, and augmented zero-shot prompting for our GLM models. Candidate selection means that of the sixteen examples returned by the API, we choose the one with the highest BLEU with the source sentence (in the default case, we just use the first returned output).

| | Acc | BLEU | PPL | l |
|--------------------------------------|------|------|-------|------|
| Cross-alignment [138] | 73.4 | 17.6 | 812 | 10.7 |
| Backtrans [115] | 90.5 | 5.1 | 424 | 9.7 |
| Multidecoder [46] | 50.3 | 27.7 | 1,703 | 10.4 |
| Delete-only [90] | 81.4 | 28.6 | 606 | 12.1 |
| Delete-retrieve [90] | 86.2 | 31.1 | 948 | 11.7 |
| Unpaired RL [168] | 52.2 | 37.2 | 2,750 | 9.4 |
| Dual RL [97] | 85.9 | 55.1 | 982 | 10.1 |
| Style transformer [33] | 82.1 | 55.2 | 935 | 10.3 |
| <u>Augmented Zero-Shot Prompting</u> | | | | |
| GPT-3 ada | 31.5 | 39.0 | 283 | 10.0 |
| GPT-3 curie | 53.0 | 48.3 | 207 | 10.3 |
| GPT-3 da vinci | 74.1 | 43.8 | 231 | 10.4 |

(a)

| | Acc | BLEU | PPL | l |
|----------------------------|-------------|-------------|-----------|------|
| <u>GLM</u> | | | | |
| Zero-shot | 69.7 | 28.6 | 397 | 9.6 |
| + cand. selection | 31.4 | 61.5 | 354 | 10.2 |
| Five-shot | 83.2 | 19.8 | 240 | 11.2 |
| + cand. selection | 61.5 | 55.6 | 306 | 10.2 |
| Augmented zero-shot | 79.6 | 16.1 | 173 | 11.3 |
| + cand. selection | 65.0 | 49.3 | 292 | 10.3 |
| <u>GLM-dialog</u> | | | | |
| Zero-shot | 59.1 | 17.6 | 138 | 17.2 |
| + cand. selection | 46.8 | 24.2 | 166 | 17.2 |
| Five-shot | 94.3 | 13.6 | 126 | 13.3 |
| + cand. selection | 81.3 | 47.6 | 345 | 10.0 |
| Augmented zero-shot | 90.6 | 10.4 | 79 | 12.8 |
| + cand. selection | 73.7 | 40.6 | 184 | 10.7 |

(b)

Second, we apply augmented zero-shot learning to GPT-3 175B; these results indicate that augmented zero-shot learning generalizes to another large language model. Third, we vary model size for GPT-3 models, finding that larger size greatly improves style transfer. Fourth, for *LLM* and *LLM-dialog*, we find that augmented zero-shot learning substantially outperforms vanilla zero-shot learning and almost reaches the accuracy of five-shot learning.

COMPARISON WITH A RANGE OF PRIOR METHODS To compare against a larger range of prior supervised methods, we used automatic evaluation, and found comparable performance with the highest-scoring method for transfer strength. The results are shown in Table ???. We were also significantly more

fluent than all other methods. Finally, our method fell short on semantic preservation compared to other methods. However, BLEU is known to penalize long sentences, and the scores do not always align with human judgements. For example, our model’s worse performance could be because it was not explicitly trained on Yelp data, so its generations are less likely to be in the style of Yelp reviews than models that were.

COMPARISON ACROSS DIFFERENT LLMS We also compared between three varieties of model: GPT-3 [19], LLM, and LLM-Dialog. We adjusted the prompt template slightly to accommodate these differences: for LLM and GPT-3, the prompt template replaced “Rewrite it to be **<style>**” with “Here is a rewrite of the text, which is **<style>**”. For our augmented zero-shot prompts we also see that the LLM-dialog version had higher accuracy than the LLM and GPT-3, but lower BLEU. Based on qualitative inspection, we believe the lower BLEU is due to the LLM-dialog adding additional detail in the generated sentences, which is consistent with an “interestingness” objective that is typically encoded into dialog training.

PROMPT CONSTRUCTION Prompt engineering can be brittle: Reynolds and McDonell [124] describe how reformulating the language of a prompt can have significant impact on performance, and that finding the right prompt is for a task is more akin to locating an already-learned task than truly learning a new one. To explore this, we compared several variations of the prompts for sentiment, varying the language of the prompt to use “*more positive/negative*,” “*happier/sadder*,” “*more optimistic/pessimistic*,” or “*more cheerful/miserable*.” As shown in Table 5.6, performance differed across the four prompts, but we found them comparable. In a real world setting, our augmented zero-shot approach allows users to effortlessly try out many different phrasings for the task until they find one that performs satisfactorily.

Reynolds and McDonell [124] further emphasize that prompt engineering is mostly about avoiding various failure cases. In this work, we use delimiters (“{” and “}”) to help avoid parsing errors, giving scores of zero when there was no valid responses with such delimiters. There are other delimiters that could be used (e.g., quotes, “(” and “)”, “<” and “>”, newlines with a colon (as used by GPT-3), etc. We chose curly braces as they were 1) likely to occur in the training data as delimiters in other contexts and 2) not frequently part of

Table 5.6: Comparing variations of augmented zero-shot learning prompt wording for sentiment style transfer.

| Model / prompt wording | Acc | Bleu | PPL |
|-------------------------------|------|------|-----|
| <u>LLM</u> | | | |
| “more positive/negative” | 76.3 | 14.8 | 180 |
| “happier/sadder” | 62.6 | 15.5 | 173 |
| “more optimistic/pessimistic” | 69.7 | 14.1 | 143 |
| “more cheerful/miserable” | 74.5 | 15.7 | 186 |
| <u>LLM-Dialog</u> | | | |
| “more positive/negative” | 90.5 | 10.4 | 79 |
| “happier/sadder” | 85.9 | 9.6 | 90 |
| “more optimistic/pessimistic” | 85.8 | 10.2 | 79 |
| “more cheerful/miserable” | 88.8 | 11.4 | 93 |

the input sentence itself. We also use a second person prompt template for the dialog, which yielded better results as it was more similar to the training data. Exploring these options more quantitatively would be an interesting direction for future work.

5.1.3 Limitations and Failure Modes

There are several limitations and failure modes with our method.

MALFORMED GENERATIONS A frequent problem that arises when using large LMs for other NLP tasks is their outputs cannot be automatically parsed into usable answers. For example, when given a prompt like “Here is some text: that is an ugly dress. Here is a rewrite of the text, which is more positive” *LLM-Dialog* might return something like “Sounds like you are a great writer!” Similar error modes exist for *LLM*, which might output something like “Here are more writing tips and tricks.” Other times, the response contains correct information, but it cannot be automatically parsed (e.g., “a good rewrite might be to say that the dress is pretty.”) In hindsight, these outputs make a lot of sense:

Table 5.7: Examples of users’ arbitrary style transfer requests for which the model suggestion was accepted.

| Input | Model suggestion |
|---|---|
| <p>You honor me,” said Bruno. As he spoke, he began to laugh. His laugh was merry sounding and melodic. He took a deep breath. When he spoke again, his voice was low and strong. “I take this great task on and with my brothers shall lead us on to victory!</p> <p>Rewrite it to be about mining.</p> | <p>You honor me, said Bruno. As he spoke, he began to laugh. His laugh was deep and gravely, as if from the bowels of the earth. He took a deep breath. When he spoke again, his voice was deep and dark. ‘I take this great task on and with my brothers shall delve into the mines!’”</p> |
| <p>I reckon,’ smiled Toe, ‘you’d have to say the same for icicles, wouldn’t you?</p> <p>Rewrite it to be more interesting.</p> | <p>‘I reckon,’ smiled Toe, ‘you’d have to say the same for icicles, wouldn’t you?’ These words were barely out of his mouth before the others realized exactly what they meant.</p> |
| <p>with its southern perfume</p> <p>Rewrite it to be more descriptive</p> | <p>A mixture of sweet honeysuckles and jasmine with a hint of tobacco and warm mesquite smoke. It was his southern perfume, with a hint of a memory of simpler times.</p> |
| <p>But Sally remained unmoved—and lonely.</p> <p>Rewrite it to be more emotional</p> | <p>Her tears rolled and rolled down her cheek, mixing with the saltwater around her, until the water was a rich mixture of the two. She wanted her poem to be cheerful like it used to be, but couldn’t find the right words.</p> |

most of the training data of large LMs is not well-formatted pairs of inputs and outputs [124]. See §5.1.2 for how we dealt with these issues.

HALLUCINATIONS Large LMs are known to hallucinate text content; we saw this happen frequently for style transfer. While this is an advantage in some contexts like creative writing, it is undesirable for applications like summarization.

Table 5.8: In black, we show the exact augmented-zero shot prompts used in our experiments, for *LLM* and GPT-3 (top), and for *LLM-Dialog* (bottom). As shown, for *LLM-Dialog*, we replaced “Here is a rewrite of the text, which is” with “Rewrite it to be”. Each line starting with “>” above was passed in as an individual dialog turn. The blue shows how an input text and goal style are concatenated to the few-shot prompt in order to produce final model output. Note that we can achieve high accuracy even though the prompt formulation resulted in some minor grammatical errors for some styles (e.g., “rewrite it to be include the word ‘snow’”). Text versions of these prompts can be downloaded at <https://bit.ly/3fLDuci>.

| |
|--|
| Augmented Zero-shot Prompt: LLM |
| <pre>Here is some text: {When the doctor asked Linda to take the medicine, he smiled and gave her a lollipop.}. Here is a rewrite of the text, which is more scary. {When the doctor told Linda to take the medicine, there had been a malicious gleam in her eye that Linda didn't like at all.} Here is some text: {they asked loudly, over the sound of the train.}. Here is a rewrite of the text, which is more intense. {they yelled aggressively, over the clanging of the train.} Here is some text: {When Mohammed left the theatre, it was already dark out.}. Here is a rewrite of the text, which is more about the movie itself. {The movie was longer than Mohammed had expected, and despite the excellent ratings he was a bit disappointed when he left the theatre.} Here is some text: {next to the path.}. Here is a rewrite of the text, which is about France. {next to la Siene} Here is some text: {The man stood outside the grocery store, ringing the bell.}. Here is a rewrite of the text, which is about clowns. {The man stood outside the circus, holding a bunch of balloons.} Here is some text: {the bell ringing}. Here is a rewrite of the text, which is more flowery. {the peales of the jangling bell} Here is some text: {against the tree}. Here is a rewrite of the text, which is include the word "snow". {against the snow-covered bark of the tree} Here is some text: {That is an ugly dress}. Here is a rewrite of the text, which is more positive."</pre> |
| Augmented Zero-shot Prompt: LLM-dialog |
| <pre>> Here is some text: {When the doctor asked Linda to take the medicine, he smiled and gave her a lollipop.}. Rewrite it to be more scary. > {When the doctor told Linda to take the medicine, there had been a malicious gleam in her eye that Linda didn't like at all.} > Here is some text: {they asked loudly, over the sound of the train.}. Rewrite it to be more intense. > {they yelled aggressively, over the clanging of the train.} > Here is some text: {When Mohammed left the theatre, it was already dark out.}. Rewrite it to be more about the movie itself. > {The movie was longer than Mohammed had expected, and despite the excellent ratings he was a bit disappointed when he left the theatre.} > Here is some text: {next to the path.}. Rewrite it to be about France. > {next to la Siene} > Here is some text: {The man stood outside the grocery store, ringing the bell.}. Rewrite it to be about clowns. > {The man stood outside the circus, holding a bunch of balloons.} > Here is some text: {the bell ringing}. Rewrite it to be more flowery. > {the peals of the jangling bell} > Here is some text: {against the tree.}. Rewrite it to be include the word "snow". > {against the snow-covered bark of the tree} > Here is some text: {That is an ugly dress}. Rewrite it to be more positive."</pre> |

Inherent style trends

We also noticed that even our “*paraphrase*” baseline, where the model was simply asked to rewrite the input sentence, was rated highly for style strength for a few styles, including “*more formal*” and “*more melodramatic*”. This implies that our method’s generations generally trend toward these styles. A direction for future work would be to see what styles and qualities of text our method (and large LMs in general) are inherently more likely to produce.

LESS RELIABLE THAN TRAINED METHODS For style transfer tasks that have available training data, prior methods that either train or finetune on that data are going to be inherently more reliable at producing

text that looks like their training data. This can be observed in the lower BLEU scores our method achieves than trained methods, despite comparable transfer accuracy. Thus, augmented zero-shot learning offers less fine-grained controllability in the properties of the style-transferred text than methods which see task-specific training data.

LARGE LM SAFETY CONCERNS Large LMs themselves come with their own host of difficulties, barriers to entry, and potential safety concerns as discussed by Bender et al. [11], which are also valid for this style transfer method. However, we also think that this method can be a useful tool in exploring and exposing the safety and boundaries of these models themselves: what happens if we try to force the large LM to make a text “more racist”, “more sexist”, or “more incendiary”? It is important to keep pushing these models to their boundaries to see where they fail and where problems arise, and specific use cases that show a broader range of the model’s capabilities also show a broader range of its failure modes.

5.2 MODELS FOR INFILLING TEXT

Natural language generation systems are increasingly being incorporated into applications where a human writer and an AI jointly collaborate to construct text. Wordcraft, The AI-assisted text processor I describe in Section 5.3 is one such application. Another is Storium, where players of a writing game have the option to accept suggestions from a natural language generation system [3]. There are also more practical domains such as email composition assistance and code synthesis [21, 166, 7]. Many of these applications are limited to generating text at the end of what has been written so far. This is because both historical n -gram language models (LMs) and state-of-the-art neural LMs are typically designed to produce text by repeatedly predicting the next word in a sequence given the previous words. However, there is a need for more powerful interactive tools which enable writers to solicit insertions at any chosen position within the existing text, a task variously referred to as fill in the blank (FITB), infilling, or the Cloze task [146]. For example, a creative writer might

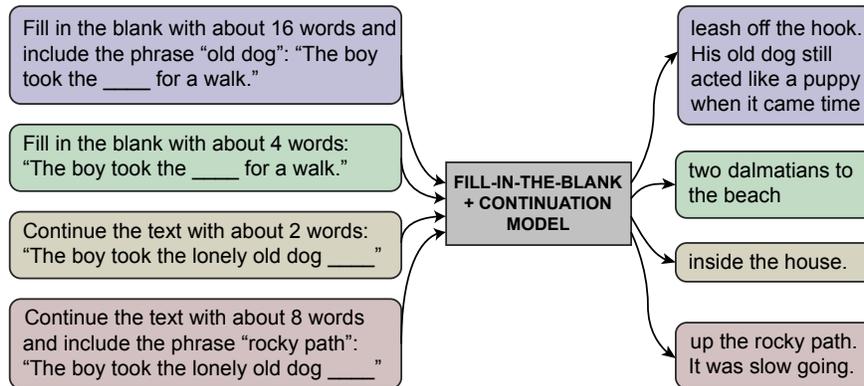


Figure 5.5: A single model that can handle a variety of related writing tasks is more efficient than separate models per task.

want a tool which can insert a description of a place or character, and a programmer might want a system that can fill in a method in the middle of their code.

Most prior work tackling FITB consider it a separate task from continuation, one to be specifically optimized for, for example training a custom model from scratch [67, 181, 104], finetuning a model trained originally for continuation [39], or using a combination of pre-trained models [64]. Having separate trained models for FITB and for continuation is inefficient for downstream applications where maintaining multiple neural networks can be prohibitive.

Any model that can do FITB can be made to do continuation simply by placing the blank at the end of the input. Thus, I describe how models trained on FITB can be employed effectively for both infilling and continuation operations. I show how T5 [119], one of the most popular pre-trained models, can reasonably handle both tasks, as it was pre-trained with a FITB-like objective. Finetuning T5 further improves its ability and also allows for the incorporation of controllability of generation length and word choice.

