

AI AS HUMANITY’S SALIERI: QUANTIFYING LINGUISTIC CREATIVITY OF LANGUAGE MODELS VIA SYSTEMATIC ATTRIBUTION OF MACHINE TEXT AGAINST WEB TEXT

Anonymous authors

Paper under double-blind review

ABSTRACT

Creativity has long been considered one of the most difficult aspect of human intelligence for AI to mimic. However, the rise of Large Language Models (LLMs), like ChatGPT, has raised questions about whether AI can match or even surpass human creativity. We present CREATIVITY INDEX as the first step to quantify the linguistic creativity of a text by reconstructing it from existing text snippets on the web. CREATIVITY INDEX is motivated by the hypothesis that the seemingly remarkable creativity of LLMs may be attributable in large part to the creativity of human-written texts on the web. To compute CREATIVITY INDEX efficiently, we introduce DJ SEARCH, a novel dynamic programming algorithm that can search verbatim and near-verbatim matches of text snippets from a given document against the web. Experiments reveal that the CREATIVITY INDEX of professional human authors is on average 66.2% higher than that of LLMs, and that alignment reduces the CREATIVITY INDEX of LLMs by an average of 30.1%. In addition, we explore variations in the CREATIVITY INDEX among different human authors and discuss the potential factors contributing to these differences. Finally, we showcase a novel application of CREATIVITY INDEX for zero-shot machine text detection, where it proves to be surprisingly effective—outperforming the strong zero-shot system DetectGPT by a substantial margin of 30.2%, and even surpassing a leading supervised system, GhostBuster, in five out of six domains.

1 INTRODUCTION

Creativity has long been considered one of the most challenging “holy grail” of human intelligence for AI to mimic (Hasselberger & Lott, 2023). However, Large Language Models (LLMs) such as ChatGPT have taken the world by storm with their creative power. From generating poetry (Sawicki et al.; Deng et al., 2024b; Sawicki et al., 2023) and composing music (Ding et al., 2024; Deng et al., 2024a; Liang et al., 2024) to designing artwork (Makatura et al., 2024; Jignasu et al., 2023; Lim et al., 2024) and crafting compelling narratives (Yuan et al., 2022; Mirowski et al., 2023a; Ippolito et al., 2022), LLMs take only seconds to produce outputs that would rival or even surpass the work of human creators. This proficiency has even sparked a growing trend of using LLMs for content creation in industrial settings. For example, major studios in Hollywood have integrated LLMs into production processes such as movie scriptwriting (Carnevale, 2023). While studio executives are optimistic about using LLMs to streamline production and reduce costs, Hollywood writers are deeply concerned about being replaced by the rapid integration of LLMs in the industry, leading to a five-month writers’ strike (Koblin & John, 2023).

While science fiction writer Ted Chiang characterizes LLMs as a blurry JPEG of the web (Hubert et al., 2024), many others wonder whether AI can indeed match or surpass the creativity of humanity. After all, LLMs have consumed orders of magnitude more works of writing than any single human could ever read, thus it may seem possible that LLMs could consequently reach a new level of literary sophistication and creativity beyond that of humanity at large.

To answer this question, the first step is to assess the level of creativity in machine texts compared to human texts. Creativity is a complex and ambiguous process that is challenging to define and

054 quantify (Csikszentmihalyi, 1997; Glaveanu et al., 2020; Eagleman & Brandt, 2017; Paeth). Several
055 previous studies have attempted to quantify creativity in writing by developing specific rubrics and
056 asking human evaluators to score the writing based on these criteria. Vaezi & Rezaei (2018) devel-
057 oped a comprehensive rubric to assess fiction writing, while Biggs & Collis (1982) used a taxonomy
058 of structural complexity to categorize creative writing. More recently, Chakrabarty et al. (2024) ap-
059 plied the Torrance Test of Creative Thinking to evaluate the creativity of short stories generated by
060 LLMs in terms of fluency, flexibility, originality and elaboration. While these rubric-based methods
061 are valuable, scaling them up to evaluate large amounts of texts generated by LLMs is impractical
062 due to the reliance on human evaluators.

063 In this work, we propose CREATIVITY INDEX, a novel statistical measure of [linguistic creativity](#)
064 in text. The key intuition underlying CREATIVITY INDEX is to quantify the degree of linguistic
065 creativity of a given text by reconstructing that text via mixing and matching of a vast amount of
066 existing text snippets on the web (See Figure 1a; 24 additional examples in Appendix Fig. 5 to
067 Fig. 30). The underlying premise of our work is that the seemingly remarkable creativity of LLMs
068 may be in large part attributable to the remarkable creativity of human-written texts on the web.
069 This contrasts with distinguished human authors such as Hemingway, whose original content and
070 unique writing style cannot be easily replicated by simply assembling snippets from other works. To
071 test this, we provide a novel computational approach to systematically attribute machine text to web
072 texts. Specifically, we introduce DJ SEARCH,¹ a novel dynamic programming algorithm that can
073 efficiently search for verbatim and near-verbatim matches of text snippets from a given document
074 against the web. Here, near-verbatim matches are defined as close paraphrases, characterized by
075 high semantic similarity. Our algorithm combines strict verbatim matching using Infini-gram (Liu
076 et al., 2024), which allows for fast retrieval of any existing sequence of words, with near-verbatim
077 semantic matching achieved through a novel application of Word Mover’s Distance (WMD) (Kusner
078 et al., 2015) computed on the word embeddings of text snippets.

078 The contribution of our work is threefold: First, we introduce the CREATIVITY INDEX to reveal
079 novel insights about machine creativity and human creativity. We find that the CREATIVITY INDEX
080 of human authors—specifically professional writers and historical figures—is on average 66.2%
081 higher than that of LLMs. This creativity gap is consistent across various domains—novel snippets,
082 modern poems, and speech transcripts—at both verbatim and semantic levels. Moreover, we no-
083 tice that Reinforcement Learning from Human Feedback (RLHF), a widely used alignment method,
084 dramatically reduces the CREATIVITY INDEX of LLMs, by an average of 30.1%. This reduction is
085 more significant at the verbatim level than the semantic level, indicating that LLMs may have con-
086 verged to certain linguistic style preferred by humans during alignment. [Furthermore, we explore](#)
087 [variations in CREATIVITY INDEX among different human authors.](#) Famous authors like Heming-
088 way and Dickens tend to have higher CREATIVITY INDEX, though this should be interpreted with
089 caution. Beyond inherent differences in creativity, CREATIVITY INDEX can be influenced by factors
090 such as writing style and the time of composition. For instance, older English writings may exhibit
091 higher CREATIVITY INDEX, as they are more difficult to reconstruct from web texts.

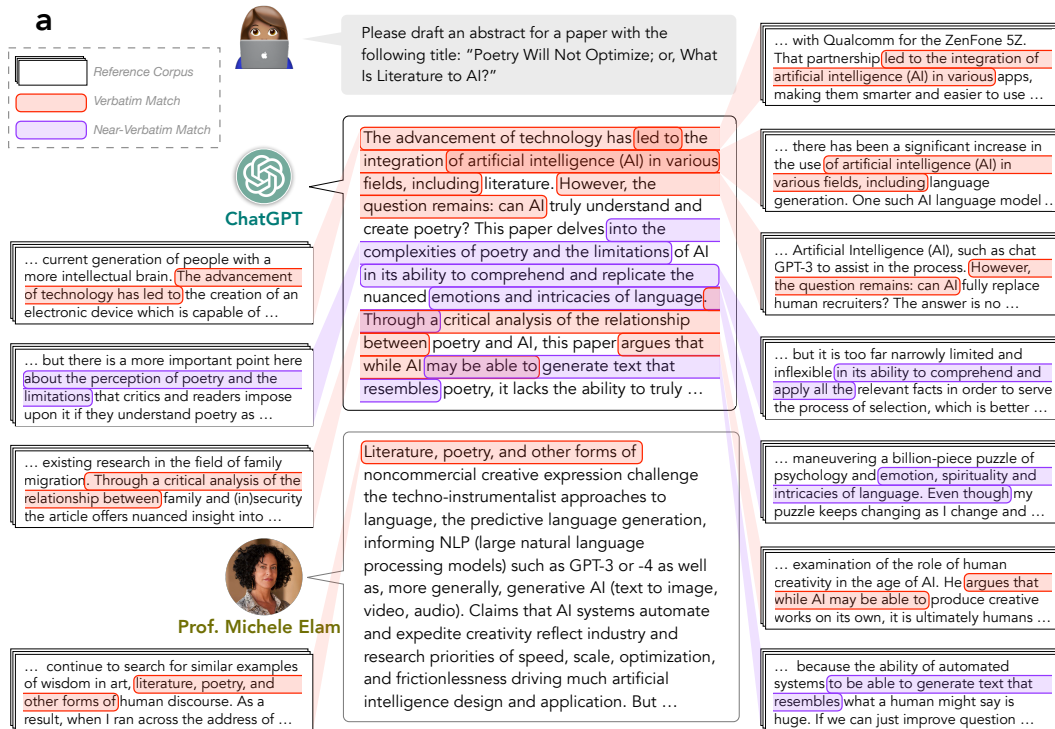
092 Second, we introduce DJ SEARCH as an efficient algorithmic tool to trace the usage of existing text
093 snippets from the web that LLMs incorporate to compose new generations. The power of LLMs
094 arises from training exhaustively on existing human-written texts on the web, and it is meaningful to
095 trace back and acknowledge the human writers whose work empowers these models’ outputs—just
096 as we credit original composers when enjoying a DJ’s remix.

097 Finally, we demonstrate a novel use of CREATIVITY INDEX as a [surprisingly effective criterion for](#)
098 [zero-shot black-box machine text detection.](#) Our method is ready to deploy out-of-the-box, requiring
099 no training or prior knowledge of the text generator. [It not only surpasses the strong zero-shot](#)
100 [baseline, DetectGPT \(Mitchell et al., 2023a\), by a significant margin of 30.2%, but also outperforms](#)
101 [a leading supervised baseline, GhostBuster \(Verma et al., 2024\)—which requires expensive data](#)
102 [collection for supervised training—in five out of six domains.](#)

103 This work also faces the following limitations. First, CREATIVITY INDEX is designed to measure
104 one specific aspect of creativity—linguistic creativity (i.e., the novelty in composing words and
105 phrases). It might not comprehensively capture other dimensions of creativity, such as rhetorical
106 complexity or structural flexibility, and is thus complementary to existing creativity measurement
107

¹The name DJ SEARCH is inspired by the way a DJ creates a remix by blending pieces of existing music.

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161



b

$$\text{CREATIVITY INDEX} = \sum_{L=a}^b L\text{-uniqueness}$$

$$L\text{-uniqueness} = \frac{\text{number of words outside of } n\text{-grams } (n \geq L) \text{ that occur in the reference corpus}}{\text{number of words in the text}}$$

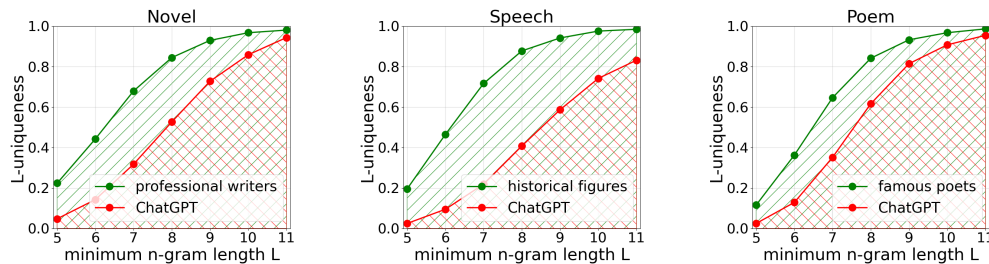


Figure 1: **a: Example outputs from DJ SEARCH.** We asked ChatGPT to generate an abstract based on the title of Prof. Michele Elam’s paper, “Poetry Will Not Optimize; or, What Is Literature to AI?” (Elam, 2023) The abstract generated by ChatGPT contains significantly more verbatim and near-verbatim matches with existing texts on the web compared to the original abstract written by Prof. Elam. **b: Definition of CREATIVITY INDEX.** CREATIVITY INDEX is mathematically equivalent to the area under the L -uniqueness curve across a range of minimum n -gram lengths L . The L -uniqueness of ChatGPT is noticeably lower than that of proficient human writers across various context granularities (i.e., n -gram lengths) in all domains, leading to a significantly higher CREATIVITY INDEX for human writers compared to ChatGPT.

162 methods. Second, while CREATIVITY INDEX is robust in reflecting statistical differences between
 163 professional human writing and seemingly remarkable outputs from LLMs, it may not fully capture
 164 nuanced differences between human writings. Beyond inherent creativity variations, factors such as
 165 writing style and the time of composition can also influence the metric. Third, CREATIVITY INDEX
 166 assumes that the input text is of sufficient quality, as our study focuses on outputs from recent LLMs
 167 that are already fluent and coherent. For less refined texts, our metric can be complemented with
 168 standard automatic quality measures, such as fluency classifiers or perplexity-based evaluations, to
 169 provide a more comprehensive assessment. Lastly, the computation of CREATIVITY INDEX is con-
 170 strained by the reference corpus used in DJ SEARCH. Without access to and inclusion of the private
 171 training data of closed-source models like ChatGPT in the reference corpus, the CREATIVITY IN-
 172 DEX for these models might be somewhat inflated. However, the research community is making
 173 progress in open-sourcing LLM research. Beyond RedPajama, newer large-scale pretraining cor-
 174 pora, such as DOLMA (Soldaini et al., 2024) and Common Corpus (PleIAs), have become available
 175 since our experiments.

176 We believe that our study will enhance the understanding of LLMs and guide informed usage of con-
 177 tent created by LLMs, by providing an interoperable and scalable measurement to assess creativity
 178 in machine texts. Additionally, we hope that the out-of-the-box machine text detection enabled by
 179 the CREATIVITY INDEX can empower individuals to discern between human texts and machine
 180 texts, fostering a more informed and critical engagement with information in the digital age.

181 2 METHOD

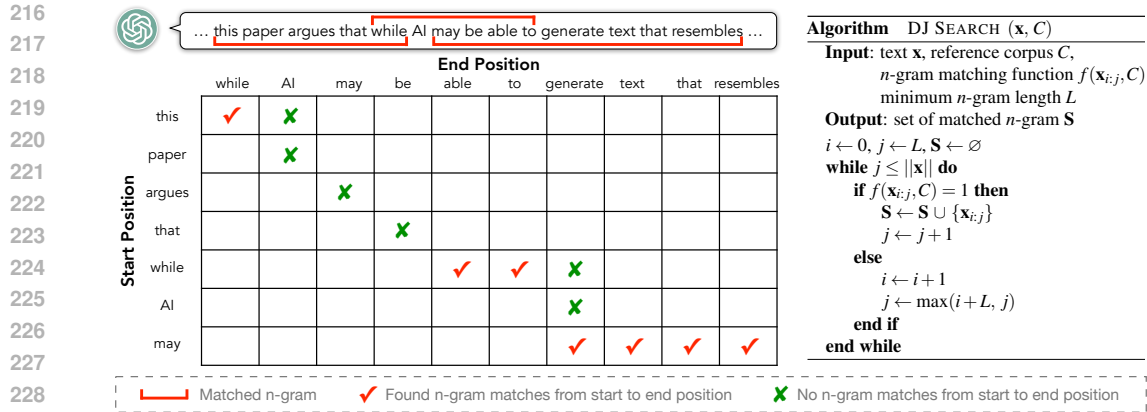
182 **CREATIVITY INDEX** The key intuition underlying CREATIVITY INDEX is to quantify the degree
 183 of linguistic creativity of a given text by estimating how much of that text can be reconstructed by
 184 mixing and matching a vast amount of existing text snippets on the web, as shown in Figure 1a.
 185 Specifically, CREATIVITY INDEX assesses the extent to which the content of the text can be traced
 186 back to similar or identical contexts found in other existing texts. This metric is grounded in the
 187 notion of originality from creative thinking in psychology literature, which is defined as the statistic
 188 rarity of a response or an idea (Torrance, 1966; Crossley et al., 2016).

189 Concretely, let \mathbf{x} be a text whose creativity we aim to quantify, such as a speech transcript or a poem,
 190 either human written or machine generated. Let an n -gram of \mathbf{x} be any contiguous sequence of n
 191 words of \mathbf{x} , and let $\mathbf{x}_{i:i+n}$ be the n -gram of \mathbf{x} starting in the i -th word. Let C be a massive reference
 192 corpus of publicly available texts on the web, and let f be a binary function that determines whether
 193 an n -gram $\mathbf{x}_{i:i+n}$ occurs anywhere in the corpus C . We define the L -uniqueness of a text \mathbf{x} as
 194 the proportion of words $w \in \mathbf{x}$ such that none of the n -grams in \mathbf{x} that include w occur in the
 195 corpus C for $n \geq L$ —denoted $\text{uniq}(\mathbf{x}, L)$. Intuitively, L -uniqueness measures the proportion of
 196 \mathbf{x} ’s words that are used in novel contexts (here, n -grams), unseen across a vast text collection C .
 197 Thus, a higher L -uniqueness implies a higher level of originality of \mathbf{x} . Formally, $\text{uniq}(\mathbf{x}, L) =$
 198 $\sum_{k=1}^{|\mathbf{x}|} \mathbb{1}\{f(\mathbf{x}_{i:i+n}, C) = 0 \ \forall i \in (k-n, k], n \geq L\} / \|\mathbf{x}\|$, where trivially $\text{uniq}(\mathbf{x}, L) \in [0, 1]$.

199 Note that when fixing \mathbf{x} , the function $\text{uniq}(\mathbf{x}, L)$ is monotonically increasing as L grows. Its im-
 200 proper integral— $\sum_{n \geq L} \text{uniq}(\mathbf{x}, n)$ —is an indicator of the overall uniqueness of \mathbf{x} across various
 201 context granularities (i.e., n -gram lengths), and because of $\text{uniq}(\mathbf{x}, L)$ ’s monotonicity it indirectly
 202 measures uniqueness growth speed. We thus define CREATIVITY INDEX as $\sum_{n \geq L} \text{uniq}(\mathbf{x}, n)$, with
 203 higher CREATIVITY INDEX indicating greater linguistic originality with respect to the corpus C , as
 204 shown in Figure 1b.

205 When a text \mathbf{x} is part of the reference corpus C , its CREATIVITY INDEX would trivially become zero.
 206 This issue often arises with works from famous authors, as their writings are widely available online.
 207 To address this, for human texts written before the cutoff date of the reference corpus, we exclude
 208 any document $\mathbf{d} \in C$ that contains copies, quotations, or citations of \mathbf{x} and compute CREATIVITY
 209 INDEX using this filtered corpus, detailed in Appendix A.3.

210 **DJ SEARCH** To enable the use of our CREATIVITY INDEX it is vital to compute it efficiently. For
 211 the efficient computation, we introduce DJ SEARCH, a dynamic programming algorithm designed
 212 to rapidly identify the set of all \mathbf{x} ’s n -grams ($n \geq L$) that occur in the corpus C .



230 Figure 2: **An illustration of DJ SEARCH algorithm.** A brute force approach would independently
 231 check if every n -gram of \mathbf{x} occurs in C , performing a quadratic number of f evaluations with respect
 232 to \mathbf{x} 's length (i.e., checking every cell in the grid). DJ SEARCH is a two-pointer method that takes
 233 only a linear number of f evaluations. By progressively analyzing n -grams starting and/or ending
 234 at a later endpoint than before, DJ SEARCH limits the total number of f evaluations to $2\|\mathbf{x}\|$. In this
 235 example, the minimum n -gram length L is set to 5.

