Note: This paper was provocatively written for a class project, and has not undergone peer review. The results are more indicative of flaws with metrics measuring similarity rather than any global finding on homogenization or lack thereof.

Evidence Against LLM Homogenization in Creative Writing

Kia Ghods **Patty Liu**

Katerina Labrou **Kincaid MacDonald** **Arjun Menon** **Addison Wu**

Abstract

LLMs are by now widely recognized as capable of producing text whose creativity rivals the average person. But what about the average *group* of people? Prior work has consistently found that collections of LLM-generated text are less diverse than equivalent collections of human writing - across lexical, syntactic, and semantic dimensions - spawning the hypothesis that, across many domains, LLMs suffer from a kind of creative 'mode collapse', and operate in a narrower space of ideas than humans. This gives rise to a concerning prognosis of human-AI collaboration: even as LLMs increase individual productivity and perceived creativity, used collectively they collapse individual viewpoints into a homogenized group-think reflective of the LLM's own worldview. Studies in this field, however, are themselves rather homogenous: most measurements of 'semantic diversity' use variations of a single metric with little empirical validation. In this paper, we develop a suite of metrics and novel datasets to benchmark the corpus-level diversity of language models. In striking contrast to previous work, we find that any measurable homogenization relative to human corpora disappears when LLMs are given even a small amount of context. This suggests that LLMs' difficulties generating diverse creative trajectories emerge from a 'cold-start problem'. Just as diverse human authors begin their stories from varied perspectives, giving LLMs even a random mimicry of this diversity can mitigate homogenization.

1 Introduction

The science fiction writer Ted Chiang has argued that LLMs are fundamentally incompatible with creativity. Creativity, he suggests, is equivalent to *choices*. Writing a ten-thousand-word short story requires at least ten-thousand choices; a hundred-word prompt can express only a fraction of this. "If an A.I. generates a ten-thousand-word story based on your prompt, it has to fill in for all of the choices that you are not making." How? This is a rather deep question. What do LLMs do when the linguistic trajectories they are prompted to generate are under-specified? The default, Chiang writes, is to "to take an average of the choices that other writers have made, as represented by text found on the Internet; that average is equivalent to the least interesting choices possible, which is why A.I.-generated text is often really bland."[4]

This is the sort of 'folk theory' about LLMs that feels intuitive to the casual user yet suspicious to experts. Indeed, computer scientists at our institution have argued against this idea of the 'average' choice on theoretical grounds. The cross-entropy loss used in causal language modeling decomposes to a KL-Divergence term between the LM's predicted distribution and the true distribution plus the entropy of natural language. At least at the token level, LLMs are trained to model exactly the

distribution of true human language – uncertainty and all.

$$\mathbb{E}_{p_i(w)} \log \frac{1}{q_i(w)} = \mathbb{E}_{p_i(w)} \log \frac{1}{p_i(w)} + \mathbb{E}_{p_i(w)} \log \frac{p_i(w)}{q_i(w)}$$

Yet, what of the *trajectory* level? Given a story prompt which could be completed in many plausible ways, does the LLM generate precisely the distribution of possible completions in human text? Answering this involves both the vagaries of sampling and some understanding of how LLMs manage to 'plan ahead' during their generation at all. Theoretically, our understanding here is quite dim. [22]

Indeed, even empirically, this is a challenging area of study. LLM benchmarks, by nature, measure the model's responses to fully-specified questions. To study LLM performance in the face of uncertainty, one must move from judging single outputs to distributions of outputs. This is the tact taken by most prior studies of LLM homogenization: define corpus-level metrics of stylistic and semantic diversity. And while strong evidence has emerged for LLMs' stylistic and lexical homogenization, the metrics used by literature to measure semantic homogenization are themselves rather homogenous – consisting of re-derivations of an 'embedding dispersion' metric that can be heavily biased by changes in style.

This paper sets out to empirically answer the question: *who's right?* - the science fiction writer, or the computer scientists? Do LLMs, in the absence of creative specification, give homogenous outputs? Is this true across a comprehensive array of metrics, lexical and semantic? And can anything be done to mitigate this loss of ideological diversity?

To study this, we begin from the *best-case scenario* for language model creativity: pure story completions from diverse human inputs. This is exactly the setting of causal language modeling, and sidesteps any problems of human-computer interaction. And by increasing the amount of context given to the model (e.g. the number of 'choices' supplied by a human), we can measure the extent to which this increases the diversity of model text.

In striking contrast to previous work, we find that LLM homogenization – whether lexical, stylistic, and semantic – disappears when the LLMs are given even a small amount of context. Given the beginning of an author's creative trajectory, our metrics suggest that LLMs can complete the trajectory with as much diversity human authors. We even find that supplying LLMs with a source of randomness increases output diversity as effectively as giving them human-written context. If, as Ted Chiang put it, creativity consists of choices, LLMs prompted with enough (possibly random) choices to escape the 'cold start' of their initial conditions do not make 'average' choices, and appear as capable of following the plot to a creative conclusion as humans.

2 Background

Within the topic of LLM homogenization of writing is a thicket of intertwingled research questions. First, what do we even mean by 'homogenization of writing'? What is the relationship between homogenization, diversity, and creativity? And how can this be measured? Beyond this is the question of context: does one care about whether using LLMs as writing assistants exerts a homogenizing influence on unwitting human authors? Or do we care about the models' intrinsic capacity for creativity, viewing the diversity of text they generate as something of a benchmark? And – more broadly yet – if LLMs seem to produce homogenous outputs, *why*? What components of their training, fine-tuning and sampling procedures might be at fault, and how could they be improved?

It speaks to the severity of the problem of homogenization that, across any collection of metrics and contexts one might care to name, the literature answers the question *Do LLMs homogenize?* affirmatively: they homogenize when used as writing assistants, and also when writing unaided; they homogenize lexically, syntactically, and semantically. And though some factors have been observed which influence this homogenization (particularly instruction tuning), none fully explain it – and there is no remedy.

First, what is creativity in creative writing? In *"The Standard Definition of Creativity"*, Runco and and Jaeger [16] describe creativity using the notions of originality and effectiveness. Originality refers to how novel or unique an idea is. Effectiveness relates to the idea's usefulness, appropriateness, or value within a certain context. Originality is not sufficient for something to be considered creative; an

idea must also fit the context providing some form of value. Without effectiveness, something may simply be random or nonsensical rather than truly creative.

If creativity is seen as stemming from the combination of *diversity* and *constraint*, it's notable that most studies only try to measure the diversity of a text or corpus of texts. As Guo, Shang, and Clavel [7] argues, modern language models are inherently constrained. Barring radical interventions, LLMs reliably generate coherent text – it's what they're trained to do. Hence, measuring the diversity of LLM generated may suffice to measure their creativity. Homogeneity is diversity's opposite. LLM homogenization of writing, then, describes a decline in diversity caused by the introduction of LLMs to some stage of the writing process.

2.1 Measuring Diversity in Writing: Style vs Semantics and the ubiquity of Embedding Dispersion

How can the diversity of a writing corpus be measured? Tevet and Berant [21] propose a hierarchical decomposition of diversity, beginning with a high-level separation of *style* (or form) from *semantics*. Within stylistic diversity, one can further consider lexical and syntactic diversity. Overlapping partially with this, Shaib et al. [18]'s taxonomy of text diversity scores distinguishes metrics based on text compression/analysis of repeated N-grams (like Type-Token Ratios, Unique-N, and simply applying 'gzip') from those based on pairwise similarity metrics (like ROUGE-L or BERTScore) whose average value across a corpus gives what they call a 'Homogenization Score':

$$\text{HS}(Z) = \frac{1}{N} \sum \text{sim}(z_i, z_j) \tag{1}$$

Where Z is a matrix of embeddings for a corpus of N texts, and z_i is the embedding of text i.