Table 5.9: Examples of the finetuning objectives. “8” is the approximate length in words of the target sequence. During finetuning, about 25% of training examples took each of these formats.

| Example Type | Input | Target |
|--------------------------|--|--|
| C4FILLBLANK no goal | fill: I love avocados. I ate a sandwich covered in them. _8_ I talked to my doctor about it later. It turned out I was allergic to avocados. | After I ate it, my mouth was itchy and tingly. |
| C4FILLBLANK with goal | fill: I love avocados. I ate a sandwich covered in them. _8_ I talked to my doctor about it later. It turned out I was allergic to avocados. Goal: mouth was itchy | After I ate it, my mouth was itchy and tingly. |
| C4FILLBLANK no goal | fill: I love avocados. I ate a sandwich covered in them. After I ate it, my mouth was itchy and tingly. I talked to my doctor about it later. _8_ | It turned out I was allergic to avocados. |
| C4FILLEND with goal | fill: I love avocados. I ate a sandwich covered in them. After I ate it, my mouth was itchy and tingly. I talked to my doctor about it later. _8_ Goal: allergic to | It turned out I was allergic to avocados. |

5.2.1 Supporting FITB and Continuation

We define filling in the blank as the task of predicting text to replace a single missing span, usually demarcated with a special token, in an input text passage. (Some prior work considers inputs with multiple blanks, but inserting text at one position at a time better matches the kinds of edits humans do.) We define continuation in the traditional language modeling sense—predicting the next token in a sequence given only the previous tokens. Donahue et al. [39] discuss how language modeling is a special case of infilling, and they use this as justification to finetune a continuation-based language model to do infilling. However, we argue that if continuation is a subtask of infilling, it makes more sense to go in the opposite direction: prioritize a model which can do infilling and check that it achieves satisfactory performance at continuation.

T5 is a model pre-trained with a “span corruption” objective very similar to FITB; the model is asked to reconstruct the missing text after randomly chosen substrings of the input are replaced with special identifiers. Thus, a pre-trained T5 model can be used without any further training to do both continuation and infilling by appropriately choosing text to mask out. The encoder-decoder architecture of T5 is also more conducive

to FITB than decoder-only architectures like GPT-2 [118] which are typically used for continuation-based language models. This is because the attention mechanism in encoder-decoder architectures allows the context on the left side of the blank to attend to the context on the right, while decoder-only architectures only support masked attention (each token can only attend to the positions to its left).

Even though T5’s pre-training objective was a form of FITB, finetuning is still advantageous. For one, our definition of FITB only includes a single masked out substring, not multiple, so finetuning improves alignment with the goal task. Finetuning also allows us to incorporate additional conditioning signals not supported by the pre-trained T5, such as being able to specify the desired length of the generated text or specify words that ought to be included in the blank, a task we refer to as “goal conditioning.” Length control, which comes by default in a traditional language model by simply sampling more or fewer tokens, is particularly necessary for FITB, where the end of the generation must fit seamlessly with the text to its right.

The biggest language models available today were largely trained in the continuation rather than the FITB paradigm [18, 12]. Since our primary goal is to have a single model for both tasks, we also address the question: if a continuation-trained model is big enough, can it handle FITB without the need for finetuning? Few-shot learning with large language models, as popularized by Brown et al. [18], has had success on many tasks in NLP. We try out this approach for FITB by designing a few-shot prompt containing several demonstrations of the FITB task, formulated in a similar “infilling by language modelling” style as Donahue et al. [39].

5.2.2 Experimental Setup

MAIN SETUP For all primary experiments, we use the 800M parameter v1.1 ‘large’ model. We also show some additional results comparing against the 3B parameter ‘XL’ T5 model. To finetune T5 for FITB, we construct training examples from documents by first partitioning the document text into a left context, gap, and right context. The input sequence is then the left and right contexts concatenated with textual representations of the additional conditioning signals. The target sequence is the true text for the blank. This formulation easily supports continuation, as the blank can be deliberately placed at the end (i.e., providing no

right context). Documents are drawn from C4, the same dataset T5 was pre-trained on. Documents are split into word sequences, and these are then randomly truncated to be between 256-512 words long. A substring of between 1 and 64 words is selected to be blanked out. For half of the training examples the blank is randomly selected, and for the other half it is always placed at the end. To support length conditioning, we follow Roberts and Raffel [126] and include a bucketed version of the target length as part of the blank. To support goal conditioning, for half the examples, a random substring of up to half the words of the target is appended to the end of the input. Examples are shown in Table 5.9.

We compare T5 against a state-of-the-art 137B parameter decoder-only language model (LLM) trained explicitly for continuation and used successfully for few-shot learning in other domains [7, 123]. This model is used (1) as a standard continuation model, prompting with only the left context of an example; and (2) in a few-shot learning paradigm.

FEW-SHOT LEARNING SETUP We experimented with prompts randomly selected from the C4, Reddit Writing Prompts, and ROC Stories training sets, as well as prompts consisting of examples handwritten by the authors with the goal of story-writing in mind. For each prompt source, we randomly generated five possible prompts, each with three examples. To simplify the task, we conditioned on desired length but did not include goal conditioning. When choosing random few-shot prompts from the three train sets, in order to keep the few-shot prompt text within the 512-token context length limit of the LLM[149] we used for inference, we only considered examples that contained 100 or fewer tokens, so that the max length of the few-shot prompt was no more than 300 tokens. This left 212 tokens for the text of the actual example we were interested in performing the FITB task on. For our hand-written prompt, we wrote the seven examples shown in Table 5.11. We generated 5 possible prompts by randomly subsampling 3 examples out of these 7 five times. Table 5.10 shows the perplexity of the generations from each few-shot prompt. We note that even leaving room for 212 tokens worth of context text, some evaluation examples did not fit in the prompt length, and these examples were skipped when doing this analysis. Based on these results, we use the best-performing prompts from ROCFILLMIDDLE and from C4FILLBLANK for comparison with the other methods.

Table 5.10: Perplexity of evaluation sets when the blank has been filled in using LLM with few-shot prompting (top) and our best fine-tuned T5 model ((bottom). Among the few-shot results, the best method for each dataset is bolded, as well as methods within one standard error.

| Few-shot source: | C4FILL | ROC FILL | RWP FILL | RWP FILL |
|-------------------------|--------------|--------------|--------------|--------------|
| | BLANK | MIDDLE | BLANK | BLANK-Sent |
| C4FILLBLANK | 15.67 | 19.72 | 19.65 | 16.82 |
| ROC FILL MIDDLE | 14.14 | 19.61 | 19.48 | 16.36 |
| RWP FILL BLANK | 24.39 | 20.29 | 32.33 | 28.13 |
| RWP FILL BLANK-Sent | 18.91 | 18.21 | 24.44 | 19.87 |
| FS CUSTOM | 17.98 | 19.80 | 21.72 | 18.38 |
| Finetuned T5 XL | 9.99 | 19.00 | 13.64 | 10.03 |
| Finetuned T5 Large | 10.33 | 20.47 | 14.08 | 10.37 |

EVALUATION We evaluate continuation and FITB on C4 as well as two story writing datasets, as creative writing assistant applications are one of the key areas we expect to benefit from multi-task models [30]. Reddit Writing Prompts (RWP) is a corpus of stories from the ‘r/WritingPrompts’ sub-Reddit [42], and we construct validation sets RWP FILL BLANK and RWP FILL END using the same method described in the previous section. C4 and RWP validation sets are capped to 5,000 examples. ROC Stories (ROC) is a crowd-sourced dataset of five-sentence commonsense stories [105]. For ROC Stories, the 2018 validation set is used to construct ROC FILL MIDDLE, where the middle sentence of each story is blanked out, and ROC FILL END, where the last sentence is blanked out. Unless otherwise noted, all evaluation is done without goal conditioning and uses random sampling with top- $k=50$ as the decoding strategy.

5.2.3 Results

FAILURE OF FEW-SHOT PROMPT FOR FILL IN THE BLANK Filling in a blank seems like a task that ought to be easy to accomplish with few-shot learning techniques. Training data for large language models often contains fill-in-the-blank style examples, as school lessons with cloze-style questions are relatively

| | Context | Target |
|---|--|---|
| 1 | An elderly man was sitting alone on a dark path. The man looked down at his feet, and realized ____ . It was a plain pine box and looked as if it had been there for a long time. The man was afraid to look inside the box. | he was holding a bright red box made of pine |
| 2 | The mantle was cluttered with objects: ____ and more than one vase of dried flowers. The bejeweled lamp was at the very back, nearly invisible. | picture frames showing grandchildren and long-ago weddings, knick-knacks collected from all over the world, |
| 3 | "We have to leave now!" Sarah shouted. ____ The only way out was up. We climbed flight after flight. The sound of the monsters banging on the door below became more distant but no less threatening. | "The zombies are going to break through any moment, and then we'll all be goners." |
| 4 | The sun was shining, and little gusts of wind brought through the window ____ shocking contrast from the stale city smells she had grown used to. | the faint scents of honeysuckle and freshly turned soil. It was a |
| 5 | I was minding my business at the park, when I was approached by a little girl who was crying because she had lost ____ so of course I helped search. | her cat, which she had just received for her birthday. She did not want her parents to know she'd already lost him. I'm a good person |
| 6 | It was a cold night, and a storm was raging out at sea. A lightning bolt lit up the sky, briefly illuminating the lighthouse ____ plummeted but just before reaching the churning water, he disappeared in a poof of purple flame! | and the young man peering hesitantly over the sheer cliff. Before the next peal of thunder he jumped. At first he |
| 7 | The magician pulled out of his pocket ____ and then a second one and a third. He didn't stop until soon the ground was covered with them. | a scarlet handkerchief |

Table 5.11: Hand-written fill-in-the-blank examples. To construct “custom” few-shot learning prompts, three of these were selected at random. In the end, the custom prompts did not result in better fill-in-the-blank performance than simply selecting randomly examples from the train set.

Table 5.12: Perplexity of evaluation sets according to LLM when the blank has been filled with approaches involving no fine-tuning (top), finetuned approaches (middle), and the groundtruth (bottom). Lower values indicate that the text was considered more fluent by the LLM.

| | C4FILL BLANK | RWPFILL BLANK | ROCFILL MIDDLE |
|---------------------|-----------------|------------------|-------------------|
| Few-shot LLM | 14.14 | 19.48 | 18.21 |
| Pre-trained T5 | 10.38 | 14.08 | 22.62 |
| Finetuned T5 | 10.33 | 14.08 | 20.47 |
| Donahue et al. [39] | N/A | N/A | 23.28 |
| Groundtruth | 9.41 | 12.99 | 16.90 |

Table 5.13: Perplexity of continuation-based evaluation sets when a continuation has been generated using approaches with no finetuning (top) and two settings of finetuning T5 (middle).

| | C4FILL END | RWPFILL END | ROCFILL END |
|------------------|---------------|----------------|----------------|
| Pre-trained T5 | 10.09 | 13.51 | 21.71 |
| T5 FILLBLANKCONT | 10.04 | 13.74 | 19.60 |
| T5 LM-ADAPTION | 10.06 | 13.71 | 19.68 |
| Groundtruth | 9.41 | 12.99 | 16.90 |

common on the internet. Furthermore, infilling ought to be an easier task than continuation since there is more information available for the model to base its prediction on. However, after conducting a large-scale study of many possible few-shot prompts, we found that this technique fell short for the fill-in-the-blank task.

One possible reason is that we did not do a sufficiently exhaustive search for a good prompt template. Zhao et al. [180] describe how one significant challenge with in-context learning is that task performance is often very sensitive to minor changes in prompt design. It is possible there exists a prompt for which in-context learning techniques would prove effective for fill-in-the-blank, but our exploration did not discover it. What we can conclude is that the process of finding an ideal prompt requires time-consuming trial-and-error and is quite difficult!

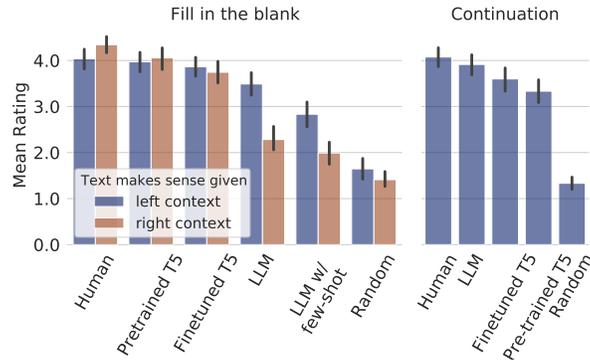


Figure 5.6: Human ratings of FITB generations (left) and continuation generations (right). Error bars are 95% confidence intervals.

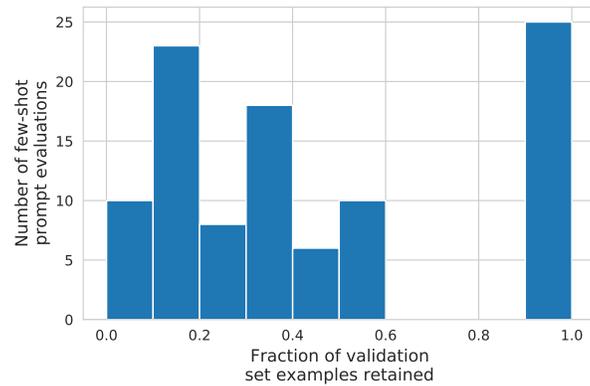


Figure 5.7: For many of the (validation set, few-shot prompt) combinations, not all validation set examples fit into the maximum sequence length for the LLM. The x-axis on this figure is the fraction of validation set examples which were retained after too-long examples were filtered out. The y-axis is the count of (validation set, few-shot prompt) pairs.

Table 5.14: Accuracy of models finetuned on FILLBLANKCONT at correctly using provided length and goal conditioning signals.

| Finetuned T5 | Context | Length |
|---------------------|---------|--------|
| C4FILLBLANK | 0.860 | 0.877 |
| RWPFILLBLANK | 0.797 | 0.881 |
| C4FILLEND | 0.858 | 0.775 |
| RWPFILLEND | 0.791 | 0.746 |

T5 GENERATES FLUENT CONTINUATIONS AND INFILLS We measure the fluency of proposed generations by evaluating the perplexity of each dataset’s examples when the predicted text is placed in the blank [39]. We use the LLM to measure perplexity¹⁴. The results are shown in Table 5.12. We see that the LLM struggles to generate fluent infills, even when used in a few-shot setting. The only exception to this is ROC Stories, a dataset with fairly simplistic, predictable language. Finetuning T5 does not result in significantly improved fluency over the pre-trained model except on ROC Stories. Lastly, for ROC Stories, we compare against Donahue et al. [39]’s finetuned GPT-2 small, which yielded less fluent predictions. Table 5.13 shows a similar analysis on our continuation-style datasets. Both T5-based models achieve roughly the same fluency.

HUMAN EVALUATION Human evaluation was conducted on 70 examples, 35 from RWPFILLBLANK and 35 from RWPFILLEND, with examples about evenly distributed across length buckets. For RWPFILLBLANK evaluation tasks, the rater was presented an input context and several possible sequences that could go in the blank. They were asked to rate each sequence first, on how well it fit the text before it, and second, on how well it fit with the text following it, according to a 5-point slider For RWPFILLBLANK, the task was almost the same, except that the rater was presented only a left context and asked to rate how well it continued the prompt. A screenshot of the Human Intelligence Task (HIT) used for annotations is shown in Figure 5.8. Workers were paid originally paid \$1.85 per HIT, but since the average HIT duration ended up being 15 minutes, we awarded each rater a bonus to raise their pay to an average of \$10 per hour. Each example was shown to three

¹⁴ Note, since this is the same model being used for generation for our continuation baseline, this metric may be biased.

Instructions for Fill-in-the-blank Evaluation Task

Your goal is to analyze how good an artificial intelligence is at generating text that makes sense with respect to the text before and after it. You will be shown the start of a passage of text where the AI's continuation has been highlighted in yellow. You will then be asked:

Does the **highlighted text** make sense?

Answer "not at all" if:

- the text is very ungrammatical OR
- though the text is grammatical, the highlighted section makes no sense with respect to the rest of the passage.

Answer "completely" if:

- the text is grammatical and smooth-flowing AND
- the contents of the highlighted section seems completely reasonable given the rest of the passage.

You must adjust every slider to be able to submit the HIT. The questions you've already worked on will be marked with a .

Question 4/8

Jackson checked his tie one last time before entering **the room. He wasn't allowed to speak,** [...]

Does the **highlighted continuation** make sense with respect to the text before it?

(not at all)  (completely) 

Now, suppose the text is continued in the following way.

Jackson checked his tie one last time before entering **the room. He wasn't allowed to speak,** now it was his turn. He glanced around the room, recognizing most of the people instantly. Everyone important was there... this was Jackson's chance. Jackson sat at the computer and opened up his AOL E-mail account, USB keys were strictly forbidden at meetings like this for being unsafe. He downloaded the PowerPoint and opened it.

Does the **highlighted section** make sense given this continuation?

(not at all)  (completely) 

SUBMIT (This will be enabled once you answer every question.)

Previous **Next**

Figure 5.8: A screenshot of the question structure for human evaluation.

raters, and annotations were rejected if the rater gave a lower overall score to the random output than to the ground-truth one. A total of 3 annotations were rejected. Overall, the Fleiss' kappa agreement of pairs of annotators giving the same numerical score to the same question was 0.26.

