Benchmarking Linguistic Diversity of Large Language Models

Yanzhu Guo*
ALMAnaCH
Inria Paris, France
yanzhu.guo@inria.fr

Guokan Shang
IFM Paris
MBZUAI, France
guokan.shang@mbzuai.ac.ae

Chloé Clavel
ALMAnaCH
Inria Paris, France
chloe.clavel@inria.fr

Abstract

The development and evaluation of Large Language Models (LLMs) has primarily focused on their task-solving capabilities, with recent models even surpassing human performance in some areas. However, this focus often neglects whether machine-generated language matches the human level of diversity, in terms of vocabulary choice, syntactic construction, and expression of meaning, raising questions about whether the fundamentals of language generation have been fully addressed. This paper emphasizes the importance of examining the preservation of human linguistic richness by language models, given the concerning surge in online content produced or aided by LLMs. We adapt a comprehensive framework for evaluating LLMs from various linguistic diversity perspectives including lexical, syntactic, and semantic dimensions. Using this framework, we benchmark several state-of-the-art LLMs across all diversity dimensions, and conduct an in-depth analysis for syntactic diversity. Finally, we analyze how the design, development, and deployment choices of LLMs impact the linguistic diversity of their outputs, focusing on the creative task of story generation.

1 Introduction

Recent Large Language Models (LLMs) have exhibited outstanding capabilities in generating both natural and formal language (Brown et al., 2020; Touvron et al., 2023), while also achieving human-level performance in language understanding, commonsense reasoning, and various other tasks (Hendrycks et al., 2020). This has led to evaluations that predominantly focus on these specific abilities (Wang et al., 2024). Meanwhile, other evaluation studies address well-recognized issues in LLMs, such as factuality (Maynez et al., 2020), safety (Zhang et al., 2024), and fairness (Gallegos et al., 2024), which remain focal points of ongoing research. However, there is a notable lack of attention paid to linguistic perspectives, particularly in diversity (Guo et al., 2024b), despite the fundamental objective of natural language generation being to produce outputs that are not only *accurate* but also *diverse* (Tevet and Berant, 2021).

Recent studies have highlighted concerns regarding the linguistic diversity of LLM outputs. By comparing human and model-generated content, researchers have shown that models frequently struggle to reflect the nuances and variations found in human expression (Shaib et al., 2024a; Giulianelli et al., 2023). Additionally, these concerns are reinforced by findings that training language models on synthetic text can lead to a further decline in linguistic diversity (Guo et al., 2024b).

In fact, LLMs tend to be inherently conservative in producing diverse content. During training, models undergo homogenization to the most frequent patterns in the training data, where creative outlier narratives, views, styles, and knowledge are often underrepresented (Kandpal et al., 2023). Unlike models, human language production involves a complex interplay of factors that go beyond merely optimizing probabilities (Holtzman et al., 2020). It is therefore crucial to emphasize evaluating output diversity in language models and systematically consider these metrics to guide future model design, development and deployment decisions.

Currently, principled and comprehensive studies on evaluating linguistic diversity are lacking in the literature (Shaib et al., 2024a). While some works on Natural Language Generation (NLG) report diversity metrics, they typically focus on a single diversity aspect (e.g., lexical diversity [Chakrabarty et al., 2022]), often experimenting within a single domain and task (e.g., news summarization [Shaib et al., 2024a]). This narrow focus is problematic since diversity varies across

* This work was partially done during the author's affiliation with École Polytechnique.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1507–1526, 2025. https://doi.org/10.1162/TACL.a.47
Action Editor: Xavier Carreras. Submission batch: 12/2024; Revision batch: 4/2025; Published 11/2025.

aspects and depends on the domain (Guo et al., 2024b). Although some efforts have been made to assess the influence of reinforcement learning from human feedback (RLHF) on diversity (Kirk et al., 2024), the impact of other key design and development stages—such as model scale, quantization, decoding strategy, and prompt formulation—remains unexplored. Additionally, there is a limited understanding of how LLMs develop the capability to generate diverse language through successive pretraining checkpoints. Ultimately, no study has benchmarked the diversity performance of state-of-the-art LLMs across different aspects and domains.

In this work, we first establish a framework for evaluating linguistic diversity of LLM outputs on a corpus level. We then benchmark six prominent LLMs on five NLG tasks, and compare the diversity of their outputs across three different aspects: lexical, syntactic, and semantic. We place particular emphasis on story generation, the most creative task where linguistic diversity plays a crucial role, conducting a deeper analysis in this context. Specifically, we examine syntactic diversity through a case study comparing the distribution of dependency trees in human-written and LLM-generated texts. Finally, we also investigate how LLM output diversity changes across different development stages, and with varying decisions of deployment.

The main research questions we address are as follows:

1. What are the key aspects of LLM output diversity, and how can they be evaluated? (See § 3)
2. How do state-of-the-art LLMs perform in terms of diversity across different tasks? (See § 5)
3. How does diversity change during each LLM development stage (e.g., pretraining, supervised fine-tuning [SFT], preference tuning)? (See § 6.1)
4. How do different design (e.g., model scale, training data) and deployment (e.g., decoding strategy, prompt formulation, quantization) choices affect diversity? (See § 6.2, § 6.3, and § 6.4)

It is worth noting that we study linguistic diversity in a monolingual context, focusing on the English language. However, the evaluation methodology is language agnostic and could easily be extended to other languages, given that employed NLP toolkits (e.g., dependency parsers, sentence embeddings) exist for the language. Furthermore, our approach to analyzing the influence of various factors on LLM outputs is adaptable to other dimensions, such as linguistic naturalness (Guo et al., 2024a).

The code for our research is available at `https://github.com/YanzhuGuo/llm-diversity`.

2 Related Work

In this section, we review methods for evaluating and analyzing linguistic diversity. We define *linguistic diversity* as the natural variation in human language across core linguistic properties, including vocabulary usage, grammatical structures, and semantic nuances. In contrast, a separate line of research focuses on *socio-linguistic diversity* (Hayati et al., 2023; Lahoti et al., 2023), which falls beyond the scope of our study.

2.1 Evaluation of Human Language

Early metrics for linguistic diversity, proposed by linguists, were developed for studies of language acquisition and language disorder detection. For example, Fergadiotis et al. (2013) employed lexical diversity metrics to identify symptoms of aphasia, while McNamara et al. (2010) showed that both syntactic complexity and lexical diversity can predict essay quality. Another study by Clercq and Housen (2017) manually annotated a small corpus of texts produced by second language learners for syntactic features such as syntactic length and clause types, considering their variation as a diversity index. However, these metrics are limited to evaluating human-written texts and either focus exclusively on lexical diversity or lack scalability due to the need for manual annotation.

The evaluation of linguistic diversity in model-generated language has emerged as a relatively recent focus of research. This development is driven, in part, by growing concerns over the increasing online prevalence of model-generated or model-influenced content (Geng and Trotta, 2024), prompting questions about whether LLMs can reflect the linguistic richness characteristic of human language (Guo et al., 2024b). However, assessing linguistic diversity is meaningful *only when the generated text meets basic standards of*

quality. For instance, a randomly initialized model might produce token sequences with high lexical diversity, but such outputs lack any practical value (Uchendu et al., 2023). Recent advances in language generation quality have brought model outputs closer than ever to human-like coherence and plausibility, making the evaluation of linguistic diversity more relevant and necessary than before.