Most researchers fuse these approaches, adopting some kind of Unique-N score as a measure of lexical diversity and an embedding-based pairwise similarity score as a measure of semantic diversity. Often the embedding model is BERT, with the assumption that the cosine similarity between BERT embeddings describes the semantic similarity of the underlying text. This similarity, averaged across a corpus, describes what [7] calls 'Embedding Dispersion' - which is also used by [12, 6, 1]; [9] also use slightly different formulations of cosine similarities between embeddings.

$$\text{ED}(Z) = 1 - \frac{1}{N} \sum_{i \neq j} \frac{z_i \cdot z_k}{\|z_i\| \|z_j\|} \tag{2}$$

These embedding approaches are easy to work with, but not without problems. In [18]'s comparison of text diversity metrics, the BERTScore-based embedding dispersion was the only metric unable to distinguish between human and AI-generated corpora. Moreover, studies have found text embedding models to be highly sensitive to lexical and syntactic variation – even to the exclusion of semantic content [Cite SCIPIO]. This suggests that embedding-based approaches are better viewed as 'global' or 'information theoretic' measures of diversity, involving some unknown combination of lexical, syntactic, and semantic features. This aligns embeddings more with text-compression methods, especially neural compression algorithms like CMIX or PPMI, which lower-bound the Kolmogorov Complexity of an input corpus.

To isolate the semantic component, special care must be taken. [12] present one such means of isolating semantic content: using an LLM to convert texts in a corpus into key point summaries, thus capturing only semantic variation - as each key point summary is written in a single LLM's voice. They perform agglomerative clustering on the BERT embeddings of key points, and count the number of clusters in each corpus as a proxy for diversity. A related approach is [9]'s "Semantic Entropy", which performs clustering over sentence [CHECK] embeddings from each corpus to identify topics, then measures the entropy of each corpus over the set of topics.

The lack of any comprehensive and standard evaluation framework for linguistic – and especially semantic – diversity in model-generated text might simply reflect the relatively recent entrance of LLMs into NLP [7]. Before models were capable of generating coherent text en masse with anything approximating the number of ideas in human-written corpora, NLP occupied itself with problems like the stylometric fingerprinting of (human) writers (with type-token ratios to measure an author's vocabulary preferences, the parsing depth of her sentences); or automated analyses of machine

translation (with comparisons between a machine-generated translation and a reference translation, using BLEU and ROUGE). Current measures of text diversity either adapt these metrics developed for different tasks, or (as with the homogenization score and semantic entropy) invent their own, and take on the burden of demonstrating that they represent meaningful features.

2.2 Homogenization as an HCI Problem

LLMs are increasingly being employed as writing 'copilots', from transformer-powered predictive autocomplete in Gmail and Apple-Intelligence equipped writing apps, to students' widespread use of ChatGPT as a brainstorming and drafting companion. Thus the question of homogenization in the context of human-computer interaction: does an author's use of LLMs lead to more homogenous creative outputs?

Padmakumar and He [12] test the affect of a predictive autocomplete interface on writing diversity. Crowdworkers who completed writing prompts while viewing autocomplete suggestions from InstructGPT produced more semantically homogenous texts than the control, as measured by embedding dispersion and the key-point count described above. Curiously, they observed no statistically-significant homogenization for crowdworkers with GPT-3-powered suggestions, even though the writers accepted the model's suggestions just as often. Unfortunately, their sample size was quite small (only 10 texts per topic), but the finding replicates earlier work that "predictive text encourages predictable writing" [3].

Predictive text might seem an especially intrusive form form of human-computer collaboration. Doshi and Hauser [6] had crowdworkers write 8-sentence 'microstories' with or without the ability to prompt an LLM for ideas; the resulting stories were rated as more creative and better-written by readers, even as the corpus of stories became more homogenous (as measured by embedding dispersion). Doshi et al. hypothesize that generative AI might provide ideas that increase individual creativity – but, as it provides similar ideas to everyone, decrease group creativity. Dell'Acqua et al. [5] found a similar pattern in LLM use by 758 employees at the Boston Consulting Group. Those randomly assigned to use GPT-4 were more productive and produced higher-rated work, but collectively produced ideas with less conceptual variation than those without LLMs. Sarkar [17] summarizes this field with McLuhanesque flair: "convergence is the message of [the medium] of AI".

Despite this consistent and robust evidence that use of LLMs in creative work reduces the diversity of ideas, an advocate of LLMs may counter that this is an HCI problem, not a model-engineering problem. We might be using it wrong! A thesaurus, misapplied, could homogenize writing styles; perhaps authors are improperly using chatbots, and with the right prompting philosophy, might they as easily serve as a fount of ideas? Designing cognitively-ergonomic HCI interfaces requires extensive knowledge of human psychology. The HCI researcher Michael Bernstein has compared the present state of human-computer interfaces to reading a textbook that was badly and erratically highlighted by its previous owner. It's possible to share information in a way that's worse than useless – what we might term the 'reverse centaur'.

2.3 Homogenization as a Benchmark

What, then, if we let the machines write by themselves? Most accept that LLMs can produce individual pieces of writing whose style and creativity rivals most human writers. But what about collections of writings? On the corpus level, can LLMs produce as many and varied ideas as a collection of human writers? – or is their apparent creativity in limited ouputs a form of 'mode collapse'?

Guo, Shang, and Clavel [7] give most comprehensive study of LLM corpus-level linguistic diversity to date. They measure LLMs' lexical (with Unique-N), syntactic, and semantic (with Embedding Dispersion) diversity across a range of datasets and model configurations, finding evidence for each type of diversity across all experimental conditions. Most relevantly, they used an older version of the Reddit Writing Prompts dataset we employ, where LLM completions of writing prompts are compared to human authors.

If writing is hard, how about summarizing? Shaib et al. [18] find that across an even wider range of metric (including ROUGE and compression-based information scores), LLM *summaries* of a corpus of human text were less diverse than either human summaries of the same, or the first three sentences

of the texts. This is particularly surprising if Shaib's reflect semantic diversity, as presumably summarizing texts only involves translating ideas from the source material, not creating them de novo.

Perhaps comparing a single model to a collection of humans is unfair. Yet Mohammadi [11] find that the outputs of distinct LLMs are more similar to each other than an equivalent human-to-human comparison.

2.4 What's causing homogenization?

Padmakumar and He [12] find that only writing suggestions from InstructGPT and not GPT-3 resulted in homogenization, suggesting that instruction-tuning resulted in some collapse of model creativity. [11] support this, finding that instruction-tuned models have lower output diversity than base models, as measured through the entropy of next-token predictions and the number of clusters formed in embedding space (during persona generation and fact retrieval tasks). They also find evidence of 'attractor states' by perturbing generation trajectories. However Guo, Shang, and Clavel [7] explicitly compare base and instruction-tuned models and find that, while instruction-tuned models have lower output diversity, even base models induce homogenization relative to humans. Instruction tuning worsens homogenization but does not explain it.

The obvious knob to adjust in search of more creative language model outputs is temperature. Yet, counterintuitively, Guo, Shang, and Clavel [7] find that temperature has no effect on semantic diversity, though it does influence syntactic and lexical diversity. Likewise, while lexical diversity increases with model size (ranging from 1.5B to 32B parameters), semantic diversity quickly plateaus. It's unclear if this merely reflects shortcomings in our ability to measure semantic diversity (through their embedding dispersion metric).