241 A brute force approach would independently check if every n -gram of \mathbf{x} occurs in C , performing a
 242 quadratic number of f evaluations with respect to \mathbf{x} 's length, and thus making it too computationally
 243 expensive. Instead, we design a two-pointer method (Laaksonen, 2020) that takes only a linear
 244 number of f evaluations, as illustrated in Figure 2. The key idea is to reduce finding all n -grams
 245 occurring in C to identifying the longest n -gram occurring in C starting at each index i : once
 246 those have been found, it is trivial to deduce all the n -gram occurring in C by computing their
 247 subsequences. Concretely, we progressively analyze the whole document \mathbf{x} by iteratively searching
 248 for the longest n -gram that starts at each index i and occurs in C , using f as the assessment. Once
 249 we have found such longest n -gram starting at i , we crucially reuse computations for $i + 1$ by noting
 250 that $f(\mathbf{x}_{i:i+n}, C) = 1$ implies $f(\mathbf{x}_{i+1:i+n}, C) = 1$. Thus, we always analyze n -grams starting
 251 and/or ending at a later endpoint than before, which upper bounds the number of analyzed n -grams
 252 (i.e., the number of f calls) to at most $2\|\mathbf{x}\|$. The implementation is detailed in Appendix A.1.

253 In addition to minimizing the number f evaluations, DJ SEARCH optimizes the time complexity of
 254 each evaluation. f determines whether a n -gram $\mathbf{x}_{i:i+n}$ occurs in the corpus C either exactly or in
 255 a semantically similar way—e.g., a paraphrase of $\mathbf{x}_{i:i+n}$ exists in C . Semantic similarity is often
 256 computed using text embeddings, which are fixed-length vector representations of text meanings.
 257 This reduces measuring text similarity to computing vector distance. Text embeddings, typically
 258 generated by complex models (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), Span-
 259 BERT (Joshi et al., 2020)) lack linearity, requiring independent computation for each n -gram in \mathbf{x}
 260 and C . To alleviate this issue we use Word Mover's Distance (WMD) (Kusner et al., 2015), an
 261 optimal transport-inspired metric that measures distance between two n -grams by combining word
 262 embedding distances between each n -gram's words. WMD enables optimizing f 's computation, as
 263 pairwise distances between word embeddings can be pre-computed for every pair of words, and then
 264 be reused in every function call of f to identify n -grams in C that are semantically similar to the
 265 ones in \mathbf{x} . The implementation is detailed in Appendix A.2.

266 To further boost efficiency, and given that occurrences of $\mathbf{x}_{i:i+n}$ are more likely in texts similar to
 267 \mathbf{x} , we estimate f by computing WMD only for the texts in C most similar to \mathbf{x} , as identified
 268 by BM25 (Robertson & Walker, 1994). Moreover, exact occurrences of $\mathbf{x}_{i:i+n}$ in C represent
 269 a less costly special case in computing f . We further optimize f 's computation by using Infini-
 gram (Liu et al., 2024), which finds exact matches of $\mathbf{x}_{i:i+n}$ in C in milliseconds; WMD is com-
 puted only if no matches are found by Infini-gram.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

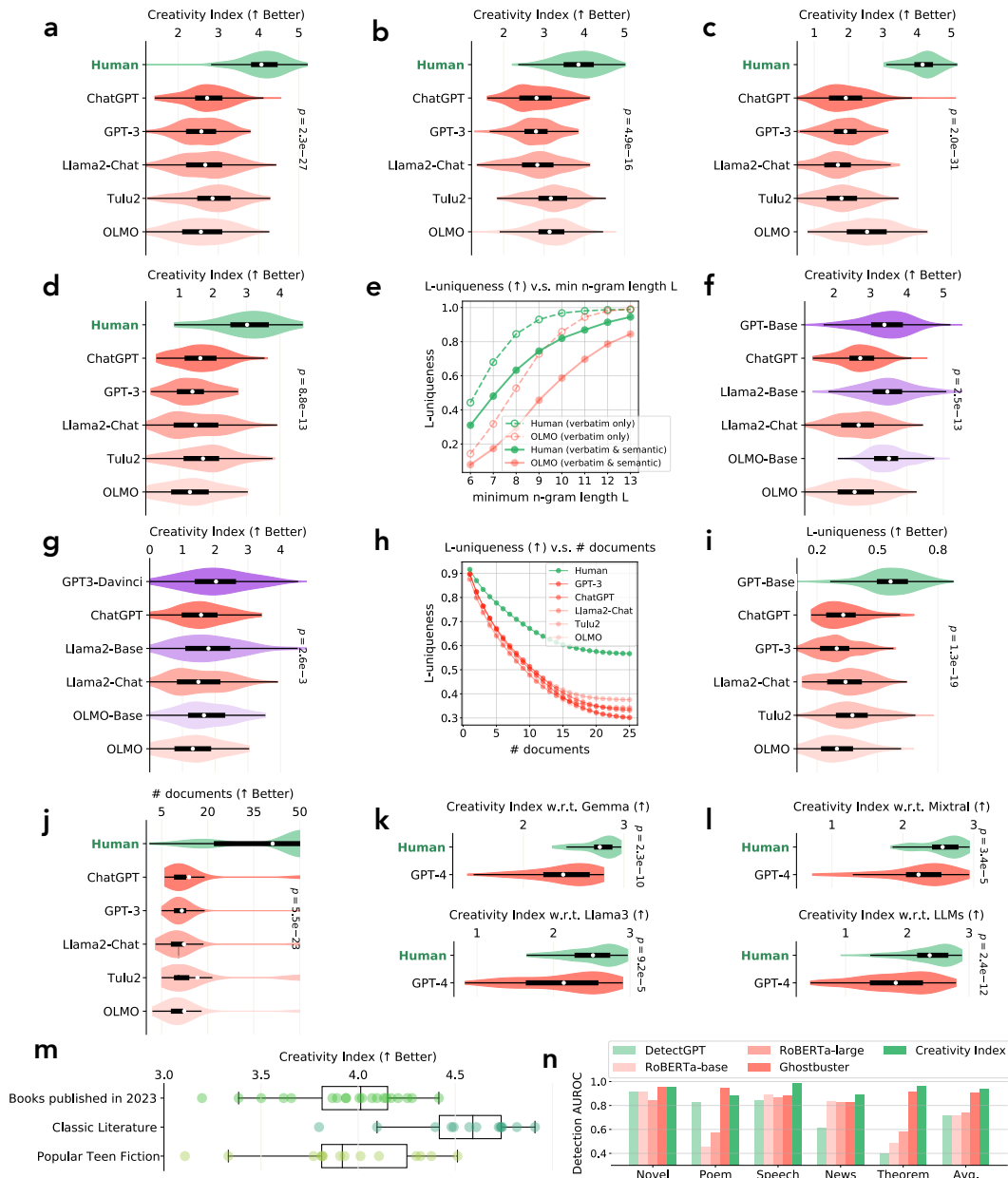


Figure 3: **a-c**: CREATIVITY INDEX in novel writing (**a**), poetry composition (**b**) and speech writing (**c**) based solely on verbatim matches. **d**: CREATIVITY INDEX in novel writing considering both verbatim and semantic matches. **e**: L -uniqueness in novel writing with respect to the minimum n -gram length L for humans and OLMo. **f-g**: CREATIVITY INDEX of LLMs before and after RLHF in novel writing, based solely on verbatim matches (**f**) and based on both verbatim and semantic matches (**g**). **h**: L -uniqueness in novel writing with respect to number of documents in the reference corpus. **i**: L -uniqueness when search over the top 50 documents in novel writing. **j**: The number of reference documents required to keep L -uniqueness below 50% in novel writing. **k-l**: CREATIVITY INDEX of GPT-4 compared to humans in novel writing based on verbatim matches, using a machine-generated reference corpus sourced from the instruction-aligned version of Gemma-7B, Llama3-8B, and Mixtral-7B, as well as a combination of all three. **m**: CREATIVITY INDEX of different groups of human writers. **n**: Detection AUROC across various domains: our approach sets a new state-of-the-art for zero-shot detection, even surpassing supervised baselines.

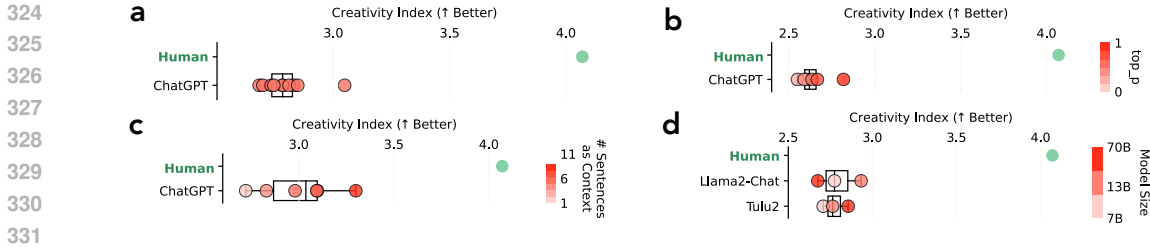


Figure 4: **a-c**: CREATIVITY INDEX of ChatGPT in novel writing based on verbatim matches, with different prompt formats (a), p values in top- p decoding (b) and prompt length (c). **d**: CREATIVITY INDEX of LLaMA 2 Chat and Tulu 2 with different model sizes.

3 EVALUATION

How does the creativity of language models compare to humans? We compute the CREATIVITY INDEX for machine texts and human texts across three creative writing tasks: novel writing, poetry composition, and speech drafting. For human texts, we use book snippets in the BookMIA (Shi et al., 2024) dataset, popular modern poems collected by PoemHunter.com, and famous speeches from the American Rhetoric speech bank. For machine texts, we prompt LLMs to generate several paragraphs of novels, poems, or speeches, starting with an initial sentence from existing human writings in each category (see Appendix B.1 for details). We experiment with state-of-the-art LLMs, including GPT-3 (Brown et al., 2020), ChatGPT (Ouyang et al., 2022), LLaMA 2 Chat (Touvron et al., 2023), Tulu 2 (Iverson et al., 2023), and OLMo Instruct (Groeneveld et al., 2024). For open-source and open-weight models, we use the largest model size available from each model family. We use RedPajama (Computer, 2023), a large-scale English corpus with 900 million web documents, as the reference corpus. The models we analyze are primarily pre-trained on the web data available before the cutoff date of the reference corpus RedPajama. We will discuss later how to handle newer models, such as GPT-4 (OpenAI et al., 2023), given that it was largely trained on more recent web data and third-party private data, both of which fall outside the reference corpus. We restrict the matching criteria to verbatim matches only in the first experiment. We will ablate the effect of different matching criteria, prompt formats, decoding strategies, context length, and model sizes in later experiments.

Our primary finding is that *humans consistently exhibit a much higher level of creativity compared to any LLM across all tasks* (Fig. 3a-c). Averaged across all models, the CREATIVITY INDEX of humans is 52.2% higher² than LLMs in novel writing ($p = 6.9 \times 10^{-27}$, by Mann-Whitney U test unless otherwise specified; $N = 600$), 31.1% higher in poetry composition ($p = 1.5 \times 10^{-15}$; $N = 600$) and 115.3% higher in speech drafting ($p = 6.1 \times 10^{-31}$, $N = 600$). This suggests that human writings are composed of far more unique combinations of words and phrases compared to model generations. On the other hand, the differences in model creativity are much smaller and show very low statistical significance ($p = 0.09$; $N = 1500$).

Furthermore, we experiment with different prompt formats on top of ChatGPT, intentionally encouraging creativity in the model’s generations by incorporating instructions such as ‘push for creative ideas, unique emotions, and original twists,’ ‘be bold and creative,’ or ‘you are a creative writer.’ (Fig. 4a) For a full list of the prompts we used, please see Appendix B.1. We found that the difference in the CREATIVITY INDEX of ChatGPT across different prompts is minimal, with no statistical significance ($p = 0.23$; $N = 600$). We also experimented with different decoding strategies by varying the p value in top- p decoding (Fig. 4b). Although a higher p value resulted in a marginally higher CREATIVITY INDEX, the difference was minimal and not statistically significant ($p = 0.23$; $N = 600$). Moreover, we ablate the effect of prompt length by varying the number of sentences from human writings included in the prompt (Fig. 4c). We found that longer prompts tended to result in a slightly higher CREATIVITY INDEX, likely due to the model copying more from the longer human text in the prompt. However, the statistical significance of these differences is very low ($p = 0.13$; $N = 600$). Lastly, we analyze the effect of different model sizes for LLaMA 2 Chat and Tulu 2, but do not observe a consistent trend ($p = 0.12$; $N = 600$) (Fig. 4d).

²The percentage difference computed using the formula: $\frac{\text{CREATIVITY INDEX (human)} - \text{CREATIVITY INDEX (model)}}{\text{CREATIVITY INDEX (model)}}$

How do different matching criteria affect creativity measurement? We experiment with restricting valid matches to verbatim only, and with allowing both verbatim and semantic matches. First, *the creativity gap between humans and LLMs becomes even larger when considering semantic matches in addition to verbatim matches* (Fig. 3d). Averaged across all models, the CREATIVITY INDEX of human, based on both verbatim and semantic matches, is 102.5% higher than LLMs in novel writing ($p = 2.6 \times 10^{-12}$; $N = 600$), whereas based on verbatim matches alone, the CREATIVITY INDEX of human is 52.2% higher than LLMs. Second, *semantic matches provide more signal for analyzing the uniqueness of longer n -grams* (Fig. 3e). For example, while the gap in L -uniqueness at $L = 11$ between human text and machine text from OLMo Instruct is 3.7% based on verbatim matches alone, this gap widens to 16.3% when considering both verbatim and semantic matches ($p = 3.1 \times 10^{-7}$; $N = 600$). This indicates that although some of the longer n -grams in machine text may appear unique at the verbatim level, they are similar to certain text snippets in the reference corpus at the content level.

What impact does RLHF have on model creativity? RLHF aims to align model’s outputs with human preferences, enhancing LLMs’ ability to follow instructions and improving their safety and adaptability. To understand the impact of RLHF on model creativity, we compare the CREATIVITY INDEX of the LLMs before and after RLHF alignment. Specifically, we experiment with GPT Base (Brown et al., 2020), LLaMA 2 Base (Touvron et al., 2023), and OLMo Base (Groeneveld et al., 2024) and compare their creativity with their counterparts post-RLHF alignment. Our main finding is that *the CREATIVITY INDEX of models after RLHF alignment is much lower than those before RLHF* (Fig. 3f-g). Based on verbatim match alone, the CREATIVITY INDEX of LLMs reduces by an average of 30.1% after RLHF ($p = 1.3 \times 10^{-12}$; $N = 600$). Based on both verbatim and semantic matches, the CREATIVITY INDEX of LLMs decreases by an average of 8.9% after RLHF ($p = 0.01$; $N = 600$). We notice that *the reduction of CREATIVITY INDEX after RLHF is noticeably larger when considering verbatim matches alone*. We speculate that models might have learned certain linguistic styles preferred by humans during RLHF, leading to a decreased surface form diversity in its outputs.

How do overlapped n -grams distribute in the reference corpus? In addition to measuring the amount of matched n -grams in a given text, we also investigate the distribution of these n -grams in the reference corpus. We aim to understand whether these matched n -grams are spread across many documents or concentrated in a few. Specifically, we identify the top N documents that contain the highest amount of matched n -grams and result in the minimum L -uniqueness for a given text. This problem can be reduced to the maximum coverage problem (Nemhauser et al., 1978) and approximated using a greedy algorithm. Here, we consider both verbatim and semantic matches. Our main finding is that *the matched n -grams in machine texts are concentrated in fewer documents compared to human texts* (Fig. 3h-j). When searching over the top 50 documents, the averaged L -uniqueness ($L = 5$) for machine texts is 32.8%, which is 73.4% lower than human texts (mean: 56.6%; $p = 3.9 \times 10^{-19}$; $N = 600$). Conversely, keeping L -uniqueness below 50% requires searching through an average of 41.2 documents for human texts, which is 213.7% more than for machine texts (mean: 13.4; $p = 1.6 \times 10^{-22}$; $N = 600$). This implies that it’s more likely to find some existing documents resemble models’ generations than human writings.

How to measure creativity in LLMs trained on data outside of the reference corpus? The CREATIVITY INDEX of GPT-4 would be significantly inflated if computed using the RedPajama corpus, as RedPajama’s cutoff date is two years earlier than GPT-4’s knowledge cutoff, and GPT-4 is additionally trained on third-party private data that we don’t have access to. We hypothesize that LLMs pre-trained on similar web data are likely to memorize and replicate similar patterns. As a result, when comparing the generations of these models, we expect them to be more similar to each other than to human texts, which often contain long-tail patterns. Therefore, to compare the creativity level of GPT-4 with humans, we use a model-generated reference corpus from newer open-weight models with knowledge cutoff dates similar to GPT-4, including the instruction-aligned versions of Gemma-7B (Team et al., 2024), Llama3-8B (AI@Meta, 2024), and Mixtral-7B (Jiang et al., 2023). Specifically, we randomly sample 150k sentences from the RedPajama corpus and prompt these models to generate document-level continuations. *Based on the model-generated reference corpus, the average CREATIVITY INDEX of humans is 30.3% higher than GPT-4 in novel writing* ($p = 2.3 \times 10^{-12}$; $N = 600$) (Fig. 3k-l). This suggests that while newer LLMs like GPT-4

432 may appear more creative when compared to public data, they still learn common patterns from their
 433 private training data and tend to emit similar patterns as other LLMs trained on comparable data.
 434

435 **How does the creativity vary among different groups of human?** Human populations are di-
 436 verse and complex, we aim to explore whether writings from different human authors exhibit varying
 437 levels of linguistic creativity. Specifically, we compare the linguistic creativity among three cate-
 438 gories of writings: books published in 2023 from the BookMIA (Shi et al., 2024) dataset, classic
 439 literature by famous authors, and popular young adult fictions, both sampled from Goodreads’ book
 440 lists. We first observe that the variation in CREATIVITY INDEX among human authors is relatively
 441 smaller compared to the difference between LLMs and humans. Additionally, we found that *clas-*
 442 *sic literature tends to exhibit higher linguistic creativity.* On average, the CREATIVITY INDEX of
 443 classic literature is 21.6% higher than young adult fictions ($p = 2.7 \times 10^{-90}$; $N = 3000$), and
 444 13.8% higher than books published in 2023 ($p = 4.3 \times 10^{-120}$; $N = 3000$). These findings should
 445 be interpreted with a grain of salt. We experimented with only a small set of writings from each
 446 category, so the results may not generalize broadly. Additionally, beyond inherent differences in
 447 creativity, CREATIVITY INDEX can be influenced by factors such as writing style and the time of
 448 composition. For instance, some classic literature are written in older English, which may result in
 449 a higher CREATIVITY INDEX because such writings are harder to reconstruct from web texts that
 450 primarily use modern English. In addition to the differences across categories, we also observed
 451 noticeable variance in creativity within each category. For example, the CREATIVITY INDEX of
 452 ‘The Hunger Games’ is 35.4% higher than ‘Twilight’ ($p = 1.5 \times 10^{-19}$; $N = 200$), even though
 453 both books belong to the category of popular young adult fiction.

454 **Can we leverage differences in creativity for detecting machine-generated text?** Based on the
 455 creativity difference between humans and LLMs, we explore a novel use case of CREATIVITY IN-
 456 DEX for zero-shot black-box machine text detection. Texts with higher creativity are more likely
 457 to be written by human. Our approach is ready to deploy out-of-the-box, requiring no training
 458 or prior knowledge of the text generator. In addition to creative writing tasks, we also test our
 459 method on detecting machine-generated fake news and theorem proofs. Detecting fake news is
 460 crucial for protecting the public from misinformation, while identifying model-generated solutions
 461 is important for regulating students’ use of LLMs in their coursework. To obtain additional test
 462 data, we prompt LLMs to generate news articles based on the fake news headlines from the Mis-
 463 info Reaction Frames (Gabriel et al., 2022) and compare them with the real news articles from the
 464 XSum (Narayan et al., 2018) dataset. Meanwhile, we prompt LLMs to generate proofs for theorems
 465 from the NaturalProofs (Welleck et al., 2022) benchmark, and compare them with the ground-truth
 466 human-written proofs. The baselines we compare against includes a widely adopted strong zero-
 467 shot detector, DetectGPT (Mitchell et al., 2023a), which uses the curvature of log probability as
 468 the detection criterion, as well as several supervised methods. These include OpenAI’s RoBERTa-
 469 based detector, fine-tuned on millions of generations from various GPT-2 sized models, as well as
 470 a more recent strong supervised detector, Ghostbuster (Verma et al., 2024), fine-tuned on thousands
 471 of generations from ChatGPT. We measure performance using the area under the receiver operating
 472 characteristic curve (AUROC), which represents the probability that a classifier correctly ranks a
 473 randomly-selected human-written example higher than a randomly selected machine-generated ex-
 474 ample. We found CREATIVITY INDEX is surprisingly effective for zero-shot machine text detection:
 475 it consistently surpasses DetectGPT and OpenAI’s detector across all domains, with significant im-
 476 provements in AUROC—30.2% and 26.9%, respectively. It also outperforms the strong supervised
 477 baseline, Ghostbuster—which requires expensive training and data collection—in five out of six
 478 domains, achieving an average AUROC improvement of 3.5% (Fig. 3n).