2.2 Evaluation of Generated Language

To the best of our knowledge, Tevet and Berant (2021) were the first authors to systematically evaluate diversity in NLG. They proposed to create diversity metrics from any two-sentence similarity measure, defining diversity as the inverse of the mean similarity score across all unordered pairs. N-gram-based metrics were used to assess form diversity, while model-based metrics like Sentence-BERT similarity measured content diversity. They concluded that a notable disparity exists between automatic metrics and human judgment, and that human evaluation of diversity becomes challenging in sets with more than ten sentences.

Since then, additional metrics have been proposed to capture linguistic diversity, including semantic diversity metrics based on natural language inference (Stasaski and Hearst, 2022) or semantic entropy (Han et al., 2022), and syntactic diversity metrics derived from *n*-grams of Part-of-Speech (POS) tags (Giulianelli et al., 2023) or graph similarity kernels of syntax trees (Guo et al., 2024b).

Another relevant research direction involves divergence-based metrics that compare the distributions of human-written and machine-generated text. Examples included MAUVE (Pillutla et al., 2021), which leveraged distributions of GPT-2 embeddings, as well as later approaches based on specific linguistic features (Guo et al., 2024a). While such metrics do not explicitly measure linguistic diversity, they can offer insights into distributional differences, of which diversity is a key component.

2.3 Impact of LLMs on Linguistic Diversity

Diverging from the above research focused on developing methods to evaluate linguistic diversity, another line of work explores how LLM-generated content impact future models or human writing patterns, often demonstrating a decline in diversity. Guo et al. (2024b) showed that iteratively training LLMs on synthetic data generated by earlier models leads to a consistent decline in lexical, syntactic, and semantic diversity, especially for tasks requiring high creativity. Similarly, Padmakumar and He (2024) reported a statistically significant reduction in linguistic diversity when humans write with InstructGPT. This reduction in linguistic diversity is also observed in other contexts: Liang et al. (2024) identified a significant frequency shift toward LLM-preferred words in academic writing, while Luo et al. (2024) reported reduced morphosyntactic diversity in machine translations compared to human translations.

Closely related to our work, Kirk et al. (2024) examined how SFT and preference tuning affect LLM generalization and diversity. They found that preference tuning substantially reduces lexical and semantic diversity compared to SFT. Our research also explores the factors that influence diversity while broadening the analysis to include a wider range of diversity aspects, models, tasks and factors. However, our findings on the impact of preference tuning differ from those of Kirk et al. (2024), likely due to differences in task domain, accentuating the importance of contextualizing conclusions.

3 Metrics for Linguistic Diversity

In this section, we present the three types of diversity central to our study: lexical, syntactic, and semantic diversity.

According to Tevet and Berant (2021), diversity can be divided into two primary dimensions: form diversity and content diversity. Lexical and syntactic diversity are sub-aspects within form diversity, whereas semantic diversity pertains to content diversity. While additional sub-aspects of form diversity, such as style diversity, exist and could potentially be measured through style representations (Soto et al., 2024), these aspects are generally less interpretable and often overlap with other dimensions of diversity. For instance, style diversity inherently intersects with lexical and syntactic diversity, as stylistic choices typically involve preferences in vocabulary and grammar. Therefore, in this study, we concentrate on the three diversity aspects (lexical, syntactic, and semantic) that are clearly defined, straightforward to interpret, and exhibit relatively low mutual correlation (further detailed in Section 5.1).

In terms of evaluation protocol, Kirk et al. (2024) distinguish between *across-input diversity* and *per-input diversity*. Across-input diversity refers to the diversity of outputs across different inputs, with only one output generated per input. In contrast, per-input diversity evaluates the capability of the model to produce diverse outputs for a single input.

In our study, we choose to measure across-input diversity, as we focus on general linguistic patterns across a broad range of generations. Formally, given a set of generated outputs $S = \{s_1, s_2, \ldots, s_n\}$, we compute $Div(S)$ differently depending on the aspect of diversity: for lexical diversity, S is treated as a set of *n*-grams, while for syntactic and semantic diversity, S is considered as a set of sentences.

We build on the linguistic diversity evaluation framework and preprocessing methods of Guo et al. (2024b), but shift the focus from studying the effects of recursive synthetic training on OPT (Zhang et al., 2022) to comparing linguistic diversity across a range of state-of-the-art LLMs. We also investigate how various design choices such as model scale and training data, and deployment factors such as decoding strategy, prompt formulation and quantization, impact diversity. In principle, the same research protocol could be extended to examine per-input diversity, allowing for the investigation of uncertainty and variability in text generation (Giulianelli et al., 2023). Although this lies beyond the scope of the current study, it represents a promising direction for future work.

In the following sections, we describe each aspect of diversity and the specific metrics used to assess them.

3.1 Lexical Diversity

Lexical diversity is a measure of the variety of vocabulary used within a text or set of texts. In essence, it assesses the richness or variability of word choices. High lexical diversity indicates a broad range of unique words, while low lexical diversity suggests repetitive or limited vocabulary.

We employ Unique-n (Johnson, 1944; Templin, 1957), established for evaluating lexical diversity. It is calculated as the ratio of unique n-grams to the total number of n-grams. When $n = 1$, it is equivalent to Type-Token Ratio (Johnson, 1944; Templin, 1957). We report the average Unique-n across unigrams, bigrams, and trigrams. Originally used in child language research, Unique-n is useful for assessing language development, where a lower value might indicate limited lexical variety (Miller, 1981). We use the global Unique-n measure rather than the moving average Unique-n because we are interested in the overall diversity capabilities of LLMs across different inputs rather than their performance on individual inputs. Moving average methods might miss global lexical repetitions due to their localized nature (Bestgen, 2023). To mitigate the influence of output length on Unique-n, we always randomly choose 40K samples to constitute the set of n-grams for each n.

3.1.1 Syntactic Diversity

Syntactic diversity refers to the range and variety of sentence structures used in a text or set of texts. It assesses how flexibly and creatively different grammatical structures, such as phrases, clauses, and sentence types, are employed. High syntactic diversity suggests varied sentence forms, while low syntactic diversity indicates repetitive or simplistic sentence structures. Syntactic diversity is a crucial but often neglected aspect of language. Exposure to a variety of syntactic structures helps language learners and models develop a richer understanding of language (Aggarwal et al., 2022). Diverse syntactic forms enhance expressiveness and subtlety in text, impacting its style and tone (Edwards and Bastiaanse, 1998). While research on syntactic diversity exists, it typically relies on manual annotation, which can be both costly and error-prone (Clercq and Housen, 2017).