Figure 5.6 shows the results. On the FITB task, the pre-trained and finetuned T5 models were indistinguishable in terms of quality. The LLM that formed continuations prompted with only the left context did somewhat better than the few-shot LLM, indicating that few-shot learning is not yet a feasible alternative to finetuning. On the continuation task, the LLM has the highest rating, which is unsurprising since it is a much larger model than T5. However, the finetuned T5 is rated almost as highly. Overall, these results suggest that

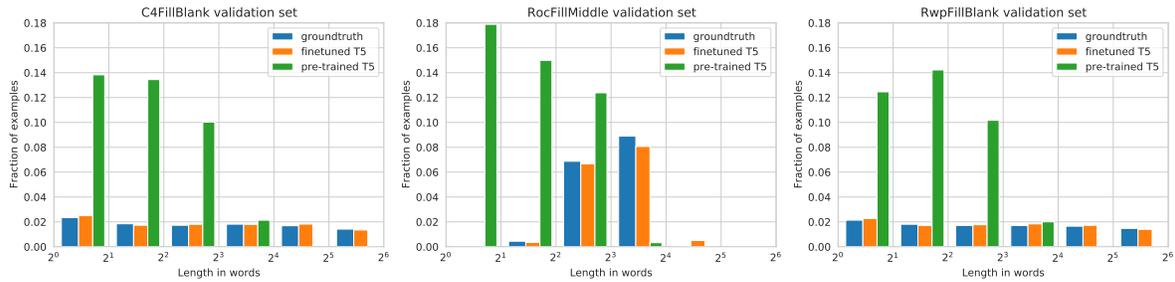


Figure 5.9: For each of the FITB validation sets, a histogram of the distribution of sequence lengths (measured in words) of the groundtruth blanked out text and the proposed infills from T5 after and before finetuning). We see that pre-trained T5 tends to produce text that is shorter than the groundtruth.

T5, unlike the LLM, can be used effectively for continuation as well as FITB. Furthermore, if one doesn't care about controllability, T5 can be used effectively for both tasks without any finetuning.

BENEFITS OF CONTROLLABILITY There are good reasons to care about controllability. For example, length conditioning is extremely important for FITB models, since it is not possible to control the generation length by simply sampling more or fewer tokens. Pre-trained T5 tends to produce infill proposals which are shorter than the groundtruth (Figure A5.9), and there is no way to ask the model to produce longer generations. In contrast, finetuned T5 was able to produce generations in the target length bucket over 74% of the time (Table 5.14). Goal conditioning, while not strictly necessary for either either task, has been shown to be useful for generative commonsense reasoning [91] and may empower users in downstream applications such as AI-assisted creative writing [127]. Finetuned T5 is able to use all of the specified goal words over 79% of the time.

FINETUNED MODELS TRANSFER ACROSS DATASETS Prior work on FITB tends to only evaluate models trained on data from the same domain as the validation set. Our results show that despite training exclusively on C4, T5 models have strong transferability to more targeted domains such as Reddit Writing Prompts. This sort of transferability is extremely important for achieving the goal of having single models which can handle many tasks and domains.

5.2.4 Conclusion

In this section, we make the case for starting with a model capable of filling in the blank when attempting to build a system that can perform both FITB and continuation. As LMs become bigger, it is unsustainable to have separately trained models per task. For example, in Wordcraft, the creative writing tool described in Chapter 5.3, over half a dozen operations are incorporated. It would be impossible to host a model for each.

Compared to the one-model-per-task paradigm, multi-task, domain-transferable models require less total training and are more efficient to store and use at inference time. In this section, we showed how T5 is easily capable of two tasks: continuation and infilling. While this is true even for the pre-trained T5, additional finetuning in the multi-task setting is still beneficial, as it allows us to carefully tailor the model to the tasks we need accomplished. We show how conditioning signals such as target length and goal text when added during finetuning allow for increased controllability at inference-time.

5.3 WORDCRAFT: AN EDITOR FOR AI-ASSISTED WRITING

In all the human evaluations discussed so far in my thesis, we recruited annotators to evaluate text in artificial settings. In Sections 3.2 and 3.3, university students were asked to label text as human-written or machine-generated in user interfaces custom-tailored to collect these sorts of annotations. In Sections 5.2 and Section 5.1, novel natural language systems were evaluated by asking raters on Amazon Mechanical Turk to view and score generated text. However, contrived annotation tasks like these give an incomplete picture as to the progress of natural language generation. Notably, the annotators were never asked to use the NLG to accomplish a task of interest; they were only asked to evaluate pre-generated outputs.

While contrived evaluation tasks have their place, it is also crucial to keep in mind why NLG systems are being built and who their target audience might be. This is especially true in the area of AI-assisted creative writing, where there is often a significant discrepancy between how novel tools are evaluated (Amazon

Mechanical Turker workers paid per annotation) and how they are intended to be used in the real world (writers seeking support or inspiration while performing their craft).

Wordcraft is an AI-augmented text processor which is intended as a real-world test bed for controllable text generation paradigms in the domain of creative writing. Our goal in developing Wordcraft was to learn how people interact with and want to interact with NLG systems—what tasks do they ask the NLG systems to do, how well do the systems deliver, and how does this feed back into the works people ultimately create. The user interface for Wordcraft consists of a traditional text processor alongside a set of NLG-powered controls that vary based on where the user’s cursor is and whether they have selected any text. The user also has access to a chatbot they can talk about their story with.

In this section, I first describe the controls implemented in Wordcraft and the motivations for including each of them. I then describe the ways in which Wordcraft offers the chance for a more realistic evaluation of state-of-the-art language generation systems through user studies with both amateur and expert writers.

5.3.1 The Wordcraft Application

Wordcraft is a web application intended for story writing with NLG assistance. The interface consists of a traditional text editor and a set of controls that prompt a large language model to perform various writing tasks. For example, Figure ?? (left) shows Wordcraft performing text infilling by suggesting alternatives for a selected passage of text, which the user can splice into their story. In Figure ?? (right), Wordcraft can be seen suggesting continuations for the user’s story. Wordcraft also includes controls for rewriting or elaborating on a selection of text. In addition to using these pre-built controls, users can also construct custom operations on the fly.

Wordcraft uses few-shot in-context learning techniques [20] to support each of its generative controls. The underlying neural language model backing all these interactions is LaMDA [149], which is described in detail in Section 5.1.1. Because LaMDA was finetuned for dialog, we frame our prompts in terms of a conversation. The remainder of this section enumerates the controls available in Wordcraft and gives an example few-shot

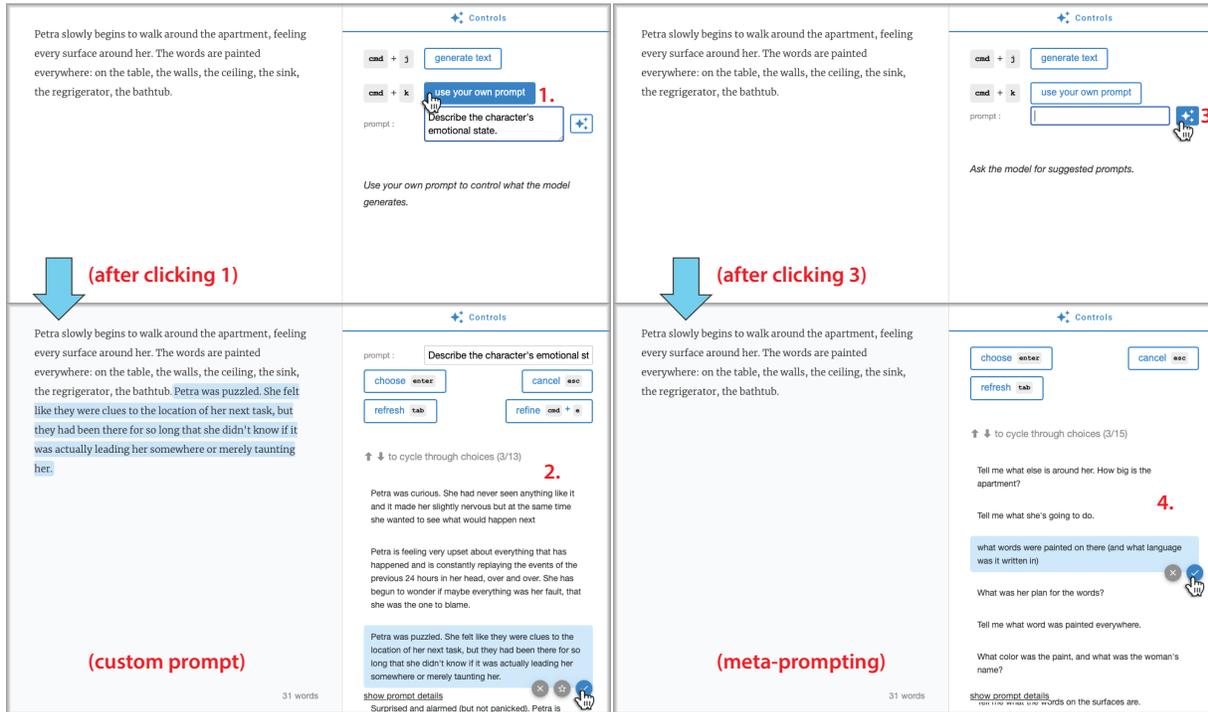


Figure 5.10: (left) Custom prompting in the Wordcraft interface. The user makes a novel request of the AI agent via a custom prompt (1), and the model attempts to fulfill the request (2). (right) Meta prompting in the Wordcraft interface. The user asks for suggestions for requests they could make of the model (3), and browses those suggestions (4).

prompt that supports each. In practice, the examples used in the few-shot prompt are randomly sampled from a small pool of examples each time the user selects a control. Each control in Wordcraft results in several candidate generations being presented to the user. When a generation is selected by the user, it is inserted into the text at the location of the user's cursor.

Continuation

As described in Section 2.4, continuation is the default action of a left-to-right neural language model such as LaMDA, achieved simply by prompting the LM with a text passage, and decoding a possible continuation. A continuation generator is useful to writers who want text appended to the end of what they have written so far. This control is available when the user's cursor is at the end of the text passage.

Initial Story Ideas

When the user has no text inputted, the user can specify the topic they want their story to be about and then have Wordcraft generate starting sentences. The intention of this control is to allow for ideation when the page is blank. It is supported with the following few-shot prompt:

- > Here's a topic: {A space monkey needs to escape from his dying planet.} Tell me the start of a new story.
- > {It was a cold, dark night on an unknown planet. In a field, a male monkey was pacing back and forth. He had been trying to figure something out, but nothing seemed to be working. He didn't have much time left.}
- > Here's a topic: {A medieval romance where a group of colorful knights do battle.} Tell me the start of a new story.
- > {The green knight rode atop his majestic, beautiful horse, across fields of pure green. In his palm was a magical, sparkling golden sword.}
- > Here's a topic: {An old man with a magical box tries to solve a mystery as his memories gradually fade away.} Tell me the start of a new story.
- > {The old man decided that he should look inside the box, and in the box there was a key. The old man couldn't quite remember why he had a key, but he knew the key was important .}
- > Here's a topic: {A coven of witches have taken up residence outside of a quiet mountain town, but they're not interested in witchcraft, they're interested in making hip-hop.} Tell me the start of a new story.
- > {An evil looking tree was growing off in the distance with no leaves and long, pointy branches. It was a sickly green and was swaying eerily without the wind.}
- > Here's a topic: {After an ancient sea-worn bottle washes up on the shores of Tripoli, a group of rag-tag pirates open it up to find a map leading to a magical treasure.} Tell me the start of a new story.
- > {A group of pirates set out to sea in search of a magical treasure - a long, sparkling blue necklace called the night's eye.}

> Here's a topic: {<USER'S TOPIC>} Tell me the start of a new story.}

Fill in the blank

When the user has some text selected in the editor, they are able to see alternatives that fit into the place of their selection. This is exactly the fill-in-the-blank task described in Section 5.2. Fill-in-the-blank is useful to writers because it allows them to insert sentences within an existing paragraph as well as see alternative phrasings for pieces they have already written. Though in Section 5.2 we showed that finetuned models perform much better at this than few-shot prompting techniques, due to the computational complexities of hosting multiple models, we ended up also supporting this operation with a few-shot learning prompt:

> {"We have to leave now!" Sarah shouted. ____ The only way out was up. We climbed flight after flight. The sound of the monsters banging on the door below became more distant but no less threatening.} Finish the following sentence by filling in the blank with a phrase. {"We have to leave now!" Sarah shouted. }

> {"Let's get out of here!"}

> Here's a new story: {The sun was shining, and little gusts of wind brought through the window ____ shocking contrast from the stale city smells she had grown used to.} Finish the following sentence by filling in the blank with a phrase. {The sun was shining, and little gusts of wind brought through the window ____ shocking contrast from the stale city smells she had grown used to.}

> {the faint scent of honeysuckle. It was a}

> Here's a new story: {I was minding my business at the park, when I was approached by a little girl who was crying because she had lost ____ so of course I helped search.} Finish the following sentence by filling in the blank with a phrase. {I was minding my business at the park, when I was approached by a little girl who was crying because she had lost ____ so of course I helped search.}

> {her new kitten.}

> Here's a new story: {It was a cold night, and a storm was raging out at sea. A lightning bolt lit up the sky, briefly illuminating the lighthouse ____ plummeted but just before

reaching the churning water, he disappeared in a poof of purple flame!} Finish the following sentence by filling in the blank with a phrase. {A lightning bolt lit up the sky, briefly illuminating the lighthouse ____ plummeted but just before reaching the churning water, he disappeared in a poof of purple flame!}

> {in a sharp silhouette.}

> Here's a new story: {The magician pulled out of his pocket ____ and then a second one and a third. He didn't stop until soon the ground was covered with them.} Finish the following sentence by filling in the blank with a phrase. {The magician pulled out of his pocket ____ and then a second one and a third. }

> {a scarlet handkerchief }

> Here's a new story: {<USER'S STORY>} Finish the following sentence by filling in the blank with a phrase. {<USER'S SELECTED SENTENCE WITH BLANK AT POSITON OF CURSOR>}

Elaboration

When the user has some text selected, they can use the elaboration button to get the model to expand upon it. This is especially useful for description generation. The elaboration operation is supported with the following few-shot prompt:

> Here's a new story: {The long shadow of the tree crept up the sidewalk} Describe the tree.

> {It was an old oak tree on the main street of a small town, the branches spread as large as the sky.}

> Here's a new story: {The car came bouncing down the hill and came to an abrupt stop in my driveway} Describe the car.

> {It was an old, old Model-T Ford, with brass fittings, and no doors}

> Here's a new story: {My pencil sits on top of the desk, left there since the events of last Saturday} Describe last Saturday.

- > {It was one of those Saturdays that starts off like any other but you have the feeling that something is about to stir. You wake up with the same amount of enthusiasm but you get up from bed a little slower than usual. You know this is going to be a busy day}
- > Here's a new story: {She wasn't really upset about what had happened at the lab the other day; she was never one to dwell on mistakes.} Describe what had happened.
- > {She had accidentally left the microorganisms alone for far too long. The lab, or what was left of it, was now filled with a strong aroma of their slime-- they had devoured all the metal, wood, and plastic they could find}
- > Here's a new story: {<USER'S STORY>} Describe `\textbf{<USER'S SELECTED TEXT>}`.

Style Transfer

Using the augmented zero-shot learning approach described in Section 5.1, we include a control that allows users to perform rewriting operations on selected text. The operations are specified in natural language. The following prompt is used.

- > Here's a new story: {An elderly man was sitting alone on a dark path. The darkness turned to light as a figure ran up to him. ____ The older man recognized the young man as a dear friend, who he hadn't seen in some time. The young man said he had had the strangest dream.} What sentence goes in the blank?
- > {The figure was holding a sword.}
- > Rewrite the sentence to use more descriptive language.
- > {The shadowy figure held in his hand a shining sword that glimmered in the darkness.}
- > Here's a new story: {There once was a man named Kentucky, who could play the guitar like nobody's business.____Kentucky was a lonely man living in a lonely house in the middle of nowhere with only his favorite guitar to keep him company. Now this guitar had been owned by many-a-man, and all of them had come to the same fate. They played the guitar until the guitar played them.} What sentence goes in the blank?
- > {He loved to play the guitar so that he could hear the guitar sing along to him, to hear the guitar's melodic hum.}

- > Rewrite the sentence to be more melancholy.
- > {He loved playing sad songs on his weathered guitar, so he could drown out his loneliness with the guitar's melodic hum.}
- > Here's a new story: {A small group of explorers landed in China, and were amazed by the beautiful, sparkling buildings. They wondered how such a beautiful country could ever be as scary as they had heard. It was at night, and the explorers were deep in the heart of the forbidden city, their flashlights not strong enough to reach the bottom of the narrow, dark hallways.----} What sentence goes in the blank?
- > {Suddenly, a terrifying sound echoed through the halls.}
- > Rewrite the sentence to use more interesting words.
- > {Out of nowhere, a thunderous cacophony reverberated through the halls.}
- > Here's a new story: {An elderly man was sitting alone on a dark path. ---- It was a plain pine box and looked as if it had been there for a long time. The man was afraid to look inside the box.} What sentence goes in the blank?
- > {The man looked down at his feet, and realized he was standing next to a box.}
- > Rewrite the sentence to be more evocative.
- > {The man nearly gasped as he peered down at his feet, suddenly realizing he was standing next to a mysterious box.}
- > Here's a new story: {Once upon a time, there was a rabbit princess. The bunny lives in the little cottage just outside of town. The story ends with her getting eaten by a fox. ---- The door hung at an angle off of its hinges.} What sentence goes in the blank?
- > {<SELECTED TEXT>}
- > Rewrite the sentence <REWRITING INSTRUCTION>.