478 4 DISCUSSION

479 This work investigates the level of linguistic creativity in texts generated by LLMs and written by
 480 humans. Our findings suggest that the content and writing style of machine-generated texts may
 481 be less original and unique, as they contain significantly more semantic and verbatim matches with
 482 existing web texts compared to high-quality human writings. We hypothesize that this limited cre-
 483 ativity in models may result from the current data-driven paradigm used to train LLMs. In this
 484 paradigm, models are trained to mimic human-written texts during the pre-training stage, and to
 485 produce outputs aligned with human preferences during the RLHF stage. As a result, models learn
 to generate fluent and coherent texts by absorbing and replicating common patterns observed in their

486 training data. This reliance on existing text patterns can restrict their originality, as their outputs are
487 inherently shaped by previously seen examples. In contrast, accomplished authors such as Hem-
488 ington go beyond simply mimicking the great writings of others; they craft their own narratives to
489 express their unique opinions, perspectives, and insights, drawing from their personal experiences,
490 emotions, and backgrounds, which translates to the more creative compositions of words and phrases
491 that our method detects. Just as a DJ remixes existing tracks while a composer creates original mu-
492 sic, we speculate that LLMs behave more like DJs, blending existing texts to produce impressive
493 new outputs, while skilled human authors, similar to music composers, craft original works.

494 This work also faces the following limitations. First, CREATIVITY INDEX is designed to measure
495 one specific aspect of creativity—linguistic creativity (i.e., the novelty in composing words and
496 phrases). It might not comprehensively capture other dimensions of creativity such as rhetorical
497 complexity or structural flexibility. Therefore, while CREATIVITY INDEX is an effective tool for
498 understanding seemingly remarkable machine texts, it is insufficient on its own to draw definitive
499 conclusions about overall creativity differences across various writings, particularly when distin-
500 guishing nuanced variations among human authors. Second, the computation of the CREATIVITY
501 INDEX is constrained by the reference corpus used for DJ SEARCH. While open-source LLMs such
502 as OLMo rely on publicly available texts from the internet for their training data, major companies
503 like OpenAI additionally curate private data to train their closed-source LLMs such as ChatGPT.
504 Without incorporating these private data into the reference corpus of DJ SEARCH, the CREATIVITY
505 INDEX of closed-source LLMs may be somewhat inflated. In addition, since the reference corpus
506 primarily consists of more recent Internet texts, it may introduce an implicit bias favoring writings
507 from older periods, as these texts are underrepresented in modern web data and are therefore harder
508 to reconstruct. Third, the overlap with existing texts identified by DJ SEARCH in models’ gener-
509 ations may not conclusively indicate memorization of a specific document. It’s possible that these
510 text fragments, or their variations, appear in multiple documents that the model has been trained
511 on, including those outside the reference corpus of DJ SEARCH. Forth, the human authors that this
512 work focuses on are those with relatively high-quality writings available in existing public datasets.
513 While some human writings can be mediocre, tedious and unoriginal, we aim to assess how the
514 creativity levels in impressive LLM outputs compare against the high-quality writings produced by
515 professional human authors. Lastly, we acknowledge that the discussion surrounding the use of
516 LLMs in social and industrial settings is highly complex, and our work here speaks only to a part
517 of it. Besides the creativity of machine-generated content, other considerations in this discussion
518 include socioeconomic factors and ethical implications, which fall beyond the scope of this paper.

518 5 RELATED WORK

519
520 **Measuring Creativity in Ideas:** Measuring creative thinking and problem solving takes root in
521 early work in psychology (Torrance, 1966), where researchers defined four pillars for creative think-
522 ing: fluency, flexibility, originality and elaboration. Crossley et al. (2016) later on developed this
523 notion and built on it to expand this to measuring creative writing in students, where they also
524 adopted n -gram novelty for a measure of originality. However, these prior work focus on creativity
525 in humans, and they also do not introduce any automated metrics or measurements.

526
527 **Measuring Creativity in Machine-generated Text Using Expert Annotators:** Closely related to
528 CREATIVITY INDEX is a recent line of work in the generative AI literature comparing the creativity
529 of human writers to that of large language models in different domains such as story telling and
530 journalism (Chakrabarty et al., 2023; 2024; Anonymous, 2024). Similar to us, the approach in this
531 direction often involves prompting an LLM to write an original story or news article, based on some
532 existing premise or press release, and then comparing the machine-generated text to the human-
533 written counterparts. These works, however, take a rather subjective approach, where they define
534 and measure creativity based on human expert annotations and whether people perceive the text to
535 be more creative, rather than an objective quantification of novelty that we provide.

536
537 **Measuring Novelty of N -grams:** Finally, closely related to our work in terms of techniques is
538 Nguyen (2024) and Merrill et al. (2024). The former attempts at finding n -gram rules that would
539 cover and predict generations from transformer models, showing that more than 70% of the times
transformers follow some pre-set patterns and rules. The latter is more similar to our work as they
also measure the novelty of generated n -grams and compare it to human-written text, however they

differ from us in two major ways: (1) they only find verbatim matches, whereas we also match to approximate, semantically similar blocks of text and (2) they compute the percentage of n -grams of a certain length in a text that can be found in the reference corpus, whereas we measure how much of the text can be reconstructed by mixing and matching a vast amount of existing text snippets of varying lengths from the web.

Machine Text Detection: Detecting machine-generated text has been explored for several years using a variety of methods (Jawahar et al., 2020; Uchendu et al., 2021). Gehrmann et al. (2019) and Dugan et al. (2023) demonstrate that even humans tend to struggle to differentiate between text written by humans and machines, highlighting the need for automated detection solutions. Some approaches involve training a classifier in a supervised manner to identify machine-generated text (Bakhtin et al., 2019; Uchendu et al., 2020), while others use a zero-shot detection method (Solaiman et al., 2019; Ippolito et al., 2020). Additionally, there is research on bot detection through question answering (Wang et al., 2023; Chew & Baird, 2003). Recently, Mitchell et al. (2023b) introduced DetectGPT, a zero-shot method based on the hypothesis that texts produced by a large language model (LLM) are located at local maxima, and thus exhibit negative curvature, in the model’s probability distribution. Follow-up work build on DetectGPT by making it faster (Bao et al., 2024) and proposing to use cross-detection when the target model is unknown (Miresghallah et al., 2024).

6 CONCLUSION

We introduce CREATIVITY INDEX, an interoperable and scalable metric designed to quantify the linguistic creativity of a given text by estimating how much of that text can be reconstructed by mixing and matching a vast amount of existing text snippets on the web. To efficiently compute the CREATIVITY INDEX, we developed DJ SEARCH, a novel dynamic programming algorithm that can search verbatim and near-verbatim matches of text snippets from a given document against the web. We find that the creativity index of professional human writers is, on average, 66.2% higher than that of LLMs. Notably, RLHF dramatically reduces the creativity index of LLMs by an average of 30.1%. Furthermore, we demonstrate that CREATIVITY INDEX can be used as a surprisingly effective criterion for zero-shot black-box machine text detection. Our method not only surpasses the strongest zero-shot baseline, DetectGPT, by a significant margin of 30.2%, but also outperforms the strongest supervised baseline, GhostBuster, in five out of six domains. We hope that this study enhances the understanding of LLMs through the lens of linguistic creativity, and fosters informed usage of content created by LLMs in real-world applications.

REFERENCES

- Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *42nd IEEE Symposium on Security and Privacy*, 2021.
- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Anonymous. Do LLMs plan like human writers? comparing journalist coverage of press releases with LLMs. In *Submitted to ACL Rolling Review - June 2024*, 2024. URL <https://openreview.net/forum?id=E3VS45jxPR>. under review.
- Mikhail J. Atallah, Victor Raskin, Michael Crogan, Christian F. Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding*, 2001. URL <https://api.semanticscholar.org/CorpusID:37687669>.
- 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, 2019. URL <https://arxiv.org/abs/1906.03351>.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria*,

- 594 *May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=Bpcgcr8E8Z>.
595
596
- 597 John B. Biggs and Kevin F. Collis. The psychological structure of creative writing. *Australian*
598 *Journal of Education*, 26:59 – 70, 1982. URL [https://api.semanticscholar.org/](https://api.semanticscholar.org/CorpusID:141381510)
599 [CorpusID:141381510](https://api.semanticscholar.org/CorpusID:141381510).
- 600 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari-
601 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agar-
602 wal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,
603 Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler,
604 Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mc-
605 Candlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-
606 shot learners. In *Advances in Neural Information Processing Systems 33: Annual Con-*
607 *ference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12,*
608 *2020, virtual*, 2020. URL [https://proceedings.neurips.cc/paper/2020/hash/](https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html)
609 [1457c0d6bfc4967418bfb8ac142f64a-Abstract.html](https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html).
- 610 Robert Carnevale. The hollywood writers’ strike may actually be aid-
611 ing ai’s takeover, May 2023. URL [https://www.thewrap.com/](https://www.thewrap.com/hollywood-writers-strike-ai-deepmind-dramatron-chatgpt/)
612 [hollywood-writers-strike-ai-deepmind-dramatron-chatgpt/](https://www.thewrap.com/hollywood-writers-strike-ai-deepmind-dramatron-chatgpt/).
- 613 Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahma, and Smaranda Muresan. Creativity
614 support in the age of large language models: An empirical study involving emerging writers.
615 *arXiv preprint arXiv:2309.12570*, 2023.
- 616 Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu.
617 Art or artifice? large language models and the false promise of creativity. In *Proceedings of the*
618 *CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May*
619 *11-16, 2024*, pp. 30:1–30:34. ACM, 2024. URL [https://doi.org/10.1145/3613904.](https://doi.org/10.1145/3613904.3642731)
620 [3642731](https://doi.org/10.1145/3613904.3642731).
- 621 Monica Chew and Henry S. Baird. Baffletext: a human interactive proof. In *IS&T/SPIE Electronic*
622 *Imaging*, 2003.
- 623 John Joon Young Chung, Shiqing He, and Eytan Adar. The intersection of users, roles, interac-
624 tions, and technologies in creativity support tools. In *Proceedings of the 2021 ACM Designing*
625 *Interactive Systems Conference*, pp. 1817–1833, 2021.
- 626 Together Computer. Redpajama: an open dataset for training large language models, 2023. URL
627 <https://github.com/togethercomputer/RedPajama-Data>.
- 628 Scott A Crossley, Kasia Muldner, and Danielle S McNamara. Idea generation in student writing:
629 Computational assessments and links to successful writing. *Written Communication*, 33(3):328–
630 354, 2016.
- 631 Mihaly Csikszentmihalyi. Flow and the psychology of discovery and invention. *HarperPerennial,*
632 *New York*, 39:1–16, 1997.
- 633 Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan,
634 Ge Zhang, Hanfeng Lin, Yizhi Li, Ying Ma, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenwu
635 Wang, Guangyu Xia, Wei Xue, and Yi-Ting Guo. Composerx: Multi-agent symbolic music
636 composition with llms. *ArXiv preprint*, abs/2404.18081, 2024a. URL [https://arxiv.org/](https://arxiv.org/abs/2404.18081)
637 [abs/2404.18081](https://arxiv.org/abs/2404.18081).
- 638 Zekun Deng, Haoxia Yang, and Jun Wang. Can ai write classical chinese poetry like humans? an
639 empirical study inspired by turing test. *ArXiv preprint*, abs/2401.04952, 2024b. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2401.04952)
640 [2401.04952](https://arxiv.org/abs/2401.04952).
- 641 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep
642 bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*
643 *the North American Chapter of the Association for Computational Linguistics: Human Language*
644 *Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational
645 Linguistics, 2019. URL <https://aclanthology.org/N19-1423>.

- 648 Shuangrui Ding, Zihan Liu, Xiao wen Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and
649 Jiaqi Wang. Songcomposer: A large language model for lyric and melody composition in song
650 generation. *ArXiv preprint*, abs/2402.17645, 2024. URL <https://arxiv.org/abs/2402.17645>.
651 17645.
- 652 Liam Dugan, Daphne Ippolito, Arun Kirubakaran, Sherry Shi, and Chris Callison-Burch. Real
653 or fake text?: Investigating human ability to detect boundaries between human-written and
654 machine-generated text. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI
655 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023,
656 Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Wash-
657 ington, DC, USA, February 7-14, 2023*, pp. 12763–12771. AAAI Press, 2023. URL <https://doi.org/10.1609/aaai.v37i11.26501>.
658
- 659 David Eagleman and Anthony Brandt. *The runaway species: How human creativity remakes the
660 world*. Catapult, 2017.
- 661 Michele Elam. Poetry will not optimize, or what is literature to ai? *American Literature*, 2023.
662 URL <https://api.semanticscholar.org/CorpusID:257606488>.
663
- 664 Tina Fang, Martin Jaggi, and Katerina Argyraki. Generating steganographic text with LSTMs. In
665 Allyson Ettinger, Spandana Gella, Matthieu Labeau, Cecilia Ovesdotter Alm, Marine Carpuat,
666 and Mark Dredze (eds.), *Proceedings of ACL 2017, Student Research Workshop*, pp. 100–106,
667 Vancouver, Canada, July 2017. Association for Computational Linguistics. URL <https://aclanthology.org/P17-3017>.
668
- 669 Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter
670 Dalsgaard. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019
671 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK,
672 May 04-09, 2019*, pp. 389. ACM, 2019. URL <https://doi.org/10.1145/3290605.3300619>.
673
- 674 Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and
675 Yejin Choi. Misinfo reaction frames: Reasoning about readers’ reactions to news headlines.
676 In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
677 (Volume 1: Long Papers)*, pp. 3108–3127. Association for Computational Linguistics, 2022. URL
678 <https://aclanthology.org/2022.acl-long.222>.
679
- 680 Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and vi-
681 sualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for
682 Computational Linguistics: System Demonstrations*, pp. 111–116. Association for Computational
683 Linguistics, 2019. URL <https://aclanthology.org/P19-3019>.
684
- 685 Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. A design space for writing support
686 tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelli-
687 gent and Interactive Writing Assistants (In2Writing 2022)*, pp. 11–24. Association for Computa-
688 tional Linguistics, 2022. URL <https://aclanthology.org/2022.in2writing-1.2>.
- 689 Katy Ilonka Gero, Tao Long, and Lydia B. Chilton. Social dynamics of AI support in creative
690 writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,
691 CHI 2023, Hamburg, Germany, April 23-28, 2023*, pp. 245:1–245:15. ACM, 2023. URL <https://doi.org/10.1145/3544548.3580782>.
692
- 693 Vlad Petre Glaveanu, Michael Hanchett Hanson, John Baer, Baptiste Barbot, Edward P Clapp, Gio-
694 vanni Emanuele Corazza, Beth Hennessey, James C Kaufman, Izabela Lebeda, Todd Lubart, et al.
695 Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative
696 Behavior*, 54(3):741–745, 2020.
697
- 698 Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya
699 Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Au-
700 thur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel,
701 Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Cryst-
al Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh

- 702 Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi,
703 Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini,
704 Noah Smith, and Hannaneh Hajishirzi. OLMO: Accelerating the science of language models. In
705 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol-*
706 *ume 1: Long Papers)*, pp. 15789–15809. Association for Computational Linguistics, 2024. URL
707 <https://aclanthology.org/2024.acl-long.841>.
- 708 William Hasselberger and Micah Lott. “where lies the grail? ai, common sense, and human practical
709 intelligence”. *Phenomenology and the Cognitive Sciences*, pp. 1–22, 2023. URL [https://](https://api.semanticscholar.org/CorpusID:264895815)
710 api.semanticscholar.org/CorpusID:264895815.
- 711 Kent F Hubert, Kim N. Awa, and Darya L. Zabelina. The current state of artificial intelligence
712 generative language models is more creative than humans on divergent thinking tasks. *Sci-*
713 *entific Reports*, 14, 2024. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:267616181)
714 [267616181](https://api.semanticscholar.org/CorpusID:267616181).
- 715 Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection
716 of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting*
717 *of the Association for Computational Linguistics*, pp. 1808–1822. Association for Computational
718 Linguistics, 2020. URL <https://aclanthology.org/2020.acl-main.164>.
- 719 Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. Creative writing with
720 an ai-powered writing assistant: Perspectives from professional writers. *ArXiv preprint*,
721 [abs/2211.05030](https://arxiv.org/abs/2211.05030), 2022. URL <https://arxiv.org/abs/2211.05030>.
- 722 Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep
723 Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels
724 in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv preprint*, [abs/2311.10702](https://arxiv.org/abs/2311.10702),
725 2023. URL <https://arxiv.org/abs/2311.10702>.
- 726 Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection
727 of machine generated text: A critical survey. In *Proceedings of the 28th International Confer-*
728 *ence on Computational Linguistics*, pp. 2296–2309. International Committee on Computational
729 Linguistics, 2020. URL <https://aclanthology.org/2020.coling-main.208>.
- 730 Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh
731 Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lu-
732 cile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,
733 Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*,
734 [abs/2310.06825](https://api.semanticscholar.org/CorpusID:263830494), 2023. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:263830494)
735 [263830494](https://api.semanticscholar.org/CorpusID:263830494).
- 736 Anushrut Jignasu, Kelly O. Marshall, Baskar Ganapathysubramanian, Aditya Balu, Chinmay
737 Hegde, and Adarsh Krishnamurthy. Towards foundational ai models for additive manufactur-
738 ing: Language models for g-code debugging, manipulation, and comprehension. *ArXiv preprint*,
739 [abs/2309.02465](https://arxiv.org/abs/2309.02465), 2023. URL <https://arxiv.org/abs/2309.02465>.
- 740 Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Span-
741 BERT: Improving pre-training by representing and predicting spans. *Transactions of the Associ-*
742 *ation for Computational Linguistics*, 8:64–77, 2020. URL [https://aclanthology.org/](https://aclanthology.org/2020.tacl-1.5)
743 [2020.tacl-1.5](https://aclanthology.org/2020.tacl-1.5).
- 744 John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A
745 watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho,
746 Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th Inter-*
747 *national Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning*
748 *Research*, pp. 17061–17084. PMLR, 23–29 Jul 2023. URL [https://proceedings.mlr.](https://proceedings.mlr.press/v202/kirchenbauer23a.html)
749 [press/v202/kirchenbauer23a.html](https://proceedings.mlr.press/v202/kirchenbauer23a.html).
- 750 Brooks Barnes Koblin and John. Hollywood writers strike: Agreement gives writers guild most
751 of what it wanted, Sep 2023. URL [https://www.nytimes.com/live/2023/09/24/](https://www.nytimes.com/live/2023/09/24/business/hollywood-writers-strike-news)
752 [business/hollywood-writers-strike-news](https://www.nytimes.com/live/2023/09/24/business/hollywood-writers-strike-news).