To address this limitation, we employ a graph-based metric for quantifying syntactic diversity (Guo et al., 2024b). This metric relies on a neural parser (Qi et al., 2020) to generate dependency trees from sentences, following the universal dependencies framework. In these trees, nodes represent words and edges capture syntactic dependencies, with nodes labeled by the corresponding part-of-speech (PoS) tags. The Weisfeiler-Lehman (WL) graph kernel (Shervashidze et al., 2011; Siglidis et al., 2020) is applied to map these trees into a reproducing kernel Hilbert space, where structurally similar graphs are positioned closer together based on the WL isomorphism test. Syntactic diversity is then computed as the average pairwise

	Instruction	Input	Output
Language Modeling (LM)	Not applicable (no instruction)	Block of 128 tokens from Wikipedia	Prediction of the next block
Machine Translation (MT)	Translate from French to English	News story sentence in French	Corresponding English translation
Summarization (Summ)	Summarize the following article	Full news article from the BBC	Summary of the news article
Next Utterance Generation (NUG)	Continue the following dialogue	Scripted dialogue on general topics	Next utterance of the dialogue
Automatic Story Generation (ASG)	Continue the following story	Story prompt shared by Reddit users	Story continuation based on the prompt

Table 1: Summary of instructions, inputs, and outputs for benchmarked NLG tasks. Task instructions are placed in the system input when supported, otherwise prepended to the user input.

distance between these graphs, formalized as: $\text{Div}_{\text{syn}}(S) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} WL(s_i, s_j)$.

3.1.2 Semantic Diversity

Semantic diversity refers to the range and variety of meanings or ideas conveyed within a text or set of texts. It evaluates how broadly and uniquely different concepts, topics, or ideas are expressed, reflecting the depth and scope of the content. Low semantic diversity often indicates repetition or a narrow focus, whereas high semantic diversity typically suggests coverage of a wide array of topics. However, texts should meet a basic quality standard before being evaluated for semantic diversity, since high semantic diversity can also arise from noisy or irrelevant content. Recent studies (Tevet and Berant, 2021; Stasaski and Hearst, 2022) have pointed out that traditional lexical metrics may not fully capture semantic diversity. Similar words can convey different meanings, and different words can convey similar meanings (Yarats and Lewis, 2018).

To address this, we first convert sentences into semantically meaningful embeddings using Sentence-BERT (Reimers and Gurevych, 2019). Semantic diversity is then quantified as the dispersion of these embeddings in the semantic space, measured by the average pairwise cosine distance (scaled to the range $[0, 1]$) between all embedding vectors: $\text{Div}_{\text{sem}}(S) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{d_{\cos}(e(s_i), e(s_j))}{2}$, where e represents Sentence-BERT embeddings.

4 Settings for Diversity Benchmarking

We outline the tasks and models used to establish our linguistic diversity benchmark. We decode the outputs for all tasks and models using a combination of nucleus sampling ($t = 0.6$) and top-k sampling ($k = 0.9$). We further analyze in Section 6.2 the impact of different decoding parameters on output diversity.

4.1 Generation Tasks

To effectively compare the linguistic diversity of LLM outputs across various scenarios, we choose five tasks with progressively increasing levels of ‘‘creativity’’. The inputs, outputs, and instructions for each task are summarized in Table 1. To maintain general model behavior and avoid overly influencing responses through prompt design, we keep the instructions minimal.

For each task, we randomly select 10K samples from the original dataset. While our experiments are conducted using a single dataset per task, we deliberately select the most representative for each. Nonetheless, the conclusions drawn from our experiments should be interpreted within the context of these datasets. We now provide a detailed introduction to each task and its associated dataset.

Language Modeling (LM) involves predicting the next token in a sequence based on the preceding tokens and is fundamental to all NLG applications. We use the Wikitext-2 dataset (Merity et al., 2017) to evaluate general purpose language modeling. Derived from Wikipedia articles, Wikitext-2 offers a rich corpus with around 2 million tokens across diverse topics. We chunk texts into blocks of 128 tokens and ask the models to predict the next 128 tokens. Language modeling serves as the basis of all other tasks, so it is considered as the least creative.

Machine Translation (MT) aims to transfer text from one language to another while maintaining the original meaning. We use the WMT-14 dataset (Bojar et al., 2014) which contains parallel corpora for multiple language pairs. For our experiments, we focus on a subset of this benchmark that includes French-to-English sentence pairs from multiple sources. We classify this task

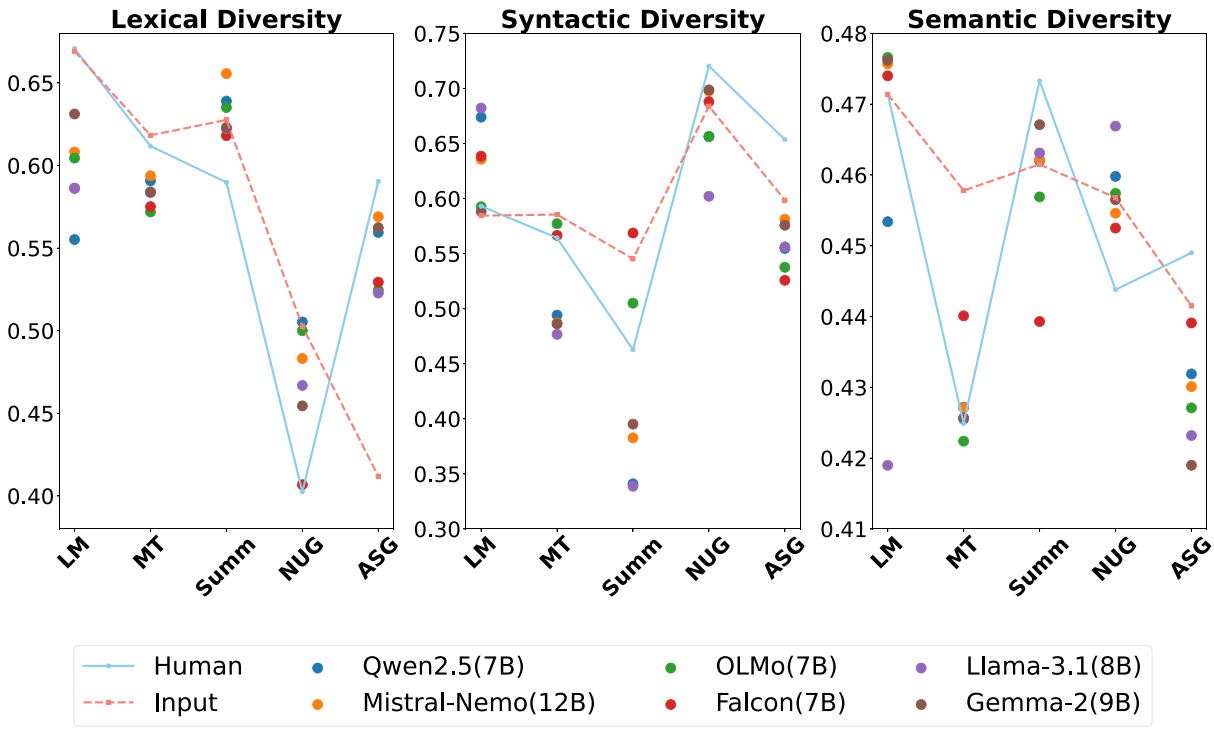

Figure 1: Linguistic diversity benchmarking results for NLG tasks detailed in Table 1.

as having a low level of creativity, as the output is expected to convey exactly the same meaning as the input.