Notably absent from this literature is much exploration of prompting strategies - a gap we hope to remedy.

3 Method

3.1 Corpus Creation

Data scraping. The primary mode of text we explore in this study is short story prose. We source text from Reddit.com, specifically the subreddits r/shortstories [14] and r/WritingPrompts [15], with posts appearing in the order on the 'Top' setting. From r/WritingPrompts, we extract 100 prompt posts along with up to 10 of their root-level comments, treating these comments as human-written completions. This dataset supports our analysis of multiple human completions per prompt. From r/shortstories, we collect 100 standalone narrative texts, which we use to evaluate global stylistic and structural similarity between human and model-generated stories.

Dataset cleaning. For both datasets, we apply filtering on the lengths of human written stories. We only keep stories that are longer than 500 words and shorter than 2,000 words. Most of the stories in the unfiltered datasets fall within this range. We control the length of stories because diversity metrics are shown to be correlated with text lengths, and our initial experiments show that LLMs usually cannot output very long texts. For the writing prompts dataset, if there are more than 10 human stories corresponding to a writing prompt, we only keep the top 10 stories with the most up-votes so that the number of stories per prompt do not vary too much while ensuring the quality of the human-written stories. Additional details on dataset cleaning and dataset statistics are in Appendix A.

Model completion generations. Unless stated otherwise, we generate model completions with fixed temperature of 0.8 and top p value of 1 with basic system prompt. Detailed experimental setup and prompts are in Appendix B.

3.2 Metrics on Homogenization

Homogenization in texts is measured through many different dimensions. We use a set of metrics to analyze homogenization holistically. We organize them into three categories detailed below.

3.2.1 Stylometric Homogenization

Stylometry attempts to capture a writer's linguistic fingerprint: their vocabulary preferences, and lexical, grammatical, and syntactic quirks which can identify an author's writing independently from the subject. Standard stylometric features include the identification of commonly-repeated n-grams (e.g. signature phrases), and linguistic patterns like the frequency of parts of speech and types of punctuation and the parsing depth of sentences.

To convert these metrics from fingerprints of single authors to measures of corpus diversity, we follow Shaib et al. [19] and Guo, Shang, and Clavel [8] in generalizing n-gram identification to a 'Unique-N' score: the fraction of repeated n-grams to total n-grams in a corpus. We also measure diversity along other stylometric axes by taking the variance across each feature dimension over a given corpus, and averaging over the dimensions.

3.2.2 Semantic Homogenization

We follow Shaib et al. [19] and Guo, Shang, and Clavel [8] in computing the **Embedding Dispersion** metric of average cosine similarity between embeddings of texts in a corpus 2. The higher the embedding dispersion of a corpus, the more dissimilar the embedding method judges its texts. To the extent that neural embedding methods capture semantic meaning over stylistic information, this indicates idea diversity.

Any story of nontrivial length contains many ideas, from sentence-by-sentence narrative details, to plot points, to some high-level moral representing the 'gist' of the story. To attempt to capture these, we perform the embedding dispersion analysis at multiple scales: sentence-level, paragraph-level, and text-level. For robustness, we also use three embedding methods: mpnet [20], all-MiniLM-L6-v2 [24], BGE [25], and E5 [23] citations.

Unfortunately, embedding methods are sensitive to style as well as substance. A high embedding dispersion might, in isolation, merely indicate a uniformity in writing style. Using multiscale embeddings provides one way of controlling for this: any stylistic influences on the embeddings should be present on all scales, starting from the sentence level. Changes in embedding dispersion between scales are then driven by purely semantic information.

3.2.3 Sentiment Homogenization

The sentiment of the story is another dimension outside of the style and the events that happen in the story. We run sentiment analysis human-written stories and LLM generated stories and compare their distributions. We use VADER, a lexicon and rule-based sentiment analysis tool. We treat each story as a text unit to compute the sentiment scores.

Metric Name	Stylometric	Semantic	Sentiment
Embedding Dispersion	?	?	?
Dispersion of Key Point Embeddings		✓	
Key-Point Cluster Count		✓	
Unique-N	✓		
Type-Token Ratio	✓		
Dislegomena Ratio	✓		
Part-of-Speech Ratios	✓		
Sentence Parsing Depth	✓		
VADER			✓

Table 1: Sampling of metrics, labeled by level of homogenization

4 Results

4.1 Stylometric Homogenization

Tables 2 and 3 show the Unique-N scores and Stylometric Feature Variance of humans and models on our Reddit Writing Prompts and Short Stories datasets.

For Writing Prompts, the difference is stark: the human texts score significantly higher diversities than any model on every stylistic metric. This replicates [8]'s findings of lexical and syntactic homogeneity in model completions of an older version of the Reddit Writing Prompts dataset.

Curiously, this pattern doesn't hold in the short stories dataset: here the human texts still have among the highest Unique-N scores (though the differences are not as stark), but the *lowest* stylometric feature variation of any model.

First, we note that the stylometric diversity scores of human writers are significantly higher in the Writing Prompts dataset than the Short Stories dataset. This may indicate that the former has a more diverse or higher-caliber community of writers.

Also observe that the models in the Writing Prompts dataset were given *context* about the human writers - by default, 50% of their story fed to the model as a prompt, as compared to a few sentences in the Writing Prompts dataset. To the degree models can modulate their default style to mimic prompted text, we should expect models in the Short Story condition to express higher lexical and syntactic diversity - which, though not evident in Unique-N, is consistent with a dramatic increase in stylometric feature variance. (Indeed, if anything, the models appear to use less diverse vocabulary and more cliched language in the Short Stories dataset, mimicking the lower diversity of the human authors.)

Table 2: Writing Prompts Stylometric Diversity Metrics

Model	Unique-1	Unique-2	Unique-3	Style Var.
human	**0.3566**	**0.8465**	**0.9784**	**2.1138**
gpt-4o	0.3080	0.7320	0.8925	0.4674
gpt-35-turbo-16k	0.2865	0.6917	0.8676	0.3379
Meta-Llama-3-1-70B-Instruct-htzs	0.2632	0.6297	0.8083	0.4821
Mistral-large-ygkys	0.2564	0.6378	0.8182	0.5036

Table 3: Short Stories Stylometric Diversity Metrics

Model	Unique-1	Unique-2	Unique-3	Style Var.
human	0.2301	**0.7375**	**0.9601**	1.4889
gpt-4o	**0.2307**	0.7189	0.9399	2.1750
gpt-35-turbo-16k	0.2062	0.6618	0.9008	1.6711
Meta-Llama-3-1-70B-Instruct-htzs	0.2220	0.6722	0.9019	**2.4032**
Mistral-large-ygkys	0.2145	0.6789	0.9090	1.9835

4.2 Semantic Homogenization

As described, we compute the embedding dispersion between embeddings of texts in a corpus: namely, given a fixed writing prompt, we compute the *intra*-human and *intra*-model similarity for their respective responses to the prompt. In Figure 1, we visualize the distribution of cosine similarity scores for multiple human-authored completions and various LLM-generated completions, using MiniLM embeddings. The human-authored responses exhibit a notably lower average cosine similarity, indicative of greater semantic diversity and broader narrative exploration within the responses to the same prompt. Conversely, completions produced by LLMs demonstrate significantly higher internal semantic similarity, suggesting that model-generated texts are more semantically constrained and tend toward narrower, more predictable narrative cluster. This finding underscores a potential homogenization effect in language models, as their outputs remain confined to more limited regions of the semantic embedding space compared to human creativity.