- 756 Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to
757 document distances. In *Proceedings of the 32nd International Conference on Machine Learning,*
758 *ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Pro-*
759 *ceedings*, pp. 957–966. JMLR.org, 2015. URL [http://proceedings.mlr.press/v37/](http://proceedings.mlr.press/v37/kusnerb15.html)
760 [kusnerb15.html](http://proceedings.mlr.press/v37/kusnerb15.html).
- 761 Antti Laaksonen. *Guide to competitive programming*. Springer, 2020.
- 762 Xia Liang, Xingjian Du, Jiaju Lin, Pei Zou, Yuan Wan, and Bilei Zhu. Bytecomposer: a human-like
763 melody composition method based on language model agent. *ArXiv preprint*, abs/2402.17785,
764 2024. URL <https://arxiv.org/abs/2402.17785>.
- 765 Jonghan Lim, Birgit Vogel-Heuser, and Ilya Kovalenko. Large language model-enabled multi-agent
766 manufacturing systems. *ArXiv preprint*, abs/2406.01893, 2024. URL [https://arxiv.org/](https://arxiv.org/abs/2406.01893)
767 [abs/2406.01893](https://arxiv.org/abs/2406.01893).
- 768 Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram:
769 Scaling unbounded n-gram language models to a trillion tokens. *ArXiv preprint*, abs/2401.17377,
770 2024. URL <https://arxiv.org/abs/2401.17377>.
- 771 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike
772 Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining
773 approach. *ArXiv preprint*, abs/1907.11692, 2019. URL [https://arxiv.org/abs/1907.](https://arxiv.org/abs/1907.11692)
774 [11692](https://arxiv.org/abs/1907.11692).
- 775 Liane Makatura, Michael Foshey, Bohan Wang, Felix Hähnlein, Pingchuan Ma, Bolei Deng, Megan
776 Tjandrasuwita, Andrew Spielberg, Crystal Elaine Owens, Peter Yichen Chen, Allan Zhao, Amy
777 Zhu, Wil J. Norton, Edward Gu, Joshua Jacob, Yifei Li, Adriana Schulz, and Wojciech Matusik.
778 Large Language Models for Design and Manufacturing. *An MIT Exploration of Generative AI*,
779 2024. <https://mit-genai.pubpub.org/pub/nmympmnhs>.
- 780 William Merrill, Noah A Smith, and Yanai Elazar. Evaluating n -gram novelty of language models
781 using rusty-dawg. *ArXiv preprint*, abs/2406.13069, 2024. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2406.13069)
782 [2406.13069](https://arxiv.org/abs/2406.13069).
- 783 Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick.
784 Smaller language models are better zero-shot machine-generated text detectors. In *Proceedings*
785 *of the 18th Conference of the European Chapter of the Association for Computational Linguistics*
786 *(Volume 2: Short Papers)*, pp. 278–293. Association for Computational Linguistics, 2024. URL
787 <https://aclanthology.org/2024.eacl-short.25>.
- 788 Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays
789 and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings*
790 *of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg,*
791 *Germany, April 23-28, 2023*, pp. 355:1–355:34. ACM, 2023a. URL [https://doi.org/10.](https://doi.org/10.1145/3544548.3581225)
792 [1145/3544548.3581225](https://doi.org/10.1145/3544548.3581225).
- 793 Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays
794 and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings*
795 *of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg,*
796 *Germany, April 23-28, 2023*, pp. 355:1–355:34. ACM, 2023b. URL [https://doi.org/10.](https://doi.org/10.1145/3544548.3581225)
797 [1145/3544548.3581225](https://doi.org/10.1145/3544548.3581225).
- 800 Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. De-
801 tectgpt: Zero-shot machine-generated text detection using probability curvature. In *International*
802 *Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol-
803 *ume 202 of Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023a. URL
804 <https://proceedings.mlr.press/v202/mitchell123a.html>.
- 805 Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. De-
806 tectgpt: Zero-shot machine-generated text detection using probability curvature. In *International*
807 *Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol-
808 *ume 202 of Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023b. URL
809 <https://proceedings.mlr.press/v202/mitchell123a.html>.

- 810 Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the sum-
811 mary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of*
812 *the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807.
813 Association for Computational Linguistics, 2018. URL [https://aclanthology.org/](https://aclanthology.org/D18-1206)
814 D18-1206.
- 815
- 816 George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations
817 for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.
818 URL <https://api.semanticscholar.org/CorpusID:206800425>.
- 819
- 820 Timothy Nguyen. Understanding transformers via n-gram statistics. *ArXiv preprint*,
821 [abs/2407.12034](https://arxiv.org/abs/2407.12034), 2024. URL <https://arxiv.org/abs/2407.12034>.
- 822
- 823 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
824 cia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red
825 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-
826 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher
827 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-
828 man, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann,
829 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis,
830 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey
831 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,
832 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila
833 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,
834 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-
835 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan
836 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-
837 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan
838 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,
839 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun
840 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-
841 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook
842 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel
843 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen
844 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel
845 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,
846 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv
847 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,
848 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,
849 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel
850 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-
851 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,
852 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel
853 Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe
854 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,
855 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,
856 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra
857 Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,
858 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-
859 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,
860 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,
861 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,
862 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-
863 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-
864 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan
865 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,
866 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-
867 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming
868 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao

- 864 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. *ArXiv preprint*,
865 abs/2303.08774, 2023. URL <https://arxiv.org/abs/2303.08774>.
866
- 867 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong
868 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kel-
869 ton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano,
870 Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human
871 feedback. *ArXiv*, abs/2203.02155, 2022. URL [https://api.semanticscholar.org/
872 CorpusID:246426909](https://api.semanticscholar.org/CorpusID:246426909).
- 873 Scott Paeth. 2 human creativity and a doctrine of creation.
874
- 875 Srishti Palani, David Ledo, George W. Fitzmaurice, and Fraser Anderson. "i don't want to feel like
876 i'm working in a 1960s factory": The practitioner perspective on creativity support tool adoption.
877 In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA,*
878 *29 April 2022 - 5 May 2022*, pp. 379:1–379:18. ACM, 2022. URL [https://doi.org/10.
879 1145/3491102.3501933](https://doi.org/10.1145/3491102.3501933).
- 880 PleIAs. Pleias/common_corpusedatsetsathuggingface.URL.
881
- 882 Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-
883 poisson model for probabilistic weighted retrieval. In *Annual International ACM SIGIR Con-*
884 *ference on Research and Development in Information Retrieval*, 1994. URL [https://api.
885 semanticscholar.org/CorpusID:2218552](https://api.semanticscholar.org/CorpusID:2218552).
- 886 Piotr Sawicki, Marek Grześ, Fabrício Góes, Daniel Brown, Max Peepkorn, Aisha Khatun, and Si-
887 mona Paraskevopoulou. On the power of special-purpose gpt models to create and evaluate new po-
888 etry in old styles. URL <https://api.semanticscholar.org/CorpusID:269589029>.
- 889 Piotr Sawicki, Marek Grzes, Fabrício Góes, Daniel Brown, Max Peepkorn, and Aisha Khatun.
890 Bits of grass: Does gpt already know how to write like whitman? *ArXiv preprint*, abs/2305.11064,
891 2023. URL <https://arxiv.org/abs/2305.11064>.
892
- 893 Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi
894 Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In
895 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria,*
896 *May 7-11, 2024*. OpenReview.net, 2024. URL [https://openreview.net/forum?id=
897 zWqr3MQUNs](https://openreview.net/forum?id=zWqr3MQUNs).
- 898 Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec
899 Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason
900 Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language
901 models, 2019. URL <https://arxiv.org/abs/1908.09203>.
- 902 Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur,
903 Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, A. Jha,
904 Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Daniel Morrison, Niklas
905 Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle
906 Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke S.
907 Zettlemoyer, Noah A. Smith, Hanna Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and
908 Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research.
909 *ArXiv*, abs/2402.00159, 2024. URL [https://api.semanticscholar.org/CorpusID:
910 267364861](https://api.semanticscholar.org/CorpusID:267364861).
- 911 Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya
912 Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard
913 Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev,
914 Alex Castro-Ros, Ambrose Slone, Amélie Hélieu, Andrea Tacchetti, Anna Bulanova, Antonia Pa-
915 terson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément
916 Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng
917 Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian
Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau,

- 918 Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Kather-
 919 ine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid,
 920 Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem,
 921 Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ra-
 922 mona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L
 923 Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham
 924 De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali
 925 Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals,
 926 Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Bar-
 927 ral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and
 928 Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL
 929 <https://arxiv.org/abs/2403.08295>.
- 930 E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*,
 931 1966.
- 932 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-
 933 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cris-
 934 tian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu,
 935 Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn,
 936 Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel
 937 Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee,
 938 Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra,
 939 Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi,
 940 Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh
 941 Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen
 942 Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey
 943 Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*
 944 *preprint*, abs/2307.09288, 2023. URL <https://arxiv.org/abs/2307.09288>.
- 945 Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text
 946 generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*
 947 *Processing (EMNLP)*, pp. 8384–8395. Association for Computational Linguistics, 2020. URL
 948 <https://aclanthology.org/2020.emnlp-main.673>.
- 949 Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A bench-
 950 mark environment for Turing test in the age of neural text generation. In *Findings of the Association*
 951 *for Computational Linguistics: EMNLP 2021*, pp. 2001–2016. Association for Computational Lin-
 952 guistics, 2021. URL <https://aclanthology.org/2021.findings-emnlp.172>.
- 953 Maryam Vaezi and Saeed Rezaei. Development of a rubric for evaluating creative writing: a multi-
 954 phase research. *New Writing*, 16:303 – 317, 2018. URL <https://api.semanticscholar.org/CorpusID:149864198>.
- 955
 956
- 957 Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghost-
 958 written by large language models. In *Proceedings of the 2024 Conference of the North American*
 959 *Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol-*
 960 *ume 1: Long Papers)*, pp. 1702–1717. Association for Computational Linguistics, 2024. URL
 961 <https://aclanthology.org/2024.naacl-long.95>.
- 962 Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. Bot or human? detecting chatgpt imposters
 963 with a single question, 2023. URL <https://arxiv.org/abs/2305.06424>.
- 964
 965
- 966 Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Natural-
 967 prover: Grounded mathematical proof generation with language models. In *Advances in*
 968 *Neural Information Processing Systems 35: Annual Conference on Neural Information Pro-*
 969 *cessing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*
 970 *2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/1fc548a8243ad06616eee731e0572927-Abstract-Conference.html.
- 971 Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with
 recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods*

972 *in Natural Language Processing*, pp. 4393–4479. Association for Computational Linguistics, 2022.
973 URL <https://aclanthology.org/2022.emnlp-main.296>.
974
975 Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: Story writing with large
976 language models. *Proceedings of the 27th International Conference on Intelligent User Interfaces*,
977 2022. URL <https://api.semanticscholar.org/CorpusID:247585187>.
978
979 Zachary Ziegler, Yuntian Deng, and Alexander Rush. Neural linguistic steganography. In Ken-
980 taro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference*
981 *on Empirical Methods in Natural Language Processing and the 9th International Joint Confer-*
982 *ence on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1210–1215, Hong Kong, China,
983 November 2019. Association for Computational Linguistics. 10.18653/v1/D19-1115. URL
984 <https://aclanthology.org/D19-1115>.
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

1026 A METHOD DETAILS

1027 A.1 IMPLEMENTATION DETAILS OF DJ SEARCH

1028 As discussed in the main text, the deployment of the CREATIVITY INDEX relies on efficiently deter-
 1029 mining whether each n -gram $\mathbf{x}_{i+i+n} \in \mathbf{x}$ can be found anywhere in the massive reference corpus C
 1030 of publicly available texts. The function $f(\mathbf{x}_{i+i+n}, C)$ is a binary indicator that determines whether
 1031 an n -gram \mathbf{x}_{i+i+n} occurs in C . In line with the definition of CREATIVITY INDEX, we only consider
 1032 the n -grams \mathbf{x}_{i+i+n} such that $n \geq L$ for some fixed constant L .

1033 While a naive approach to checking whether \mathbf{x}_{i+i+n} appears in C for every n -gram $\mathbf{x}_{i+i+n} \in \mathbf{x}$
 1034 would take $O(|\mathbf{x}|^2)$ calls³ to f (see Algorithm 1), using a two-pointer approach we can radically
 1035 reduce this to $O(|\mathbf{x}|)$ calls (see Algorithm 2). Note that a two-pointer approach does $O(|\mathbf{x}|)$ calls to
 1036 f since in each iteration we advance at least one of the two pointers i and j by 1, and $0 \leq i, j \leq |\mathbf{x}|$.

1040 Algorithm 1 Naive Computation

```

1041 NGramsFoundi,j ← False  ∀ i ∈ [0..|x|) and j ∈ [0..|x|)  ▷ matrix to store n-gram occurrence
1042 for i ∈ [0, 1, ..., |x| - L] do
1043   for j ∈ [i + L, ..., |x|] do
1044     NGramsFound(i, j) ← f(xi:j, C)
1045   end for
1046 end for
1047 return NGramsFound

```

1049 Algorithm 2 Efficient computation of DJ SEARCH(**x**, *C*)

```

1051 NGramsFoundi,j ← False  ∀ i ∈ [0..|x|) and j ∈ [0..|x|)  ▷ matrix to store n-gram occurrence
1052 i ← 0, j ← L
1053 while j < |x| do
1054   NGramsFound(i, j) = f(xi:j, C)
1055   if NGramsFound(i, j) then
1056     j ← j + 1  ▷ we will search for xi:j+1 next
1057   else
1058     i ← i + 1  ▷ since xi:j was not found, xi:j+k will not be found for all k > 0
1059     j ← max(i + L, j)  ▷ we only explore L-grams and beyond
1060   end if
1061 end while
1062 return NGramsFound

```

1063 A.2 IMPLEMENTATION DETAILS OF WORD MOVER’S DISTANCE

1064 Let w be an n -gram. Let $f(w, C)$ be the function that determines whether w appears in any text
 1065 $\mathbf{d} \in C$, either exactly or as a phrase that is highly similar in meaning to w (e.g., a paraphrase of w).
 1066 Trivially, $f(w, C) := \bigcup_{\mathbf{d} \in C} f(w, \mathbf{d})$, and here on we will only discuss how to compute $f(w, \mathbf{d})$.

1067 An established approach for finding semantically similar phrases to a given n -gram w is to compute
 1068 its embedding— $\text{embedding}(w)$ —and then independently compute its similarity to the embeddings
 1069 of all other n -grams to be analyzed. An *embedding* of a n -gram is a vector that represents the mean-
 1070 ing of such n -gram in an k -th dimensional space of fixed size, enabling the comparison of similarity
 1071 between concepts expressed in different surface forms. This comparison is typically done using
 1072 *cosine similarity*, the scaled dot product between the two embeddings being compared. Text embed-
 1073 dings are generated by models specifically trained to this effect (e.g., BERT (Devlin et al., 2019),
 1074 RoBERTa (Liu et al., 2019), SpanBERT (Joshi et al., 2020)) making their computation expensive at
 1075 a large scale. Notably, text embeddings usually do not possess linearity, i.e. the embedding of con-
 1076 catenating n -grams w and v cannot be deduced from knowing $\text{embedding}(w)$ and $\text{embedding}(v)$,
 1077 and instead needs to be computed from scratch.

1078 ³There are $(|\mathbf{x}| - L)(|\mathbf{x}| - L + 3)/2$ spans to analyze if L is the minimum n -gram length to be considered.

Since our goal is to find the n -grams of \mathbf{d} that are highly similar to w , using the traditional approach would entail comparing embedding(w) with the embeddings of all n -grams in C , which are approximately $\sum_{d \in C} |d|^2$ in number. Note that this also implies independently computing $\approx \sum_{d \in C} |d|^2$ embeddings, which increases the computation costs significantly. Instead we use Word Mover’s Distance (Kusner et al., 2015) (WMD), a method to estimate similarity between two n -grams by combining comparisons between pairs of *word* embeddings. This enables lifting the requirement to independently computing the embedding for each n -gram in C . Concretely, the Word Movers’ Distance between two n -grams w and v is defined as follows:

$$\begin{aligned} D_{w \rightarrow v} &:= \frac{1}{|w|} \sum_{i \in [0..|w|)} \min_{j \in [0..|v|)} 1 - \text{cosine_similarity}(\text{embedding}(v_j), \text{embedding}(w_i)) \\ &= 1 - \frac{1}{|w|} \sum_{i \in [0..|w|)} \max_{j \in [0..|v|)} \text{cosine_similarity}(\text{embedding}(v_j), \text{embedding}(w_i)) \end{aligned}$$

$$\text{WMD}(w, v) := \max(D_{w \rightarrow v}, D_{v \rightarrow w})$$

WMD also pre-filters the words considered in w and v to only include the *content words* in the analysis (i.e. discards *stop-words*, such as *the, a, an, it, on, ...*).

Note that $D_{w \rightarrow v}$ ’s definition is asymmetric ($D_{w \rightarrow v} \neq D_{v \rightarrow w}$). Thus, we consider the Word Movers’ Distance of two n -grams w and v as the maximum of $D_{w \rightarrow v}$ and $D_{v \rightarrow w}$: w and v are highly similar if their distance is below a threshold δ for both $D_{w \rightarrow v}$ and $D_{v \rightarrow w}$ (See Algorithm 3):

$$\text{WMD}(w, v) = \max(D_{w \rightarrow v}, D_{v \rightarrow w}) < \delta$$

Algorithm 3 Conceptual writeup of $f(w, \mathbf{d})$ using Word Mover Distance (WMD) to find the n -grams of a single text $\mathbf{d} \in C$ that are highly similar to the n -gram w and are of length $\geq L$.

```

procedure DIRECTIONALWMD( $w, v$ )
  return  $1 - \frac{1}{|v|} \sum_{j \in [0..|v|)} \max_{i \in [0..|w|)} \text{cosine\_similarity}(\text{embedding}(w_i), \text{embedding}(v_j))$ 
end procedure
for  $a \in [0, 1, \dots, |\mathbf{d}|)$  do
  for  $b \in [a + L, \dots, |\mathbf{d}|]$  do
     $\text{symmetricWMD} \leftarrow \max(\text{directionalWMD}(\mathbf{d}[a : b], w), \text{directionalWMD}(w, \mathbf{d}[a : b]))$ 
    if  $\text{symmetricWMD} < \delta$  then
      return True
    end if
  end for
end for
return False

```

Avid readers may notice that Algorithm 3 repeatedly computes the maximum over the same set, and sums of contiguous similarity scores; these can be pre-computed. Algorithm 4 shows these optimizations, resulting in an algorithm of time complexity $O(|d| \cdot |w| + |d|^2 |w|) = O(|d|^2 |w|)$, assuming already computed word embeddings. Note that because there is a fixed vocabulary, all word embeddings as well as cosine similarities of word embedding pairs can be pre-computed.

We described how to compute $f(w, \mathbf{d})$ for a single document $\mathbf{d} \in C$, as we have already established that $f(w, C) = \bigcup_{\mathbf{d} \in C} f(w, \mathbf{d})$. To accelerate computation, and given that similar n -grams to $\mathbf{x}_{i:i+n}$ are more likely to occur in texts similar to \mathbf{x} , we select C ’s top most likely documents to contain w using a BM25Robertson & Walker (1994) index, denoted C' . We then approximate $f(w, C) \approx \bigcup_{\mathbf{d} \in C'} f(w, \mathbf{d})$.

As a final optimization, we note that it is unnecessary to compute the costly $f(w, C)$ for finding semantically similar matches for w in the case where w appears exactly in C . To check if w appears exactly in C , we can leverage the existing, less expensive approach Infini-Gram (Liu et al., 2024) and search for the semantic similar matches only if Infini-Gram could not find any exact matches.