Summarization (Summ) is the process of generating concise summaries of lengthy texts, preserving key information while minimizing redundancy. We use the XLSUM dataset (Hasan et al., 2021), which features news articles in various languages along with their summaries. Our experiments focus on the English portion of the dataset. While we also categorize this task as low in creativity, it allows slightly more flexibility than machine translation, as the model must decide which information to prioritize and include in the summary.

Next Utterance Generation (NUG) aims to produce natural utterances in conversations while maintaining contextual relevance. For this task, we use the DailyDialog dataset (Sai et al., 2020), a human-curated multi-turn dialogue corpus designed to cover a broad range of topics relevant to everyday interactions. In our setup, the model is always prompted to predict the final utterance based on all preceding dialogue turns. We consider next utterance generation to be a creative task, as there is a large space of possible and coherent utterances in response to a certain dialog context. However, the everyday nature and structure of the dataset place some limits on the level of creativity.

Automatic Story Generation (ASG) centers on producing engaging and coherent narratives from story prompts or initial contexts. We employ the WritingPrompts dataset (Fan et al., 2018), which comprises prompts and corresponding stories contributed by Reddit users. It includes a wide variety of prompts in different formats, encouraging diverse and creative responses. Among our tasks, we consider story generation to be the most creative, as the prompts typically impose minimal constraints on narrative structure and content, allowing for maximal expressive freedom.

4.2 Language Models

We evaluate the following families of models: Llama (Dubey et al., 2024), Mistral (Jiang et al., 2023), Olmo (Groeneveld et al., 2024), Gemma (Team et al., 2024), Qwen (Yang et al., 2024), and Falcon (Almazrouei et al., 2023). The comparison of these models across various key characteristics is provided in Table 4 in Appendix A.

To ensure comparability, we select the latest version of each model family that is closest in scale to 7 billion parameters. The scale selected for each model is specified in the legend of Figure 1. We purposefully include models developed by

organizations from different countries to be culturally inclusive. For language modeling, we use base models. For all other tasks, we employ instruction-tuned versions.

5 Results of Diversity Benchmarking

Figure 1 visualizes the benchmarking results of linguistic diversity across various tasks. Round dots represent the diversity of model outputs, while solid lines represent human reference outputs. Dashed lines depict the diversity of task-specific inputs (as detailed in Table 1), reflecting the conditions under which the outputs were generated. Tasks are organized in ascending order of creativity level. The detailed numerical results are provided in Table 5 in Appendix B.

For the machine translation task, the inputs are in French; hence, semantic diversity is measured using a multilingual SentenceBERT (Reimers and Gurevych, 2020), and syntactic diversity is evaluated with a French-specific dependency parser. As a result, these scores may not be directly comparable to those for English. The diversity of human reference outputs serves as a baseline for interpreting whether the model under or over represents the diversity for each task.

In this section, we first analyze metric correlations in Section 5.1, then compare diversity scores across tasks and models in Section 5.2. Finally, in Section 5.3, we perform a case study on syntactic diversity in story generation, comparing human and model outputs.

5.1 Correlation Study

Correlation Between Diversity and Quality. We manually verify that all models produce plausible and coherent text that meets the basic requirements for diversity evaluation across all tasks. Building on this, we examine more specific qualities of the model outputs. Figure 2 illustrates the correlation between diversity and quality in model outputs, using task-specific automatic metrics as quality indicators. For the language modeling task, perplexity is used to evaluate the model's performance on reference text continuations. For machine translation, we use COMET (Rei et al., 2020), which takes into account both the source text and reference translation. For the remaining three tasks, BERTScore (Zhang et al., 2020) is used to measure the relevance between inputs and outputs. Due to the inherently subjective nature of these tasks, automatic metrics generally exhibit weak correlations with human judgments (Liu et al., 2023) and should be interpreted cautiously. Nonetheless, we adopt BERTScore as an approximate quality indicator, as embedding-based metrics of this kind demonstrate the strongest system-level correlation with human evaluations among available automatic measures (Chhun et al., 2022).

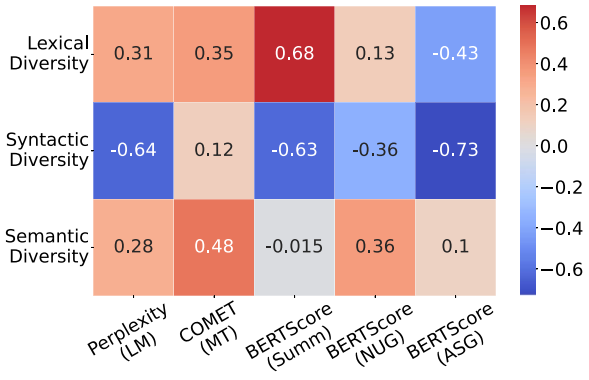

Figure 2: Pearson correlation matrix between diversity metrics and quality metrics.

Our results show a positive correlation between quality and lexical as well as semantic diversity in model outputs. In contrast, syntactic diversity often exhibits negative correlations, where higher syntactic diversity is associated with lower quality scores. This may be attributed to the tested domains inherently exhibiting low ground-truth syntactic diversity (e.g., in language modeling) or to the limitations of quality metrics in recognizing the value of syntactic variation (e.g., in summarization, automatic story generation, and next utterance generation). *These findings highlight the need to report diversity metrics alongside quality metrics for comprehensive evaluation, as the relationship between the two is not consistent across tasks or aspects.*

Correlation Between Diversity Aspects. The correlations between different diversity aspects are shown in Figure 3, revealing a moderate positive correlation between syntactic and semantic diversity (0.55). However, lexical diversity shows a weak positive relationship with syntactic diversity (0.13) and a slight negative correlation with semantic diversity (-0.14), indicating that *the richness of vocabulary is independent from the variety of grammatical structures and meaning*.

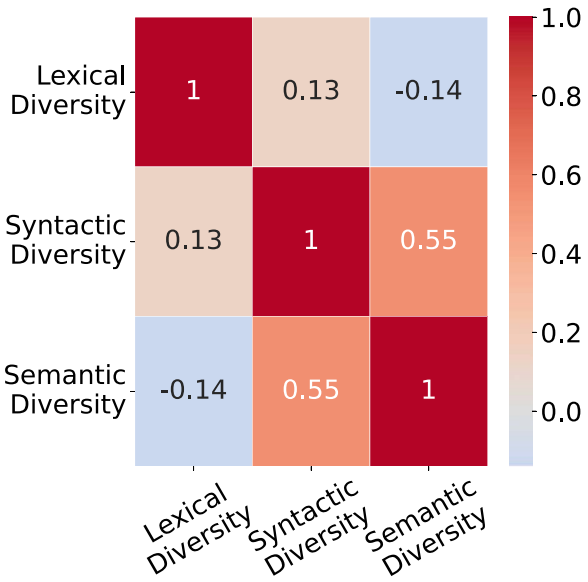

Figure 3: Pearson correlation matrix between different diversity metrics.

5.2 Comparison Across Tasks and Models

We now examine the results in Figure 1 to assess human diversity results across tasks, compare model diversity against human diversity, and finally evaluate the diversity performance across different models.