An important caveat is that the MiniLM model used for these embeddings has a max input length of 256 tokens, and any input exceeding this limit is truncated. This truncation could potentially omit significant portions of longer responses, possibly affected the similarity measurements. To assess the impact of this limitation, we also conducted analyses using BGE and E5 embedding models, which have a maximum input length of 512 tokens. These models produce higher-dimensional embeddings (BGE: 1024 dimensions, E5: 1024 dimensions) compared to MiniLM's 384 dimensions.

Interestingly, while the overall trend of higher intra-model similarity compared to intra-human similarity persists across these models, the absolute similarity values are notably higher. (Figures 15, **??**). This observation suggests that higher-dimensional embeddings may inherently yield higher cosine similarity scores, though the exact interplay between embedding dimensionality and similarity metrics remains an open question. Further investigation is warranted to disentangle the effects of embedding dimensionality from genuine semantic similarity in such analyses.

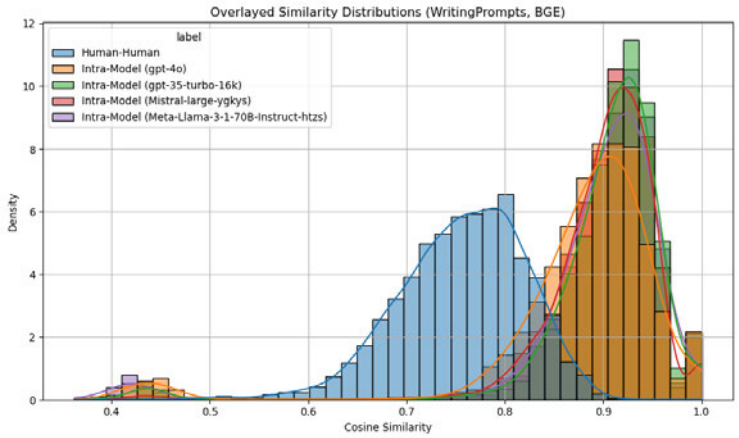

Figure 1: Distribution of intra-prompt semantic similarity among human-written responses (blue) and various LLM-generated completions (orange, green, red, purple) for the WritingPrompts dataset, measured using BGE embeddings. Human completions exhibit lower average cosine similarity, indicating greater semantic diversity compared to the model completions, which demonstrate higher internal semantic homogenization.

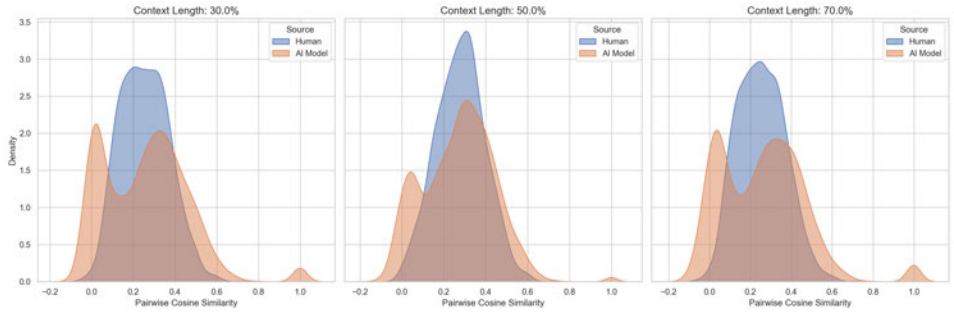

Figure 2: Distribution of cosine similarities between BGE text embeddings of human vs model ShortStories corpora. The LLM corpora vary in the percentage of the story they complete: 30% -70%.

But while the embedding dispersion clearly separates the human and model completions of WritingPrompts, this is markedly false of our second dataset, ShortStories. Here, the default condition of LLMs completing 50% of the text shows complete overlap between the model distribution and human distribution. This accords with the previous results that stylometric features do not robustly distinguish LLMs from humans in ShortStories: LLMs given the extra context of ShortStories no longer homogenize stylistically. This now suggests *either* that the LLMs given the context of ShortStories don't homogenize semantically, or that the embedding dispersion is failing to capture semantic information.

4.2.1 Isolating Semantic Diversity with Key-Point Summaries

Another means of separating style from semantics in embeddings is to have a language model rewrite each text. [13] prompted an LLM to summarize texts from their corpora into key points, then computed embeddings from those key points, and performed agglomerative clustering on the embeddings to approximate the number of unique key points per corpus.

We performed the same experiment with the Reddit Short Stories dataset. We prompted GPT-4o to summarize each story into key points, embedded these key points with BGE, and performed Louvain and DBSCAN clusterings over these embeddings with a variety of powers. Figure 3 shows the results for Louvain; the results for DBSCAN are similar. Both show wild variance in the number of clusters between parameter choices, with no robust discrepancy between the number of clusters for human vs model corpora.

We then performed our embedding dispersion analysis on the key-point embeddings, to assess whether the key point summaries of either corpora were significantly more self-similar. Here again, there is no difference between the model and human key-point summaries.

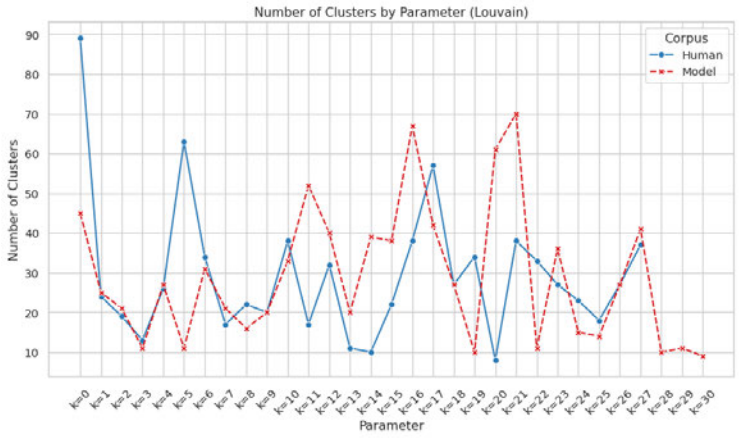

Figure 3: Numbers of clusters discovered by Louvain on a k-nearest neighbors graph for different k, for the model and human corpora from the Reddit Short Stories dataset.

This negative result is in contrast to [13]'s finding that LLM-generated corpora had fewer key-point clusters. We note that the Short Story setting is a best-case scenario for model creativity, involving substantial human text in the prompt as a strong, semantically-diverse basepoint from which the model generations can begin. The lack of key-point separation suggests that the model completions of human stories didn't supply semantically similar endings to the human-written beginnings – or, if they did, that the key-point summaries and embeddings operated on too coarse a level to notice it.

4.3 Sentiment Homogenization

Figure 11 plot the sentiment score distribution of human-written short stories vs. LLM generated short stories in the short stories dataset (top) and the writing prompt dataset (bottom). The sentiment scores $s \in [-1, 1]$, with scores $s > 0.05$ indicating positive sentiment, scores $s < -0.05$ indicating negative sentiment, and scores $s \in [-0.05, 0.05]$ indicating neutral sentiment. We observe that although the majority of human stories have positive sentiment, approximately 30% of human-written stories have negative sentiment. In contrast, the stories generated by LLMs are more skewed towards having positive sentiment. This is true across both datasets and across all different LLMs.

5 Methods to Decrease Homogenization

Our results in Section 4.1 show that LLM outputs in the Short Stories dataset have stylistic diversity comparable to that of human writings. This observation prompts us to hypothesize that providing longer writing context in input help improve model output diversity. We want to further study how much context is needed in the prompt to make LLMs' writings as diverse as human writings, and what kind of context is required. We run the following two sets of experiments to study these.