Algorithm 4 Efficient Computation of $f(w, \mathbf{d})$ (optimization of Algorithm 3)

```

1134 token_similarityi,j ← cosine_similarity(embedding(wi), embedding(dj))  ∀ i ∈ [0..|w|) and j ∈
1135 [0..|d|)
1136 for j ∈ [1, ..., |d|] do
1137   doc_prefix_similarityj ← doc_prefix_similarityj-1 + maxi ∈ [0..|w|) token_similarityi,j-1
1138 end for
1139
1140 for a ∈ [0, 1, ..., |d|] do
1141   for b ∈ [a + L, ..., |d|] do
1142     computed_WMD(d[a : b], w) ← 1 - (doc_prefix_similarityb -
1143 doc_prefix_similaritya) / (b - a)
1144     computed_WMD(w, d[a : b]) ← 1 -  $\frac{1}{|w|} \sum_{i \in [0..|w|)} \max_{j \in [a..b)} \text{token\_similarity}_{i,j}$ 
1145     symmetric_WMD ← max(computed_WMD(d[a : b], w), computed_WMD(w, d[a : b]))
1146     if symmetric_WMD < δ then
1147       return True
1148     end if
1149   end for
1150 end for
1151 return False

```

A.3 DEDUPLICATION OF THE REFERENCE CORPUS

When a text \mathbf{x} is part of the reference corpus C , its CREATIVITY INDEX would trivially become zero. This issue often arises when analyzing the works of famous authors, as their writings are frequently copied, quoted, or cited online. To address this, when analyzing human texts written before the cutoff date of the reference corpus, we exclude any document $\mathbf{d} \in C$ that contains copies, quotations, or citations of the text \mathbf{x} from the reference corpus C , and compute CREATIVITY INDEX of \mathbf{x} using this filtered reference corpus.

Specifically, we measure the degree of overlap between \mathbf{x} and \mathbf{d} by calculating the length of the longest common subsequence (LCS) between them, normalized by the length of \mathbf{x} . Formally, $S(\mathbf{x}, \mathbf{d}) = \frac{||\text{LCS}(\mathbf{x}, \mathbf{d})||}{||\mathbf{x}||}$. If \mathbf{x} and \mathbf{d} have a high degree of overlap (i.e., $S(\mathbf{x}, \mathbf{d}) \geq \alpha$), it's very likely that \mathbf{d} contains an exact copy of \mathbf{x} . If \mathbf{x} and \mathbf{d} show a moderate amount of overlap (i.e., $\beta \leq S(\mathbf{x}, \mathbf{d}) < \alpha$), we prompt a LLM to determine whether \mathbf{d} contains copies or quotations of \mathbf{x} using in-context examples provided below. Additionally, if \mathbf{d} includes the author name or title of \mathbf{x} , it is highly likely that \mathbf{d} contains a citation of \mathbf{x} . In practice, we set the values of α and β to 0.9 and 0.3, respectively, and use LLaMA 2 Chat as the LLM to check for copies and quotations.

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

Please check if paragraph A contains any copies or quotations from paragraph B.

Here are some examples:

Paragraph A: In the end though, I did the required reading, complained bitterly about being bored, wrote the requisite essay, and promptly forgot all about it. "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days ...

Paragraph B: He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao ...

Answer: Yes

Paragraph A: He was an old man who fished alone in a lobster boat off the Maine coast and he had gone 117 days without taking a crustacean. His luck was not bad, rather his judgment was good (don't fish the Atlantic in winter). Then he met us and for all I know his luck changed. El Campion is due for a change of luck ...

Paragraph B: He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao ...

Answer: No

Paragraph A: Santiago, the "old man who fished alone," in Hemingway's "The Old Man and the Sea" appears as one who has an undefeatable character, a loving, cheerful character, and very humble. The writer describes him in this way: "Everything about him was old except his eyes, and they were the same color as the sea ...

Paragraph B: He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao ...

Answer: Yes

Paragraph A: He was an old man who could see the form of his god, and a monk, moreover. Izzie had limited ability to communicate directly with her own deity. Much of her life she had proceeded by vague impressions and only glimpsed the great god's image briefly in the depths of meditation ...

Paragraph B: He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao ...

Answer: No

Here is the test example:

Paragraph A: [A]

Paragraph B: [B]

Answer:

B EVALUATION

B.1 MACHINE TEXT GENERATION

We experiment with state-of-the-art LLMs: GPT-3 (Brown et al., 2020) (`text-davinci-003`), ChatGPT (Ouyang et al., 2022) (`gpt-3.5-turbo`), LLaMA 2 Chat (Touvron et al., 2023), Tulu 2 (Iverson et al., 2023) and OLMo Instruct (Groeneveld et al., 2024) along with their base model before RLHF: GPT Base (Brown et al., 2020) (`davinci-002`), LLaMA 2 Base (Touvron et al., 2023) and OLMo Base (Groeneveld et al., 2024). These models are primarily pre-trained on the web data available before the cutoff date of the reference corpus RedPajama (Computer, 2023). We additionally discuss how to handle newer models, such as GPT-4 (OpenAI et al., 2023), which are largely trained on more recent web data and third-party private data, both of which fall outside the reference corpus RedPajama.

To obtain machine texts, we prompt LLMs to generate several paragraphs of novels, poems, or speeches, starting with an initial sentence taken from existing human writings in each category. To construct test data for machine text detection, we further prompt LLMs to generate news articles based on the fake news headlines from the Misinfo Reaction Frames (Gabriel et al., 2022) and to generate theorem proofs for questions from the NaturalProofs (Welleck et al., 2022) benchmark. The prompts used for each task are illustrated below. For all generations, we use nucleus sampling with $p = 0.9$ and set the maximum length of the generated texts to 288 tokens.

```
Please write a few paragraphs for a novel starting with the following
prompt: [PROMPT SENTENCE]
```

```
Please write a poem starting with the following line: [PROMPT LINE]
```

```
Please write a speech starting with the following sentence: [PROMPT
SENTENCE]
```

```
Please write a news article based on the given headline: [NEWS
HEADLINE]
```

```
Please provide a proof for the following theorem: [THEOREM QUESTION]
```

To obtain model-generated reference corpus to compare the CREATIVITY INDEX of GPT-4 with humans, we randomly sample 150k sentences from the RedPajama corpus and prompt open-weight LLMs with knowledge cutoff dates similar to GPT-4 to generate document-level continuations. The models we use are the instruction-aligned versions of Gemma-7B (Team et al., 2024) (`gemma-7b-it`), Llama3-8B (Meta-Llama-3-8B) (AI@Meta, 2024), and Mixtral-7B (`Mistral-7B-v0.1`) (Jiang et al., 2023). The prompt used to generate continuations is illustrated below. We use nucleus sampling with $p = 0.9$ and set the maximum length of the generated texts to 2048 tokens.

```
Please generate a continuation for the following sentence: [PROMPT
SENTENCE]
```

We additionally experiment with different prompt formats, intentionally encouraging creativity in models' generations by incorporating instructions such as 'push for creative ideas, unique emotions, and original twists,' 'be bold and creative,' or 'you are a creative writer.' Please see below for a full list of the prompts we tried.

```
Write a few paragraphs for a novel from the following prompt, pushing
for creative ideas, unique emotions, and original twists.
Prompt: [PROMPT SENTENCE]
```


1296 Use the following prompt to write a few paragraphs for a novel with
1297 creative, unique perspectives or twists. Let your originality shine.
1298 Prompt: [PROMPT SENTENCE]

1299

1300 Create a few paragraphs from the following prompt for a novel, focusing
1301 on novel ideas, emotions, or perspectives. Be as creative as possible.
1302 Prompt: [PROMPT SENTENCE]

1303

1304 Write a few paragraphs for a novel based on the following prompt,
1305 exploring unexpected twists, emotions, or unique perspectives. Be bold
1306 and creative.
1307 Prompt: [PROMPT SENTENCE]

1308

1309 Based on the following prompt, and write a few paragraphs for a novel
1310 that explore unexpected twists, deep emotions, or unique perspectives.
1311 Let your creativity flow, and don't be afraid to experiment with
1312 unconventional ideas or characters
1313 Prompt: [PROMPT SENTENCE]

1314

1315 As a creative agent, write a few paragraphs for a novel based on the
1316 following prompt, bringing your novel ideas and original emotions to
1317 life.
1318 Prompt: [PROMPT SENTENCE]

1319

1320 You are a creative writer, write a few paragraphs for a novel based
1321 on the following prompt. Explore unique perspectives and unexpected
1322 twists, and let your creativity guide you.
1323 Prompt: [PROMPT SENTENCE]

1324

1325 You are a creative agent, free to shape this story in any direction.
1326 Write a few paragraphs for a novel based on the following prompt, using
1327 your imagination to uncover surprises and depth.
1328 Prompt: [PROMPT SENTENCE]

1329

1330 As a creative writer, your task is to write a few paragraphs for a
1331 novel based on the following prompt. Dive into original ideas, explore
1332 emotions, and surprise yourself.
1333 Prompt: [PROMPT SENTENCE]

1334

1335 You are a creative writer who brings stories to life. Write a few
1336 paragraphs for a novel based on the following prompt, letting your
1337 imagination take bold, unexpected turns.
1338 Prompt: [PROMPT SENTENCE]

1337 B.2 DATASET DETAILS

1339

1340

1341 **Reference Corpus:** We use RedPajama (Computer, 2023), the largest web data collection avail-
1342 able at the time of this study, as our reference corpus. RedPajama contains 100 billion text docu-
1343 ments with 100+ trillion raw tokens from 84 CommonCrawl dumps.

1344

1345 **Novel:** For human-written novels, we use book snippets from the BookMIA (Shi et al., 2024)
1346 dataset. The BookMIA dataset contains approximately 10k book snippets, with an average length of
1347 around 650 words per snippet. We randomly sample 100 book snippets from the BookMIA dataset
1348 and select the first K sentences of each snippet such that their total length exceeds 256 words, to
1349 use as human text. Since novels we use were published after the cutoff date of RedPajama, there's
no need for deduplication before DJ SEARCH.

1350 **Speech:** For the transcripts of human speeches, we randomly sample 100 speeches from the fa-
 1351 mous speeches available in the American Rhetoric speech bank. For each speech, we randomly
 1352 sample continuous K sentences such that their total length exceeds 256 words, to use as human
 1353 text. Since these speeches were made before the cutoff date of RedPajama, deduplication is needed
 1354 before DJ SEARCH.

1355
 1356 **Poem:** For human-written poems, we randomly sample 100 poems from the popular modern po-
 1357 ems collected by PoemHunter.com. Since these poems were published before the cutoff date of
 1358 RedPajama, deduplication is needed before DJ SEARCH.

1359
 1360 **News Article:** We use news articles from the XSum (Narayan et al., 2018) dataset as the human
 1361 text for the machine text detection task. The Xsum dataset contains around 200k new articles, with
 1362 an average length of around 380 words per article. We randomly sample 500 articles to use as human
 1363 text. Since these news articles were released before the cutoff date of RedPajama, deduplication is
 1364 needed before DJ SEARCH. For machine-generated fake news, we randomly sample 500 fake news
 1365 headlines from the Misinfo Reaction Frames (Gabriel et al., 2022), and based on these headlines,
 1366 LLMs are asked to generate corresponding news articles.

1367
 1368 **Theorem Proof:** We use the ground-truth human-written proofs from the NaturalProofs (Welleck
 1369 et al., 2022) dataset as the human text for the machine text detection task. The NaturalProofs
 1370 dataset contains approximately 24k theorems and their corresponding proofs. We randomly sam-
 1371 ple 500 theorem-proof pairs and use the ground-truth proofs as human text. Since the NaturalProofs
 1372 dataset was curated after the cutoff date of RedPajama, there’s no need for deduplication before
 1373 DJ SEARCH. For machine-generated math proofs, we prompt LLMs to write proofs for the 500
 1374 theorems we sampled.

1375
 1376 B.3 PARAMETERS OF DJ SEARCH

1377
 1378 We set the minimum n -gram length L in DJ SEARCH to 5, and set the threshold for Word Mover’s
 1379 Distance to 0.95 for semantic matches. We observe that the L -uniqueness is close to zero for most
 1380 human and machine texts when $L \leq 5$ and close to one when $L \geq 12$. Therefore, in practice, we
 1381 sum up the L -uniqueness for $5 \leq L \leq 12$ when computing CREATIVITY INDEX.

1382 The only experiment with slightly different parameters is to compare the creativity of GPT-4 with
 1383 humans. We observed that the L -uniqueness is close to one when $L \geq 7$ based on the model-
 1384 generated reference corpus. Therefore, we sum up the L -uniqueness for $5 \leq L \leq 7$ when computing
 1385 CREATIVITY INDEX.

1386
 1387 C RELATED WORK

1388
 1389 **Measuring Creativity in Ideas:** Measuring creative thinking and problem solving takes root in
 1390 early work in psychology (Torrance, 1966), where researchers defined four pillars for creative think-
 1391 ing: fluency, flexibility, originality and elaboration. Crossley et al. (2016) later on developed this
 1392 notion and built on it to expand this to measuring creative writing in students, where they also
 1393 adopted n -gram novelty for a measure of originality. However, these prior work focus on creativity
 1394 in humans, and they also do not introduce any automated metrics or measurements.

1395
 1396 **Measuring Creativity in Machine-generated Text Using Expert Annotators:** Closely related to
 1397 CREATIVITY INDEX is a recent line of work in the generative AI literature comparing the creativity
 1398 of human writers to that of large language models in different domains such as story telling and
 1399 journalism (Chakrabarty et al., 2023; 2024; Anonymous, 2024). Similar to us, the approach in this
 1400 direction often involves prompting an LLM to write an original story or news article, based on some
 1401 existing premise or press release, and then comparing the machine-generated text to the human-
 1402 written counterparts. These works, however, take a rather subjective approach, where they define
 1403 and measure creativity based on human expert annotations and whether people perceive the text to
 be more creative, rather than an objective quantification of novelty that we provide.

1404 **Measuring Novelty of N -grams:** Finally, closely related to our work in terms of techniques is
1405 Nguyen (2024) and Merrill et al. (2024). The former attempts at finding n -gram rules that would
1406 cover and predict generations from transformer models, showing that more than 70% of the times
1407 transformers follow some pre-set patterns and rules. The latter is more similar to our work as they
1408 also measure the novelty of generated n -grams and compare it to human-written text, however they
1409 differ from us in tow major ways: (1) they only find verbatim matches, whereas we also match to
1410 approximate, semantically similar blocks of text and (2) they compute the percentage of n -grams of
1411 a certain length in a text that can be found in the reference corpus, whereas we measure how much
1412 of the text can be reconstructed by mixing and matching a vast amount of existing text snippets of
1413 varying lengths from the web.

1414 **Machine Text Detection:** Detecting machine-generated text has been explored for several years
1415 using a variety of methods (Jawahar et al., 2020; Uchendu et al., 2021). Gehrmann et al. (2019)
1416 and Dugan et al. (2023) demonstrate that even humans tend to struggle to differentiate between
1417 text written by humans and machines, highlighting the need for automated detection solutions.
1418 Some approaches involve training a classifier in a supervised manner to identify machine-generated
1419 text (Bakhtin et al., 2019; Uchendu et al., 2020), while others use a zero-shot detection method (So-
1420 laiman et al., 2019; Ippolito et al., 2020). Additionally, there is research on bot detection through
1421 question answering (Wang et al., 2023; Chew & Baird, 2003). Recently, Mitchell et al. (2023b)
1422 introduced DetectGPT, a zero-shot method based on the hypothesis that texts produced by a large
1423 language model (LLM) are located at local maxima, and thus exhibit negative curvature, in the
1424 model’s probability distribution. Follow-up work build on DetectGPT by making it faster (Bao
1425 et al., 2024) and proposing to use cross-detection when the target model is unknown (Miresghallah
1426 et al., 2024).

1427 Various strategies have been developed to detect machine-generated text in real-world settings. One
1428 notable approach is watermarking, which embeds algorithmically detectable patterns into the gen-
1429 erated text while maintaining the quality and diversity of the language model’s outputs. Initial
1430 watermarking techniques for natural language were proposed by Atallah et al. (2001) and have been
1431 adapted for neural language model outputs (Fang et al., 2017; Ziegler et al., 2019). Recent advance-
1432 ments include Abdelnabi & Fritz (2021) work on an adversarial watermarking transformer (AWT)
1433 for transformer-based language models. Unlike methods dependent on specific model architectures,
1434 Kirchenbauer et al. (2023) introduce a watermarking technique applicable to texts generated by any
1435 common autoregressive language model.

1436 **Application of LLMs in Creative Writing:** Recent advancements have highlighted the potential
1437 of LLMs in supporting various creative writing endeavors, ranging from short stories (Yang et al.,
1438 2022) to screenplays (Mirowski et al., 2023b). Enhancing LLMs to produce text that aligns more
1439 closely with human preferences has made them adept at following user instructions, thereby turn-
1440 ing them into valuable tools for individuals without technical expertise. This progress has boosted
1441 the commercial viability of LLMs as writing aids, which can continue a narrative, describe scenes,
1442 or offer feedback. Chung et al. (2021) conducted a review of literature on creativity support tools
1443 across various arts, leading to the development of a taxonomy that includes roles, interactions, and
1444 technologies. In contrast, Frich et al. (2019) and Palani et al. (2022) focused on how creative prac-
1445 titioners select new tools, highlighting their emphasis on functionality, workflow integration, and
1446 performance, and noting that personal recommendations often guide their choices. Additionally,
1447 Gero et al. (2022) created a space based on the cognitive process model of writing, influencing inter-
1448 face design decisions. Gero et al. (2023) further explored the social dynamics of AI in creative tasks,
1449 revealing a disconnect between writers’ objectives and the support provided by computer tools.