Custom Prompting

In the custom prompting control, users have the ability to write any request to the system they want. For example, they could ask "Give me an origin story for the ogre" or "Tell me what happens next." We use a similar approach to style transfer, with a prompt demonstrating a variety of requests:

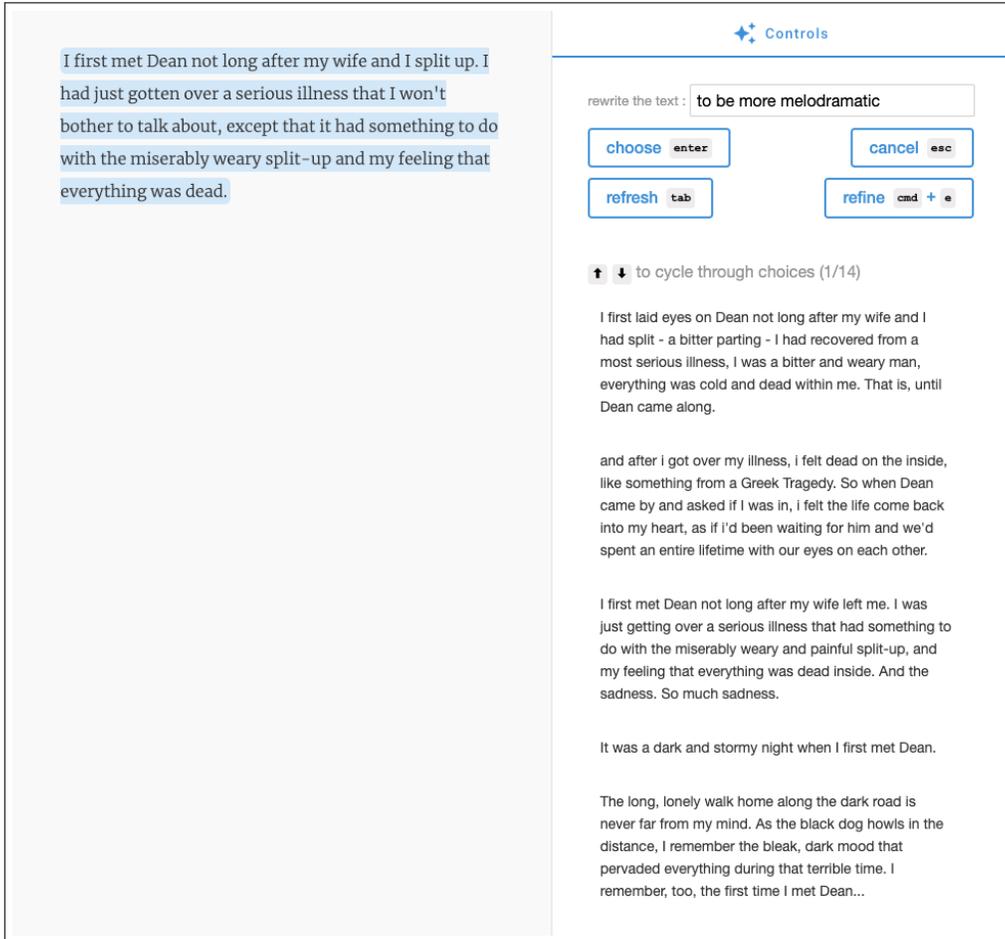


Figure 5.11: Screenshot of Wordcraft with the custom prompting feature.

- > Here is some text: {next to the path} Rewrite it to be about France.
- > {next to la Siene}
- > Here's a new story: {A young boy was sitting on the sidewalk.} What comes next?
- > {He was holding a lantern and looking ahead up the road.}
- > Here is some text: {The man stood outside the grocery store, ringing the bell.} Rewrite it to be about clowns.
- > {The man stood outside the circus, holding a bunch of balloons.}
- > Here's a new story: {My pencil sits on top of the desk, left there since the events of last Saturday.} Describe last Saturday.
- > {It was one of those Saturdays that starts off like any other but you have the feeling that something is about to stir. You wake up with the same amount of enthusiasm but you get up from bed a little slower than usual. You know this is going to be a busy day}
- > Here's a new story: {<USER'S STORY>} <CUSTOM PROMPT>

Suggested Prompts

Finally, we implemented a control that allows users to request suggestions for what to write about next. We expected this feature to be useful for novice writers to overcome writer's block. It uses the following prompt:

- > Here's a new story: {A long time ago, in the beginning of time, beautiful crystalline frog people inhabited the earth. Their king, the crystal king, had mysteriously disappeared on a trip to the sea.} What would you tell me about next?
- > {Tell me more about the frog people.}
- > Here's a new story: {The green knight rode atop his majestic, beautiful horse, across fields of pure green. In his palm was a magical, sparkling golden sword.} What would you tell me about next?
- > {Tell me what happened next.}
- > Here's a new story: {Ninja schools are where the ninjas are trained for their missions. Some are trained in the forests and snowy mountains and others are given the ultimate training in the hottest palace in the land.} What would you tell me about next?

```

> {Tell me about the ninjas}
> Here's a new story: {The witch doctor had a large pile of magic stones. All the magic
stones had a power - healing magic, shapeshifting, seeing into the past.} What would
you tell me about next?
> {Tell me what the witch doctor did with them.}
> Here's a new story: {The desert sun beat down on the sand, and tiny bits of sand flew
with every step. The two young travelers trudged onward. They were tired, but they knew
they had to go forward.} What would you tell me about next?
> {Tell me more about the two travelers.}
> Here's a new story: {A prince and princess from the future set out in search of the
mythical land of the unicorns. They had heard tale of magical unicorn dust that could
heal any wound.} What would you tell me about next?
> {Tell me what their journey was like.}
> Here's a new story: {A giant monster was chasing after a woman and yelling at her to get
away. The woman ran, knowing that she was in horrible danger.} What would you tell me
about next?
> {Tell me why the monster was chasing her.}
> Here's a new story: {<USER'S STORY>} What would you tell me about next?"

```

5.3.2 User Study with Novice Writers

To evaluate Wordcraft's effectiveness, we conducted a user study in which 25 hobbyist writers (whom we refer to as U1-U25) were asked to write stories with and without Wordcraft. The goal of this user study was to understand which controls writers preferred using and whether the assistance provided by Wordcraft was valuable to writers.

- directly compare Wordcraft against simpler systems
- learn what features worked well and what features were missing

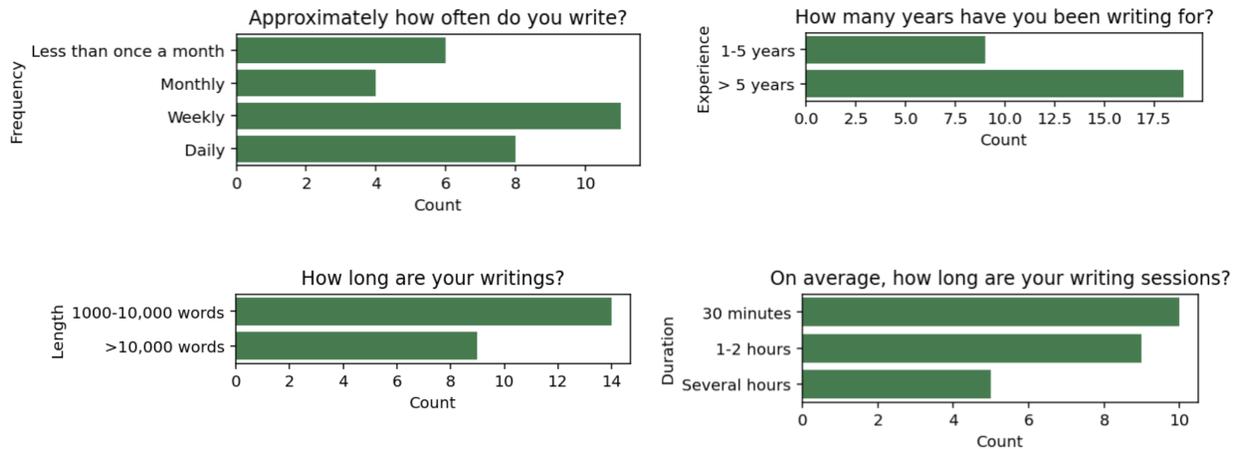


Figure 5.12: Writing habits survey results.

- fix any major bugs

Methodology

We recruited participants via advertisements on Google-internal mailing lists. We then screened for individuals who practice creative writing on a regular basis, but who had not yet published their writing. Participants volunteered for the study and were not compensated. Most of the participants (23 out of 25) did not consider themselves to be machine learning practitioners and had not interacted with a generative language model previously. We asked participants to complete a pre-study questionnaire about their writing habits (Figure 5.12). Then we carried out a within-subjects study, giving each user three writing prompts and asking them to write 100-300 word stories under the following three experimental conditions (illustrated in Figure 5.13):

1. **full** the full Wordcraft tool.
2. **cont** (baseline) a text editor with a single control: LaMDA will propose continuations to the text written so far. The *cont* condition enables us to evaluate Wordcraft against existing AI-assisted writing applications which most often feature continuation as a single control.
3. **chat** (baseline) a plain text editor shown alongside a chat dialog window. Users can converse with the LaMDA-powered chatbot, but the chatbot only “knows” what the user types to it. The *chat* condition

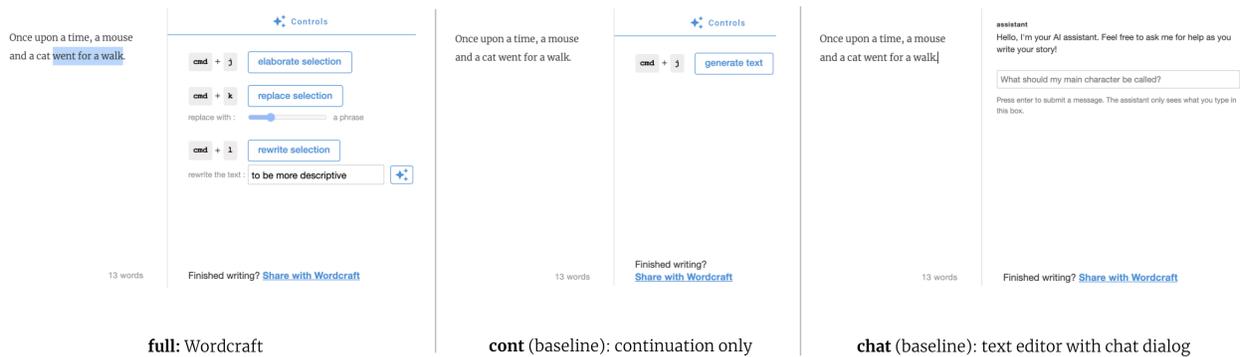


Figure 5.13: Experimental conditions for the user study. The stories written by participants as part of the study can be viewed here: <https://storage.googleapis.com/wordcraft-stories/index.html>. The website also includes purely machine-generated baselines.

enables us to evaluate the utility of the prompt and UX scaffolding we designed for Wordcraft against giving users straightforward access to the underlying model, without any scaffolding. An omniscient user could theoretically reproduce the functionality of the *full* condition by replicating Wordcraft’s prompts.

Participants were asked to write stories based on the following three prompts ¹⁵:

1. You arrive at Grandma’s funeral to find thousands of people from around the world also in mourning. You are entirely unaware that Grandma had 16.4m followers on Twitter.
2. All of the ‘No. 1 Dad’ mugs in the world change to show the actual ranking of Dads suddenly.
3. You lost your sight - along with everyone else on Earth - in The Great Blinding. Two years later, without warning, your sight returns. As you look around, you realize that every available wall, floor and surface has been painted with the same message - Don’t Tell Them You Can See.

Participants were given ten minutes to write each story. We felt that ten minutes was enough time for users to acquaint themselves with the interface and write 100-300 words, while managing users’ expectations for how much time they would need to spend on the study in total. To control for writing ability and prompt difficulty, for each user, the three conditions were randomly paired with the three prompts, and the user was asked to

¹⁵ Prompts were selected from the [Writing Prompts subreddit](#).

Table 5.15: Usage statistics.

| Property | <i>chat</i> Chat | <i>cont</i> Continuation | <i>full</i> Wordcraft | Overall |
|-----------------------------------|------------------|--------------------------|-------------------------|-----------------------|
| Requests made avg | 6.3 \pm 1.3 | 4.3 \pm 0.52 | 7.3 \pm 0.74 | 6.0 \pm 0.53 |
| Accepted suggestions avg | N/A | 0.17 \pm 0.08 | 1.3 \pm 0.25 | 0.51 \pm 0.12 |
| Story word count avg | 233 \pm 18 | 237 \pm 16 | 267 \pm 21.6 | 247 \pm 11.1 |
| Model word count avg (% of story) | N/A | 2.9 \pm 2 (1.3%) | 42.3 \pm 14.2 (13.2%) | 16.2 \pm 5.7 (5.2%) |
| Time considering suggestions avg | N/A | 67.1s \pm 8.7s | 41s \pm 4.1s | 44.5s \pm 3.2s |
| Time to complete avg | 11m \pm 62.5s | 11.52m \pm 131.5s | 9.97m \pm 37.1s | 10.8m \pm 48.6s |

Table 5.16: Usage statistics broken down by request type (Wordcraft only).

| Request type | Requests made | Suggestions accepted |
|-------------------|---------------|----------------------|
| Rewrite | 27 | 5 (18.5%) |
| Story seed | 22 | 12 (54.5%) |
| Suggest a prompt | 40 | 2 (5%) |
| Fill-in-the-blank | 4 | 0 (0%) |
| Continue | 36 | 4 (11.1%) |
| Next sentence | 7 | 1 (14.3%) |
| Elaborate | 3 | 0 (0%) |
| Custom | 51 | 9 (17.6%) |

write a story for each setting. We also randomized the order in which the conditions were presented. Users were not given any training for the various conditions—they were simply given a website link and asked to write a story with the interface. Users were told that they were participating in a study of AI assisted writing, but they were not explicitly asked to solicit help from the AI agent, as we were interested in learning how often users would want to make use of AI-assisted controls.

Results

This section describes the overall successes and failures of the NLG-powered assistive writing features incorporated into Wordcraft, before comparing Wordcraft to the baseline conditions in depth.

Table 5.17: Rewrite requests (user completions of ‘Rewrite this...’)

to be a little less angsty • to be about mining • to be better written • to be less diabolical • to be more absurd • to be more adventurous • to be more Dickensian • to be more emotional • to be more magical • to be more melodramatic • to be more philosophical • to be more revolutionary • to be more surprising • to be more suspenseful • to be more technical • to be more whimsical • to be warmer • to fit better grammatically with the rest of the story • to make more sense

Table 5.18: Custom prompts written by users.

Tell me more about her twitter account. • More about the dad please. • Tell me about Elaine’s amazing twitter account. • Tell me about the father. • Tell me about the funeral home, grandma, the punk kid, and the crowd. • Tell me how the man reacted as he found out he could see again. • Tell me more about Daniel. • Tell me more about what it’s like to have to pretend to be blind when you can see. • Tell me what happens next. • Tell me what the letter says. • What are the words on the floor? What language are they in? • What would happen if we could quantify love? • Why were they watching me? • More about the dad please. • So this guy was tight with your grandma? • Tell me about Elaine’s amazing twitter account. • Tell me about the father. • Tell me about the funeral! • Tell me about the funeral home, grandma, the punk kid, and the crowd. • Tell me how the man reacted as he found out he could see again. • Tell me more about Daniel. • Tell me more about what it’s like to have to pretend to be blind when you can see. • Tell me what happens next. • Tell me what the letter says. • What are the words on the floor? What language are they in? • What would happen if we could quantify love? • Who is telling me not to tell them? • Why were they watching me?