5.1 Varying Context Length in Input

We generate model outputs with the short stories dataset by providing varying amount of the human written stories in the prompt to study how much story context is sufficient to prompt diverse model

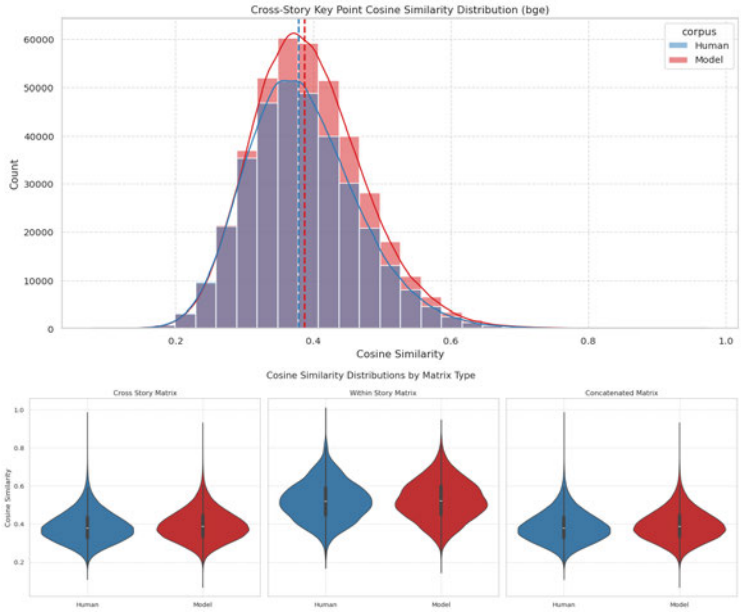

Figure 4: Top: Cosine Similarities of model and human key point summaries, embedded with BGE. Bottom: Violin plots of the distribution of cosine similarities between embeddings of key-point summaries of the Reddit Short Stories dataset. Negligible difference is observed between the model and human key points, whether comparing sentence level key point embeddings across stories (Cross-Story Matrix), sentence-level embeddings within stories (Within-Story Matrix) or concatenating sentence-level key-point embeddings for each story and comparing between stories (Concatenated Matrix).

outputs. Our base experiments generate model completions using the first 50% of human-written stories. We generate two additional sets of model completions with the first 30% and 70% of human-written stories.

Table 4 and 5 show results on stylometrics on 30% and 70% cut lengths respectively. We observe that cut length variations between 30% to 70% do not seem to have significant impact on stylometric diversity. Cut length also does not appear to significantly affect semantic diversity. Figures regarding this can be found in

Table 4: Short Stories Cut Length 30% Stylometric Diversity Metrics

Model	Unique-1	Unique-2	Unique-3	Style Var.
gpt-4o	0.2213 (-0.0094)	0.6969 (-0.0220)	0.9184 (-0.0215)	2.5142 (+0.3392)
gpt-35-turbo-16k	0.2106 (+0.0044)	0.6664 (+0.0046)	0.8951 (-0.0068)	2.5684 (+0.1652)
Meta-Llama-3-1-70B-Instruct-htzs	0.2018 (-0.0137)	0.6321 (-0.0468)	0.8706 (-0.0384)	3.5268 (+1.5433)

Table 5: Short Stories Cut Length 70% Stylometric Diversity Metrics

Model	Unique-1	Unique-2	Unique-3	Style Var.
gpt-4o	0.2734 (+0.0427)	0.7535 (+0.0356)	0.9299 (-0.0100)	3.2672 (+1.07783)
gpt-35-turbo-16k	0.2216 (+0.0154)	0.6709 (-0.0091)	0.8953 (-0.0066)	2.6994 (+1.0283)
Meta-Llama-3-1-70B-Instruct-htzs	0.2319 (+0.0099)	0.6611 (-0.0111)	0.8726 (-0.0293)	3.8679 (+1.8844)

Figure 5: **Sentiment score distribution on human-written short stories vs LLM generated short stories.** LLM generated stories sentiment scores are skewed more towards positive sentiments.

5.2 Injecting Randomness in Input

Another method we explore is including random words for inspiration in the system prompt. We hypothesize that this will improve output diversity because we are injecting randomness into the generation process. In this experiment, we use *google-10000-english-no-swears* word list from google-10000-english [10]. We do POS tagging on the words and only keep words in the noun, adjective, adverb, and verb categories to form our word list. The reason is that words in other categories, such as prepositions, may not contain enough information, and words from the above categories should provide sufficient randomness. For each LLM generation, we randomly sample 5 words, and append them to the prompt "here is a list of random words to take inspiration from". We run the experiment on the Writing Prompt dataset.

Table 6 shows stylometrics on the new dataset. We observe that while the LLM generated outputs are still less diverse measured by these metrics compared to the human-written stories, the diversity scores improve for all models across all metrics. This suggests that injecting randomness into system prompt does help improve output stylometric diversity.

Table 6: Writing Prompts with Random Words Stylometric Diversity Metrics

Model	Unique-1	Unique-2	Unique-3	Style Var.
human	**0.3566**	**0.8465**	**0.9784**	**2.1138**
gpt-4o	0.3245 (+0.0165)	0.7716 (+0.0394)	0.9333 (+0.0408)	0.5157 (+0.0483)
gpt-35-turbo-16k	0.3077 (+0.0212)	0.7358 (+0.0441)	0.9124 (+0.0448)	0.3795 (+0.0416)
Meta-Llama-3-1-70B-Instruct-htzs	0.2898 (+0.0334)	0.6815 (+0.0437)	0.8621 (+0.0441)	0.9291 (+0.4470)

6 Discussion

The advent of LLMs has both broadened the range of questions worth asking about machines and creativity, and introduced new tools with which we might answer them. Sadly, the tools we have are as yet no match for our questions. Especially to the question of semantic homogenization, the metrics we find in the literature - the refraction of LLM texts through LLM-derived embeddings, possibly funneled through LLM-written key-point summaries - leave it unclear whether the measurement is actually semantic, or how much is lost through the biases of translation. (Recall Anderson, Shah, and Kreminski [2]'s finding that even in summarizing texts, LLMs homogenized more than human summarizers.)

The strongest interpretation of our results is that adding context is sufficient to mitigate homogenization, perhaps by eliminating the homogenizing influence LLM's uniform starting point. Given some context, either the beginning of an author's creative trajectory, or even a source of randomness, LLMs can escape this uniform default condition and generate corpora as diverse as can human authors.

However, this places a lot of faith in the metrics. A more accurate descriptor: adding context is sufficient to reduce homogenization below a level detectable by our metrics. And while the field of stylometry is well-developed, there's a cavity in the literature of corpus-level measures of semantic information. Quite likely the cosine similarity between embeddings measures some combination of stylistic and weakly semantic information. The development and validation of stronger semantic metrics could confirm the extent to which adding context can eliminate ideological homogenization.

7 Limitations and Future Works

The first future direction is to finish running all experiments and computing all current metrics on both the short stories and writing prompts datasets. Due to time constraints, some experiments are only on one of the datasets, but it will be valuable to have results on both.

Although we compute various metrics to try to measure diversity across stylometrics, semantic, and sentiment, it is still unclear how comprehensive they are. We plan to supplement our current metrics with using LLM as a judge or run a human evaluation test. We will present three texts to the LLM or human judge, two human written stories and one LLM story, or the other way around. We will ask the judge to pick out the story that is the least similar to the other two. This method will ideally take all three categories of diversity into account.

We plan to further investigate how much context and what type of context in the prompt is sufficient to make LLM outputs as diverse as human-written short stories. We have results on preliminary exploration of this in Section 5, but we plan to run additional experiments. We plan to use shorter context length to study when LLM outputs become less diverse than human-written stories. Additionally, we plan to study whether including more random words in the prompt alone without providing the actual start of the stories can also make LLM generations as diverse as human-written stories.