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

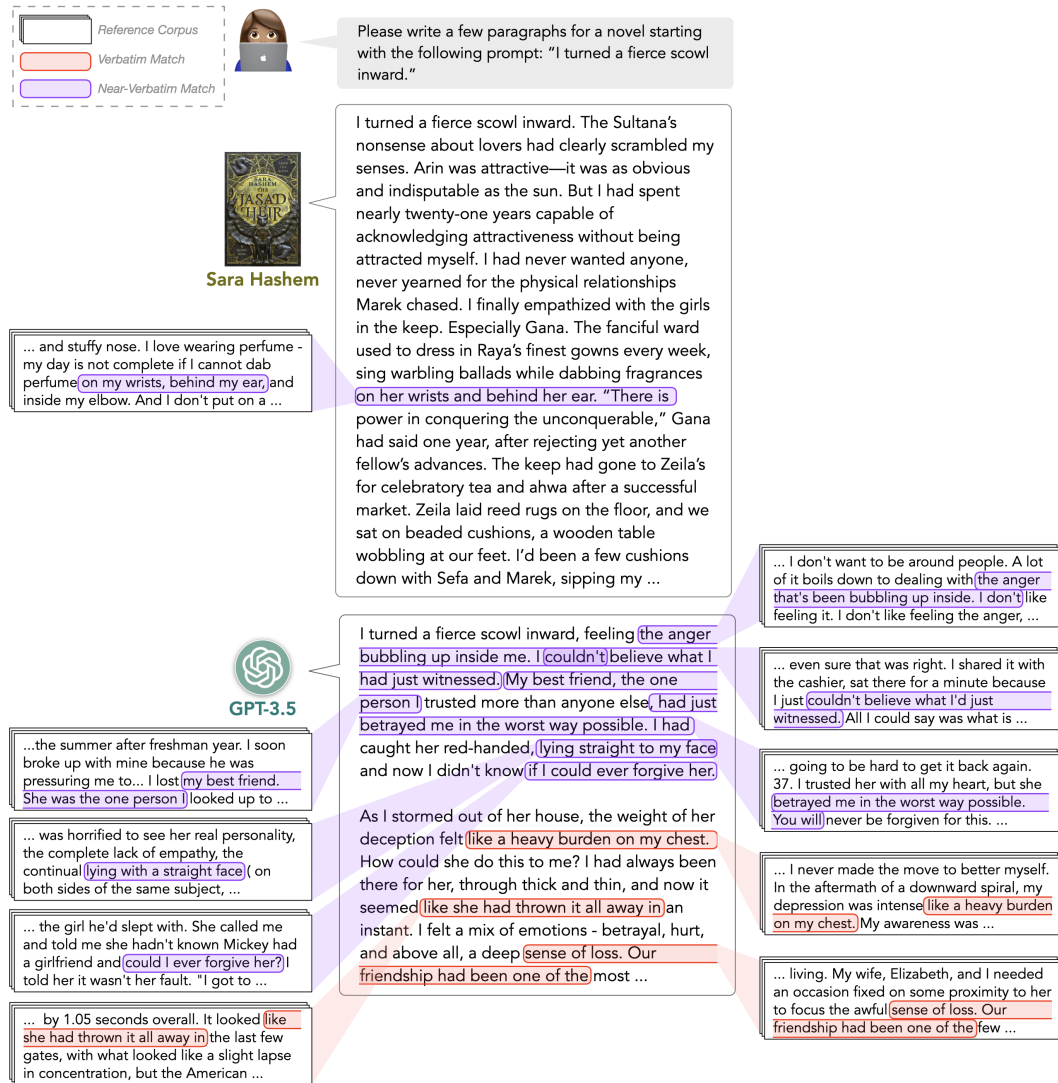


Figure 5: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565



Figure 6: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

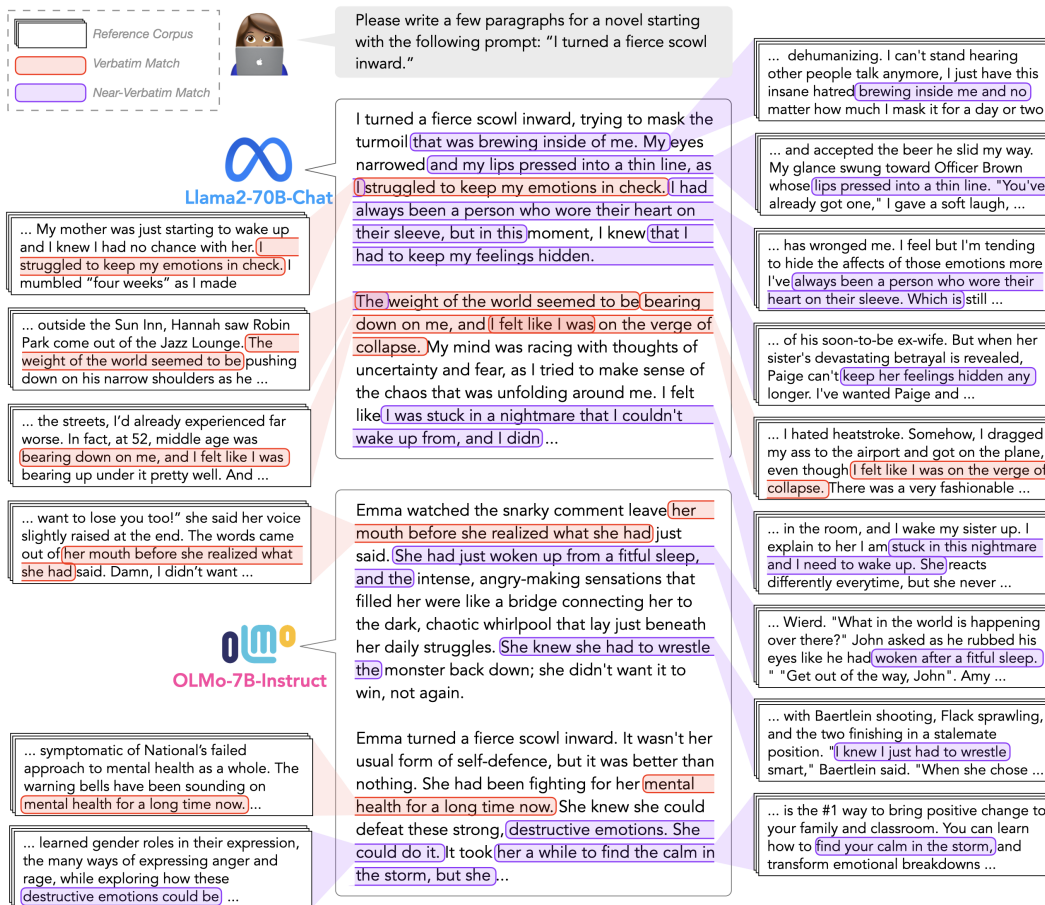


Figure 7: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673



Figure 8: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727



Figure 9: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

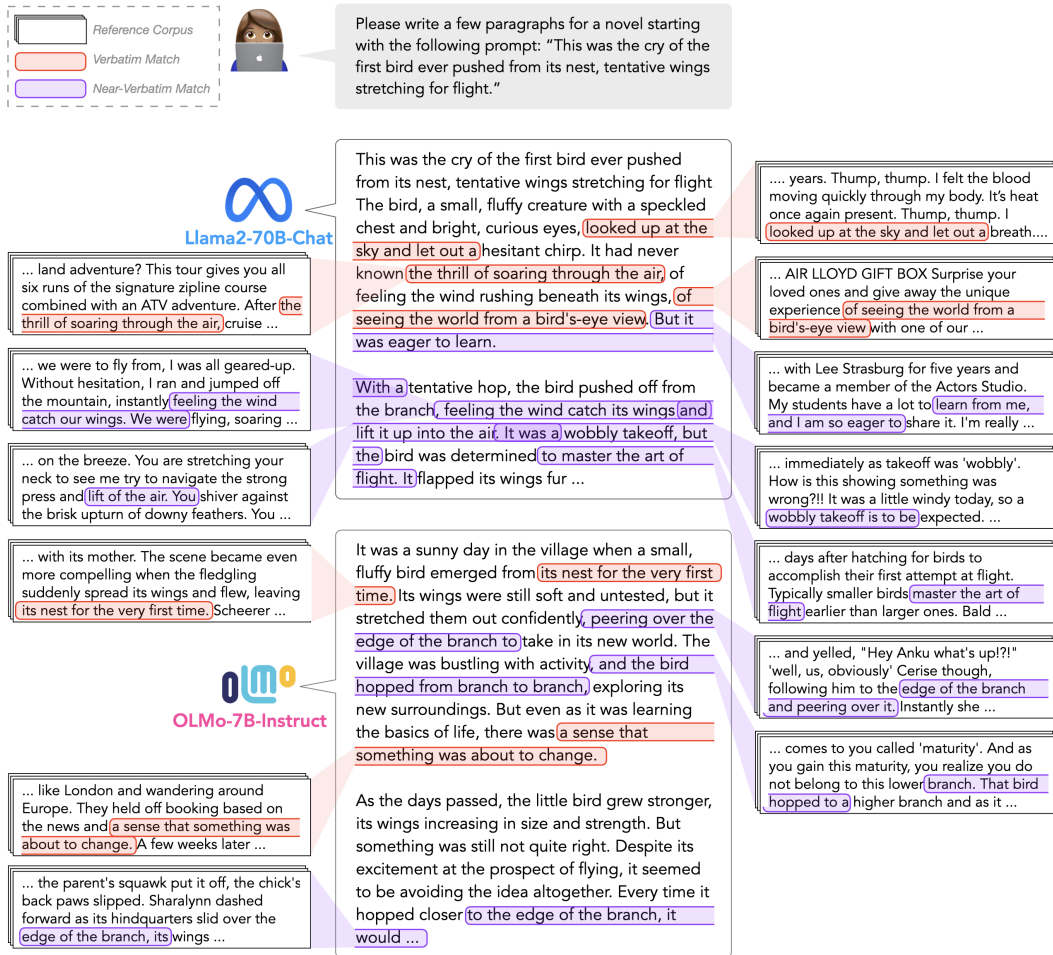
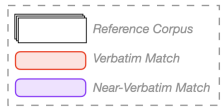


Figure 10: Example outputs from DJ SEARCH based on both verbatim and semantic matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835



Please write a few paragraphs for a novel starting with the following prompt: "A sickening thunk as a fist crashed into his temple; he slumped to the ground."



Hannah Whitten

A sickening thunk as a fist crashed into his temple; he slumped to the ground. Someone grabbed Lore's braid, fallen from its jet pins, and jerked it backward. She snarled, but the Presque Mort's arms closed around her, kept her confined. It took two to do the same to Bastian; the Sun Prince thrashed, shouting curses that echoed through the slowly darkening atrium. One of the monks struck out with a dagger; the sharp edge sliced through Bastian's eyebrow, sheeting blood and shocking him into enough stillness to be subdued, arms twisted behind his back. The shadow of the moon moved closer to the low-hanging sun. The Presque Mort who held Lore steered her toward August's throne. The Sainted King stood motionless and aloof, hands behind his back. Another Presque Mort—the one from the leak, walking almost normally on a prosthetic foot—approached the dais and handed the King a dagger, cast in silver and scrolled over ...



A sickening thunk echoed through the air as a fist crashed into his temple, sending a searing pain through his skull. He felt his legs give out from underneath him, and he slumped to the ground, his vision blurring as his head hit the hard earth. He could feel his pulse racing in his temples, and his breath coming in ragged gasps. As he lay there, he could feel the weight of his attacker's footsteps approaching him, the ground trembling beneath his feet. He tried to lift his head to see who had attacked him, but the pain was too much, and he fell back onto the ground, his eyes slipping closed ...

... Sandstream staggered a few tail-lengths towards the pond. His knees collapsed under under him, and he slumped to the ground, his tail dangling in the water. ...

... leaped to his feet along with the other men as Laiman burst into the clearing, his eyes wide and his breath coming in ragged gasps. "Patrol," he wheezed, ...

... stretching in a row from Alton to East Tisted. As he lay there, he could feel the back of his shirt, damp from sweat, even though the sun was pale and barely ...

... of treatments to no avail. Songaila tried to participate in the first training camp practice last month in Richmond, but the pain was too much, and he has not ...

... When this happens, he seems a little disoriented, his legs give out from underneath him, and he falls or almost falls. How should he deal with this? ...

... Poe had snuck liquor from his father's cupboard when they were far too young to be drinking it. He slumped to the ground, his vision blurring as tears formed at ...

... # _Chapter 59_ "Don't move," said Jude. Sebastian froze, his gaze locking with Hero's. He could feel his pulse racing in his neck, feel the wind buffeting his ...

... the door. I tried to ignore them, glanced at Lugh slumped in the opposite corner. No hands. No feet. He tried to lift his head and I choked back a cry as I saw the ...

... His hand slipped from Fenrir's and he fell to the ground. He tried to stand, tried to look back to see who had attacked him, but a boot pressed against his temple ...

... it was before. He felt the world starting to go dark, and had no energy to fight it...and so he fell back onto the ground, his conscious slipped away from him. When ...

Figure 11: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

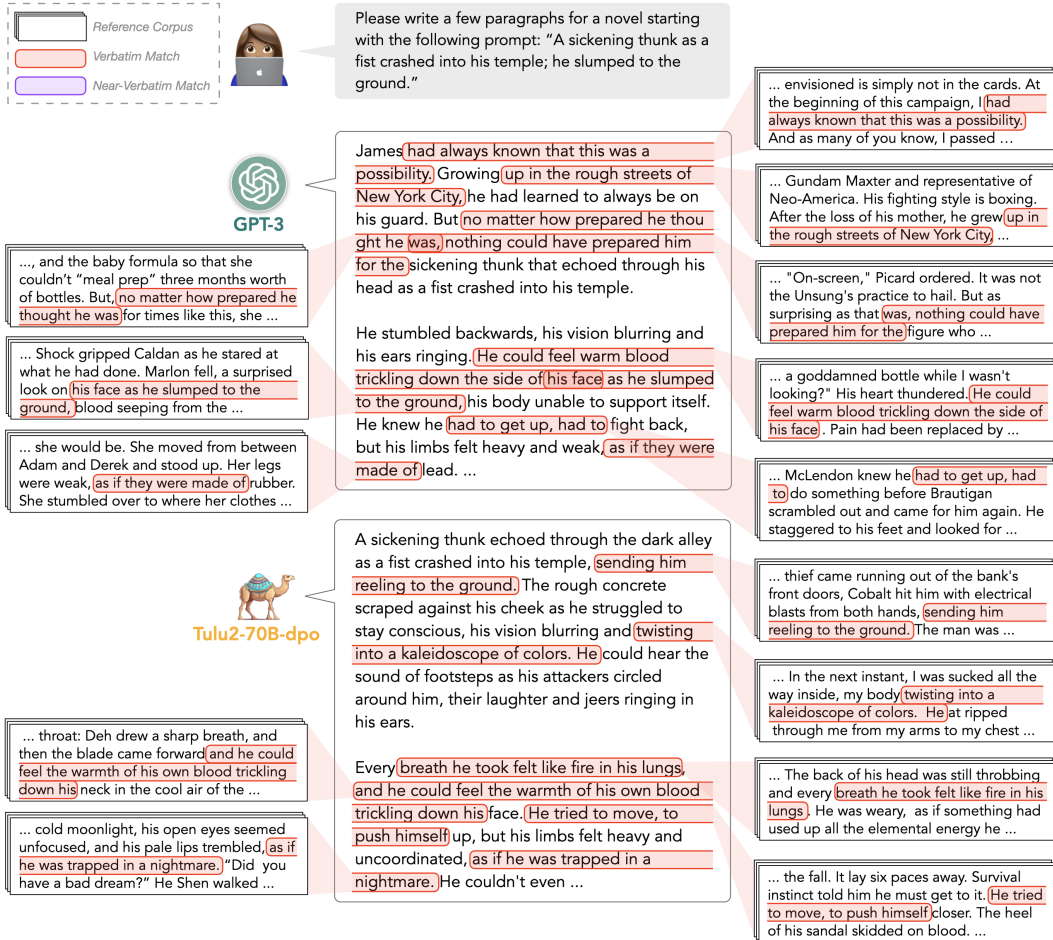


Figure 12: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943



Figure 13: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

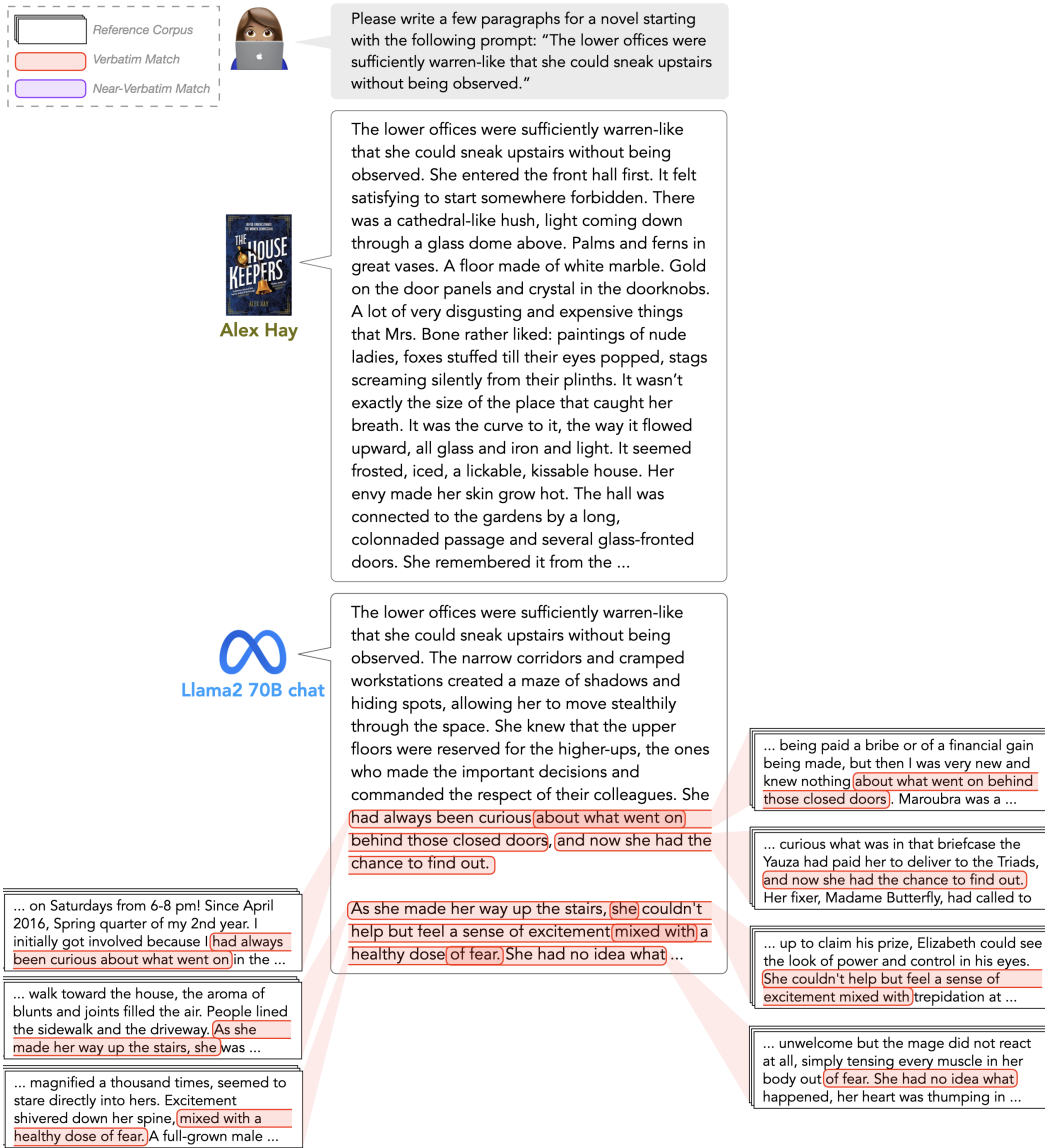


Figure 14: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

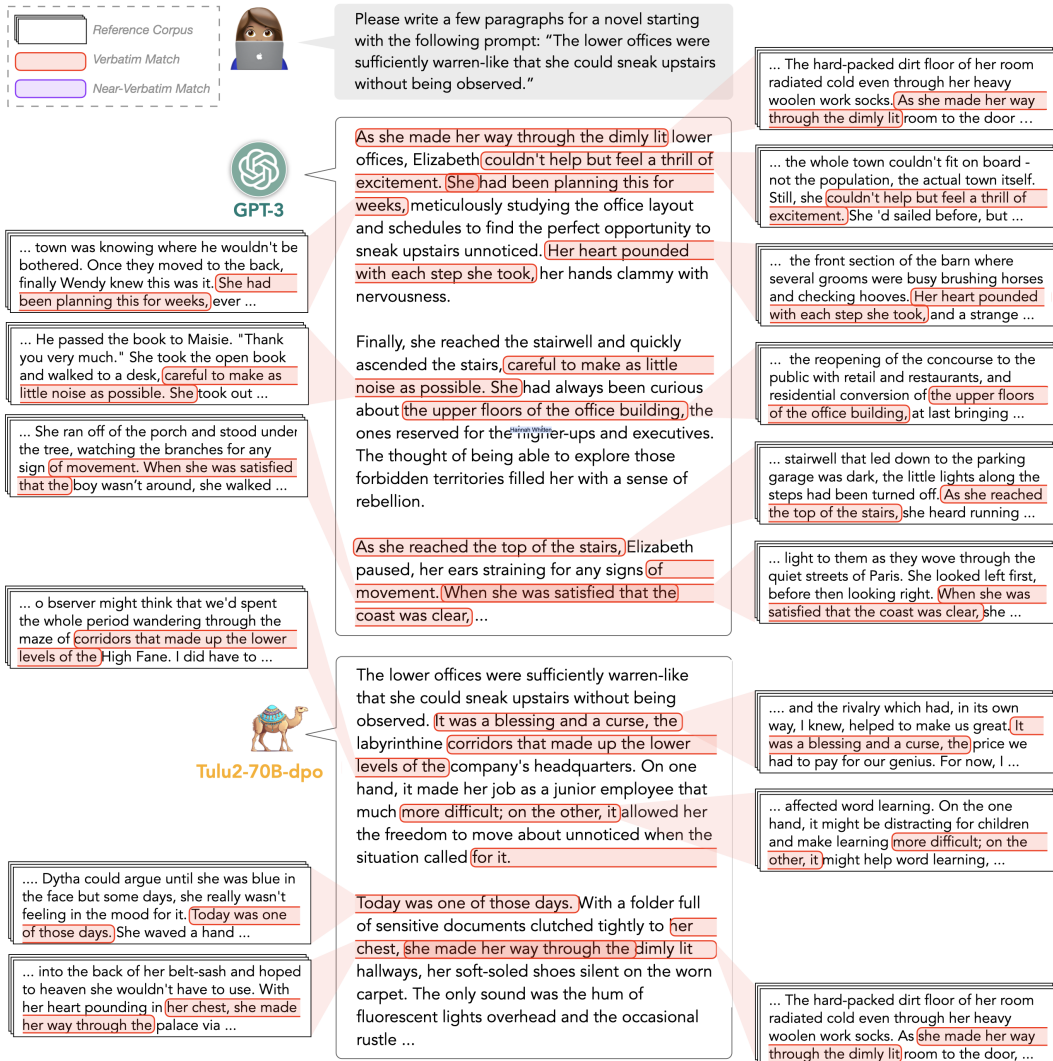


Figure 15: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105

Reference Corpus
 Verbatim Match
 Near-Verbatim Match

Please write a few paragraphs for a novel starting with the following prompt: "The lower offices were sufficiently warren-like that she could sneak upstairs without being observed."