Human Output Diversity. *Human-level diversity varies across tasks, with no clear correlation observed among different aspects.* Notably, utterances in human dialogs exhibit the lowest lexical diversity and the highest syntactic diversity, unlike the written text present in the remaining four tasks. The low lexical diversity may be attributed to the conversations being specifically scripted for English learners to practice daily-life dialog. These dialogs focus on generic topics, leading to a limited range of vocabulary. In contrast, the high syntactic diversity can be explained by the inherent spontaneity of conversational language, where different speakers tend to vary significantly in their use of syntactic structures (Healey et al., 2014; Dubuisson Duplessis et al., 2017). Human summaries show limited lexical and syntactic variation but exhibit the highest semantic diversity, suggesting a narrow range in form but a broad range in content. In contrast, human translations score lowest in semantic diversity, reflecting the restricted topical scope of the source texts. Wikipedia-based language modeling displays high topic diversity, while human-written stories tend to be diverse across all three dimensions.

Model Output Diversity. *LLMs lack diversity compared to humans for tasks demanding high levels of creativity, such as story generation.* Overall, the scores of different LLMs across tasks and diversity aspects tend to resemble each other, potentially due to the use of similar development procedures, architectures, and datasets. However, this remains an assumption, as most LLM developers do not fully disclose their training data or protocols, even when the models themselves are open-weight. The extent to which LLMs under- or over-represent diversity compared to humans varies significantly by the task domain. For the task of story generation, which demands the highest levels of creativity and freedom of expression, LLMs consistently lag behind humans in all three diversity aspects. In contrast, for tasks like next utterance generation, LLMs surpass human references in both lexical and semantic diversity. This discrepancy arises because the DailyDialog dataset focuses on generic, everyday topics designed for English language learning, while LLMs, unconstrained by this context, frequently steer conversations toward more complex topics.

LLM Comparisons. While the overall performance of the models appears to be similar, in-depth comparisons showcase notable differences. *Models pretrained on fewer tokens, such as Falcon and OLMo, consistently generate outputs with lower lexical diversity.* Specifically, Falcon and OLMo are pretrained on 1.5T and 2.7T tokens, respectively, compared to Llama-3.1, which is trained on 15T tokens. However, this effect is not observed for syntactic or semantic diversity. *Models with less strict data filtration exhibit greater diversity in creative tasks*, such as story generation. For example, Qwen2.5, which filters data exclusively for quality, exhibits significantly higher diversity in story generation across all aspects compared to Llama-3.1, Gemma-2, and OLMo, whose data is extensively filtered for quality, privacy, and safety.

5.3 Comparing Syntactic Diversity Between Humans and Models

To further compare humans and models, we conduct a case study on syntactic diversity using dependency tree distribution. Syntactic diversity is chosen as it is less explored than lexical and semantic diversity. Moreover, syntactic patterns

reflected by POS tag n-grams are more generalizable than lexicon n-grams and more interpretable than semantic embeddings.

We adopt the Precision-Recall framework proposed by Le Bronnec et al. (2024). This framework relies on GPT-2 embeddings, followed by Principal Component Analysis (PCA) and K-means clustering, to estimate the supports of text distributions. In our study, we replace the original GPT-2 embeddings with the implicit distribution of dependency tree embeddings induced by the WL graph kernel. Precision is defined as the proportion of dependency trees from model-generated text that lie within the support of dependency trees from human-written text. A high precision indicates that the model-generated structures are more plausible and human-like, thus reflecting their quality. Recall, on the other hand, measures the proportion of dependency trees from human-written text that fall within the support of the model-generated distribution. A high recall suggests that the model captures the full diversity of human-written structures. The method for computing pairwise distances between dependency trees is described in Section 3.1.1, which serves as the basis for constructing the distance matrix. All other hyper-parameters remain consistent with the original work (Le Bronnec et al., 2024).

Table 2 presents the precision and recall scores for all evaluated models on the story generation task. The results reveal that all models exhibit near-perfect precision, indicating that almost all generated sentences are syntactically plausible. In contrast, recall scores are substantially lower than precision scores across all models, revealing their limited capacity to capture the full breadth of human syntactic diversity. *This points to a notable gap between models and humans in syntactic diversity for the story generation task where high creativity is required.*

To further illustrate these findings, Table 3 lists examples of syntactic patterns (POS tag n-grams) that are frequently found in human dependency trees but are missing from the model-generated ones. Conversely, we also identify syntactic patterns that models over-generate but are less common in human outputs. Recent studies (Shaib et al., 2024b) indicate that models often memorize syntactic templates encountered during pretraining, which are rarely overwritten during SFT and preference tuning. This suggests that the observed gap in syntactic patterns may stem from a mismatch between pretraining and downstream task domains. For instance, in the pretraining corpus of OLMo, over 80% of the data originates from web pages in Common Crawl, while less than 0.3% comes from Project Gutenberg books, one of the only sources potentially aligned with the narrative style required for story generation (Soldaini et al., 2024).

	Llama	Mistral	Qwen	Gemma	Falcon	OLMo
Precision	99.20	99.20	99.47	99.07	99.63	99.73
Recall	35.20	65.87	75.27	37.97	75.00	39.40

Table 2: Comparison of dependency tree distributions between humans and models for the story generation task.

6 Factors Influencing LLM Diversity

In this section, we explore key factors that may influence the diversity of LLM outputs. The factors under consideration include pretraining token counts, instruction tuning, decoding parameters, prompt formulation, model scale, and quantization. For decoding parameters, prompt formulation and instruction tuning, we conduct experiments across all models. We employ OLMo for assessing the impact of pretraining token counts since it provides full access to its pretraining datasets and model weights at various checkpoints throughout its development. Since OLMo models are available in only two sizes, we additionally leverage Qwen2.5 models (Yang et al., 2024) to investigate the effects of model scale and quantization.

All experiments in this section are conducted on the story generation task, where linguistic diversity plays a central role. Its minimal input constraints and strong emphasis on creativity make it an ideal benchmark for evaluating linguistic diversity. Moreover, as shown in Figure 1, all models fall significantly short of human performance in terms of diversity on this task, highlighting the importance of identifying which factors contribute to this gap. *We emphasize that the conclusions drawn in this section are specific to the story generation task and should not be generalized to broader LLM behavior without further investigation.*

6.1 Impact of Training Stages

Pretraining. We choose OLMo, pretrained on the Dolma corpus (Soldaini et al., 2024), to study

	Human		Language Models	
	POS tag *n*-gram	Example	POS tag *n*-gram	Example
n=3	(ADV, ADV, ADP)	right along with	(PRON, NOUN, ADJ)	her voice soft
n=4	(VERB, ADP, DET, NOUN)	picking up the pieces	(NOUN, CCONJ, NOUN, PRON)	carvings and symbols that
n=5	(DET, NOUN, ADP, DET, NOUN)	the cackling of the fire	(PRON, NOUN, VERB, ADP, NOUN)	its feathers stained with blood
n=6	DET, ADJ, NOUN, ADP, DET, NOUN)	the old woman down the street	(ADJ, NOUN, ADP, NOUN, CCONJ, NOUN)	particular focus on time and space

Table 3: Examples of syntactic patterns favored by either humans or models are illustrated using *n*-grams of POS tags. Human patterns are derived from human dependency trees that are not within the model dependency tree neighborhoods, while model patterns have high frequency in model dependency trees and low frequency in human dependency trees.