5.3.3 The role of AI in co-writing

We observed that users solicited help from the AI agent at every stage of the writing process—from high-level story concepting down to rewriting and editing. Thus, we found that the AI agent played many different roles in collaborative writing.

Users asked the AI agent for help in story ideation and brainstorming. For example, nine users at one point presented the AI agent with their story and asked simply: ‘*What happens next?*’. Another user solicited help developing the premise for their story: ‘*What would happen if we could quantify love?*’.

Another typical use case for soliciting help from the AI was when just starting a story. The story seed control, in which the AI agent provides opening sentences for a story given a writing prompt, had the highest success rate of any control: 55% (Table 5.16) of suggestions were accepted by users. In these cases, the AI

served to kick-start the writing process for users who might have been blocked. Users also found the AI agent helpful for generating smaller scale details for their story, such as names for characters and locations.

Many users remarked on the usefulness of the AI agent's suggestions, even if they didn't end up using them verbatim: *'Multiple suggestions around the highlight or next phrasing were very helpful, even if I didn't use the whole phrase ... it was like having someone suggest things that I might have thought of myself'* (U9). Another user commented: *'It was good at generating a bunch of relevant ideas that inspire my next lines and get me unstuck. I was never tempted to use any of the lines verbatim, but it was fun inspiration'* (U4). Some also noted the AI agent's tendency to provide offbeat suggestions as a strength: *'the off the wall suggestions were fun to play around with and helped shape how the story took form'* (U5).

We also observed users having ideas for events before knowing how they fit into an existing story - and in such cases asking the AI agent to fill in gaps. For example, users would build a scene and then ask the AI agent to provide plot points that would contextualize the scene. One user in their story described a character being watched, and then asked the AI agent *'Why were they watching me?'* Many of the custom prompts (Table 5.18) we collected fall under this use case. These included prompts such as *'Tell me what the letter says.'* and *'Tell me about the funeral home, grandma, the punk kid, and the crowd.'*

We also observed users asking the AI agent for help in smaller scale edits, for example: *'Rewrite this sentence to fit better grammatically with the rest of the story'*, or *'Rewrite this sentence to make more sense.'* Many of the requests in Table 5.17 fall under this category.

Shortcomings of AI in co-writing

The user study revealed many shortcoming in AI co-writing. Users' observations of the AI agent's shortcomings mostly center on its lack of contextual awareness. For example, though the assistant might provide several fluent, well-written alternatives to a sentence as part of the rewrite control, its suggestions do not necessarily make sense given the rest of the story. One user whose story mentioned numbers moving on a coffee mug received suggestions from the AI agent which implied that *'live animals (snakes specifically) were moving'*. Users also noted many grammatical issues, for example that the AI agent's suggestions were often not in the

same tense as the rest of the story. Some also noted that the AI did not seem aware of their story’s established point of view (first person versus third person).

Wordcraft versus baseline 1: continuation-only

In this section we compare Wordcraft to the continuation-only baseline (*cont*). This baseline allows us to measure the utility of Wordcraft’s prompting methods and UX patterns for the story writing task against the typical experience of *continue-my-text* seen in existing LLM-powered writing tools. We analyzed the activity logs from each user’s writing session, and extract quantitative findings based on the following metrics (results in Table 5.15):

- *Requests made avg*: On average, how many times the user requested assistance from the AI while writing a story.
- *Accepted suggestions avg*: On average, how many of the AI’s suggestions the user accepted.
- *Time considering suggestions avg*: The average time users spent between soliciting help from the AI, and accepting a suggestion or dismissing the suggestions.
- *Model word count avg*: The average number of words in the final story that came directly from the AI agent.
- *Time to complete avg*: The average time spent to produce the final story.

Participants made significantly more requests of the AI agent using Wordcraft (7.31 ± 0.74) than the continuation-only baseline (4.35 ± 0.52) according to a paired-sample T-test ($p = 0.003$). Participants also *accepted* significantly ($p = 0.0003$) more of the AI’s suggestions using Wordcraft (1.27 ± 0.25 vs 0.17 ± 0.079). Accordingly, the stories written with Wordcraft contained significantly ($p = 0.0068$) more text from the AI agent ($13.2\% \pm 3.8\%$) than stories written with continuation-only ($1.3\% \pm 0.92\%$).

Users found Wordcraft more helpful. The results from the Likert-scale questions in the exit interviews are in Figure 5.14. According to a Mann-Whitney U test, users found Wordcraft significantly more helpful than the continuation-only baseline ($p = 0.0266$).

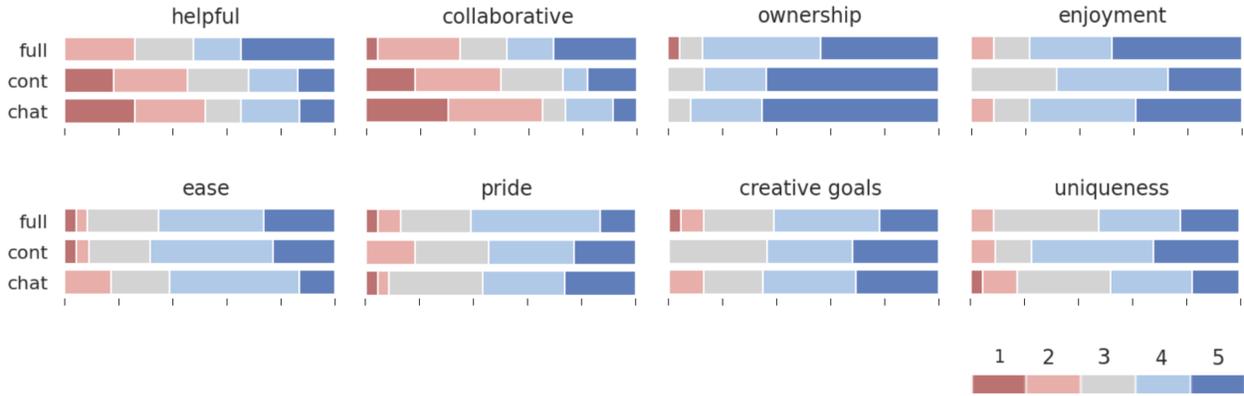


Figure 5.14: Exit interview results. 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree.

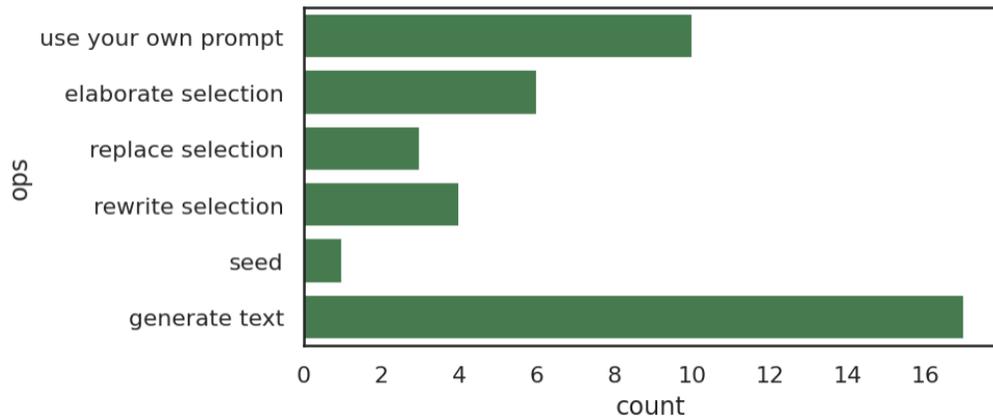


Figure 5.15: Results from exit interview question ‘Which controls did you enjoy using?’

We found that participants enjoyed using controls aside from continuation. We note in Figure 5.15 that more than half of the responses to the question ‘Which controls did you enjoy using’ indicate controls *other* than ‘generate text’ (aka continuation). ‘Use your own prompt’ was the most used control and the second most highly voted control, suggesting users enjoyed having the ability to design their own prompts. Table 5.16 lists the number of times each control was used and accepted in Wordcraft, further demonstrating that participants found utility in a variety of controls other than continuation.

Wordcraft versus baseline 2: chat

In this section we compare Wordcraft to the *chat* baseline. This again allows us to measure the importance of Wordcraft’s prompting methods and UX patterns, but also to observe how writers might make use of a powerful open-domain LLM-powered chatbot in their creative process. As the two conditions present significantly different user experiences, many of the metrics from the previous section are not applicable. Thus we also analyze the chat transcripts and free-response exit interview questions to extract qualitative themes.

According to a Mann-Whitney U test, users found Wordcraft significantly more helpful than the chat baseline ($p = 0.0134$), and also found it to be significantly more collaborative ($p = 0.00668$). Nevertheless, we observed high user engagement with the chatbot assistant - on average users sent 6.3 ± 1.3 messages to the chatbot (Table 5.15). One user remarked: *The AI agent asked good probing questions that gave me inspiration* (U22). Many talked to the chatbot the way a writer might discuss a project with a friend, asking for suggestions about overall direction as well as story details. For example (U1):

user: How should I start a story about being blind?

AI agent: It’s a unique perspective, so focus on description. What can you hear, smell, taste, and touch? These should all be used in your story.

Another user (U4) asked about specific plot elements and character details:

user: what is grandma famous for on twitter?

AI agent: She is known on Twitter for her love for cooking with the great recipe ideas she shares

and her love for gardening and giving advice for what you can grow in yours.

user: what cake recipe is grandma famous for

AI agent: She is known for her classic Victoria sponge cake recipe. Her recipe has over 5,000 likes which makes it one of her most popular.

user: what is grandma's twitter handle

AI agent: Her twitter handle is @mylovelygrandma where she regularly tweets her cooking tips and photos of her bountiful garden.

Finally, participants found the chatbot to be effective at retrieving information. Some treated the chatbot as a search box, asking questions such as: *how many people are fathers around the world*. In the exit interview, one user commented about the chatbot: *"I loved that it felt like a more targeted version of Google, and I could do background research while writing"* (U21).

5.3.4 User Study with Expert Writers

For the Wordcraft Writers Workshop, we commissioned 12 published writers from diverse stylistic backgrounds, ranging from poetry to fantasy, to craft stories using Wordcraft. The writers were instructed to work on pieces of about 1,500 words. They were also asked to keep informal journals of their observations during the process. Finally, we conducted entrance and exit interviews with each writer before the workshop started and after they had finished with their stories.

{TODO: Finish inserting this section. }

5.4 CONCLUSION

In this chapter, I have presented two ways—natural language-specified style transfer and fill-in-the-blank—that language models can be applied to more specific tasks than simply continuing a prompt. I then incorporated these tasks into Wordcraft, a tool for human-AI collaboration for creative writing. User studies with Wordcraft revealed both the strengths and challenges of incorporating NLG into the creative writing process and suggested several avenues for future research. These include **{TODO: }**.

5.5 SUMMARY OF CONTRIBUTIONS

The work on models for doing both fill-in-the-blank and continuation generation tasks was published as “The Case for a Single Model that can Both Generate Continuations and Fill in the Blank” in the 2022 Findings of the North American Association of Computational Linguistics [66]. The work was performed with my collaborators Daphne Ippolito, Liam Dugan, Emily Reif, Ann Yuan, Andy Coenen, and Chris Callison-Burch. I led this project, designed and ran all experiments, and performed most of the analysis.

The work on style transfer was published as “A Recipe For Arbitrary Text Style Transfer with Large Language Models ” in the 2022 Proceedings of the Association of Computational Linguistics [123]. The project was completed with my collaborators Emily Reif, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. I worked with Emily Reif to come up with the premise for the project: of formulating style transfer as an arbitrary rewriting operation. I helped to design the experiments and contributed significantly to analysis of the results.

The Wordcraft tool was built jointly with Ann Yuan, Andy Coenen, and Emily Reif. Ann Yuan led the user study with amateur writers, and I led the user study with professional writers. I contributed significantly to the design and implementation of the LM-powered controls in Wordcraft, and I contributed to the user interface design.

6 | CONCLUSION

This thesis focuses on (1) analyzing neural language models to better understand the text they are able to generate, and (2) studying the feasibility of applying large language models to tasks that could benefit from human-AI writing collaboration.

Over the course of my PhD, the field of language generation with neural language models has progressed from character-level LSTMs with tens of millions of learned weights that are only capable of generating passages such as “An ICBM, the [[gurt and land]] has registered \$155 billion in U.S. and August 1688, and makes sure the US-transplantation disbanded backwards in the County by authorizing disputes that tend to carry over this peninsula” [55] to non-recurrent networks with hundreds of billions of learned weights that are capable of outputting text even skilled humans can have trouble distinguishing as machine-generated. Despite these advancements, even state-of-the-art language models tend to generate text that fails in both subtle and unpredictable ways. They also struggle to understand context which should influence their generations, and they fall short on supporting the kinds of fine-grained controllability that are crucial for meaningful human-AI collaboration. However, the future of the field is promising; **{TODO: say something positive}**.

I conclude by summarizing the main takeaways from my research and the future work suggested by them.

Detection of generated text is getting harder but no less important.

Chapter 3, I present the challenge of detecting machine-generated text. As neural language models get better, it is becoming only more challenging for humans to notice they are interacting with a bot. Indeed, in a recent controversy, a YouTuber inundated 4chan with GPT-Neo generated text ¹⁶. While some users eventually

¹⁶ <https://youtu.be/efPrtcLdcdM>

caught on, many continued to be fooled even after the YouTuber revealed the scheme. In computer vision [128, 121], it is standard to watermark generated images, but thus far, this has not been applied to large language model outputs. The increasing pervasiveness of generated text on the internet is problematic not just because of its potential societal impacts but because it sullies our future training sets. In machine translation, it is a well-known problem that automatically translated text could corrupt training sets, and some watermarking techniques have been proposed [155]. This is a problem that those who build large neural language models (and their training sets) need to start concerning themselves with as well. My research on automatic detection was performed in 2018 on 768M parameter models. It would be very valuable to reconsider the automatic detection problem on state-of-the-art generation systems and in more realistic contexts (such as on documents where only a portion of the text may be generated).

Text generation involves tradeoffs.

In Chapters 2 and 3, I describe how there exists a tradeoff between generating diverse text that is easier for humans to detect because it contains obvious errors—and generating mundane text that is harder to detect, but lacks the lexical diversity of a real human writer. This tradeoff continues to be important, both for academic research—because we need to ensure that comparisons between different NLG systems are fair—and for practitioners—because the setting chosen can have a significant impact on user experience. For example, Wordcraft users complained that the text was in too dull a style, a problem that might have been resolved very simply by increasing the sampling temperature. Further research is needed into techniques for sampling from the long tail of low-likelihood words without causing semantic errors.

Memorization is a serious concern but can be hard to define.

In Chapter 4, I focus on the memorization problem; language models are capable of regurgitating text from their train sets. Memorization is most often caused by examples being over-represented in the training data, but in recent work [176], we show that language models also memorize rare sequences. (We show this by measuring counterfactual memorization—how much more likely an example is according to models that saw

it during training compared to models that never saw the example.) More detailed studies of what kinds of content are more susceptible to memorization and the training dynamics behind memorization will be important subjects of future work.

In the research discussed in this thesis, I have considered *all* instances of memorization as problematic. While memorization is often a sign of poor generalization, and at its worst, it can divulge private information, memorization can also be a good thing (e.g., we might want our language model to be able to accurately quote famous speeches). It will be important to develop more advanced techniques for controlling memorization, allowing models to quote verbatim when there is a good reason to.

More attention needs to be paid to dataset quality.

Chapter 4 shows that duplicate text in language model training data is a crucial source of memorization. However, duplication is only one of many traits that can influence how well models pre-trained on a large dataset perform on a variety of tasks of interest. Research answering questions on the effect of pre-training dataset composition—including how stringent toxicity filters should be, what languages to include, and the balancing of different sources (news, Wikipedia, books, etc.)—could have important ramifications for how new datasets are built and how trained models are used. It will be impossible to recommend a single set of data composition rules that will be optimal for all downstream tasks (for example, filtering out all hate speech will result in a model that is less likely to generate hate speech but is also less useful as part of a hate speech detection system), but at least further research should allow us to answer questions about the tradeoffs involved in different dataset composition decisions.

Supporting many tasks from fewer models is valuable.

As neural language models increase in number of parameters, it is becoming increasingly infeasible to create one custom-tailored model per task that needs to be supported. In Chapter 5, I show how a single pre-trained language model can be made to support a large variety of style transfer tasks that previous work would have typically trained several separate models for. I also argue that we should be pre-training large language models

for a fill-in-the-blank-style objective, rather than a continuation one, because filling in the blank is a strictly more versatile task. There has been a significant focus in recent work on developing training objectives to support a variety of downstream tasks with minimal additional task-specific adaption [162, 131]. However, these approaches require a substantial amount of annotated training data, and self-supervised pre-training objectives which yield multi-task-capable models is an important subject for future research.