8 Contributions

Addison and Kia were our 'scrapers in chief', compiling the 100-1000s of posts that form our Reddit Short Stories and Writing Prompts datasets. Patty was our 'API warrior', collecting and cleaning LLM completions of each dataset, as well as key-point summaries and other experimental conditions. She also fended off passive aggressive inquiries from the overlords of Princeton's AI Sandbox.

Katerina helped refine our questions, performed literature review on the broader creativity research, and drafted parts of the introduction and background. Kincaid wrote up the current 'related work' section of the background, and any faults therein are his own.

Patty and Kincaid developed the stylistic metrics and applied them in experimental settings. Kia developed the initial 'distribution of cosine similarities' approach, and applied it across models to obtain a comprehensive suite of first results, which he refined throughout the semester. Kincaid ran experiments with multiscale embeddings, both with some weird 'Von Neumann Entropy' metrics that didn't end up being used, and on the key-point summaries setup from Padmakumar.

Patty and Menon also explored interventions to mitigate homogenization. Patty explored the effect of giving models access to a source of randomness during generation. Menon ran our experiments on differing cut lengths - the 'Sanjeev' setting - computing embedding dispersions and compression ratios across varying context sizes while controlling for the confounding factor of text length.

All members of the group met for dozens of hours over the semester to collectively make sense of our tangled research questions and each other's sometimes contradictory results.

Acknowledgments and Disclosure of Funding

The authors acknowledge that *Machine Behavior* was a stellar class and beacon of light in their academic lives. Those authors with multiple projects from other courses acknowledge that this Homogenization project was their unconditional favorite. They acknowledge the inestimable help provided by Manoel in guiding this project, providing the initial idea of completions across varying cut-lengths as an experimental setting, and challenging our early results with follow-up experiments like the random word modulation. Some of the authors also acknowledge helpful conversations about the research with Peter Henderson and Ben Loufer. Finally, the authors gratefully acknowledge the doors of the Princeton Friend Center, which never failed to hold themselves open.

References

[1] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. "Evaluating Creativity Support Tools via Homogenization Analysis". In: CHI EA '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–7. ISBN: 979-8-4007-0331-7. DOI: 10.1145/3613905.3651088. (Visited on 03/03/2025) (cit. on p. 3).

[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. "Homogenization Effects of Large Language Models on Human Creative Ideation". In: C&C '24. New York, NY, USA: Association for Computing Machinery, June 2024, pp. 413–425. ISBN: 9798400704857. DOI: 10.1145/3635636.3656204. URL: https://dl.acm.org/doi/10.1145/3635636.3656204 (visited on 03/03/2025) (cit. on p. 12).

[3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. "Predictive Text Encourages Predictable Writing". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 128–138. ISBN: 978-1-4503-7118-6. DOI: 10.1145/3377325.3377523. (Visited on 05/06/2025) (cit. on p. 4).

[4] ed Chiang. *Why A.I. Isn't Going to Make Art*. https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art. (Visited on 05/06/2025) (cit. on p. 1).

[5] Fabrizio Dell'Acqua et al. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4573321. (Visited on 05/06/2025) (cit. on p. 4).

[6] Anil R. Doshi and Oliver P. Hauser. "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content". In: *Science Advances* 10.28 (July 2024), eadn5290. DOI: 10.1126/sciadv.adn5290. (Visited on 05/06/2025) (cit. on pp. 3, 4).

[7] Yanzhu Guo, Guokan Shang, and Chloé Clavel. *Benchmarking Linguistic Diversity of Large Language Models*. Dec. 2024. DOI: 10.48550/arXiv.2412.10271. arXiv: 2412.10271 [cs]. (Visited on 04/29/2025) (cit. on pp. 3–5).

[8] Yanzhu Guo, Guokan Shang, and Chloé Clavel. *Benchmarking Linguistic Diversity of Large Language Models*. arXiv:2412.10271 [cs] version: 1. Dec. 2024. DOI: 10.48550/arXiv.2412.10271. URL: http://arxiv.org/abs/2412.10271 (visited on 04/29/2025) (cit. on pp. 6, 7).

[9] Seungju Han, Beomsu Kim, and Buru Chang. "Measuring and Improving Semantic Diversity of Dialogue Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 934–950. DOI: 10.18653/v1/2022.findings-emnlp.66. (Visited on 04/29/2025) (cit. on p. 3).

[10] Josh Kaufman. *google-10000-english*. https://github.com/first20hours/google-10000-english. GitHub repository. 2021 (cit. on p. 11).

[11] Behnam Mohammadi. *Creativity Has Left the Chat: The Price of Debiasing Language Models*. June 2024. DOI: 10.48550/arXiv.2406.05587. arXiv: 2406.05587 [cs]. (Visited on 03/03/2025) (cit. on p. 5).

[12] Vishakh Padmakumar and He He. *Does Writing with Language Models Reduce Content Diversity?* July 2024. DOI: 10.48550/arXiv.2309.05196. arXiv: 2309.05196 [cs]. (Visited on 03/03/2025) (cit. on pp. 3–5).

[13] Vishakh Padmakumar and He He. *Does Writing with Language Models Reduce Content Diversity?* arXiv:2309.05196 [cs]. July 2024. DOI: 10.48550/arXiv.2309.05196. URL: http://arxiv.org/abs/2309.05196 (visited on 03/03/2025) (cit. on pp. 8, 9).

[14] *r/shortstories*. URL: https://www.reddit.com/r/shortstories/ (cit. on p. 5).

[15] *r/WritingPrompts*. URL: https://www.reddit.com/r/WritingPrompts/ (cit. on p. 5).

[16] Mark A. Runco and Garrett J. and Jaeger. "The Standard Definition of Creativity". In: *Creativity Research Journal* 24.1 (Jan. 2012), pp. 92–96. ISSN: 1040-0419. DOI: 10.1080/10400419.2012.650092. (Visited on 05/06/2025) (cit. on p. 2).

[17] Advait Sarkar. "Intention Is All You Need". In: () (cit. on p. 4).

[18] Chantal Shaib et al. *Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores*. Mar. 2024. DOI: 10.48550/arXiv.2403.00553. arXiv: 2403.00553 [cs]. (Visited on 03/03/2025) (cit. on pp. 3, 4).

[19] Chantal Shaib et al. *Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores*. 2024. arXiv: 2403.00553 [cs.CL]. URL: https://arxiv.org/abs/2403.00553 (cit. on p. 6).

[20] Kaitao Song et al. *MPNet: Masked and Permuted Pre-training for Language Understanding*. 2020. arXiv: 2004.09297 [cs.CL]. URL: https://arxiv.org/abs/2004.09297 (cit. on p. 6).

[21] Guy Tevet and Jonathan Berant. "Evaluating the Evaluation of Diversity in Natural Language Generation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 326–346. DOI: 10.18653/v1/2021.eacl-main.25. (Visited on 05/04/2025) (cit. on p. 3).

[22] Keyon Vafa et al. *Evaluating the World Model Implicit in a Generative Model*. arXiv:2406.03689 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2406.03689. URL: http://arxiv.org/abs/2406.03689 (visited on 04/30/2025) (cit. on p. 2).

[23] Liang Wang et al. *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. 2024. arXiv: 2212.03533 [cs.CL]. URL: https://arxiv.org/abs/2212.03533 (cit. on p. 6).

[24] Wenhui Wang et al. *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. 2020. arXiv: 2002.10957 [cs.CL]. URL: https://arxiv.org/abs/2002.10957 (cit. on p. 6).