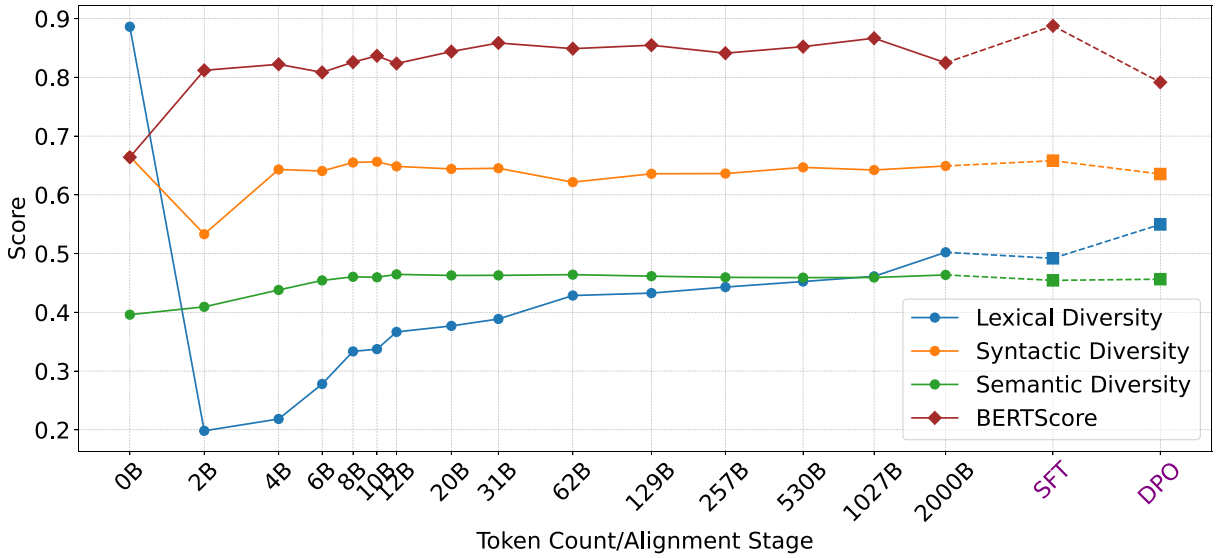

Figure 4: Linguistic diversity metrics after different LLM training stages. The pretraining stage is broken into various steps with increasing token counts, which are presented on a log scale for visualization. Experiments are conducted with the OLMo model on the story generation task.

the evolution of linguistic diversity during pretraining. This is because OLMo is the only model in our benchmark with publicly available intermediate checkpoints during pretraining. The results are presented in Figure 4. Intitially, lexical diversity is exceptionally high, as expected for an untrained model that generates random tokens. This metric drops sharply after the first checkpoint (2B tokens) but then gradually increases throughout the pretraining process, without reaching saturation. In contrast, syntactic diversity also experiences a sharp decline early on; however, it saturates much more quickly, fluctuating within a narrow range afterward. Semantic diversity shows a steady increase from the beginning but also saturates relatively quickly. *These observations suggest that while increasing training data generally improves lexical diversity, alternative strategies are needed to enhance syntactic and semantic diversity.*

Instruction Tuning. We now move on to study the impact of instruction tuning on linguistic diversity. After pretraining, OLMo underwent SFT on Tulu v2 (Ivison et al., 2023) and direct preference optimization (DPO) (Rafailov et al., 2023) on Ultrafeedback (Cui et al., 2024), with DPO applied on top of SFT. We observe that SFT has minimal impact on any diversity metric, while DPO leads to a decrease in syntactic diversity and an increase in lexical diversity, potentially reflecting characteristics of the SFT and DPO datasets.

Since all models in our benchmark provide both base and instruction-tuned versions, we extend our analysis to assess the impact of instruction tuning across the full set. As shown in Figure 5, the results mirror those observed for OLMo: Instruction-tuned versions show higher lexical diversity compared to their base counterparts but exhibit reductions in syntactic and semantic diversity. Notably, the decline in syntactic diversity

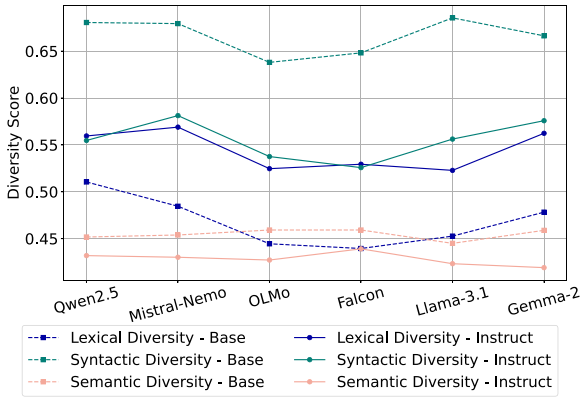

Figure 5: Impact of instruction tuning.

is more pronounced than that in semantic diversity. *These findings indicate that while additional training—regardless of the stage—enhances vocabulary richness, aligning models with human preferences tends to constrain them to a narrower range of grammatical structures and meanings.*

6.2 Impact of Decoding Parameters

Achieving a balance between quality and diversity in LLM outputs is a known challenge, as there is often a trade-off between these two aspects (Caccia et al., 2020). The choice of decoding strategy plays a crucial role in controlling this trade-off (Zhang et al., 2021). Here, we investigate how varying the decoding temperature affects the outputs in the story generation task, with results visualized in Figure 6. Output quality is estimated based on their relevance to the inputs, using BERTScore as a metric.

The results show that increasing the temperature—making decoding less restrictive—leads to greater lexical diversity, with only a minor reduction in relevance to the inputs. It might be due to the creative nature of the story generation task that the quality-diversity trade-off is so subtle. For syntactic diversity, while most models show fluctuating performance within a certain range, some exhibit a clear downward trend, specially OLMo and Falcon, which are trained on significantly fewer tokens compared to the other models. However, no consistent trends are observed for semantic diversity metric as decoding parameters change. This aligns with the observations of Tevet and Berant (2021), which indicate that adjusting decoding parameters tends to affect the form of the text rather than its meaning.

Furthermore, we note that, across most models, the relative ranking of diversity scores remains stable as the temperature varies. *This suggests that conducting experiments with a fixed temperature is sufficient for consistent evaluation*. Based on our findings, we set the temperature to 0.6 for all other experiments. Figure 6 shows that at a temperature of 0.6, the relevance to the inputs remains relatively high while diversity scores significantly improve compared to lower temperatures.

6.3 Impact of Prompt Formulation

Previous studies have established that LLMs exhibit considerable sensitivity to prompt formulations, particularly affecting their performance on discriminative downstream tasks (Sclar et al., 2024; Wahle et al., 2024). Here, we explore whether the linguistic diversity of stories generated by LLMs is similarly influenced by variations in the formulation of prompts. We conduct experiments across the full range of models, and the results are depicted in Figure 7. The solid lines represent results obtained using the standard prompt, consisting of the task-specific instruction ‘‘Please continue the following story’’ combined with sample-specific inputs from the WritingPrompts dataset. To evaluate prompt sensitivity, we modify the prompt to explicitly encourage creativity by changing the instruction to ‘‘Please continue the following story and be as creative as possible’’. Results from these modified prompts are shown as dash-dot lines in Figure 7.