Evaluation of NLG systems should happen in real-world settings.

In Chapter 5, I present Wordcraft, a text editor with NLG-powered writing assistance intended for creative writers. Through user studies with both novice and professional writers, we explore the strengths and weaknesses of state-of-the-art natural language generation. Studies of NLG use by real users can lead to different and more nuanced conclusions than those from more contrived human evaluation schemes. For example, when evaluating the use of augmented zero shot learning for style transfer with Amazon Mechanical Turk-based evaluation, we saw that annotators preferred our approach's generations over other approaches. However, use of this feature within Wordcraft revealed just how much of gap there still is between the types of transfers writers want to do and the capability of our approach.

Fiction writing is only one domain where NLG-powered tools could be impactful, and it would be valuable to see the types of studies I ran with Wordcraft be conducted in other domains. In particular, NLG has many possible uses in tools that assist people in learning how to write (both children and new-language learners). It could also have impact in domains that require writers to produce a lot of text quickly, such as script writing for immersive video games.

LIST OF TABLES

| | |
|-----------|---|
| Table 2.1 | Examples of the string “A hippopotamus ate my homework.” tokenized using three different vocabularies. With the subword tokenizer, the rare word “hippopotamus” gets broken up into multiple tokens. For word-level tokenizers, if the word “hippopotamus” occurred very infrequently in the corpus used to build the vocabulary (or perhaps the writer of the sentence misspelled it), it would typically get replaced with an out-of-vocabulary token (row 4). 11 |
| Table 2.2 | A survey of datasets which have been used to train large general-purpose neural language models. 23 |
| Table 2.3 | Examples of pre-training objectives used in popular general-purpose models. In these examples, the original training sequence is “The hippopotamus ate my homework. It made me very mad.” 24 |
| Table 3.1 | The number of excerpts used for training, validation, and testing. Three decoding strategies—top- k with $k=40$, nucleus sampling with $p=0.96$, and full random sampling ($p=1.0$)—were employed. The language model was either prompted with a single word (1wordcond) or used without any prompt sequence (nocond). 30 |
| Table 3.2 | Performance (accuracy and AUC) of the fine-tuned BERT classifier and several simple baselines on detecting length-192 sequences generated with one word of priming (1worcond). Note that p1.0 refers to untruncated random sampling, where we sample from 100% of the probability mass. The last column shows human performance on the same task where accuracy with a 50% baseline is computed by randomly pairing samples from each decoding strategy with a human-written sample. 31 |
| Table 3.3 | The number of human annotations collected. In total, there were 50 examples from each sampling strategy and 150 examples of web text. Each example was shown to at most three raters. 34 |
| Table 3.4 | The 10 examples that “expert” raters were guided through before they were asked to perform the detection task. These are hand-selected to showcase the spectrum of generated text and human-written text. 35 |
| Table 3.5 | Accuracy of BERT fine-tuned discriminator when trained on samples from one strategy (rows) and evaluated on another (columns). Trained on samples with 192 tokens. The ‘mixed’ dataset is one containing an equal portion of samples from each strategy. 40 |
| Table 3.6 | Average probability of ‘machine-generated’ according to each length-192 discriminator. The expected in-domain probability is 0.5. One token of conditioning. 40 |

| | |
|------------|---|
| Table 3.7 | Some 192-token examples where at least two expert raters agreed with each other, but were not in agreement with the automatic discriminators. The first row shows examples where the ground-truth was human-written, the second shows machine-generated examples where the corresponding discriminator guessed incorrectly, and the third shows machine-generated examples where the discriminator was correct, but raters got it wrong. 41 |
| Table 3.8 | Statistics on the annotation tasks (game rounds) available in our system. The second column shows the number of game rounds available for each system. The discrepancies in number of annotations per dataset is partially due to the fact that players were able to choose which domain they performed annotations in. 48 |
| Table 3.9 | Statistics on the students who were invited to complete annotations on ROFT. “Avg Ann / Part” is the average number of annotations per participating student, while “Avg Score / Part” is the average score. “Avg Time” is the average time it took a participant to read one sentence. Standard error is shown. 48 |
| Table 3.10 | Average points earned is the main metric reported in the Results section. This table shows the Spearman’s rank correlation between average points per user and several other possible metrics: (a) the fraction of times the user correctly guessed on or after the boundary; (b) the fraction of times the user guessed exactly on the boundary; and (c) the average number of sentences after the boundary of the user’s guess (giving new score for guesses before the boundary). 49 |
| Table 3.11 | The mean scores for each domain on annotations involving XL-sized models for $p=0.4$. Asterisk denotes generation by CTRL. Interval is $\alpha = 0.95$ confidence. 54 |
| Table 3.12 | The Spearman’s rank correlation coefficient between the number of annotations performed before the current annotation and the score on the current annotation, for all n players who have performed k or more annotations. Players in Group B, who were given extra instruction and incentives, improved over time while those in Group A did not. 56 |
| Table 3.13 | The text of the exit survey questions given to players after completing their annotations 60 |
| Table 3.14 | (top) The possible reasons players could select for why text was machine generated, and (bottom) several examples of custom reasons players wrote. 61 |
| Table 3.15 | The number of times each reason was given for text being machine-generated, and the mean score over those annotations. We see that when players select reasons like “grammar” or “generic,” they are much less likely to be correct than when selecting “common_sense” or “irrelevant.” 63 |
| Table 4.1 | The number of sequences memorized by one model, and not memorized by another. Not all sequences memorized by a small model are also memorized by a larger model. As a model gets larger, it memorizes more unique sequences. 82 |

| | |
|-----------|--|
| Table 4.2 | Qualitative examples of near-duplicates identified by NEARDUP from each dataset. The similarity between documents is highlighted. Note the small interspersed differences that make exact duplicate matching less effective. Examples ending with “[...]” have been truncated for brevity. More data available in Appendix. 98 |
| Table 4.3 | The fraction of examples identified by NEARDUP as near-duplicates. 102 |
| Table 4.4 | The fraction of tokens (note Table 4.3 reports the fraction of <i>examples</i>) identified by EXACTSUBSTR as part of an exact duplicate 50-token substring. 102 |
| Table 4.5 | On the left, we show the URLs that had the greatest proportion of examples marked as near-duplicates by NEARDUP (filtered to URLs which occurred at least 10 times). On the right, we show the 20 most frequent URLs in C4 for which all examples were marked as near-duplicates by NEARDUP. 103 |
| Table 4.6 | When generating 100k sequences with no prompting, over 1% of the tokens emitted from a model trained on the original dataset are part of a 50-token long sequence copied directly from the training dataset. This drops to 0.1% for the deduplicated datasets. 107 |
| Table 4.7 | A selection of substrings identified by EXACTSUBSTR as being in C4 multiple times. The number of times this exact substring occurs in C4 is also given. 113 |
| Table 4.8 | A selection of substrings generated by XL-ORIGINAL with no prompting (and top- k with $k=50$) that were identified by EXACTSUBSTR as being in C4 multiple times. The number of times each substring was found in C4 is given. We observe that most memorized generations tend to be from advertisements. 114 |
| Table 4.9 | For each model, the perplexity of the official validation set (<i>Orig</i>), valid set examples which were identified by NEARDUP as matches of train set examples (<i>Dups</i>), and valid set examples identified by NEARDUP as unique (<i>Unique</i>). Due to the size of the RealNews validation set, we evaluated on only the first 25k examples meeting each condition. 115 |
| Table 5.1 | Example style transfer outputs from augmented zero-shot learning for non-standard styles. 120 |
| Table 5.2 | Comparing augmented zero-shot prompting with supervised style transfer methods on the Yelp sentiment style transfer dataset using automatic evaluation. Acc: accuracy; PPL: perplexity. The inference-only table shows our method applied to 3 different sizes of GPT-3, plus our own LLM. 125 |
| Table 5.3 | Sentiment style transfer results with candidate selection (cand. select.). Candidate selection means that of the sixteen examples returned by our model, we choose the one with the highest BLEU with the source sentence. 126 |

| | |
|------------|---|
| Table 5.4 | For sentiment style transfer, we show accuracy, BLEU, perplexity (PPL), and sequence length (l) reported as the average of positive \rightarrow negative and negative \rightarrow positive. In (a), augmented zero-shot prompting of GLM and GPT-3 are compared with prior supervised style transfer methods. In (b), we compare zero-shot, five-shot, and augmented zero-shot prompting for our GLM models. Candidate selection means that of the sixteen examples returned by the API, we choose the one with the highest BLEU with the source sentence (in the default case, we just use the first returned output). 127 |
| Table 5.6 | Comparing variations of augmented zero-shot learning prompt wording for sentiment style transfer. 129 |
| Table 5.7 | Examples of users' arbitrary style transfer requests for which the model suggestion was accepted. 130 |
| Table 5.8 | In black, we show the exact augmented-zero shot prompts used in our experiments, for <i>LLM</i> and GPT-3 (top), and for <i>LLM-Dialog</i> (bottom). As shown, for <i>LLM-Dialog</i> , we replaced "Here is a rewrite of the text, which is" with "Rewrite it to be". Each line starting with ">" above was passed in as an individual dialog turn. The blue shows how an input text and goal style are concatenated to the few-shot prompt in order to produce final model output. Note that we can achieve high accuracy even though the prompt formulation resulted in some minor grammatical errors for some styles (e.g., "rewrite it to be include the word 'snow'"). Text versions of these prompts can be downloaded at https://bit.ly/3fLDuci . 131 |
| Table 5.9 | Examples of the finetuning objectives. "8" is the approximate length in words of the target sequence. During finetuning, about 25% of training examples took each of these formats. 134 |
| Table 5.10 | Perplexity of evaluation sets when the blank has been filled in using LLM with few-shot prompting (top) and our best fine-tuned T5 model ((bottom). Among the few-shot results, the best method for each dataset is bolded, as well as methods within one standard error. 137 |
| Table 5.11 | Hand-written fill-in-the-blank examples. To construct "custom" few-shot learning prompts, three of these were selected at random. In the end, the custom prompts did not result in better fill-in-the-blank performance than simply selecting randomly examples from the train set. 138 |
| Table 5.12 | Perplexity of evaluation sets according to LLM when the blank has been filled with approaches involving no fine-tuning (top), finetuned approaches (middle), and the groundtruth (bottom). Lower values indicate that the text was considered more fluent by the LLM. 139 |
| Table 5.13 | Perplexity of continuation-based evaluation sets when a continuation has been generated using approaches with no finetuning (top) and two settings of finetuning T5 (middle). 139 |
| Table 5.14 | Accuracy of models finetuned on FILLBLANKCONT at correctly using provided length and goal conditioning signals. 141 |

| | | |
|------------|--|-----|
| Table 5.15 | Usage statistics. | 157 |
| Table 5.16 | Usage statistics broken down by request type (Wordcraft only). | 157 |
| Table 5.17 | Rewrite requests (user completions of ‘Rewrite this...’) | 158 |
| Table 5.18 | Custom prompts written by users. | 158 |

LIST OF ILLUSTRATIONS

- Figure 2.1 Each point corresponds to the outputs from one decoding strategy. The x-axes give the dist-2, ent-4, and perplexity scores of the generated text. The y-axes give the human-judged fluency, coherence, and interestingness of the outputs on a scale from 0 to 1. The Pearson Correlation coefficients between each statistic and the average of fluency, coherence, and interestingness are shown in parentheses. 19
- Figure 3.1 The interface of the task used for human evaluation. Each time the user presses next, the passage's length is doubled. On the left, we show the first step of evaluation, on the right, the second to last. 33
- Figure 3.2 For some of the questions, the text "Dear AMT Worker: to show you're reading, please select definitely [X] for this one." was inserted into the last text segment, and "Did you read carefully?" was appended to the end. 34
- Figure 3.3 In (a), the average (over sequences in the test set) k chosen at each step during generating with nucleus sampling is plotted. Adding a single word of priming strongly impacts the k s chosen for the first few positions, but this difference quickly dissipates. In (b), we consider the first token generated in each sequence by top- k , and plot what fraction of these are captured by the k most common unique tokens from the vocabulary. Overall, at its first step, top- k concentrates 80% of its probability mass in the 500 most common tokens from the vocabulary. 37
- Figure 3.4 (a) and (b) show human rater accuracy of correctly identifying an excerpt as human-written or machine-written, shown with 80% confidence internals, in (a), broken up by decoding strategy and in (b), overall. Accuracy increases as raters observe more tokens. (c) shows that for short excerpts, most rater mistakes are them incorrectly thinking machine-generated text is human written. The two errors types become more balanced at longer lengths. 38
- Figure 3.5 In (a), accuracy increases as the length of the sequences used to train the discriminator is increased. In (b), we see that the BERT fine-tuned discriminator predicts about the same number of false-positives as false-negatives when trained with samples generated using top- p sampling. However, for top- k , it more often mistakes machine-generated text to be human-written, while for untruncated random sampling the opposite is the case. 39
- Figure 3.6 In the boundary detection task, players see one sentence at a time and try to guess when they transition from human-written to machine-generated. 44

- Figure 3.7 The distribution of boundary sentence positions over all game rounds available on RoFt (top), all game rounds that received annotations (middle), and all game rounds included in this paper’s analysis after filtering out problematic annotations (bottom). 51
- Figure 3.8 **(left)** Comparison of mean player score across different genres with GPT-2 XL $p=0.4$ (and CTRL $p=0.4$ for speeches). **(right)** Comparison of mean player score across different values of p for nucleus sampling (GPT-2 XL), as well as a “sanity-check” baseline. 52
- Figure 3.9 **(left)** For Stories, as model size increases (using $p=0.4$), detection becomes harder. **(middle)** For Recipes, extra finetuning does not significantly impact detectability. **(right)** For Speeches, using a “[Politics]” control code (with the CTRL model) has no impact on detectability compared to using a random control code. 53
- Figure 3.10 Performance over time for the two player groups (§3.3.2). Players in Group B, who were given extra instruction and incentives, improved over time while those in Group A did not. 56
- Figure 3.11 Violin plots showing results of our mandatory exit survey. A violin plot is a box plot that also provides a density estimation. Results shown are filtered to only include players who did at least 20 rounds. We see that reading the help guide, being a native English speaker, and providing a custom response for your familiarity with NLG all contribute to a higher mean score while high domain expertise does not seem have an affect (except in the case of short stories, where variance is lower for domain experts). 57
- Figure 3.12 We see that human sentences tended to have a different number of named entities than generated sentences. Players picked up on the correct trend in Stories, but not in News or Speeches. 57
- Figure 3.13 **(left)** Histogram showing the relationship between points earned and the number of seconds an annotation took. Annotators tended to earn more points on annotations they spent longer on. **(right)** Among players who completed at least 100 annotations, average annotation speed decreased with increased experience at the task. 58
- Figure 3.14 The reasons provided by players as to why a given example was generated broken up per model that generated the text 62
- Figure 4.1 We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. Larger models memorize a larger fraction of their training dataset, following a log-linear relationship. This is not just a result of better generalization, as shown by the lack of growth for the GPT-2 baseline models. The dark shaded region is one std away from the mean, and the lighter shaded region represents the min and max over all document lengths. 73

- Figure 4.2 We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. Examples that are repeated more often in the training set are more likely to be extractable, again following a log-linear trend (baseline is GPT-2 XL). 75
- Figure 4.3 We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. As the number of tokens of context available increases, so does our ability to extract memorized text. 76
- Figure 4.4 Percentage of sequences extracted as a function of model scale where we sample randomly from the training set. 77
- Figure 4.5 Percentage of sequences extracted as we vary the length of the prompt. For each sequence length n , $n-50$ tokens are used as the prefix, and we check for extraction of the remaining 50 tokens. 78
- Figure 4.6 Text examples that are memorized by the 6B model, but not by smaller models. Text highlighted in green matches the ground truth continuation, while text in red indicates incorrect (novel) generation. 80
- Figure 4.7 **(left)** Using beam search with $b=100$ slightly increases the data extracted. **(right)** We observe considerably more memorization when checking whether the generated sequence occurs anywhere in the entire training set (Section 4.2.2). However, this approach is very computationally expensive so we do not use it for experiments. 81
- Figure 4.8 Text examples that are memorized by all the models: given 50-token prompts on the left, the next 50 tokens generated by all the models match the groundtruth continuation. 83
- Figure 4.9 Text examples that are memorized by the 125M model (according to true-continuation match), but not memorized by larger models (the generated texts do not match the true continuation, nor any other training examples). The first column shows the prompt. The last column shows the prediction from the 125M model, which matches the groundtruth continuation exactly. 83
- Figure 4.10 Text examples that are memorized but are not heavily duplicated in the training set. Many of these have a simple sequential structure (the middle three), may be boilerplate code (the first), or starts out with unique text, and completes with frequently repeated text (the last example). Overall, these are easily completed sequences. 84
- Figure 4.11 Text examples that are heavily replicated in the training set but not memorized. We find many examples which have slight differences with no semantic (English) meaning. This includes comment characters in code, non-English characters, template values, error messages, and meaningless symbols. We also surprisingly find a large number of slightly different but heavily repeated documents about dumpsters. 84