... the reality show Undiscovered. Dawn was nervous for her daughter. This e-mail contained news that could potentially change her life forever. Dawn opened ...

... Hyacinth whispered as they approached the corner. The Sun Is Also a Star I'm sorry? He peered around the corner, making sure the coast was clear. How did you know ...

... right path again, but with the impending trap, that was not very likely. Details details! She had a mission to accomplish, and she was already going to do it well, ...

... by contrast, made a terrible racket, even through his thick fur coat. Narmora was the first to reach the gates. She pressed herself against the wall, listening intently to ...

... her just as powerless. A whispered request from his sensual lips and she'd melt into him and beg for more. Her heart pounded in her chest as she stared into ...

... Alice said, then fell in behind the other woman as she led her to Mr Davenport's office, and the interview that could change her life forever. It was hardly surprising ...

... learning in the Deepblue consumed huge volumes of resources. Even though he had the subsidy of 30,000 coins, he knew that it came with a price. It looked ...

... Fuck time travel, fuck Setsuna, fuck Island. Loser: UzaMaid! - We don't understand how Kamoi walks freely under the light of the sun. She should be in a ...

... as a provincial premier. With his swearing-in, Mr Mabuza at least has a finger in the jamb of the door to the highest office in the land. Ms ...

The lower offices were sufficiently warren-like that she could sneak upstairs without being observed. Ivy knew this well, as she had been working in the building for years. The dimly lit corridors and maze-like layout made it easy for her to slip away unnoticed. As she made her way up the stairs, her heart raced with anticipation. She was about to do something risky, something that could potentially change her life forever.

Reaching the top floor, Ivy paused to catch her breath. She peered around the corner, making sure the coast was clear before continuing on. She had a mission to accomplish, and she was determined to see it through. As she crept closer to her destination, the sound of muffled voices grew louder, causing her to freeze in her tracks. She pressed herself against the wall, listening ...

She snuck through the hallway, careful not to make a sound. She had to be quiet, she had to be quick, or she would be caught. Her heart pounded in her chest as she made her way up to the higher offices of the building, her destination hidden from view in the dark corners of the lower floors.

She had a plan, a risky plan, that could change her life forever. It was a plan that could make her rich and give her the power to do good in the world. But she knew that it came with a price. If she was caught, she would be arrested, her dreams of wealth and power shattered like a glass prism under the light of the sun.

She reached the top of the stairs and pushed open the door to the highest office in the ...

... "excited to start something new," and felt more comfortable making the transition since she had been working in the building for years and she was ...

... oblivious to her and her bookish self, so much so that they let her wonder away in a ball, which made it easy for her to slip away with Nathaniel. Anyway, the ...

... night, Cord. "Abigail," was all he said, but she could feel his eyes on her as she left the kitchen. As she made her way up the stairs, the men's voices carried to ...

... body tight against the trunk in an attempt to make herself as invisible as possible as she took the time to catch her breath. She peered around to check if ...

... Whatever the reason, an odd pricking sensation ran the length of Black's spine, causing her to freeze in her tracks. Clucking her tongue against the roof of ...

... Henry crossed the bridge into Oakland. No lights in the windows of the house. Inside, down the hallway, careful not to make a sound, Thomas was asleep on ...

... me at all. No matter the outcome, it didn't work the way I wanted it to: If Jax won, it had to be quick, or she would look bad for taking so long. If Bliss won, ...

... It would have been cruel to puncture the boy's illusion, to remind him of those who lived in obscurity, hidden from view in the dark corners of the metropolis, the ...

... technical/sectoral bridge," Zeichner says. Aside from being good for business, Zeichner believes IoT has the power to do good in the world. He emphasises, ...


... French soil, she thought with a shiver of fear; enemy territory. Now she was a criminal, a terrorist, a spy. If she was caught, she would be executed. She put ...

... the words. "We are," she said. "He's in a room at the top of the stairs." Or so she hoped. She reached the top of the stairs and pushed open the double doors. To ...

Figure 16: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a few paragraphs of a novel, beginning with a first sentence taken from a human-written novel snippet.

2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159


Reference Corpus
 Verbatim Match
 Near-Verbatim Match



Please draft a speech with the opening sentence:
"It is with this and much more in mind that I come before you today."

... One editorialist in the trade press wrote that "the FCC of the New Frontier is going to be one of the toughest FCC's in the history of broadcast regulation." If he meant that we intend to enforce the law in the public interest, let me make it perfectly clear that he is right: We do. If he meant that we intend to muzzle or censor broadcasting, he is dead wrong. It wouldn't surprise me if some of you had expected me to come here today and say to you in effect, "Clean up your own house or the government will do it for you." Well, in a limited sense, you would be right because I've just said it. But I want to say to you as earnestly as I can that it is not in that spirit that I come before ...


carbon-cutting goals. "Our building sector is going to be one of the toughest sectors to fully decarbonize," said Deborah Donovan, Massachusetts director...



Newton Minow

... >>>> The police have a responsibility to enforce the law in the public interest.> Actually, and interestingly, you're mistaken. And, actually, but far less...

...and the Liberals didn't do anything about it. But I want to say to the government, and this is very serious, that we have been living through some very...



GPT3

... As I stand before you, I am filled with a deep sense of responsibility and purpose. For I understand that every word I speak here today holds the power to shape our future, to inspire change and to bring about progress.

I am honored to have this opportunity to address such a distinguished audience, and I want to take this moment to express my gratitude for your presence and attention.

We are living in a time of great uncertainty and challenges. Our world is constantly evolving, and with it, so are the issues we face. From climate change to social inequality, from political divisions to economic disparities, our society is grappling with complex problems that require immediate attention.

However, amidst all the chaos and confusion, there is...

... I am delighted to have this opportunity to address such a distinguished audience, at such a stellar event, and on such a lovely evening...

... The holidays are upon us, and I want to take this moment to express my gratitude for so many meaningful moments and exchange...

... the appointment in a letter dated 29th January, 2019. "I would like to express my gratitude for your presence and valuable input and leadership role at the first

... The crypto world is constantly evolving, and with it, so are the trading platforms. LeverFi is one such platform that is making waves in the industry [...] Share your videos

... they found no problems that require immediate attention. However, the dam has been given a "poor" rating by the EPA. See below for the original documents...

...Regulate the industry, or the government will do it for you. With that ultimatum on the table, several major players in the media industry...

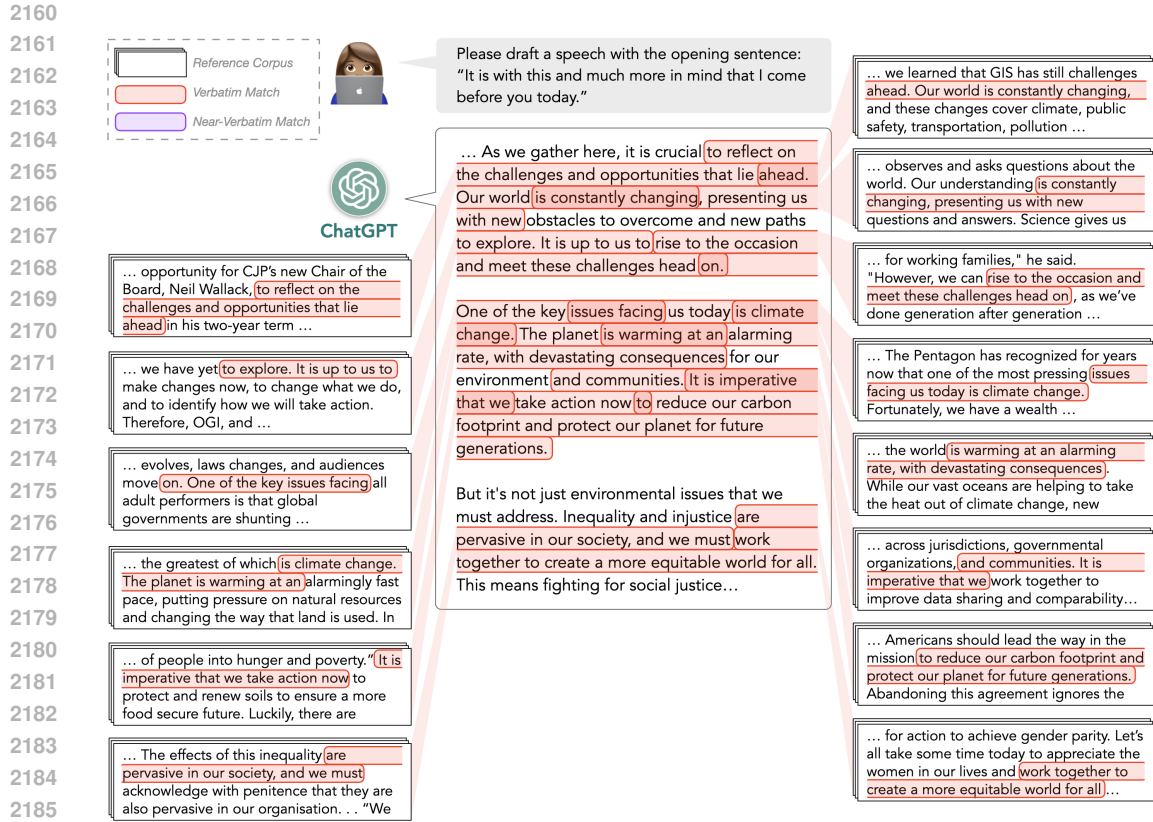
... their trust in me and our Liberal team," McNeil said. "It is with a deep sense of responsibility and purpose that I will make certain that our plan is delivered...

... this goal. Once again, I am honored to have this opportunity to address you and look forward to the rest of the sessions today...

... pandemic has put everyone on edge. We are living in a time of great uncertainty and almost a year later, there's still no guarantee of when this pandemic will end.

...Murphy to tie it all together in the end. Even with all the chaos and confusion, there is one sure thing about the returning characters -- we are glad to have them

Figure 17: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.



2187 Figure 18: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

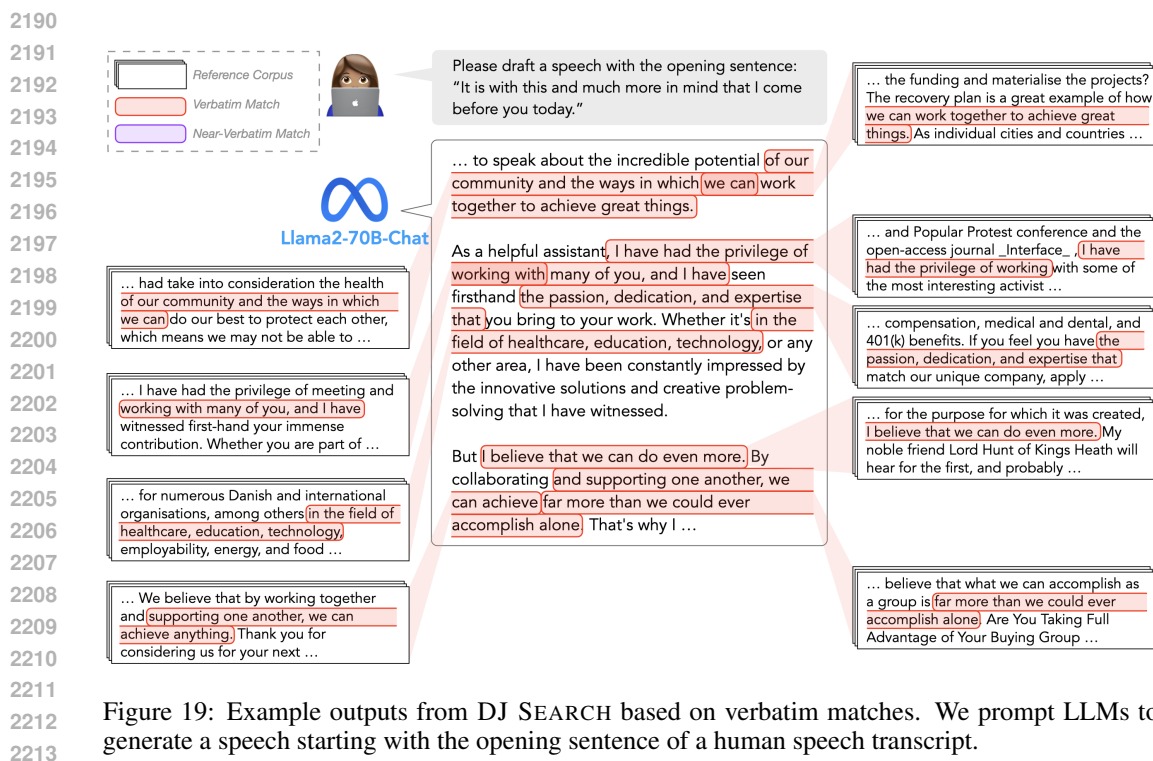


Figure 19: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239

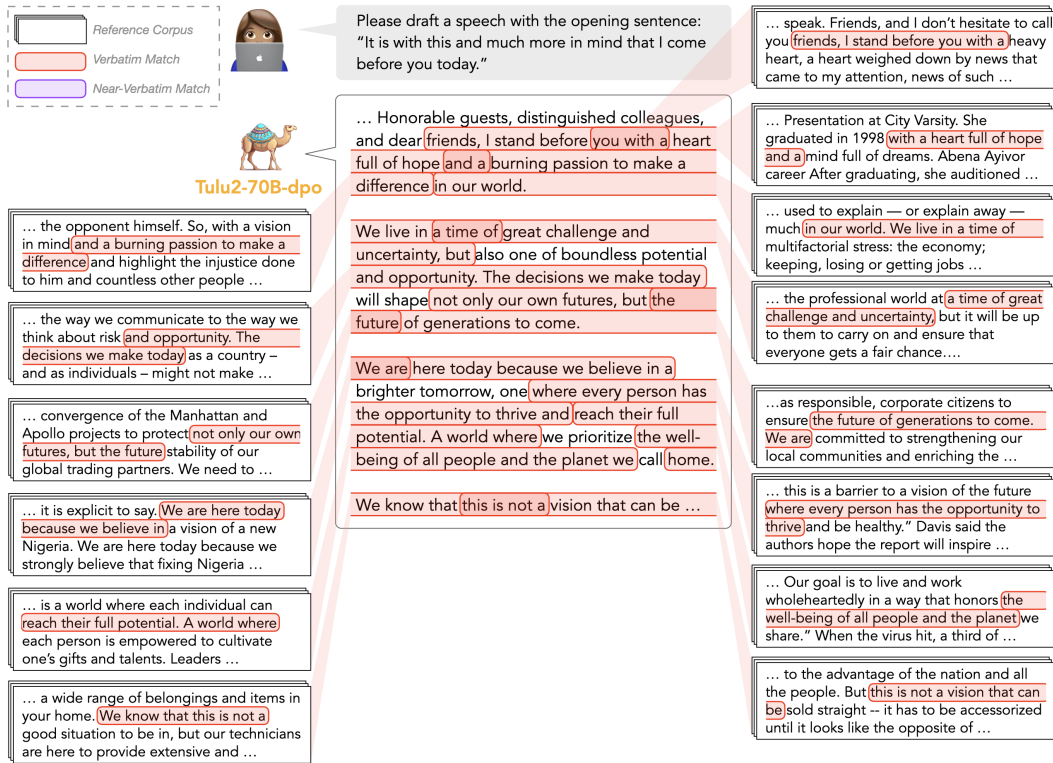


Figure 20: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265

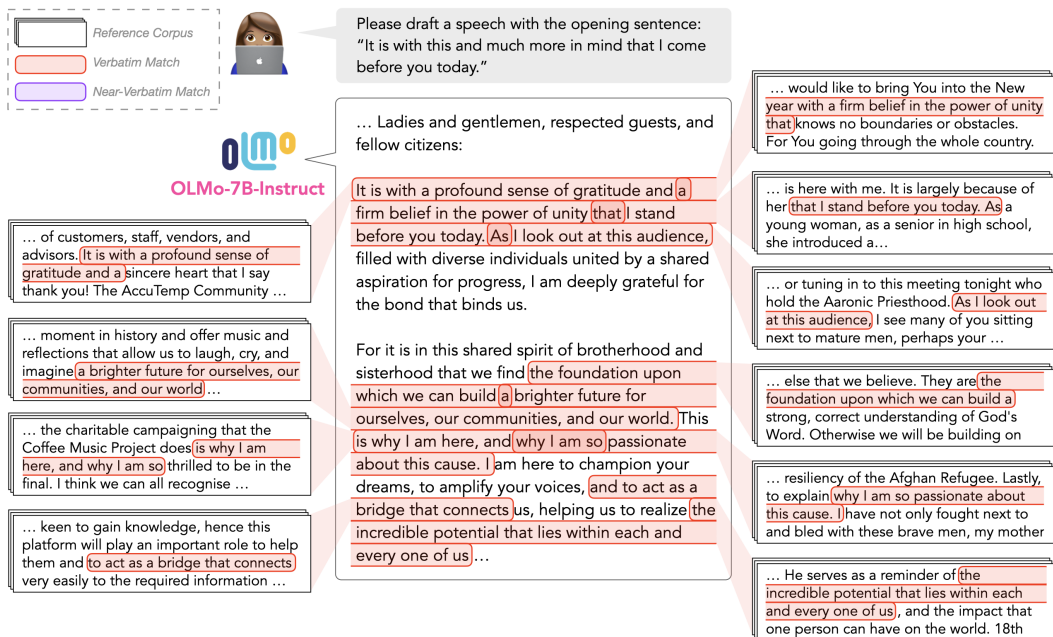


Figure 21: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321

Reference Corpus
 Verbatim Match
 Near-Verbatim Match

Please draft a speech with the opening sentence:
"Green fields and dense forests are disappearing."

Lyndon Johnson

... A few years ago we were greatly concerned about the "Ugly American." Today we must act to prevent an ugly America. For once the battle is lost, once our natural splendor is destroyed, it can never be recaptured. And once man can no longer walk with beauty or wonder at nature his spirit will wither and his sustenance be wasted. A third place to build the Great Society is in the classrooms of America. There your children's lives will be shaped. Our society will not be great until every young mind is set free to scan the farthest reaches of thought and imagination. We are still far from that goal. Today, 8 million adult Americans, more than the entire population of Michigan, have not finished 5 years of school. Nearly 20 million have not finished 8 years of school. Nearly 54 million -- more than one quarter of all ...

... the farthest reaches of thought and imagination. We are still far from that goal. Today, 8 million adult Americans, more than the entire population of ...

... THE MODERATOR: Ladies and gentlemen, distinguished guests, and dear friends, media friends, very warm welcome to everybody, and welcome to the ...

... Ladies and gentlemen, distinguished guests, and fellow nature enthusiasts, it is with a heavy heart that I stand before you today to address a pressing issue that threatens the very fabric of our planet. The green fields and dense forests that once adorned our landscapes are rapidly disappearing, and it is our duty to take action before it's too late.

... Vue enthusiasts, it is with a heavy heart that I must remind you that Vue is headed to that great console in the sky at the end of January. Here's" the new round ...

... CHURCHILL: "My dear friends, I stand before you today to address a matter of grave importance. The 1929 Reapportionment Act, also known ...

... control, tearing the X-Men family apart and threatening to destroy the very fabric of our planet. The film is the most intense and emotional X-Men movie ever ...

... of government affairs, Robert McCaw. "White supremacist violence is an issue that threatens the very fabric of our nation and must be addressed in ...