Our analysis indicates that altering the prompt formulation has minimal impact on the diversity of generated stories across all three evaluated aspects. This suggests that the linguistic patterns exhibited by LLMs across creative generations represent inherent model characteristics that are less sensitive to prompt variations compared to accuracy-based performance on discriminative tasks. Consequently, enhancing linguistic diversity in LLM outputs through straightforward prompt engineering alone would be challenging.

6.4 Impact of Model Scale and Quantization

We now study the impact of model scale on linguistic diversity with the Qwen2.5 model. Qwen2.5 has been released in various sizes, ranging from 0.5B to 72B parameters. Due to computational resource constraints, we limit our exploration of linguistic diversity to models up

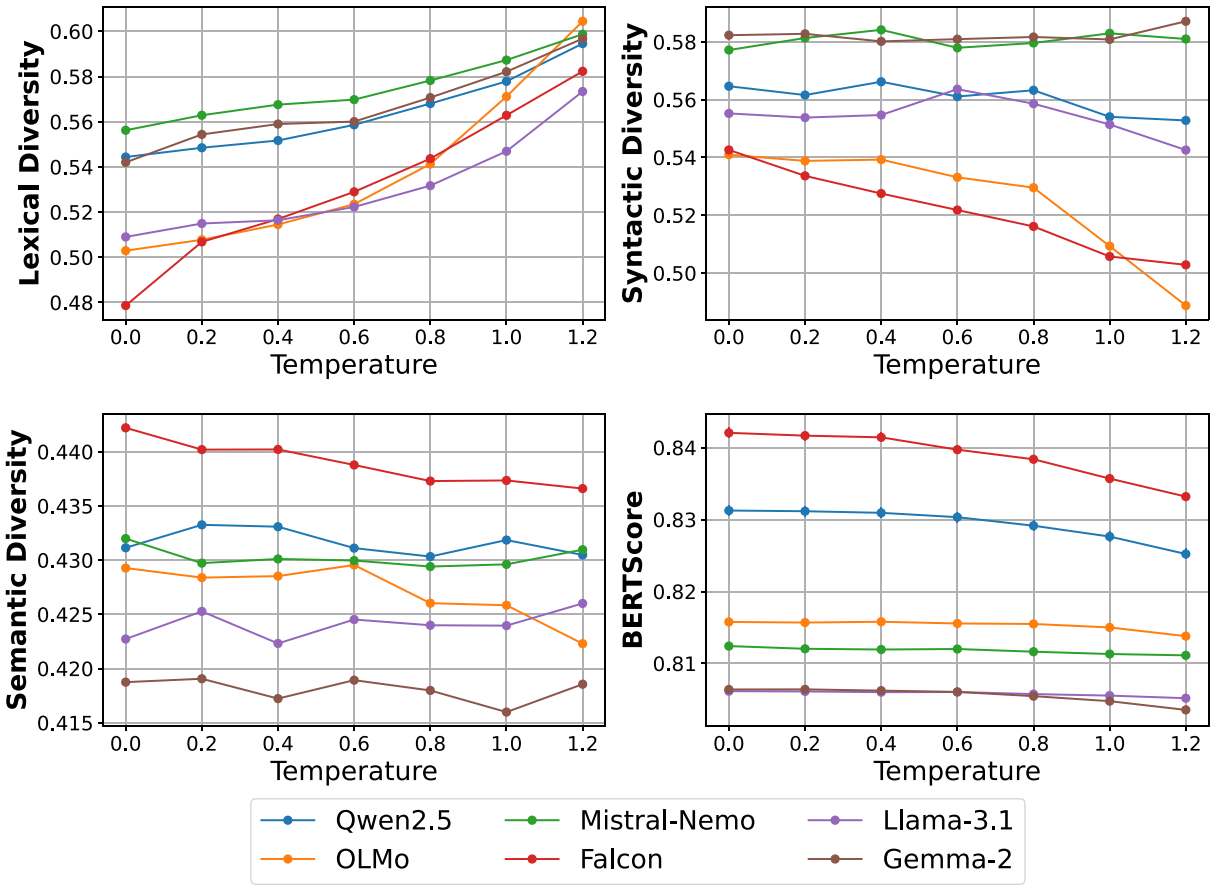

Figure 6: Impact of decoding parameters. Experiments are conducted on the story generation task.

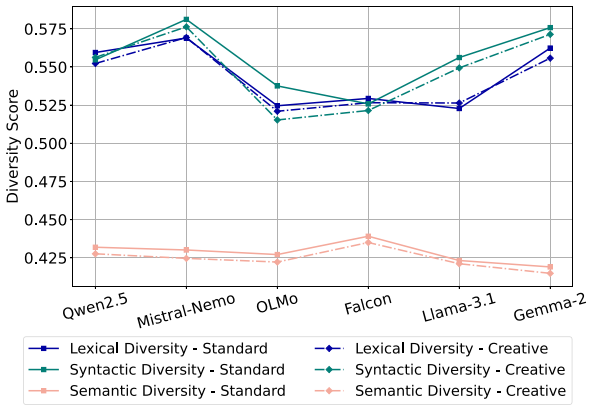

Figure 7: Impact of prompt formulation.

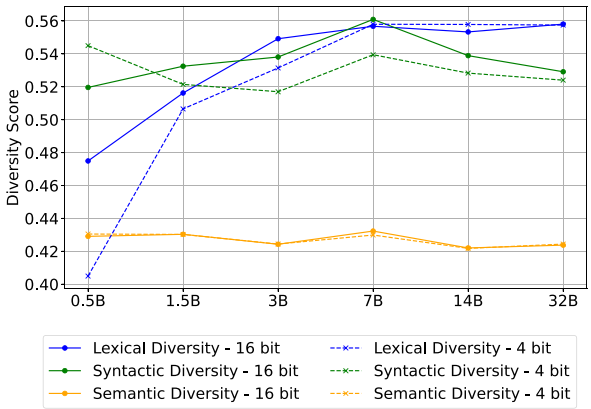

Figure 8: Impact of model scale and quantization.

to 32B parameters. The results are presented in Figure 8. *We observe that lexical diversity consistently increases with model size, while semantic diversity remains stable throughout*. In contrast, syntactic diversity remains relatively stable overall but exhibits an initial increase followed by a decline, peaking at 7B parameters, *indicating that scaling up is not always the solution to higher linguistic diversity*.