[25] Shitao Xiao et al. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. 2023. arXiv: 2309.07597 [cs.CL] (cit. on p. 6).

A Dataset Cleaning Statistics

Figure 6 plots the word counts of human-written short stories we collect from r/ShortStories and comments in the posts we collect from r/WritingPrompts. The distributions are computed on raw dataset before any cleaning. Most short stories from r/ShortStories are between 500 and 2000 words, which motivates setting that as the word count range of stories we keep. There are many really short comments to posts in r/WritingPrompts and those are likely not short stories, so setting a lower bound of 500 words will filter those out. Additionally, there are some posts in r/WritingPrompts that are not writing prompts, and we filter those out as well. Figure 7 plots the number of comments in each post we collect from r/WritingPrompts before any dataset cleaning. Most posts have below 50 comments, and some of these comments are not short stories. There are a few popular posts with many comments, and some of those are not post on writing prompt. We choose to restrict the maximum number of human-written short stories per writing prompt to 10 most up-voted stories because we want prompts to have relatively similar number of stories and want to stories to be high quality.

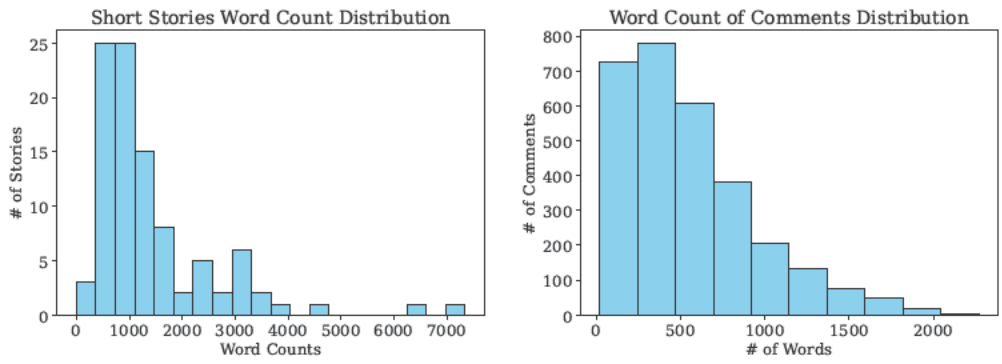

Figure 6: **Human-written stories word count distributions.** Most human written stories in the two datasets we collect are between 500 words and 2000 words.

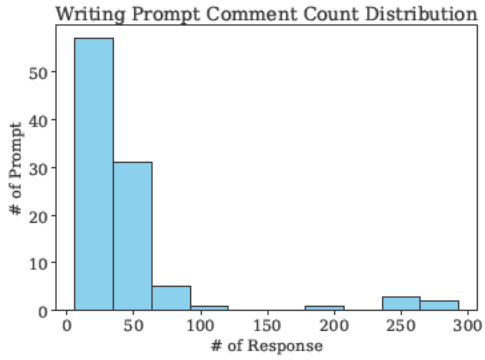

Figure 7: **Number of comments in writing prompt posts.** Most comments have less than 50 responses.

Figure 8 plots the difference in word count between human-written stories and LLM generated stories. The difference is measured on the short stories dataset and calculated by of words in human written story - of words in corresponding LLM generated story. The distribution concentrate around 0 and is relatively symmetrical around 0, which is desirable.

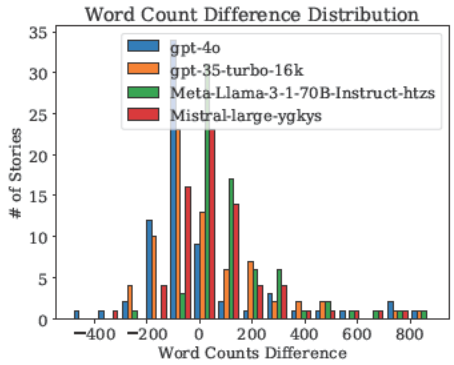

Figure 8: **Difference in word count between human-written stories and LLM generated stories.** The difference is measured on the short stories dataset and calculated by of words in human written story - of words in corresponding LLM generated story.

B Additional Experimental Setups

The base system prompts we use to generate experimental results are shown below. The random words system prompts include an additional sentence: Please use the following words as inspirations: {*list of random words*}.

Base Prompt for Short Stories

You are a creative writing assistant. Complete the following story in a compelling way.

What follows is the first half of a story. Please write the second half.
What you write should be as long as the first half (around {*target word count*} words).

{*first half of the story*}

Base Prompt for Writing Prompts

You are a creative writing assistant. Complete the following story in a compelling way.

What follows is the beginning of a story. Please complete the rest of the story with around {*target word count*} words.

{*writing prompt*}

C Additional Experimental Results

C.1 Semantic Homogenization

Figure 15 and 10 show additional results on cosine similarity metrics on the writing prompt dataset with other embedding models.

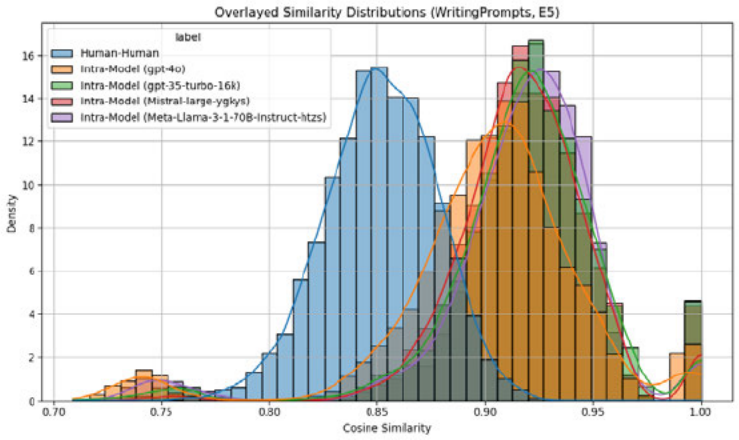

Figure 9: Distribution of intra-prompt semantic similarity among human-written responses (blue) and various LLM-generated completions (orange, green, red, purple) for the WritingPrompts dataset, measured using E5 embeddings.

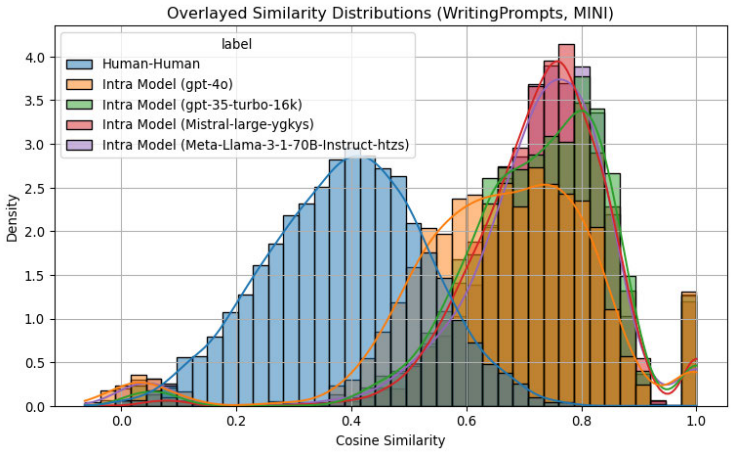

Figure 10: Distribution of intra-prompt semantic similarity among human-written responses (blue) and various LLM-generated completions (orange, green, red, purple) for the WritingPrompts dataset, measured using MiniLM embeddings.