- Figure 4.12 Text examples that are from The Pile and memorized by GPT2-XL. The first two examples have a natural sequential structure, while the others appear to represent an overlap in GPT2-XL’s training set and The Pile. 85
- Figure 4.13 (a) Masked language model objective: Larger models have a higher fraction of sequences extractable on T5; with one standard deviation of variance shaded in dark and the minimum and maximum shaded light. (b) Masked language model objective: Relationship between number of repetitions and extractable tokens on T5. (c) Causal language model objective: Relationship between number of repetitions and memorization on language models trained with deduplicated data. 86
- Figure 4.14 For each substring of length k , we plot the probability that there exists a second identical length- k substring in the same train set. Matches with length under 10 subword tokens are common, and account for 90% of tokens. We choose a threshold of 50 for experiments. 95
- Figure 4.15 The distribution of near-duplicate cluster sizes from running NEARDUP on C4. 101
- Figure 4.16 Impact of deduplicating the training set on validation perplexity. In (a), we plot the results from T5 base (110M parameters) across three training runs with different random initializations. The black bar represent the lowest perplexity to the highest perplexity, and the colored bar the median perplexity. In (b), we plot the results from T5 XL (1.5B parameters). For C4, we evaluate on *C4 Original*, the original validation set; *C4 Unique*, a subset of the validation set identified by NEARDUP as having zero matches across C4; and *C4 Duplicates*, a subset of the validation set identified by NEARDUP as having a match in the C4 train set. 106
- Figure 4.17 Memorized continuations distribution 107
- Figure 4.18 The proportion of generations which have edit similarity above 0.8 with the groundtruth continuation when using the LM to generate continuations for 32-token prompts identified by NEARDUP as either duplicated or unique. 108
- Figure 5.1 Zero-shot, few-shot, and augmented zero-shot prompts for style transfer. The boldface text is the zero-shot prompt, and the plain text is the additional priming sequence. The full prompts used in this paper are shown in Table 5.8. We encourage readers to examine the outputs of our model at <https://bit.ly/3fLDuci>. 118
- Figure 5.2 The rating UI used for human evaluation. The user may be shown a number of blue squares at once with the same original text and different outputs. 122
- Figure 5.3 Human evaluation of style transfer for six atypical styles. Our method is rated comparably to the human-written ground truth. Error bars are mean standard error. 123
- Figure 5.4 Human evaluation of sentiment and formality transfer. Our method is rated comparably to human-written ground truth as well as prior methods. Error bars show Standard Error of the Mean. Unsup. MT is Prabhume et al. [115]; Dual RL is Luo et al. [97]. 124
- Figure 5.5 A single model that can handle a variety of related writing tasks is more efficient than separate models per task. 133

- Figure 5.6 Human ratings of FITB generations (left) and continuation generations (right). Error bars are 95% confidence intervals. 140
- Figure 5.7 For many of the (validation set, few-shot prompt) combinations, not all validation set examples fit into the maximum sequence length for the LLM. The x-axis on this figure is the fraction of validation set examples which were retained after too-long examples were filtered out. The y-axis is the count of (validation set, few-shot prompt) pairs. 140
- Figure 5.8 A screenshot of the question structure for human evaluation. 142
- Figure 5.9 For each of the FITB validation sets, a histogram of the distribution of sequence lengths (measured in words) of the groundtruth blanked out text and the proposed infills from T5 after and before finetuning). We see that pre-trained T5 tends to produce text that is shorter than the groundtruth. 143
- Figure 5.10 (left) Custom prompting in the Wordcraft interface. The user makes a novel request of the AI agent via a custom prompt (1), and the model attempts to fulfill the request (2). (right) Meta prompting in the Wordcraft interface. The user asks for suggestions for requests they could make of the model (3), and browses those suggestions (4). 146
- Figure 5.11 Screenshot of Wordcraft with the custom prompting feature. 152
- Figure 5.12 Writing habits survey results. 155
- Figure 5.13 Experimental conditions for the user study. The stories written by participants as part of the study can be viewed here: <https://storage.googleapis.com/wordcraft-stories/index.html>. The website also includes purely machine-generated baselines. 156
- Figure 5.14 Exit interview results. 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree. 161
- Figure 5.15 Results from exit interview question ‘Which controls did you enjoy using?’ 161

BIBLIOGRAPHY

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. “Deep learning with Differential Privacy.” In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.
- [2] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. “Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection.” In: *International Conference on Advanced Information Networking and Applications*. Springer. 2020, pp. 1341–1354.
- [3] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. “STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation.” In: *arXiv preprint arXiv:2010.01717* (2020).
- [4] Miltiadis Allamanis. “The adverse effects of code duplication in machine learning models of code.” In: *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. 2019, pp. 143–153.
- [5] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election.” In: *Journal of economic perspectives* 31.2 (2017), pp. 211–36.
- [6] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. “Large-scale differentially private BERT.” In: *arXiv preprint arXiv:2108.01624* (2021).
- [7] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. “Program synthesis with large language models.” In: *arXiv preprint arXiv:2108.07732* (2021).
- [8] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc’Aurelio Ranzato, and Arthur Szlam. “Real or Fake? Learning to Discriminate Machine from Human Generated Text.” In: *arXiv preprint arXiv:1906.03351* (2019).
- [9] Jack Bandy and Nicholas Vincent. *Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus*. 2021. arXiv: [2105.05241](https://arxiv.org/abs/2105.05241) [cs.CL].
- [10] Emily M. Bender and Batya Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 587–604. DOI: [10.1162/tacl_a_00041](https://doi.org/10.1162/tacl_a_00041). URL: <https://www.aclweb.org/anthology/Q18-1041>.
- [11] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.

- [12] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*. Version 1.0. If you use this software, please cite it using these metadata. Mar. 2021. DOI: [10.5281/zenodo.5297715](https://doi.org/10.5281/zenodo.5297715). URL: <https://doi.org/10.5281/zenodo.5297715>.
- [13] Burton H Bloom. “Space/time trade-offs in hash coding with allowable errors.” In: *Communications of the ACM* 13.7 (1970), pp. 422–426.
- [14] Gwern Branwen. “GPT-3 Creative Fiction.” In: (2020). URL: <https://www.gwern.net/GPT-3>.
- [15] Andrei Z Broder. “On the resemblance and containment of documents.” In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE. 1997, pp. 21–29.
- [16] D. W. Brown. *Corpus of Presidential Speeches*. Retrieved from <http://www.thegrammarlab.com>. 2016.
- [17] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramèr. *What Does it Mean for a Language Model to Preserve Privacy?* 2022. arXiv: [2202.05520](https://arxiv.org/abs/2202.05520) [stat.ML].
- [18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. *Language Models are Few-Shot Learners*. 2020. arXiv: [2005.14165](https://arxiv.org/abs/2005.14165) [cs.CL].
- [19] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models are Few-Shot Learners.” In: *CoRR* abs/2005.14165 (2020). arXiv: [2005.14165](https://arxiv.org/abs/2005.14165). URL: <https://arxiv.org/abs/2005.14165>.
- [20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “Language Models are Few-shot Learners.” In: *Advances in Neural Information Processing Systems 33*. 2020.
- [21] Daniel Buschek, Martin Zürn, and Malin Eiband. “The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers.” In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–13.
- [22] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. “Quantifying Memorization across Neural Language Models.” In: *arXiv preprint arXiv:2202.07646* (2022).
- [23] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. “The secret sharer: Evaluating and testing unintended memorization in neural networks.” In: *USENIX Security Symposium*. 2019.

- [24] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. *Extracting Training Data from Large Language Models*. 2020. arXiv: [2012.07805](https://arxiv.org/abs/2012.07805) [cs.CR].
- [25] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. “One billion word benchmark for measuring progress in statistical language modeling.” In: *arXiv preprint arXiv:1312.3005* (2013).
- [26] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. “Evaluating large language models trained on code.” In: *arXiv preprint arXiv:2107.03374* (2021).
- [27] Hung Chim and Xiaotie Deng. “A New Suffix Tree Similarity Measure for Document Clustering.” In: *Proceedings of the 16th International Conference on World Wide Web. WWW '07*. Banff, Alberta, Canada: Association for Computing Machinery, 2007, pp. 121–130. ISBN: 9781595936547. DOI: [10.1145/1242572.1242590](https://doi.org/10.1145/1242572.1242590). URL: <https://doi.org/10.1145/1242572.1242590>.
- [28] Kyunghyun Cho. “Noisy parallel approximate decoding for conditional recurrent language model.” In: 2016.
- [29] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text.” In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 7282–7296.
- [30] Andy Coenen, Luke Davis, Daphne Ippolito, Ann Yuan, and Emily Reif. “Wordcraft: a Human-AI Collaborative Editor for Story Writing.” In: (2021).
- [31] Edith Cohen. *Min-Hash Sketches: A Brief Survey*. 2016. URL: <http://www.cohenwang.com/edith/Surveys/minhash.pdf>.
- [32] Nicole A Cooke. *Fake News and Alternative Facts: Information literacy in a Post-Truth Era*. American Library Association, 2018.
- [33] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. “Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5997–6007. DOI: [10.18653/v1/P19-1601](https://doi.org/10.18653/v1/P19-1601). URL: <https://www.aclweb.org/anthology/P19-1601>.
- [34] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. “Transformer-xl: Attentive language models beyond a fixed-length context.” In: *arXiv preprint arXiv:1901.02860* (2019).
- [35] Jan Milan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Álvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. “Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 3971–3984.

- [36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423>.
- [37] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. “Documenting the English Colossal Clean Crawled Corpus.” In: *arXiv preprint arXiv:2104.08758* (Apr. 2021). arXiv: [2104.08758](https://arxiv.org/abs/2104.08758) [cs.CL].
- [38] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. *Documenting the English Colossal Clean Crawled Corpus*. 2021. arXiv: [2104.08758](https://arxiv.org/abs/2104.08758) [cs.CL].
- [39] Chris Donahue, Mina Lee, and Percy Liang. “Enabling Language Models to Fill in the Blanks.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 2492–2501.
- [40] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. “RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 189–196.
- [41] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “Calibrating noise to sensitivity in private data analysis.” In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284.
- [42] Angela Fan, Mike Lewis, and Yann Dauphin. “Hierarchical Neural Story Generation.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 889–898. URL: <https://www.aclweb.org/anthology/P18-1082>.
- [43] William Fedus, Barret Zoph, and Noam Shazeer. *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. 2021. arXiv: [2101.03961](https://arxiv.org/abs/2101.03961) [cs.LG].
- [44] Vitaly Feldman and Chiyuan Zhang. “What neural networks memorize and why: Discovering the long tail via influence estimation.” In: *Advances in Neural Information Processing Systems*. 2020.
- [45] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures.” In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333.
- [46] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. “Style Transfer in Text: Exploration and Evaluation.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018. URL: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17015>.
- [47] Rodney A. Gabriel, Tsung-Ting Kuo, Julian McAuley, and Chun-Nan Hsu. “Identifying and characterizing highly similar notes in big clinical note datasets.” In: *Journal of Biomedical Informatics* 82 (2018), pp. 63–69. ISSN: 1532-0464. DOI: <https://doi.org/10.1016/j.jbi.2018.04.009>. URL: <https://www.sciencedirect.com/science/article/pii/S153204641830073X>.

- [48] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. “Property inference attacks on fully connected neural networks using permutation invariant representations.” In: *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018, pp. 619–633.
- [49] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. “The Pile: An 800GB Dataset of Diverse Text for Language Modeling.” In: *arXiv preprint arXiv:2101.00027* (2020).
- [50] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. “The Pile: An 800GB Dataset of Diverse Text for Language Modeling.” In: *arXiv preprint arXiv:2101.00027* (2020).
- [51] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. *Datasheets for Datasets*. 2020. arXiv: [1803.09010](https://arxiv.org/abs/1803.09010) [cs.DB].
- [52] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. “GLTR: Statistical Detection and Visualization of Generated Text.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 111–116.
- [53] Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. “A systematic exploration of diversity in machine translation.” In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1100–1111. URL: <https://ttic.uchicago.edu/~gregory/papers/emnlp2013diversity.pdf>.
- [54] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. “English gigaword.” In: *Linguistic Data Consortium, Philadelphia* 4.1 (2003), p. 34.
- [55] Alex Graves. “Generating sequences with recurrent neural networks.” In: *arXiv preprint arXiv:1308.0850* (2013).
- [56] Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. “Wiki-40B: Multilingual Language Model Dataset.” In: *LREC 2020*. 2020. URL: <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.296.pdf>.
- [57] Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. “Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings.” In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 901–910.
- [58] Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. “Unifying Human and Statistical Evaluation for Natural Language Generation.” In: *CoRR* abs/1904.02792 (2019). arXiv: [1904.02792](https://arxiv.org/abs/1904.02792). URL: <http://arxiv.org/abs/1904.02792>.
- [59] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. *Ethical Challenges in Data-Driven Dialogue Systems*. 2017. arXiv: [1711.09050](https://arxiv.org/abs/1711.09050) [cs.CL].
- [60] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory.” In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [61] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. “The Curious Case of Neural Text Degeneration.” In: *CoRR* abs/1904.09751 (2019).

- [62] Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 328–339.
- [63] Zhiqiang Hu, Roy Ka-Wei Lee, and Charu C. Aggarwal. “Text Style Transfer: A Review and Experiment Evaluation.” In: *CoRR abs/2010.12742* (2020). arXiv: 2010.12742. URL: <https://arxiv.org/abs/2010.12742>.
- [64] Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. “INSET: Sentence Infilling with INter-SENTential Transformer.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 2502–2515.
- [65] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. “Automatic Detection of Generated Text is Easiest when Humans are Fooled.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 1808–1822.
- [66] Daphne Ippolito, Liam Dugan, Emily Reif, Ann Yuan, Andy Coenen, and Chris Callison-Burch. “The Case for a Single Model that can Both Generate Continuations and Fill in the Blank.” In: (2022). DOI: 10.48550/ARXIV.2206.04812. URL: <https://arxiv.org/abs/2206.04812>.
- [67] Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. “Unsupervised hierarchical story infilling.” In: *Proceedings of the First Workshop on Narrative Understanding*. 2019, pp. 37–43.
- [68] Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. “Comparison of Diverse Decoding Methods from Conditional Language Models.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3752–3762.
- [69] Paul Jaccard. “The distribution of the flora in the alpine zone.” In: *New phytologist* 11.2 (1912), pp. 37–50.
- [70] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. “Auditing differentially private machine learning: How private is private sgd?” In: *arXiv preprint arXiv:2006.07709* (2020).
- [71] Bargav Jayaraman and David Evans. “Evaluating differentially private machine learning in practice.” In: *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 2019, pp. 1895–1912.
- [72] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. “Deep Learning for Text Style Transfer: A Survey.” In: *CoRR abs/2011.00416* (2020). arXiv: 2011.00416. URL: <https://arxiv.org/abs/2011.00416>.
- [73] Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. “IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and Translation.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3097–3109. DOI: 10.18653/v1/D19-1306. URL: <https://www.aclweb.org/anthology/D19-1306>.
- [74] Anjali Kannan and Oriol Vinyals. “Adversarial evaluation of dialogue models.” In: *arXiv preprint arXiv:1701.08198* (2017).