We further investigate the impact of post-training quantization on linguistic diversity. We quantize the Qwen2.5 models of various scales to 4-bit precision with the bitsandbytes library,[1] whereas the original models were run with bf16. As shown in Figure 8, *quantization does not affect semantic diversity but reduces both syntactic and*

[1] `https://huggingface.co/docs/bitsandbytes/index`.

lexical diversity. The reduction in lexical diversity is more pronounced in smaller models, while the effect on syntactic diversity becomes more evident in larger models. *This finding suggests that quantization has greater impact on the diversity of form rather than content.*

7 Conclusion

Our study offers crucial insights into the linguistic diversity of current LLMs. By leveraging a comprehensive evaluation framework focused on lexical, syntactic, and semantic diversity, we provide a fresh perspective beyond traditional quality metrics. Our analysis reveals that, despite the impressive capabilities of LLMs in generating coherent and plausible text, there is a significant gap when it comes to replicating the linguistic richness of human language for creative tasks such as story generation. Furthermore, we find that factors like pretraining data volume, instruction tuning, decoding strategies, model scale, and quantization significantly influence diversity metrics. In particular, while instruction tuning improves lexical diversity, it constrains syntactic and semantic diversity, indicating a narrowing of expressive flexibility. These findings raise an important concern: as LLMs become more prevalent in content creation, their outputs may trend towards homogenization, risking a loss of linguistic richness. Our research highlights the necessity of a more holistic and forward-looking approach in developing language models, one that prioritizes the preservation of linguistic diversity alongside optimizing performance metrics.

Acknowledgments

We thank Professor Michalis Vazirgiannis for providing the computational resources that supported this project. This research was partially funded by the ANR-23-CE23-0033-01 SINNet project and the ANR-TSIA HELAS chair.

References

Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.445

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Yves Bestgen. 2023. Measuring lexical diversity in texts: The twofold length problem. *Language Learning*. https://doi.org/10.1111/lang.12630

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3302, https://doi.org/10.3115/v1/W14-33

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language gans falling short. In *International Conference on Learning Representations*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem - instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.460

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bastien De Clercq and Alex Housen. 2017. A cross-linguistic perspective on syntactic complexity in l2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334. https://doi.org/10.1111/modl.12396

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2024. Ultrafeedback: Boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia. et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 71–81, Saarbrücken, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5510

Susan Edwards and Roelien Bastiaanse. 1998. Diversity in the lexical and syntactic abilities of fluent aphasic speakers. *Aphasiology*, 12(2):99–117. https://doi.org/10.1080/02687039808250466

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

G. Fergadiotis, H. H. Wright, and T. M. West. 2013. Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-language Pathology*, 22(2):S397–408. https://doi.org/10.1044/1058-0360(2013/12-0083), PubMed: 23695912

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179. https://doi.org/10.1162/coli_a_00524

Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.887

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. https://doi.org/10.18653/v1/2024.acl-long.841

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2024a. Do large language models have an English accent? Evaluating and improving the naturalness of multilingual LLMs. *arXiv preprint arXiv:2410.15956*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024b. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.228

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.413

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? Criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS One*, 9(6):e98598. https://doi.org/10.1371/journal.pone.0098598, PubMed: 24919186

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with tulu 2.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15. https://doi.org/10.1037/h0093508

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalization and diversity. In *The Twelfth International Conference on Learning Representations*.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.643

Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre Allauzen. 2024. Exploring precision and recall to assess the quality and diversity of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11418–11441, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.616

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.153

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371. https://doi.org/10.1162/tacl_a_00645

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86. https://doi.org/10.1177/0741088309351547

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

J. F. Miller. 1981. *Assessing Language Production in Children: Experimental Procedures*. Assessing communicative behavior. University Park Press.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity?

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations, pages 101–108, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827. https://doi.org/10.1162/tacl_a_00347

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024a. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024b. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.368

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(null):2539–2561.

Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. 2020. Grakel: A graph kernel library in python. *Journal of Machine Learning Research*, 21(54):1–5.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.840

Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y. Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of

machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.6

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, ned - new edition edition, volume 26. University of Minnesota Press. https://doi.org/10.5749/j.ctttv2st

Guy Tevet and Jonathan Berant. 2021. 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*, pages 326–346, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.25

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee. 2023. Does human collaboration enhance the accuracy of identifying LLM-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174. https://doi.org/10.1609/hcomp.v11i1.27557

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.617

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui,

Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5591–5599. PMLR.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.830

A Comparison of Benchmarked LLMs

We provide a comprehensive comparison of the LLMs included in our benchmark in Table 4, highlighting several key characteristics relevant to their design, development, and deployment.

	Llama-3.1-8B	Mistral-NeMo-12B	Qwen2.5-7B	Gemma-2-9b	Falcon-7b	OLMo-7B
Organization	Meta	Mistral	Alibaba	Google	TII	Ai2
Country	USA	France	China	USA	UAE	USA
Open weights	yes	yes	yes	yes	yes	yes
Open data	no	no	no	no	partially	yes
Tokenization	BPE (Tiktoken)	BPE (Tiktoken)	BPE	SentencePiece	BPE	BPE
Vocabulary size	128K	128K	151K	256K	65K	50K
#tokens	15T	unknown	18T	8T	1.5T	2.7T
Data filter	quality, privacy, safety safety	unknown	quality	quality, privacy, safety safety	quality	quality, privacy, safety safety
Synthetic data	post-training	unknown	pre/post-training	post-training	unknown	post-training
Multilinguality	yes	yes	yes (over 29 languages)	not in particular	yes (Latin alphabet)	no
Alignment	rejection sampling SFT, DPO	SFT	SFT, DPO	SFT, PPO	SFT	SFT, DPO
Release date	July 2024	July 2024	September 2024	June 2024	May 2023	February 2024

Table 4: Comparison of benchmarked LLMs.

B Detailed Results of Linguistic Diversity Benchmarking

We present the detailed results of our linguistic diversity benchmarking experiments in Table 5. These results are also visualized in Figure 1.

		Human	Input	Qwen2.5	Mistral-Nemo	OLMo	Falcon	Llama-3.1	Gemma-2
Lexical Diversity	**LM**	67.08	66.92	55.52	60.82	60.45	58.63	58.62	**63.12**
	MT	61.18	61.83	59.07	**59.38**	57.19	57.51	58.36	58.40
	Summ	58.98	62.76	63.90	**65.56**	63.51	61.81	62.23	62.30
	NUG	40.25	50.27	**50.53**	48.32	50.00	40.68	46.69	45.45
	ASG	59.04	41.19	55.95	**56.90**	52.46	52.94	52.28	56.24
Syntactic Diversity	**LM**	59.31	58.45	67.39	63.57	59.26	63.85	**68.22**	58.83
	MT	56.43	43.47	49.42	48.71	**57.72**	56.66	47.67	48.62
	Summ	46.27	54.52	34.10	38.27	50.50	**56.87**	33.88	39.52
	NUG	72.03	68.39	65.63	69.76	65.63	68.80	60.21	**69.87**
	ASG	65.35	59.86	55.47	**58.12**	53.76	52.58	55.62	57.58
Semantic Diversity	**LM**	47.14	47.14	45.34	47.57	**47.66**	47.40	41.90	47.62
	MT	42.49	33.58	42.72	42.71	42.24	**44.01**	42.55	42.57
	Summ	47.33	46.15	46.20	46.22	45.69	43.93	46.31	**46.71**
	NUG	44.38	45.68	45.98	45.46	45.74	45.25	**46.69**	45.65
	ASG	44.90	44.15	43.19	43.01	42.71	**43.91**	42.32	41.90

Table 5: Linguistic diversity benchmarking results for NLG tasks detailed in Table 1. For each type of diversity, the highest model score for each task is highlighted in **bold**.