... on Health Care Financing. "The safety of our children is paramount, and it is our duty to take action to protect them from toxic substances found in flavored ...

... to public education, including our school system. We have just three weeks to take action before it's too late. The bills are:" of those who live there, many ...

... already lost 85% of their forest cover – in fact, every minute, an area of forest equivalent to 27 soccer fields is lost globally.1 Facts like these are hard ...

... The fight to be recognized and accepted has been an ongoing battle for decades, perhaps longer, and while we have made progress, today there are ...

... can vary greatly. Land-use and land-cover changes such as deforestation, agricultural expansion, urbanization, and other human activities have a direct ...

... to a better technology, year after year. Electronics have always produced waste, but in recent years, the rate at which we use and discard devices has rapidly ...

But why should ...


Figure 22: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375

Reference Corpus

Verbatim Match

Near-Verbatim Match



ChatGPT

Please draft a speech with the opening sentence:
"Green fields and dense forests are disappearing."

... at an alarming rate, and it is up to us to take action before it is too late. The loss of these vital natural habitats not only impacts the environment but also has far-reaching consequences for our health, well-being, and the future of our planet.

As trees are cut down and land is cleared for agriculture and development, we are losing valuable carbon sinks that help regulate our climate. Deforestation contributes to greenhouse gas emissions and exacerbates the effects of climate change, leading to more frequent and severe natural disasters such as wildfires, floods, and droughts.

Furthermore, the destruction of forests and fields threatens biodiversity and disrupts ecosystems, putting countless species at risk of extinction. These natural habitats provide essential services such as clean air and water, pollination, and soil fertility, which are ...

Llama2-70B-Chat

... at an alarming rate, and it's a trend that we must reverse. The impact of deforestation and land degradation is not only felt by the environment, but also by communities and individuals who rely on these ecosystems for their livelihoods.

Forests provide us with clean air and water, fertile soil, and a wealth of biodiversity. They are also a critical source of income for many communities, providing wood, fruit, and other forest products. But sadly, these valuable resources are being lost at a rate of tens of thousands of hectares per day.

The consequences of deforestation are far-reaching and devastating. Soil erosion ...

... and it is up to us to take action on those words. Worshiping together in the context of a larger body provides benefits and protections to us that would ...

... are taken to protect and preserve the rainforest before it is too late. The loss of the tropical rainforest would be the biggest catastrophe for mankind possible. In ...

... did yesterday and expect a different result." Delayed action not only impacts the environment but also has financial implications. The longer occupants ...

... and duration of heatwaves. Heat also has far-reaching consequences for our health, causing or contributing to dozens of debilitating or deadly illnesses ...

... top of the pile call the shots. We are paying for these profits with our health, our well-being, and the future of our planet. The trickle down of environmental ...

... I be an increase in natural disasters, such as wildfires, floods, and droughts. In Chico, as a whole, the economy will be affected by all of this. Northern ...

... humans, our welfare and economy, but it's putting countless species at risk of extinction. My artwork portrays the dilemma created by climate change ...

... the smartest deals of the last few years. League of Legends is spreading across the globe at an alarming rate, and it's a prime mover in the increasingly lucrative ...

... into the sea. Effect On Human Health The effects of global warming are not only felt by the environment, but by humans as well. Higher temperatures can lead to ...

... and Lake Turkana as well as the people who rely on these ecosystems for their livelihoods. According to Terri Hathaway, director of International Rivers' ...

... also work to protect and enhance farmland so that future generations will have clean air and water, fertile soil, and a source for healthy, delicious food ...

... personally? A: The forests represent a national heritage and a wealth of biodiversity. They are also spaces for the release of oxygen and they provide a ...

... who pay thousands of dollars to digitize books at a rate of tens of thousands of volumes a year?" It seems like those who ask such questions have little ...

Figure 23: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429

Reference Corpus
 Verbatim Match
 Near-Verbatim Match

Please draft a speech with the opening sentence:
"Green fields and dense forests are disappearing."

Tulu2-70B-dpo

... at an alarming rate. Our once lush and thriving landscapes are transforming into barren and desolate wastelands, devoid of life and color. The unchecked expansion of industries, urbanization, and deforestation have resulted in the devastating loss of our natural heritage.

But the loss of greenery is not just an aesthetic issue. The consequences of this depletion are dire and far-reaching. Forests and fields are the lungs of our planet, providing us with the oxygen we need to breathe and absorbing the carbon dioxide that threatens our existence. Without them, we face the danger of climate ...

... have the opportunity to transform the nature of public participation in government. Without them, we face the danger of cutting the people out ...

OLMo-7B-Instruct

... Ladies and gentlemen, distinguished guests, and everyone watching or listening to us today: As we stand here, on this grand and important day, we are faced with a challenge that demands our attention and calls upon our shared responsibility for the future.

Green fields and dense forests are disappearing. The sights that once defined our planet with beauty and abundant biodiversity are dwindling, and the natural world we once knew is being supplanted by cities, highways, and urban sprawl. This disappearance, a cause for alarm and concern, holds not just ecological and environmental implications but also social, cultural, and economic impacts that touch all of us, irrespective of our nationality, ethnicity, or place of residence.

... overburdened. Green fields and dense forests are disappearing. A few years ago we were greatly concerned about the 'Ugly American'. Today we must act to ...

... populations in the name of conservation can result in severe social, cultural, and economic impacts that often affect the most marginalized groups among ...

The causes of this phenomenon are well-known, and they are multifaceted and interconnected: deforestation ...

... for chihuahua stolen in SF smash-and-grab The subsidence is not just an aesthetic issue. The sinking sidewalks have created danger for pedestrians ...

... for between 70-80% of the Amazon rainforest loss. Forests are the lungs of our planet, providing us with oxygen and also sequestering planet-warming ...

... "Wherever we are, we are connected to the sea it provides us with the oxygen we need to breathe and supports the climate that we live off. We need a healthy ...

... may remember from school, plants photosynthesize, producing oxygen for us to breathe and absorbing the carbon dioxide that's dangerous for ...

... THE MODERATOR: Ladies and gentlemen, distinguished guests, and dear friends, media friends, very warm welcome to everybody, and welcome to the press

... goal of curling an extra five pounds for three additional reps we are faced with a challenge that demands our full attention. Losing focus under such conditions ...

... Sea region, thus involving not only spatial and economic implications but also social, cultural, and ecological implications. Especially, the logistics and the ...

... campaign and raise awareness of the fact that all of us, irrespective of our nationality, are citizens of the same planet earth. Werner Wintersteiner and ...

... to individuals based on their race, ethnicity, or place of residence. The systematic refusal to provide goods or services is not based ...


Figure 24: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a speech starting with the opening sentence of a human speech transcript.

2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483


Reference Corpus

Verbatim Match


Near-Verbatim Match



Please write a poem starting with the following line: Like a long arrow through the dark the train is darting,



Henry Van Dyke



GPT-3

... Bearing me far away, after a perfect day of love's delight:
Wakeful with all the sad-sweet memories of parting,
I lift the narrow window-shade and look out on the night.
Lonely the land unknown, and like a river flowing,
Forest and field and hill are gliding backward still athwart my dream;
Till in that country strange, and ever stranger growing,
A magic city full of lights begins to glow and gleam.
Wide through the landscape dim the lamps are lit in millions;
Long avenues unfold clear-shining lines of gold across the green;
Clusters and rings of light, and luminous pavilions, —
Oh, who will tell the city's name, and what these wonders mean?
Why do they beckon me, and what have they to show me?
Crowds in the blazing street, mirth where the feasters meet, kisses and wine: ...

... Racing through the night with a rhythmic heart beating.
Passing by towns and fields, it's always departing,
A journey of endless possibilities, always repeating.

With each passing mile, it's a new adventure,
A chance to escape, to leave behind the mundane.
As the wheels spin, they create a symphony of texture,
A song that soothes the soul, a journey that's far from plain.

Through tunnels and over bridges, it travels on,
Carrying passengers to their destinations unknown.
It's a symbol of freedom, of being unbound,
A fleeting moment of escape, before returning to reality's throne.

As the train moves forward, the darkness fades,
And the world outside slowly comes to light.
It's a reminder ...

... Come on, come in, and like a river flowing, In volume irresistible toward the raging sea; Let all the nations see that Britons now are willing ...

... commissioners. I have a wonderful team of about 6,000 employees, and, every day, it's a new adventure, so we really try to make sure that our community is ...

... to a tone which successfully blends the surreal and the mundane. As the poet writes, she evokes several emotions at once. In 'The Horse' she cleverly ...

... I get new directions and head along Calle Mexico toward destinations unknown. It's a sunny warm afternoon and the city is still quiet from the ...

... that will gain power when the scandal of Kate allegedly killing Pope comes to light. It's a pretty sound partnership, Callahan gets rid of his challenger and gets ...

Figure 25: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.

2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537

Reference Corpus
 Verbatim Match
 Near-Verbatim Match



ChatGPT

Please write a poem starting with the following line: Like a long arrow through the dark the train is darting,

... Racing against the night, relentless and unyielding.
 Its iron wheels thundering on the tracks, relentless,
 Pushing forward through the darkness, relentless.

The world outside blurs into a streak of lights, As the train hurtles forward, unstoppable might. Through valleys and mountains, it forges ahead, Leaving behind a trail of smoke and dread.

The passengers inside are but fleeting shadows, Caught in the whirlwind of this metal beast's echo.
Their faces illuminated by the glow of screens,
Lost in their own worlds, in their own dreams.

Yet, through the windows, the world whizzes by, A blur of landscapes, under the vast sky.
 Each moment passing, each second fleeting,
 As the train hurtles forward, its ...



Llama2 70B chat

... Through the night it speeds, a pathway charting,
 Carrying dreams and hopes, it hastens by,
 With a rhythm swift, a symphony in the sky.

The passengers, a diverse throng,
 Each with a story, a life, a song,
 Their faces lit by the moon's pale glow,
 Their spirits high, as the journey grows.

The engine roars, a mighty sound,
 As the wheels clack, a rhythmic all around,
 Through tunnels deep, through valleys wide,
 The train speeds on, with a steady stride.

The ...

Figure 26: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.

2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591

Reference Corpus

Verbatim Match

Near-Verbatim Match

Please write a poem starting with the following line: Like a long arrow through the dark the train is darting,

Tulu2-70B-dpo

... Its metal **body slicing through the night**, never halting.
With each rhythmic pulse of its mighty heart, The wheels spin and churn, never **to fall apart**.

Through the tunnels deep, and the valleys wide, The train speeds along, its progress untried. With each passing mile, a **new story unfolds**, Of adventure and dreams, **of secrets untold**.

In the cabin's quiet, the passengers stir, Lost in their thoughts, and a world they prefer. For the train, you see, **is not just a mode of transport**, But a vessel of hope, a ...

... back to her apartment. Nora pedaled swiftly, her **body slicing through the night**, her hat shuddering in the wind. She knew the way well; even in the darkness ...

... the pain continued, but so did my determination **not to fall apart**. **Through** out the last weeks of 1975, I spent as much time as I could bringing Ebenezer's ...

... generous welcome or with a hostile attack is left **untold**. **In the** film 2001, at different points in the story, rectangular monoliths (not a pyramid as ...

... wide-awake and thrumming hotly through her blood, despite the nagging awareness **of secrets untold**. "Darling?" Henry sounded incredulous. "What ...

... certified for 60 passengers and three cars. **Øysang ferry** **is not just a mode of transport**, it is also a tourist attraction. The ferry runs between Risør city and ...

OLMo-7B-Instruct

... The train darted with a brave snap, Its passengers rushing to their fate, **As the lights flashed and the cable ran**.

The wind was high as it howled through the rails, As the train approached its destination, A sense **of excitement and jeopardy**, **In the shadows of the night**.

Through tunnels **that seemed like an eternity**, The train danced and weaved in sight, With each passing car, a little lighter, As the light of day seemed to delay.

The passengers, all with their own story, Saw **the world pass by in a blur**, As they rushed towards their destination, **In the dawn of a new day**.

And as the train arrived at last ...

... any book. **As the lights flashed and the alarm** sounded with a decorum fitting for the library, the staff told us to exit immediately and leave everything ...

... car runs, but it is that shot of adrenalin, and feeling **of excitement and jeopardy**, that reminds us that we have missed the action so much. The nature of ...


... Suffice to say that, after a lapse of time **that seemed like an eternity**, we both rolled into the sunken road and were, for a moment, under cover. Here, as I was ...

... in a hurry or just get a buzz from watching **the world pass by in a blur** the modern generation of fast trains is a travel pleasure. Here are the world's 10 ...

Figure 27: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.


2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645

Reference Corpus
 Verbatim Match
 Near-Verbatim Match



Please write a poem starting with the following line: Why, when our sun shines clearest,

... Why, when our hopes seen nearest,
 Why, when our life feels dearest,
 Rises a secret pain—
 Hope's perfect mirror broken—
 Shadows of things unspoken—
 Why will not some sure token
 Calm us to rest again?
 Mixed with all earthly blessing
 Lingers the fear distressing—
 -Conscience within confessing
 Nothing of ours is pure.
 Still must such thoughts upbraid us,
 Seeking our own to aid us;
 God, not ourselves, hath made us;
 Trusting in Him we're sure.
 Thus, from our sorrows gleaming
 Thoughts of the world's deep meaning,
 Let us rejoice while leaning
 Firm on our Father's arm.
 Now are we one for ever,
 Joined so that none may sever,
 Souls, so united, never
 Faint through mischance or harm ...



James Clerk Maxwell

... When the birds sing their sweetest songs,
 Do we still feel so alone,
 As if we don't belong?


Do we let our worries cloud,
 The beauty of each day?
 Or do we carry past pains,
 That we can't seem to shake away?

Perhaps it's our own perception,
 That keeps us feeling small,
 We forget to see the blessings,
 And let the darkness fall.

But if we pause and look around,
 With open hearts and eyes,
 We'll see the love that surrounds us,
 And feel less lonely inside.

For in the midst of nature's symphony,
And the warmth of the sun,
We'll find peace and solace,
 In the moments we are one.

So let us not dwell on the shadows,
 That try to dim our light,
 But bask in the beauty of the present ...



GPT-3

... mountains, it is easy to find things to celebrate, it's our own perception, more than any lack of beauty, that gets in our way. Citrus trees are generous trees ...

... your picture. it is special. it is unique. it is you. And the warmth of the sun will feel heavy on your skin. And the smell of fresh grass will come to your mind ...

... was is as good of time as any to start anew again. So let us not dwell on the existential meaningless of the Sunday loss and instead let's dwell generally ...

... than ever? It's often said that we don't bask in the beauty of the present moment because we're too busy staring at a screen. But is technology the problem or it is ...

... among the wild flowers that bloom, or where the birds sing their sweetest songs, these must all be made a part of that gentle life of devotion that you are to live ...

... cities and the wildlife that inhabits them. Should it worry us that we are building on places we don't belong? No, well, it shouldn't, because we are going to be ...

... Piano on Gem by Jimmy Vargas. Album art by Dave Dycus and Steve Chab. ... can't seem to shake away the fame of the crooked arrow. Would all my friends ...

... to see the world. When we love, we can more easily see the love that surrounds us, and recognize acts and expressions of love. Love can often be overlooked ...

... Towers, "Shock Corridor" Under the Blog Watch, Out of the sun, We'll be having some fun, People walkin' above, We'll be makin' love... Ted Barlow ...

Figure 28: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.

2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699

Reference Corpus

Verbatim Match

Near-Verbatim Match

Please write a poem starting with the following line: Why, when our sun shines clearest,

ChatGPT

... Does the darkness seem to linger?

Why, when our hearts are lightest,
Do shadows creep in to make us shiver?

In the brightest of days,
There's always a hint of night,
A reminder of the balance,
Between darkness and light.

But fear not the shadows,
For they too have their place,
In the tapestry of life,
Adding depth and grace.

Embrace the contrast,
And welcome the unknown,
For in the darkest moments,
Our true strength is shown.
So when our sun shines clearest,
And the shadows start to fall,
Remember, it's in the contrast,
That we find beauty in it all ...

Llama2 70B chat

... Do we often feel the most unease?
Is it because the light reveals,
All the things we'd rather not see?

The shadows hide the imperfections,
The flaws that we'd rather not face,
But when the sun shines bright and bold,
It illuminates the darkest place.

It's then we see the things we fear,
The doubts and worries that we hold dear,
The insecurities that plague our mind,
The fears that we cannot unwind.

But perhaps, this is a chance,
To face the things that we must advance,
To embrace the light and let it guide,
To ...

... cnr 2nd Ave & Princess St; h11am-11pm Mar-Nov) There's always a hint of pleasures to come swirling around the tables of Dawson's most inviting bar. Enjoy ...

... para bellum #ThisTaserGoesTo11 Between darkness and light. (I'll comment on the others later once others have commented. Madre can't exactly ...

... Burt Young and Tina Louise Star in Newly Released Faith-Based Family Drama "Tapestry" (In the tapestry of life, only the weaver sees the full design. Tapestry ...

... I'll feel the earth grow warm beneath my feet I got to move beyond what you've begun (And the shadows start to fall) behind Hope begins to rise I can see the ...

... and living high, rather the turning away from the things we'd rather not see and deal with. There are no words that can articulate the Mask of Fury ventilating ...

... and dark under-eye circles, amongst other flaws that we'd rather not face. While there are several cosmetic options to hide these many skin issues, you might ...

... that to the darkness...even the smallest candle flame illuminates the darkest place. Our partner is lighting those candles little by little and so is everyone else working ...

... under 7 years, and save yourself £1,354.74 in interest. Remember, it's in the interest of credit card companies for you to pay off as little as possible each month ...


... mighty and unpredictable; finally, it's usually windy/foggy/wear your fleece weather. But when the sun shines bright and you find a good spot on the Bay ...


... on the water... (It's then we see the true horror of the men in the water... in the burning black oil... The transport boat silently slips into the " heavy ...


... Couldn't Fit into a Tailored Bridesmaid Dress (The insecurities that plague) my daily muddling through life, my lifelong weight issues always seem to take ...


Figure 29: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.

2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753


 Reference Corpus

 Verbatim Match

 Near-Verbatim Match



Please write a poem starting with the following line: Why, when our sun shines clearest,



Tulu2-70B-dpo

... Do shadows still manage to creep?
Through corners they meander,
And in our hearts, they leap.

In moments of pure gladness,
When joy seems never-ending,
A pang of sadness finds us,
Our smiles, a sorrow blending.


But is this not the balance,
The yin and yang of living?
A dance between the lightness,
And the darkness we are giving.

So let us welcome shadows,
As sun sets on our days,
Embrace the night's caresses,
And in its secrets, find our ways.

For life is not just sunshine,
It ...

... study means what you think it means
The 2023 transition The yin and yang of
climate science The faux urgency of the
climate crisis is giving us no time or ...

... And I think that's one lie. We humans
need to tell and a story they need to tell
that life is not just sunshine, that in when
we say that, but we get 24/7 system ...



OLMo-7B-Instruct

... The shadows take a staggered step,
The light that warms us so kind,
Saves just enough to share.

In tranquility, the clouds alight,
Dancing the most elegant,
A symphony of powdery hues,
A canvas of unmatched beauty.

The world beneath, a serene,
Revealing the truth, the art,
Of light and shadows at play,
A melodious tale.

But as the night advances fast,
The shadows, together again,
Battle for our vision's space,
Trying to overtake.

Yet, despite the warmth they steal,
Our world remains intact,
A harmonious harmony,
Our ...

... and show admiration for the efforts the
country is making to restore security.
Revealing the truth, the Mozambique
government mentioned ...

... tales had been factual, then in his
drunken heat the man had blindly selected
a jewel of unmatched beauty. The
dressmaker gestured for them to follow ...

... latest edition here. The Tallons in the
Times, 1897 Yet, despite the warmth of the
tale told by the pub's owner, when my
parents' party walked out of the doors ...

Figure 30: Example outputs from DJ SEARCH based on verbatim matches. We prompt LLMs to generate a poem starting with the first line of a human-written poem.