- [75] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. “Scaling laws for neural language models.” In: *arXiv preprint arXiv:2001.08361* (2020).
- [76] Juha Kärkkäinen and Peter Sanders. “Simple linear work suffix array construction.” In: *International colloquium on automata, languages, and programming*. Springer. 2003, pp. 943–955.
- [77] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. “Ctrl: A conditional transformer language model for controllable generation.” In: *arXiv preprint arXiv:1909.05858* (2019).
- [78] Pang Ko and Srinivas Aluru. “Space efficient linear time construction of suffix arrays.” In: *Annual Symposium on Combinatorial Pattern Matching*. Springer. 2003, pp. 200–210.
- [79] Kalpesh Krishna, John Wieting, and Mohit Iyyer. “Reformulating Unsupervised Style Transfer as Paraphrase Generation.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 737–762. DOI: [10.18653/v1/2020.emnlp-main.55](https://doi.org/10.18653/v1/2020.emnlp-main.55). URL: <https://www.aclweb.org/anthology/2020.emnlp-main.55>.
- [80] Taku Kudo and John Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.” In: *CoRR abs/1808.06226* (2018). arXiv: [1808.06226](https://arxiv.org/abs/1808.06226). URL: <http://arxiv.org/abs/1808.06226>.
- [81] Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. “Importance of a Search Strategy in Neural Dialogue Modelling.” In: 2018.
- [82] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. “Best practices for the human evaluation of automatically generated text.” In: *Proceedings of the 12th International Conference on Natural Language Generation*. 2019, pp. 355–368.
- [83] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. “Deduplicating Training Data Makes Language Models Better.” In: 2022.
- [84] Brian Lester, Rami Al-Rfou, and Noah Constant. “The Power of Scale for Parameter-Efficient Prompt Tuning.” In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 3045–3059.
- [85] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880.
- [86] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. “A Diversity-Promoting Objective Function for Neural Conversation Models.” In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 110–119. DOI: [10.18653/v1/N16-1014](https://doi.org/10.18653/v1/N16-1014). URL: <https://www.aclweb.org/anthology/N16-1014>.

- [87] Jiwei Li and Dan Jurafsky. “Mutual information and diverse decoding improve neural machine translation.” In: 2016.
- [88] Jiwei Li, Will Monroe, and Dan Jurafsky. “A simple, fast diverse decoding algorithm for neural generation.” In: 2016.
- [89] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. “Adversarial learning for neural dialogue generation.” In: *arXiv preprint arXiv:1701.06547* (2017).
- [90] Juncen Li, Robin Jia, He He, and Percy Liang. “Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer.” In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1865–1874. DOI: [10.18653/v1/N18-1169](https://doi.org/10.18653/v1/N18-1169). URL: <https://www.aclweb.org/anthology/N18-1169>.
- [91] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. “CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020, pp. 1823–1840.
- [92] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. “Adversarial ranking for language generation.” In: *Advances in Neural Information Processing Systems*. 2017, pp. 3155–3165.
- [93] Stephanie Lin, Jacob Hilton, and Owain Evans. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. 2021. arXiv: [2109.07958](https://arxiv.org/abs/2109.07958) [cs.CL].
- [94] Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. “Revision in continuous space: Unsupervised text style transfer without adversarial learning.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8376–8383. URL: <https://arxiv.org/abs/1905.12304>.
- [95] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.” In: *arXiv preprint arXiv:2107.13586* (2021). URL: <https://arxiv.org/pdf/2107.13586.pdf>.
- [96] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. “Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses.” In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 1116–1126.
- [97] Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. “A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer.” In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 5116–5122. DOI: [10.24963/ijcai.2019/711](https://doi.org/10.24963/ijcai.2019/711). URL: <https://doi.org/10.24963/ijcai.2019/711>.
- [98] Udi Manber and Gene Myers. “Suffix arrays: a new method for on-line string searches.” In: *siam Journal on Computing* 22.5 (1993), pp. 935–948.

- [99] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. “Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images.” In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).
- [100] Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. “Simultaneous translation and paraphrase for language education.” In: *Proceedings of the Fourth Workshop on Neural Generation and Translation*. 2020, pp. 232–243.
- [101] R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. “How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN.” In: *CoRR abs/2111.09509* (2021). arXiv: [2111.09509](https://arxiv.org/abs/2111.09509). URL: <https://arxiv.org/abs/2111.09509>.
- [102] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” In: *arXiv preprint arXiv:1301.3781* (2013).
- [103] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. “Evaluating Style Transfer for Text.” In: *CoRR abs/1904.02295* (2019). arXiv: [1904.02295](http://arxiv.org/abs/1904.02295). URL: <http://arxiv.org/abs/1904.02295>.
- [104] Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. “Finding and Generating a Missing Part for Story Completion.” In: *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2020, pp. 156–166.
- [105] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. “A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories.” In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 839–849. DOI: [10.18653/v1/N16-1098](https://doi.org/10.18653/v1/N16-1098). URL: <https://www.aclweb.org/anthology/N16-1098>.
- [106] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. *WebGPT: Browser-assisted question-answering with human feedback*. 2021. arXiv: [2112.09332](https://arxiv.org/abs/2112.09332) [cs.CL].
- [107] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond.” In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 2016, pp. 280–290.
- [108] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. “Adversary instantiation: Lower bounds for differentially private machine learning.” In: *arXiv preprint arXiv:2101.04535* (2021).
- [109] Colin Raffel, Nikhil Kandpal, Eric Wallace. “Deduplicating Training Data Mitigates Privacy Risks in Language Models.” In: *arXiv preprint* (2021).
- [110] Ge Nong, Sen Zhang, and Wai Hong Chan. “Linear suffix array construction by almost pure induced-sorting.” In: *2009 data compression conference*. IEEE. 2009, pp. 193–202.

- [111] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. “Why We Need New Evaluation Metrics for NLG.” In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2241–2252.
- [112] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: a Method for Automatic Evaluation of Machine Translation.” In: *Association for Computational Linguistics*. 2001. URL: <https://www.aclweb.org/anthology/P02-1040>.
- [113] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. *Carbon Emissions and Large Neural Network Training*. 2021. arXiv: 2104.10350 [cs.LG].
- [114] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation.” In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: <http://www.aclweb.org/anthology/D14-1162>.
- [115] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. “Style Transfer Through Back-Translation.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 866–876. DOI: 10.18653/v1/P18-1080. URL: <https://www.aclweb.org/anthology/P18-1080>.
- [116] Raul Puri and Bryan Catanzaro. “Zero-shot text classification with generative language models.” In: *arXiv preprint arXiv:1912.10165* (2019). URL: <https://arxiv.org/abs/1912.10165>.
- [117] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving language understanding by generative pre-training.” In: (2018).
- [118] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language models are unsupervised multitask learners.” In: *OpenAI Blog* 1.8 (2019).
- [119] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- [120] Swaroop Ramaswamy, Om Thakkar, Rajiv Mathews, Galen Andrew, H Brendan McMahan, and Françoise Beaufays. “Training production language models without memorizing user data.” In: *arXiv preprint arXiv:2009.10031* (2020).
- [121] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text-conditional image generation with clip latents.” In: *arXiv preprint arXiv:2204.06125* (2022).
- [122] Sudha Rao and Joel Tetreault. “Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer.” In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 129–140. DOI: 10.18653/v1/N18-1012. URL: <https://www.aclweb.org/anthology/N18-1012>.

- [123] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. “A Recipe for Arbitrary Text Style Transfer with Large Language Models.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022.
- [124] Laria Reynolds and Kyle McDonell. *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm*. 2021. arXiv: [2102.07350](https://arxiv.org/abs/2102.07350) [cs.CL].
- [125] Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C. Uthus, and Zarana Parekh. “TextSETTR: Label-Free Text Style Extraction and Tunable Targeted Restyling.” In: *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) (2021)*. arXiv: [2010.03802](https://arxiv.org/abs/2010.03802). URL: <https://arxiv.org/abs/2010.03802>.
- [126] A Roberts and C Raffel. “Exploring Transfer Learning with T5: The Text-To-Text Transfer Transformer.” In: *Google AI Blog* (2020). URL: <https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html>.
- [127] Melissa Roemmele. “Inspiration through Observation: Demonstrating the Influence of Automatically Generated Text on Creative Writing.” In: (2021), pp. 52–461. URL: https://computationalcreativity.net/iccc21/wp-content/uploads/2021/09/ICCC_2021_paper_32.pdf.
- [128] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. “Photo-realistic Text-to-Image Diffusion Models with Deep Language Understanding.” In: *arXiv preprint arXiv:2205.11487* (2022).
- [129] Keisuke Sakaguchi and Benjamin Van Durme. “Efficient Online Scalar Annotation with Bounded Support.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 208–218. DOI: [10.18653/v1/P18-1020](https://doi.org/10.18653/v1/P18-1020). URL: <https://www.aclweb.org/anthology/P18-1020>.
- [130] Evan Sandhaus. “The new york times annotated corpus.” In: *Linguistic Data Consortium, Philadelphia* 6.12 (2008), e26752.
- [131] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. “Multitask Prompted Training Enables Zero-Shot Task Generalization.” In: *International Conference on Learning Representations*. 2021.
- [132] Timo Schick and Hinrich Schütze. “It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners.” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2339–2352. URL: <https://www.aclweb.org/anthology/2021.naacl-main.185>.
- [133] Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. “Are We Safe Yet? The Limitations of Distributional Features for Fake News Detection.” In: *arXiv preprint arXiv:1908.09805* (2019).
- [134] Abigail See, Peter J. Liu, and Christopher D. Manning. “Get To The Point: Summarization with Pointer-Generator Networks.” In: *CoRR* abs/1704.04368 (2017). arXiv: [1704.04368](https://arxiv.org/abs/1704.04368). URL: <http://arxiv.org/abs/1704.04368>.

- [135] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. “Do Massively Pretrained Language Models Make Better Storytellers?” In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 2019, pp. 843–861.
- [136] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units.” In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: [10.18653/v1/P16-1162](https://doi.org/10.18653/v1/P16-1162). URL: <https://www.aclweb.org/anthology/P16-1162>.
- [137] Noam Shazeer and Mitchell Stern. “Adafactor: Adaptive learning rates with sublinear memory cost.” In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4596–4604.
- [138] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. “Style Transfer from Non-Parallel Text by Cross-Alignment.” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/2d2c8394e31101a261abf1784302bf75-Paper.pdf>.
- [139] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership inference attacks against machine learning models.” In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.
- [140] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. “Release Strategies and the Social Impacts of Language Models.” In: *arXiv preprint arXiv:1908.09203* (2019).
- [141] Jonghyuk Song, Sangho Lee, and Jong Kim. “Crowdtarget: Target-based detection of crowdturfing in online social networks.” In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2015, pp. 793–804.
- [142] Emma Strubell, Ananya Ganesh, and Andrew McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. 2019. arXiv: [1906.02243](https://arxiv.org/abs/1906.02243) [cs.CL].
- [143] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. ““Transforming” Delete, Retrieve, Generate Approach for Controlled Text Style Transfer.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3269–3279. DOI: [10.18653/v1/D19-1322](https://doi.org/10.18653/v1/D19-1322). URL: <https://www.aclweb.org/anthology/D19-1322>.
- [144] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to Sequence Learning with Neural Networks.” In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 3104–3112. URL: <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.

- [145] Yik-Cheung Tam, Jiachen Ding, Cheng Niu, and Jie Zhou. “Cluster-based Beam Search for Pointer-Generator Chatbot Grounded by Knowledge.” In: *Dialog System Technology Challenges 7 at AAAI 2019*. 2019.
- [146] Wilson L Taylor. ““Cloze procedure”: A new tool for measuring readability.” In: *Journalism Bulletin* 30.4 (1953), pp. 415–433.
- [147] Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Françoise Beaufays. *Understanding Unintended Memorization in Federated Learning*. 2020. arXiv: [2006.07490 \[cs.LG\]](https://arxiv.org/abs/2006.07490).
- [148] Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. “Investigating the impact of pre-trained word embeddings on memorization in neural networks.” In: *International Conference on Text, Speech, and Dialogue*. Springer. 2020, pp. 273–281.
- [149] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. “LaMDA: Language Models for Dialog Applications.” In: *arXiv preprint arXiv:2201.08239* (2022).
- [150] Trieu H Trinh and Quoc V Le. “A simple method for commonsense reasoning.” In: *arXiv preprint arXiv:1806.02847* (2018).
- [151] Alan Turing. “Computing machinery and intelligence-AM Turing.” In: *Mind* 59.236 (1950), p. 433.
- [152] Chris J Vargo, Lei Guo, and Michelle A Amazeen. “The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016.” In: *New media & society* 20.5 (2018), pp. 2028–2049.
- [153] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need.” In: *CoRR* abs/1706.03762 (2017). arXiv: [1706.03762](https://arxiv.org/abs/1706.03762). URL: <http://arxiv.org/abs/1706.03762>.
- [154] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [155] Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Josef Och, and Juri Ganitkevitch. “Watermarking the outputs of structured prediction with an application in statistical machine translation.” In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 1363–1372.
- [156] Yannick Versley and Yana Panchenko. “Not just bigger: Towards better-quality Web corpora.” In: *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. 2012, pp. 44–52.
- [157] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models.” In: 2016.
- [158] Oriol Vinyals and Quoc Le. “A Neural Conversational Model.” In: *arXiv preprint arXiv:1506.05869* (2015).
- [159] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online.” In: *Science* 359.6380 (2018), pp. 1146–1151.

- [160] Ben Wang and Aran Komatsuzaki. *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. <https://github.com/kingoflolz/mesh-transformer-jax>. May 2021.
- [161] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. “Serf and turf: crowdturfing for fun and profit.” In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 679–688.
- [162] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. “Finetuned Language Models are Zero-Shot Learners.” In: *International Conference on Learning Representations*. 2021.
- [163] Peter Weiner. “Linear pattern matching algorithms.” In: *14th Annual Symposium on Switching and Automata Theory (swat 1973)*. IEEE. 1973, pp. 1–11.
- [164] Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. “Learning from Task Descriptions.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1361–1375. DOI: [10.18653/v1/2020.emnlp-main.105](https://doi.org/10.18653/v1/2020.emnlp-main.105). URL: <https://www.aclweb.org/anthology/2020.emnlp-main.105>.
- [165] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. “Transformers: State-of-the-Art Natural Language Processing.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [166] Yonghui Wu. “Smart compose: Using neural networks to help write emails.” In: *Google AI Blog* (2018).
- [167] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation.” In: *arXiv preprint arXiv:1609.08144* (2016).
- [168] Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. “Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 979–988. DOI: [10.18653/v1/P18-1090](https://doi.org/10.18653/v1/P18-1090). URL: <https://www.aclweb.org/anthology/P18-1090>.
- [169] Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. “On Variational Learning of Controllable Representations for Text without Supervision.” In: *Proceedings of the International Conference on Machine Learning (ICML)* abs/1905.11975 (2020). arXiv: [1905.11975](https://arxiv.org/abs/1905.11975). URL: <http://arxiv.org/abs/1905.11975>.

- [170] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. “mT5: A massively multilingual pre-trained text-to-text transformer.” In: *arXiv preprint arXiv:2010.11934* (2020).
- [171] Mikio Yamamoto and Kenneth W Church. “Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus.” In: *Computational Linguistics* 27.1 (2001), pp. 1–30.
- [172] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. “Privacy risk in machine learning: Analyzing the connection to overfitting.” In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.
- [173] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. “Wordcraft: Story Writing With Large Language Models.” In: *27th International Conference on Intelligent User Interfaces*. 2022, pp. 841–852.
- [174] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. “Defending Against Neural Fake News.” In: *arXiv preprint arXiv:1905.12616* (2019).
- [175] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyang Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. “PanGu- α : Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation.” In: *arXiv preprint arXiv:2104.12369* (2021).
- [176] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. “Counterfactual Memorization in Neural Language Models.” In: *arXiv preprint arXiv:2112.12938* (2021).
- [177] Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. “Trading Off Diversity and Quality in Natural Language Generation.” In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. 2021, pp. 25–33.
- [178] Xiang Zhang, Junbo Zhao, and Yann LeCun. “Character-Level Convolutional Networks for Text Classification.” In: *Proceedings of the Conference on Neural Information Processing Systems* (Sept. 2015). arXiv: [1509.01626](https://arxiv.org/abs/1509.01626) [cs].
- [179] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. “Generating informative and diverse conversational responses via adversarial information maximization.” In: *Advances in Neural Information Processing Systems*. 2018, pp. 1815–1825. URL: <https://papers.nips.cc/paper/7452-generating-informative-and-diverse-conversational-responses-via-adversarial-information-maximization.pdf>.
- [180] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. *Calibrate Before Use: Improving Few-Shot Performance of Language Models*. 2021. arXiv: [2102.09690](https://arxiv.org/abs/2102.09690) [cs.CL].
- [181] Wanrong Zhu, Zhiting Hu, and Eric Xing. “Text infilling.” In: *arXiv preprint arXiv:1901.00158* (2019).

- [182] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.
- [183] Zheming Zhu, Delphine Bernhard, and Iryna Gurevych. “A Monolingual Tree-based Translation Model for Sentence Simplification.” In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 1353–1361. URL: <https://www.aclweb.org/anthology/C10-1152>.
- [184] Albert Ziegler. *GitHub Copilot: Parrot or Crow?* <https://docs.github.com/en/github/copilot/research-recitation>. 2021.
- [185] Jakub Łacki, Vahab Mirrokni, and Michał Włodarczyk. *Connected Components at Scale via Local Contractions*. 2018. arXiv: [1807.10727](https://arxiv.org/abs/1807.10727) [cs.DC].