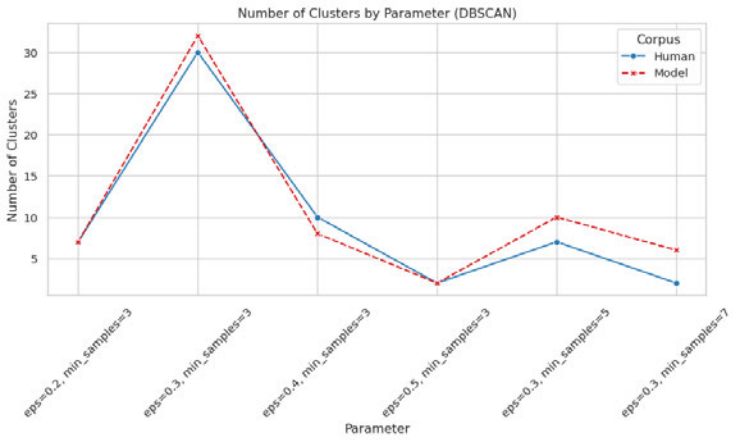

Figure 11: Number of clusters inferred by DBSCAN, across clustering parameters. Note that the number varies widely and shows no clear pattern of model or human dominance.

C.2 Semantic Homogenization across Cut Lengths

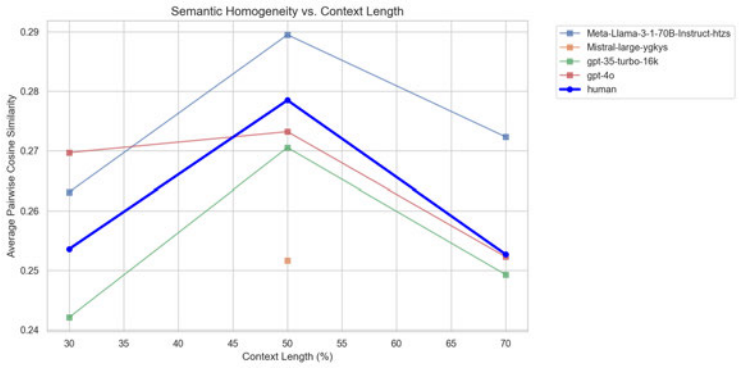

Figure 12: Trends in semantic homogeneity as context length increases for various LLM (light blue, orange, greed, salmon) and human-generated (dark blue) completions for the WritingPrompts dataset.

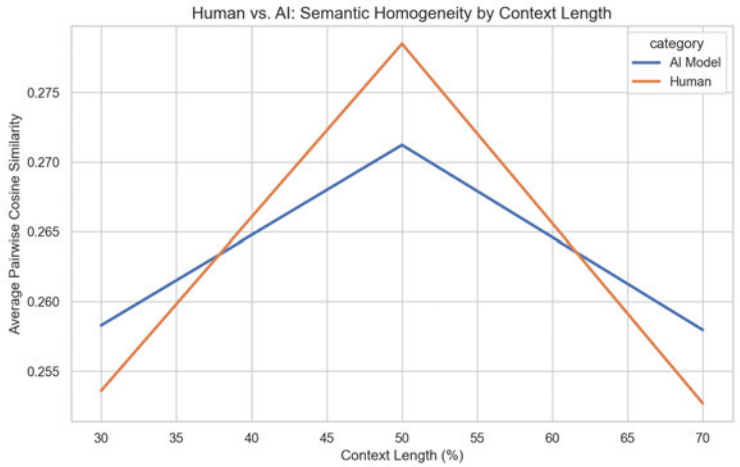

Figure 13: Trends in semantic homogeneity as context length increases for LLM (aggregated) and human-generated completions for the WritingPrompts dataset.

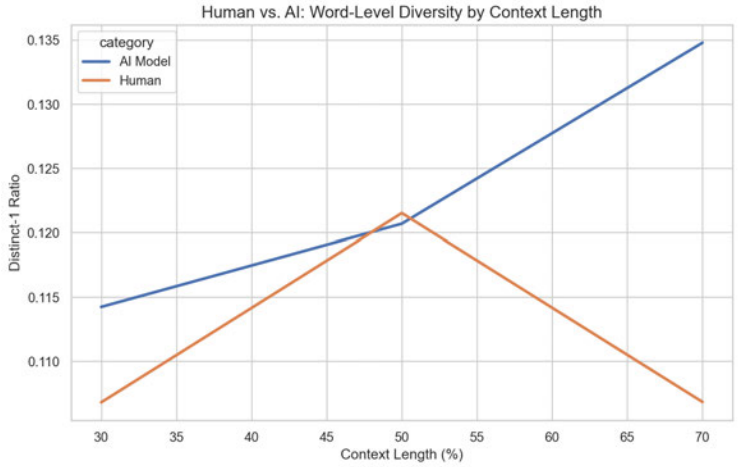

Figure 14: Trends in distinct unigram diversity as context length increases for LLM (aggregated) and human-generated completions for the WritingPrompts dataset.

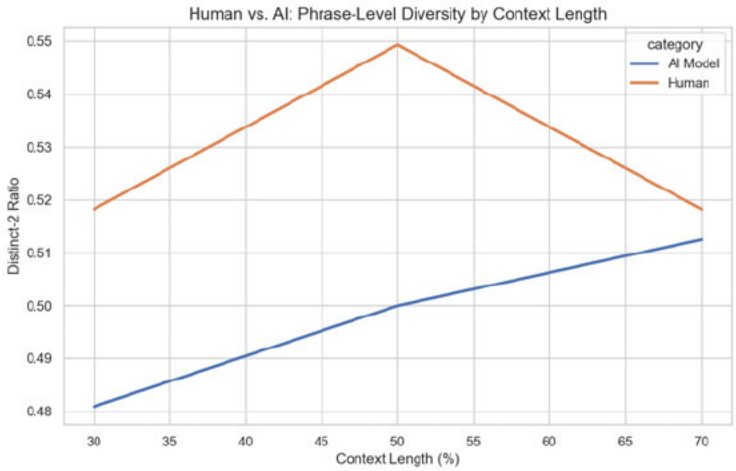

Figure 15: Trends in distinct bigram diversity as context length increases for LLM (aggregated) and human-generated completions for the WritingPrompts dataset.

D Sentiment Diversity on Random Words Dataset

Figure 16 plots sentiment scores distributions across models on the dataset generated with random words in the prompt. We observe that including random words in the prompt did not help improve sentiment diversity in stories. Similar to the stories generated with base prompt, almost all of the LLM-generated stories have positive sentiment.

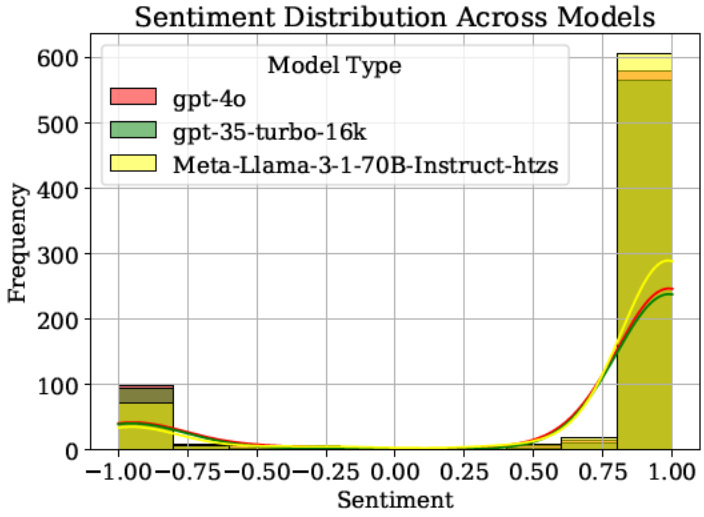

Figure 16: