Domain Regeneration:
How well do LLMs match syntactic properties of text domains?

Da Ju
Meta AI
daju@meta.com

Hagen Blix
New York University
hagen.blix@nyu.edu

Adina Williams
FAIR, Meta AI
adinawilliams@meta.com

Abstract

Recent improvements in large language model performance have, in all likelihood, been accompanied by improvements in how well they can approximate the distribution of their training data. In this work, we explore the following question: which properties of text domains do LLMs faithfully approximate, and how well do they do so? Applying observational approaches familiar from corpus linguistics, we prompt commonly used, opensource LLMs to regenerate text from three domains of permissively licensed English text which are often contained in LLM training data—Wikipedia, news text, and ELI5. In a fairly semantically-controlled setting, this regeneration paradigm allows us to investigate whether LLMs can faithfully match original human text domains. We investigate varying levels of syntactic abstraction, from simpler properties like sentence length, and article readability, to more complex and higher order properties such as dependency tag distribution, parse depth, and parse complexity. We find that the majority of the regenerated distributions show a shifted mean, a lower standard deviation, and a reduction of the long tail, as compared to the human originals.

1 Introduction

The question of whether models can transfer capabilities across different domains of texts, or **domain transfer** has a long history in NLP. Domain considerations have contributed greatly to the establishment of the pretrain-finetune paradigm (Devlin et al., 2019; Liu et al., 2019) used for training LLMs (Ruder et al., 2019). Since the rise in prevalence of LLMs, however, there has been little work explicitly verifying whether state-of-the-art LLMs can actually generate text that matches different, well-described, and well-delineated human-generated text domains.

One might presume that LLMs should be generally competent in matching text distributions, insofar as one could, from a zoomed out perspective, describe the entire process of pretraining itself as a process of fitting a model to a distribution. However, it is also possible that the increasing number of post-training interventions, such as instruction tuning, automatic preference alignment or other kinds of interventions, could affect this.

Other model design decisions can also impact a models' ability to match a human-generated distribution. LLMs can suffer from **model collapse** (Dohmatob et al., 2024; Hamilton, 2024; Lanchantin et al., 2025), whereby the model has trained on outputs from previous models, which can negatively affect the diversity and quality of generations. Even without being trained on synthetic data, models can fail to match the diversity of human-generated data. LLMs often generate similar data patterns (Hupkes et al., 2023). They can repeat words/tokens (Juzek and Ward, 2024), use less diverse topics (Bache et al., 2013; Alihosseini et al., 2019) or both (Padmakumar and He, 2023).

However, most work investigating the ability of LLMs to match the diversity of human written text does so in the context of an unspecified and uncontrolled "neutral" domain. Given that there are many, widely described differences in lexical content, style, syntactic structure etc. across domains (Biber, 1991; DiMarco and Hirst, 1993; Dewdney et al., 2001; Lee, 2002; Williams et al., 2018; Li et al., 2019), one might wonder whether models can match human diversity in domains with consistent and well controlled style.

We explore this question with a paradigm that we call **LLM-regeneration**: using the beginning of articles from a well-described domain (e.g. Wikipedia, Fan and Gardent 2022), we prompt the LLM to complete the article, and then compare the regenerated article to the original. This setting allows us to exert more control over the content of the generations, thus making for a clearer picture of distribution match.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2367–2388
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

We use this regenerated data to explore the diversity of model outputs in a number of interrelated syntactic measurements. Beyond diversity, we also explore two other **signatures** of imperfect domain matching in this work: difference in the means of the LLM and human distributions, and a reduced long tail where present in the human distribution. When present, we take each of these three signatures to be evidence that the LLMs we study fail to perfectly match the human text.

While past work has indicated that some LLMs generate more homogeneous syntax than humans (Shaib et al., 2024), such investigations have thus far been restricted to part-of-speech tags, and have only been explored for "neutral" domain LLM generations. Here, we combine investigations of text domain with syntactic complexity metrics in an attempt to delimit a reasonably sized problem space; in addition to text readability and sentence length, we explore more abstract metrics for syntactic complexity, including parse depth, unique dependency tag and constituency label count, and Yngve scores.

Overall, the contributions of this work are: (i) We define an experimental setting called LLM-regeneration that enables us to measure how well LLMs match human text with some controls over semantics and domain. (ii) We investigate the extent to which LLMs can match the distribution of text from three well-described human domains. (iii) We illustrate three signatures of domain mismatch—diversity, mean shift, and reduction of the long tail—and show how the three are present for several syntactic complexity metrics, as calculated on generations from opensource models from two model families. These results are important, as they can inform us about whether models can genuinely match text domains. Information about whether LLMs differ in syntactic complexity and variability from human-generated text may additionally be useful for detection of synthetic text, or to guide model improvement.

2 Methods

2.1 Models

In this work, we mostly utilize the Llama family of models, as they are strongly performing models with open weights. For the majority of our experiments, we utilize Llama-V2 (Touvron et al., 2023) 70B instruction finetuned. We also utilize Llama-V3.3 instruction finetuned (Llama Team, 2024) 70B and Llama-V3.1-8B for a subset of our experiments as an additional comparison for the model-specificity of our results. We additionally investigate two models from the Mistral family: Mistral-Small-24B-Intruct[1] and Ministral-8B-Instruct[2] to verify that our findings hold across model family and size.

2.2 Data

We are interested in whether SOTA or near-SOTA LLMs can match properties of text corresponding to its domain, and thus need to select datasets that typify domains. However, some practical complications arose when we embarked on selecting datasets. First of all, we needed to consider which text the model was trained on. In principle, if a model was trained on text from a domain, it should be better at matching the distributional properties of text from that domain. However, the precise nature of the training data for LLMs is generally proprietary information, though it is likely that LLMs were trained on many domains. Clearly, it would not be scalable to investigate all of them (nor to determine where their boundaries lay).

Additionally, we could, in principle, select a set of data and train an LLM from scratch on it. However, due to budgetary constraints, we would doubtless have to focus on a smaller, and likely less performant model. In that case, if we observed differences between the human and model distributions, those differences could just be due to the model being weak, not to anything interesting about the distributions the model had learned.

Given these considerations, we made the practical decision to focus on Wikipedia, a datasource known to be used in open training datasets (BigScience Workshop et al., 2023; Soldaini et al., 2024) and to be a popular and well-studied data source for many NLP applications in English (Wu and Weld, 2010; Horn et al., 2014; Ni and Florian, 2016; Yang et al., 2018; Dinan et al., 2019; Klang and Nugues, 2019; Ein-Dor et al., 2019; Dinan et al., 2020; Nie et al., 2020; Calixto et al., 2021; Eisenschlos et al., 2021; Petroni et al., 2021; Semnani et al., 2023). As additional domains, we also investigate the standard news articles dataset, CC-News[3], and "Explain Like I'm 5" dataset (ELI5; Fan et al. 2019) of simply written questions and answers from an online forum.

Despite their difference in size (our Wikipedia

---

[1] mistralai/Mistral-Small-24B-Instruct-2501

[2] mistralai/Ministral-8B-Instruct-2410

[3] https://huggingface.co/datasets/vblagoje/cc_news

datasets contains roughly 10x more data than CCNews, and CCNews is larger than ELI5), all datasets are representative of consistent and fairly well circumscribed domains. All three datasets are characterized by internally enforced stylistic standards (e.g., due to editors or moderators), and any competent reader could easily match text to the relevant domain. That is, their differences should allow us to determine whether some of the trends we find for Wikipedia are specific to that data source or more general.

2.3 Data Processing

Data Cleaning. Given that Wikipedia data contains a significant amount of structured text, such as lists, titles, urls and citations, we perform a data cleaning stage using the parsing results we collected to enhance signal quality. First, we removed all sentences with fewer than 3 words or more than 500 words. Second, we eliminated all sentences that contained neither a verb nor an auxiliary verb, as identified by POS tagging, to ensure that our data consists of full English sentences. To verify that these filtering steps didn't drive our main results, we plot data ablation results in Appendix A, which show the same trends as our main results.

Note that additional data cleaning is an implicit part of our process. We calculate all metrics based on successful Stanza dependency and constituency parsing. Metrics will not be available if an article is empty in the source data, or contains non-English content. Some metrics, such as the depth score and Yngve score, may encounter errors if the tree parsing is excessively deep. In cases where articles pass the parsing stage but have a valid article length of zero (meaning the article is not empty in the beginning, but all sentences have been removed due to prior length and POS filtration), we filter out all depth and Yngve scores of zero. For sentence-level metrics, we aggregate results from all successfully parsed sentences to complete the calculations. For metrics aggregated at the article level, we exclude any article if any of its sentences fail in parsing or metric calculation.

Regenerating the Data with LLMs. To generate text from the models that replicates our domains of interest, we adopt a similar approach to that used in Ju et al. (2024) to ensure that the article topic and content do not vastly differ. We isolate the first 256 words from a Wikipedia article, and the first 180 words from a CCNews article respectively (since news articles are shorter on average). For ELI5, we prompt the models only with the title of the thread (e.g., *What is an ETF?*). We then feed these into the model using the prompts in Appendix B. We use vLLM (Kwon et al., 2023) for generation with its default coding configurations, including a temperature of 1.0, which is considered a "medium" temperature. The resulting articles will be approximately matched to the original articles in their topic and content. We collect all the articles in each domain, which then serve as our regenerated corpus for downstream analysis. We also performed the two data cleaning steps described above on the model-regenerated data as well.

2.4 Parsing

We employ the data processing pipeline outlined by Williams et al. (2021) and used in Ju et al. (2024) for our analysis. Our pipeline uses the Stanza tool (Qi et al., 2020) to process the sentences and generate dependency and constituency parses for later analysis. The pipeline consists of:

1. **Tokenization & Sentence Segmentation**
2. **Dependency Parsing:** We use the default parser (Chen and Manning, 2014) to dependency parse all text in our experiments.
3. **Constituency Parsing:** In addition to dependency parsing, we construct a constituency parse for each sentence, which is subsequently used to compute our metrics.

Rarely, a portion of the parsing pipeline would fail (for example, if the sentence in question was merely a set of hyperlinks in the case of some of the Wikipedia data). In that case, the data point would be excluded from our analysis. Details on exclusions can be found in Appendix A.

2.5 Metrics

For the majority of our experiments, we plot the distribution of binned scores so that we can compare the regenerated data to the human data for both domains. For all metrics, we observe approximately Gaussian distributions for both the original human data and the regenerated data. We plot the overall normal fit line, as calculated by the defaults in Seaborn (Waskom, 2021) using `matplotlib` (Hunter, 2007) for easy visual inspection.

Comparing the human data and the LLM-regenerated data, we isolate three observational signatures of domain shift that recur across metrics: the human and the regenerated distributions can differ in **mean**, they can differ in **variance**, and they

Metric		Signature		
Type	Domain	μ	σ	Long Tail
Flesch-Kincaid	news	$\approx$	$\searrow$	reduced
	wiki	$\nearrow$	$\searrow$	reduced
	ELI5	$\searrow$	$\searrow$	reduced
Dependency	news	$\nearrow$	$\searrow$	n/a
	wiki	$\approx$	$\searrow$	n/a
	ELI5	$\nearrow$	$\searrow$	n/a
Depth	news	$\nearrow$	$\searrow$	$\approx$
	wiki	$\nearrow$	$\searrow$	$\approx$
	ELI5	$\approx$	$\approx$	$\approx$
Yngve	news	$\searrow$	$\searrow$	reduced
	wiki	$\nearrow$	$\searrow$	reduced
	ELI5	$\searrow$	$\searrow$	reduced
Constituency	news	$\nearrow$	$\searrow$	reduced
	wiki	$\nearrow$	$\searrow$	reduced
	ELI5	$\approx$	$\searrow$	reduced

Table 1: Schematic description of Llama's distribution shift for our five investigated metrics relative to the human baseline for all text domains. μ refers to mean shift ($\nearrow$ refers situations when the mean of the distribution is higher for Llama than for the human), σ to the standard deviation of the distribution($\searrow$ refers to situations where the distribution is narrower for Llama than for the original), and 'long tail' to whether a heavy right tail that was present in the human distribution is reduced for the Llama distribution ('n/a' marks situations with no long tail in the human distribution of the metric).

can differ in the presence of a **long tail** (a heavy right tail). A schematic summarizing our results is in Table 1.

Flesch-Kincaid. First, we measure the Flesch-Kincaid grade level score, following Flesch (1948). This score pertains to an article overall, and is a standard metric in the education field and in NLP to estimate the reading level of a piece of text, with higher scores being more difficult. The Flesch-Kincaid score relies on words per sentence and syllables per word to derive an estimate of the ease of reading the text snippet. We calculate the Flesch-Kincaid scores using the `py-readability-scores` library[4], which relies on the Natural Language Toolkit (Bird et al., 2009). For each dataset, we take each article and calculate its Flesch-Kincaid grade level score, then we consider the scores for all articles as a distribution.[5]

2.5.1 Syntactic Metrics

For the other four metrics, we relied on syntactic parses, generated following the procedure we described above in §2.4. Unlike for the Flesch-Kincaid score, for all syntactic metrics, we calculate the result per sentence. We consider each syntactic metric as a distribution relative to domain and generation source (human vs. LLM).

Dependency Tags. Dependency tags provide a description of the relation between units in a sentence. For each sentence, we count the number of unique dependency tags.

Parse Depths. For each sentence, we count the depth of a constituency parse.

Yngve Scores. As a first qualification on parse depth, we also explore a measure of left vs. right branching parse trees. We measure each sentence's Yngve score (Yngve, 1960) following Roark et al. (2007), which argues that the Yngve metric is useful for diagnosing cognitive impairments. The score roughly corresponds to the deviation of a parse tree from a completely right-branching tree—it is the average number of left branches on the path from the root node to each leaf.

Constituency Labels. As a further qualification of parse depth, we also calculate the number of unique constituency labels in a sentence's constituency parse.

3 Results

We generally report the results for Llama-70B models in our figures, but more detailed figures that provide additional data for smaller models and/or models from the Mistral family are available in a number of appendices.

3.1 Descriptive Results

To situate our results, we first measure some basic dataset statistics. In Table 2, we present the statistics for the parse tree depth experiment. Recall that some datasets may slightly differ in size based on the success of metric calculation as described above in §2.5. Also, note that complexity metrics can be correlated with sequence length (Salkar et al., 2022).

Across the board, the regenerated data is similar to the original data in terms of words per sentence. For CCnews and Wikipedia, the regenerated

---

[4] https://github.com/cdimascio/py-readability-metrics/tree/master

[5] We observed that removing sentences and restructuring an article during the data cleaning stage leads to issues with readability score calculation. Therefore, we calculated readability scores without data cleaning for articles exceeding 100 words, as shorter articles lack sufficient content for accurate readability assessment.

Datasets	Articles	Sentences	Words	S/A	W/S	W/A
CCNews	0.6M	12.6M	0.3B	21.8	24.6	535.0
Llama-2-70B	0.7M	19.9M	0.5B	28.1	25.6	718.4
Llama-3.3-70B	0.7M	23.9M	0.7B	34.0	27.7	941.4
Mistral-24B	0.7M	22.7M	0.5B	32.1	23.9	765.1
Mistral-8B	0.7M	22.1M	0.5B	31.2	24.2	756.7
Llama-3.1-8B	0.7M	19.8M	0.5B	28.7	26.5	759.2
Wikipedia	6.4M	114.6M	2.9B	17.7	24.9	441.9
Llama-2-70B	6.6M	234.0M	5.4B	35.4	23.2	821.0
Mistral-24B	6.5M	307.7M	7.0B	47.5	22.6	1076.1
ELI5	0.6M	34.2M	0.6B	56.2	18.6	1047.3
Llama-3.3-70B	0.6M	9.3M	0.2B	15.2	20.4	310.6
Mistral-24B	0.6M	6.6M	0.1B	10.8	18.5	200.3
Mistral-8B	0.6M	6.1M	0.1B	10.2	18.7	191.3

Table 2: Descriptive statistics (average) on the datasets used for the parse tree experiments. Eligible sentences meet the following criteria: sentence length falls within 3-500 words, and contains 1 $\leq$ verb or auxiliary verb.

data contains more sentences per article, and hence more words per article than the original data. For ELI5, conversely, the regenerated data contains fewer sentences per article, and hence fewer words per article than the original. Due to our data cleaning (§2.3), there is some difference in the number of articles preserved for analysis between generation sources (LLM v. human), with slightly more regenerated articles being analyzed. We plot the distribution over sentence lengths in Figure 14 in Appendix C. We observe that the regenerated data shows a shifted mean, a reduction in variability and a reduced long tail, when compared to the original human data for each domain.

3.2 Flesch-Kincaid Scores

For this simplification metric, shown in Figure 1, we observe that all distributions are roughly normal and the three signatures are present. We observe that Flesch-Kincaid readability scores for the human data deviate slightly from the normal distribution in that they have a right tail, as indicated by the fact that the bars around the center are somewhat above the fit curve on the left, and somewhat below the fit curve on the right. As compared to their human-generated variants, each regenerated dataset has a narrower distribution, and a reduced long tail. For CCNews and Wikipedia, we observe an upward shifted mean, while the mean for ELI5 is shifted downwards. In Appendix D, we plot the full results for all tested models in Figure 15; we also report the means, medians, and standard deviations for this metric in Table 5, alongside other readability metrics.

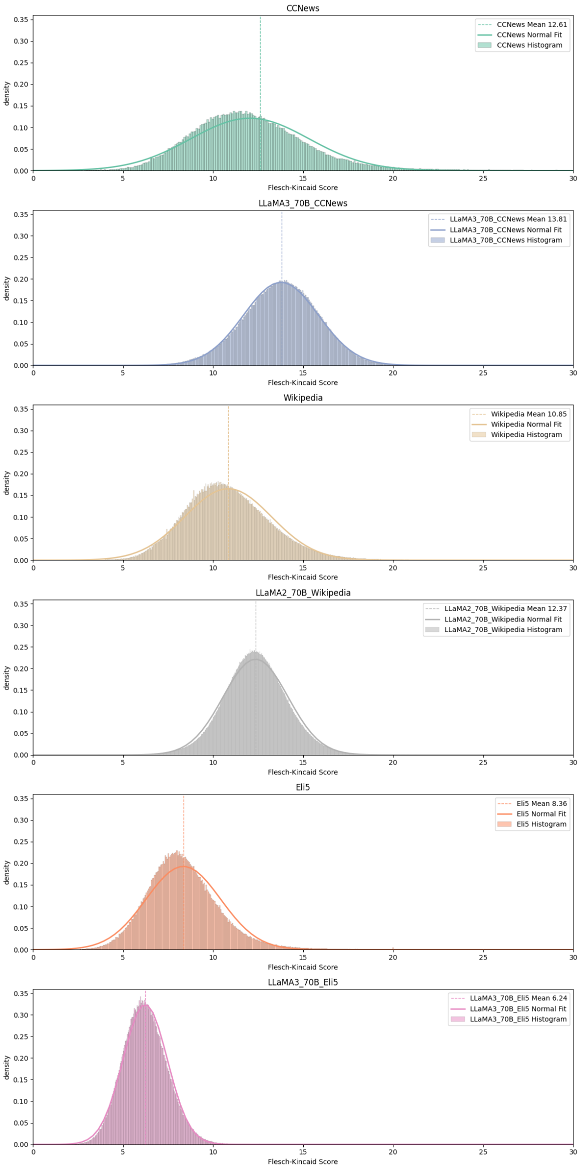

Figure 1: Flesch-Kincaid readability score distributions. Normal fit curves informally illustrate the fact that regenerated data appears to be normally distributed, with narrower distributions and a reduced long tail on the right relative to the human datasets.

3.3 Dependency Tags

As Figures 2–4 show, the regenerated data in each domain is more narrow and has a slightly upward shifted mean relative to the human data. The original human data is close to normally distributed, so we do not report a reduced long tail for this metric.

3.4 Depth Scores

Normal fits for depth scores are provided in Figure 5. For Wikipedia, and CCNews, the parse tree depth of the regenerated data shows a narrowing effect. Interestingly, the regenerated data for ELI5 is

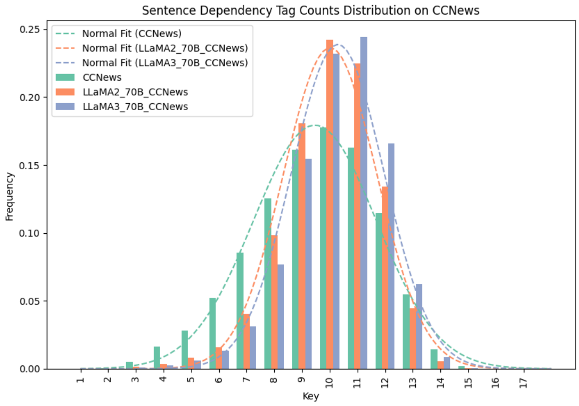

Figure 2: Proportion of sentences in CCNews that have a particular number of unique dependency tags.

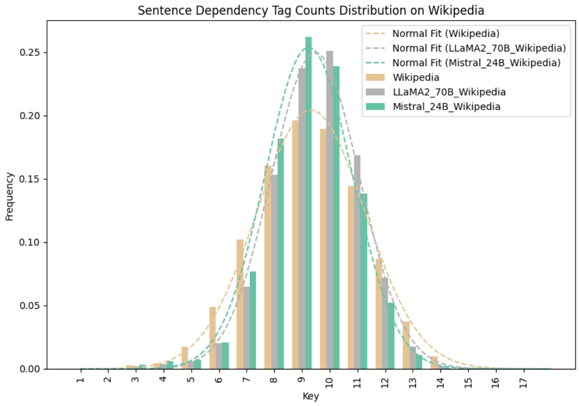

Figure 3: Proportion of sentences in Wikipedia that have a particular number of unique dependency tags.

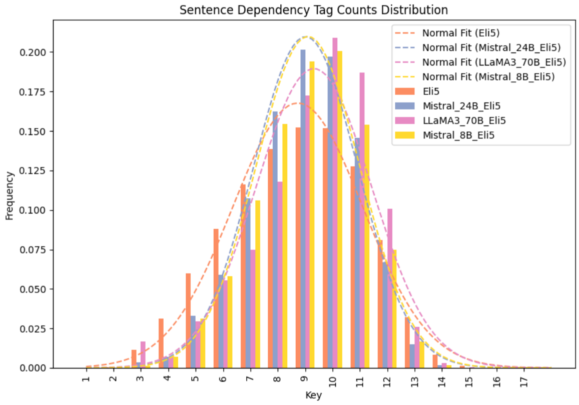

Figure 4: Proportion of sentences in ELI5 that have a particular number of unique dependency tags.

close in variance to the original data across models. This is the only time in our study that we do not find a clear narrowing effect. For the Llama-70B-regenerated data, we find a higher mean across all domains. Across domains and models (with the exception of Mistral-24B's CCNews and Wikipedia data), the slight right tail is reproduced in the regenerated data (see Figure 16 in Appendix E).

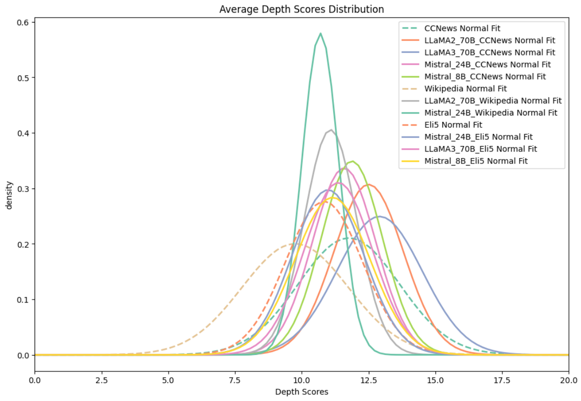

Figure 5: Average depth score normal fits. Dotted lines indicate human original domains, and solid lines indicate model regenerations.

3.5 Yngve Scores

Yngve scores for Llama-regenerated data are shown in Figure 6. In all three domains, we see considerably more narrow distributions in the Llama-generated data, and a shorter, less heavy long right tail. Nonetheless, the Llama-regenerated data shows a considerable right tail. The mean is increased relative to the human mean in the case of Wikipedia, while in the CCNews data, and ELI5, the mean is slightly lower. In Appendix F, we plot the full results for all tested models in Figure 17, where data regenerated with other models show the same trends.

3.6 Constituency Labels

Figures 7–9 show the distribution of unique constituency labels per sentence. Across domains and models, we find that the regenerated data has a narrower distribution. The human distributions exhibit a slight right tail that is largely absent with in the regenerated data. In the regenerated CCNews, the smaller Mistral models have a slightly downshifted mean, while the Llama models shift their mean upwards. For Wikipedia, and ELI5, the models all shift their mean upwards.

3.7 Summary and Interpretation of Results

Across the majority of our metrics and datasets, we observe that models generate distributionally less diverse data, as evidenced by both a decrease in variability and a reduced long tail. Both of these signatures are compatible with the interpretation that the models are simplifying relative to the human domain: if they were generating syntactically

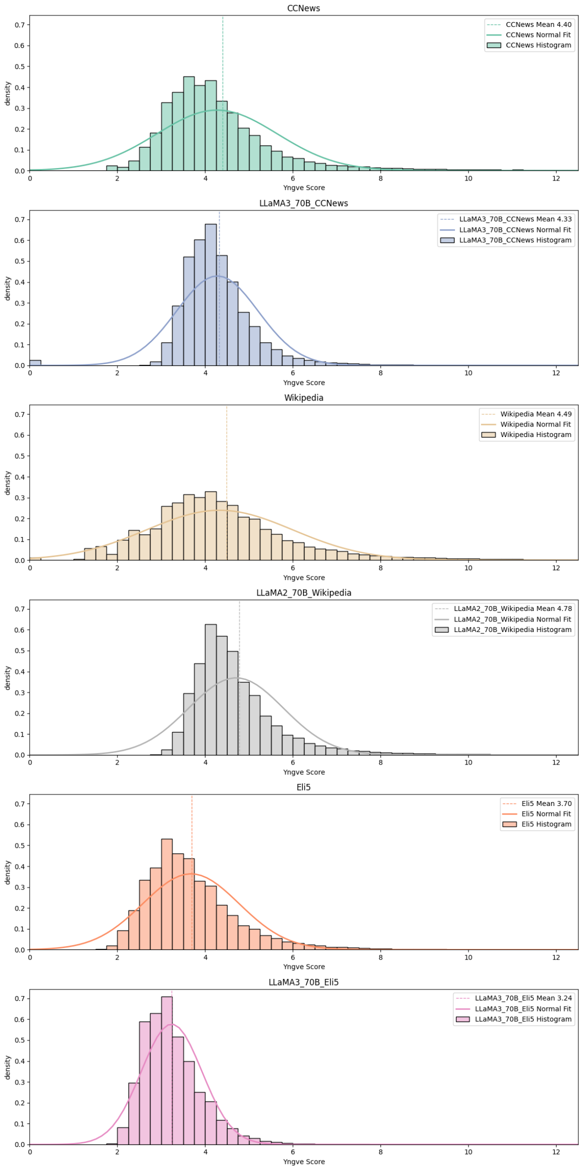

Figure 6: Yngve parse complexity score distribution for each dataset. Normal distribution curves fit to the data show that both human and Llama-regenerated datasets are not particular normally distributed. Llama-regenerated datasets show a narrower distribution than human with a heavy right tail that is reduced in comparison to the human datasets (but still visibly present).

simpler sentences overall, this could be underlying the lower variability, and if they were failing to capture rarer syntactic phenomena, or rarer combinations of syntactic phenomena, this might account for the reduced long tail.

For example, the long tail of Yngve scores (Figure 6) would contain strongly left branching structures that are grammatical but rare in English (e.g., clausal subjects). Many linguists take such structures to be evidence of the recursive nature of the combinatorial system that underlies human languages, and as such a loss of the long tail is compatible with syntactic simplification by the LLMs.

Interestingly, even for the unique dependency tags metric, where we report no long tail reduction, because the human data is very close to normally distributed, the regenerated data deviates from the human data. Across our other metrics, the human data is more left-leaning (exhibits a right tail) than the regenerated data. Even here, where the human data is very close to normal, the regenerated data leans right of that, thus showing the same directional difference as elsewhere.

However, while the model-regenerated data is uniformly less diverse than the corresponding human-generated texts across our variety of measures (reduced variability and a reduced long tail), the mean shifts vary across domains: For CCNews and Wikipedia, the Llama-generated data is mostly shifted upwards, while the mean of the regenerated ELI5 tends to match or be shifted downwards relative to the original data. Since ELI5 is presumably more simple than the overall average training data (while CCNews and Wikipedia may either match it or be more complex), the direction of mean shifts in the regenerated data suggests that models overshoot their domain complexity. In other words, in terms of the mean of our complexity measures (but not in terms of diversity) the models appear to regenerate domain data that is a more extreme form of the human original (simplifying in the case of the simpler ELI5 domain, and shifting the mean complexity upwards in the case of Wikipedia and CCNews).

What we don't see is all regenerated data consistently landing at some kind of middle ground—a "neutral" domain, that would reflect the properties of some average of all its training data—regardless of which domain the model is prompted to match. Since this isn't present, we can conclude that the models do have some notion of domain and encode the fact that domains differ. However, the data that models regenerate are consistently less diverse, and in terms of the mean of our complexity measures, they tend to be more extreme than the human originals, suggesting that the models' notion of domain is not fully humanlike.

4 Qualitative Reflections on LLM-Regenerated Data

Llama-V2 Regenerated Wikipedia. We observed several trends when manually inspecting

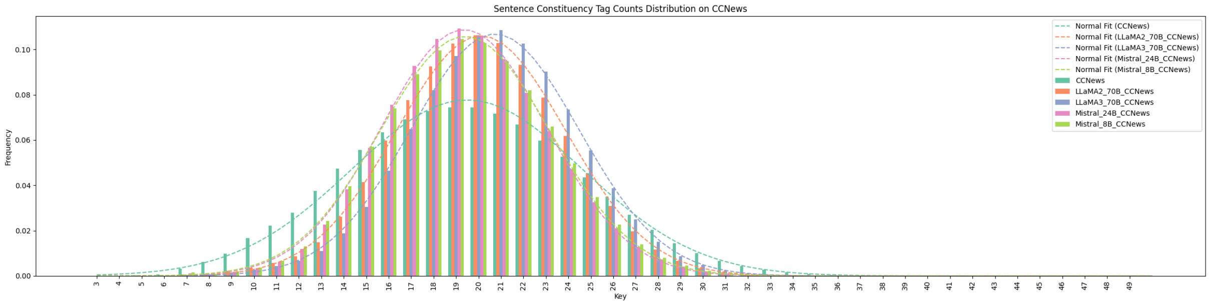

Figure 7: Proportion of sentences in CCNews (y-axis) that have a particular number of unique constituency labels (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3.

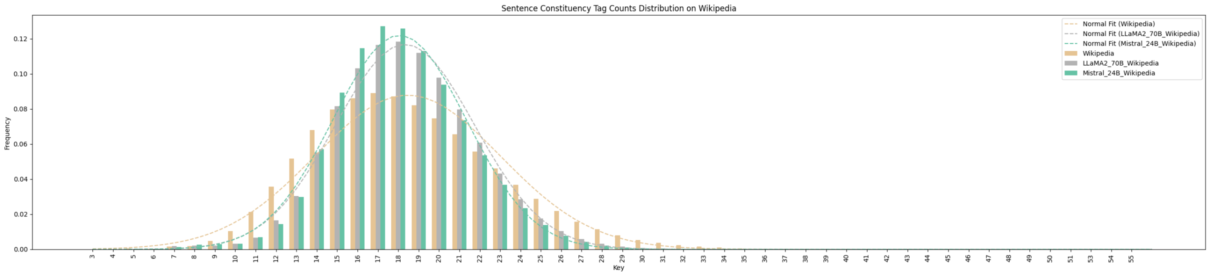

Figure 8: Proportion of sentences in Wikipedia (y-axis) that have a particular number of unique constituency labels (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3.

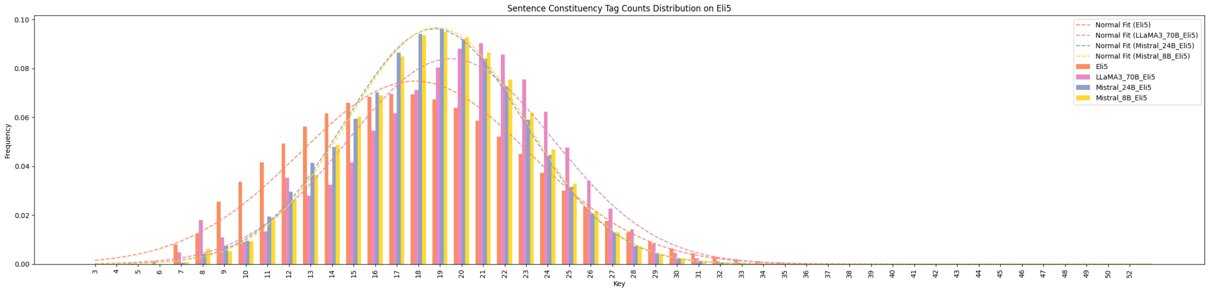

Figure 9: Proportion of sentences in ELI5 (y-axis) that have a particular number of unique constituency labels (x-axis).

the data regenerated by the LLMs. When comparing Llama-V2-regenerated Wikipedia articles to the original human ones, we observed spelling normalizations (e.g. British *-ise* becomes American *-ize*), increased inclusion of value judgments (which go against Wikimedia's editorial guidelines called the "Neutral Point of View"[6])—and an increased prevalence of essay-like wrap-up sentences.

For example, the final sentences of the Llama-V2 regenerated article on "A" are both explicitly concluding and unusually complimentary: *In conclusion, the letter* a *is an important and versatile letter in the English language. It is used as an indefinite article, a pronoun, a prefix, a suffix, and in many abbreviations and acronyms.*[7] As may be clear from this example, the models introduce stylistic elements that are not generally in keeping with the style of the original Wikipedia domain. More examples are provided in Appendix G. These observations point the way to future work that explores the consequences of the domain regeneration paradigm on stylistic elements.

---

[6] https://meta.wikimedia.org/wiki/Neutral_point_of_view

[7] Note also that it wrongly asserts that the letter can be used as a pronoun (a property that indefinite articles like English "a" do indeed have in other languages like German, but not in English) and a suffix, of which the text alleges elsewhere that it denotes the performer of an action (peculiarly, in non-rhotic variants of English, the Latinate plural *-a* may be homophonous with the agent nominalization *-er*, which does indeed denote the performer of an action).

Llama-Regenerated CCNews. Llama-V2 and Llama-V3 were both prone to inserting a higher number of quotations attributed to famous or influential people than the original articles, which largely described an event. For example, we saw novel inclusions of quotations attributed to Jeff Gundlach, Warren Buffet, Mark Schneider, Zhang Yuhua, Chen Qi, and David Cameron. In the rare case where the original CCNews articles contained quotations, usually only one quote was present. It would be fairly long in comparison to Llama-V2 regenerated CCNews quotes. There were also a number of cases of Llama-V2 inserting acronyms where none had existed in the source (e.g. *Albuquerque Little Theatre (ALT)*).

CCNews regenerated by Llama-V2 and Llama-V3 also displayed unusual wrap-up sentences, except, unlike for Wikipedia, they appeared to be more PR or sales related. For example, the article on “ARKit 1.5” ended with *Whether you’re a tech enthusiast, a developer, or simply someone interested in the future of technology, ARKit 1.5 demos are certainly worth keeping an eye on.* More examples and comparisons between Llama-V2 and Llama-V3 are present in the Appendix H.

5 Related Work

Closest to our work is Shaib et al. (2024), which explored sequences of part-of-speech tags in training data and model generations. They reported that several LLMs generated more syntactically homogeneous text, as compared to human ground truth. Our work differs from theirs in that we focus on different datasets and models, and perform distinct experiments. We take a distributional view and are interested in the domain match setting, exploring additional signatures of model-and-human difference and more syntactic metrics. They explore neutral domain text, diving deeper into the effect of decoding temperature, and also exploring the additional summarization setting.

6 Conclusion

Using our regeneration paradigm, we have uncovered systematic syntactic differences between human-generated and model-regenerated text. Across a variety of syntactic complexity metrics, the regenerated text showed lower variability as well as a reduced long tail, when compared against the human-generated text in the same domain, while the mean of the measurements was often shifted in a way that suggests that models overshoot when trying to match domain properties.

Our results may have practical implications (i.e. on decisions about whether or not to utilize LMs as components in domain transfer systems) and theoretical implications (e.g. about the empirical status of syntactic long tail effects).

Acknowledgments

We would like to thank Chantal Shaib, Jessica Forde, Candace Ross, and Levent Sagun for conversations relating to model collapse. We’d also like to thank Sebastian Ruder for feedback on an early draft, and the audience of the Organized Session on LLMs, Linguistics, and Psycholinguistics at the 2025 Annual Meeting of the Linguistic Society of America in Philadelphia, PA for comments and questions that helped us shape the final paper.

7 Limitations

Tooling and Pipeline. While our visual inspections didn’t surface any immediate issues, we acknowledge the possibility of tooling failures when we try to calculate metrics or parse sentences that are extremely long or complicated. However, since we uniformly apply our tools across domains and generation sources, we expect any errors to be comparable, and thus not to have an outsized impact on our results.

Decoding Temperature. In this work, we used the default temperature from vLLM. We presume that lowering the temperature would decrease randomness, presumably further reduce diversity, and higher temperature could increase diversity, but it is not immediately clear what the effect would be on the reduction of the long tail. A more thorough exploration of decoding temperature could be explored in future work.

Syntactic Complexity Metrics. In this work, we utilized existing complexity metrics from prior literature. However, we have anecdotally observed some additional changes to the style and content, which one could devise metrics to specifically target. Future work could perform more data analysis to help guide the creation of additional informative syntactic complexity metrics, which, in turn, could help us gain more insights into the type of simplification LLMs affect, and inspire architectural or training improvements.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Bache, David Newman, and Padhraic Smyth. 2013. Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 23–31.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gall'e, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, G'erard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Mu noz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Tas ar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Peri nán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Incorporated.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3651–3661, Online. Association for Computational Linguistics.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: Classification of genres in text. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*.

Chrysanne DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–500.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. 2024. Strong model collapse.

Liat Ein-Dor, Ariel Gera, Orith Toledo-Ronen, Alon Halfon, Benjamin Sznajder, Lena Dankin, Yonatan Bilu, Yoav Katz, and Noam Slonim. 2019. Financial event extraction using Wikipedia-based weak supervision. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 10–15, Hong Kong. Association for Computational Linguistics.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Angela Fan and Claire Gardent. 2022. Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Robert Gunning. 1952. The technique of clear writing.

Sil Hamilton. 2024. Detecting mode collapse in language models via narration. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 65–72, St. Julian's, Malta. Association for Computational Linguistics.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 458–463, Baltimore, Maryland. Association for Computational Linguistics.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Da Ju, Karen Ullrich, and Adina Williams. 2024. Are female carpenters like blue bananas? a corpus investigation of occupation gender typicality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4254–4274, Bangkok, Thailand. Association for Computational Linguistics.

Tom S. Juzek and Zina B. Ward. 2024. Why does chatgpt "delve" so much? exploring the sources of lexical overrepresentation in large language models.

Marcus Klang and Pierre Nugues. 2019. Docria: Processing and storing linguistic data with Wikipedia. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 400–405, Turku, Finland. Linköping University Electronic Press.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization.

David Lee. 2002. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. In *Teaching and learning by doing corpus analysis*, pages 245–292. Brill.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

AI@Meta Llama Team. 2024. The Llama 3 Herd of Models.

Jian Ni and Radu Florian. 2016. Improving multilingual named entity recognition with Wikipedia entity type mapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1275–1284, Austin, Texas. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text.

Nikita Salkar, Thomas Trikalinos, Byron Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 341–350, Online only. Association for Computational Linguistics.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. 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.

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.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

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.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

A Ablation of data cleaning

In this section, we also present our results for less filtered data on two of our domains, Wikipedia and CCNews. Overall, we see the same trends as for the filtered data presented in the main paper. For dependency tags and constituency labels for both datasets: we see mean shift (mostly for CCNews) and narrowing (for both datasets). Descriptive statistics following the length and POS filtration are presented in Table 2. Our overall cleaning and processing pipeline (including length filtering) excluded on average less than 10% in the case of CCNews datasets, and less than 15% in the case of Wikipedia datasets.

Datasets	Sentences	Words	S/A	W/S	W/A
CCNews	13.8M	0.3B	23.9	23.2	554.3
LLaMA2 70B	20.6M	0.5B	29.1	24.9	724.9
LLaMA3 70B	24.5M	0.7B	34.6	27.3	946.4
Wikipedia	129.5M	2.9B	20.0	22.5	450.4
LLaMA 2 70B	277.5M	5.6B	42.0	20.3	854.9

Table 3: Descriptive statistics on raw data.

Datasets	Sentences	Words	S/A	W/S	W/A
CCNews	13.6M	0.3B	23.4	23.5	551.3
LLaMA2 70B	20.4M	0.5B	28.8	25.1	724.4
LLaMA3 70B	24.4M	0.7B	34.5	27.4	946.2
Wikipedia	122.8M	2.9B	19.0	23.6	448.1
LLaMA 2 70B	257.3M	5.6B	39.0	21.8	850.2

Table 4: Descriptive statistics after length filtration.

Readability scores were consistently calculated on entire articles without any data cleaning. We compute readability scores for articles containing more than 100 words, as shorter articles do not provide sufficient indicators of readability. Metrics such as the depth score and Yngve score, which were initially aggregated at the article level, showed minimal variation upon inspection. Therefore, they are not included in the ablation results presented here. For reference, we provide results on dependency and constituency parsing using the raw data below.

Main Results Figures Ablations. Figures 10–13 present data ablations on our main results. Overall, we see the same rough trends as described in the main paper.

For Figure 11, we can see a difference in the dependency parses for fewer than three words between the human-generated data and the Llama-V2-generated data. We decided to filter out these lengths, because we expected the data to be noisy and uninformative about syntactic structure (there are very parses available for complete sentences with 3 or fewer words). We observe a similar trend for Figure 13, except that the noisy section extend to approximately 6 constituency labels, which is reflective of the same underlying fact that there are not many available parses for 3 words.

B Prompts

We prompted the LLMs using the two prompts below, one for each text domain. We retrieved the {topic} and {title} respectively from the original data sources and fed them into the model prior to including the initial section of text for the model to attempt to regenerate. Next, we included instructions that match standard instruction-tuning prompts, which also included a target article length in words, which we set to the average length of article from each domain. We observed that the regenerations were never word-for-word identical to the human versions, and we also observed that the models did not generate exactly the average lengths provided in the prompts.

B.1 Wikipedia

```
NUM_FIRST_PARA_LENGTH = 256
TEXT_PROMPT = """
Generate a Wikipedia article on the topic of {topic}.
Use the following first paragraph from the original Wikipedia article as a starting point:

{first_paragraph}

Now, expand upon the provided paragraph by providing additional details ,
historical context , notable events , key figures , and any relevant subtopics.
Aim for a well-structured and informative Wikipedia style article with a minimum length of 700 words.
Ensure that the content is factually accurate , well-written , and on Wikipedia writing style.
"""
```

B.2 CCNews

```
NUM_FIRST_PARA_LENGTH = 180
TEXT_PROMPT = """
Generate a news article on the topic of {title}.
Use the following first paragraph from the original news article as a starting point:
```

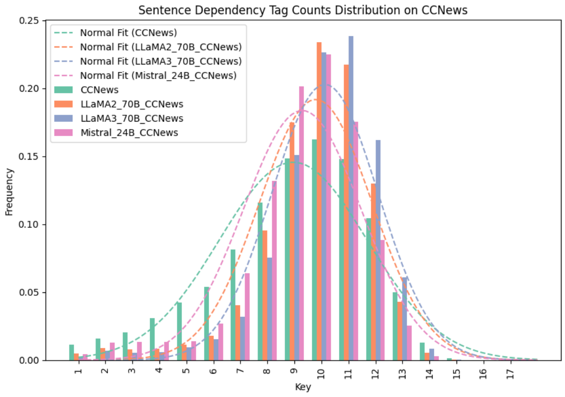

Figure 10: Proportion of sentences in CCNews (y-axis) that have a particular number of unique dependency tags (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3. This comparison was made on data without any cleaning.

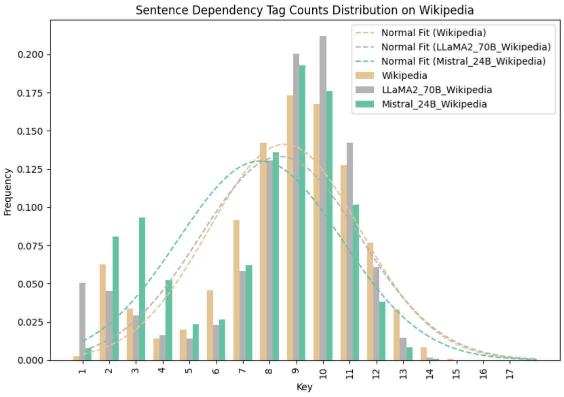

Figure 11: Proportion of sentences in Wikipedia (y-axis) that have a particular number of unique dependency tags (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3. This comparison was made on data without any cleaning.

```

{first_paragraph}

Now, expand upon the provided paragraph by providing additional details, context, notable events, key figures, and any relevant discussions. Aim for a well-structured and informative news style article with a minimum length of 500 words. Ensure that the content is factually accurate, well-written, and on news writing style.
"""
```

B.3 Eli5

```
TEXT_PROMPT = """
Generate a reddit reply to this thread {title}.

Aim for an Explain Like I'm Five style reply with a minimum length of 100 words. Ensure that the content is factually accurate, well-written, and on Explain like I'm Five writing style.
"""
```

C Sentence Lengths

As Figure 14 illustrates, no length distribution is perfectly normal. When compared to the Llama-70B regenerations, We observe that the original human distributions have shorter sentences on average (i.e., the regenerated distributions have upward shifted means) for for CCNews and ELI5, but longer sentences for Wikipedia. We also observe that the original human distributions are also wider (i.e., the regenerated distributions have less variance) for all data sources (with ELI5 being the weakest effect, likely because more generations are at the length floor). Finally, we observe that the original distributions appear to have a longer and heavier right tail than their model generated counterparts for all data sources. For the other models, the mean shift is inconsistent across models, but all models show reduced variability, and a reduced long tail.

D More Information on Readability

D.1 Full results for the Flesh-Kincaid Readability Scores for all tested models.

We present full results for Flesh-Kincaid scores in Figure 15.

CCNews. For the CCNews datasource, all model regenerated datasets have upward shifted means and narrowed distributions, in keeping with the subset presented in the main paper. They are all additionally more left than right tailed, when compared to the original human distribution.

Wikipedia. For the Wikipedia datasource, Llama-V2-70B and Mistral-24B both had upward shifted mean, a narrower distribution, and a reduced right tail. Interestingly, the Mistral model has a near-perfectly normal distribution, while Llama-V2-70B retains a slight right tail.

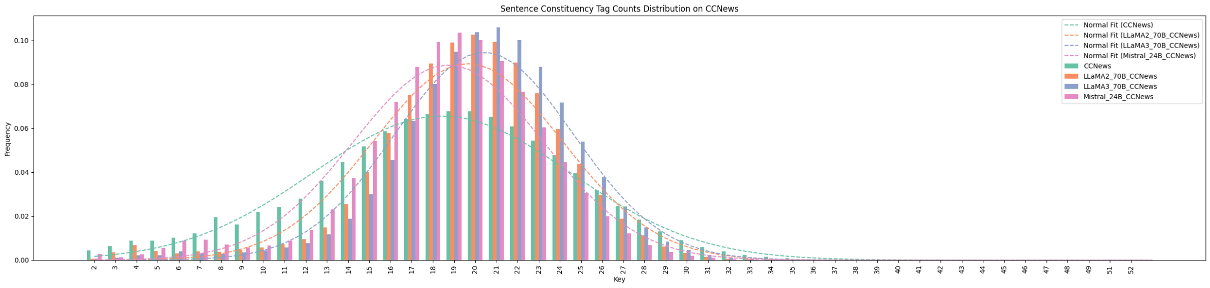

Figure 12: Proportion of sentences in CCNews (y-axis) that have a particular number of unique constituency tags (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3. This comparison was made on data without any cleaning.

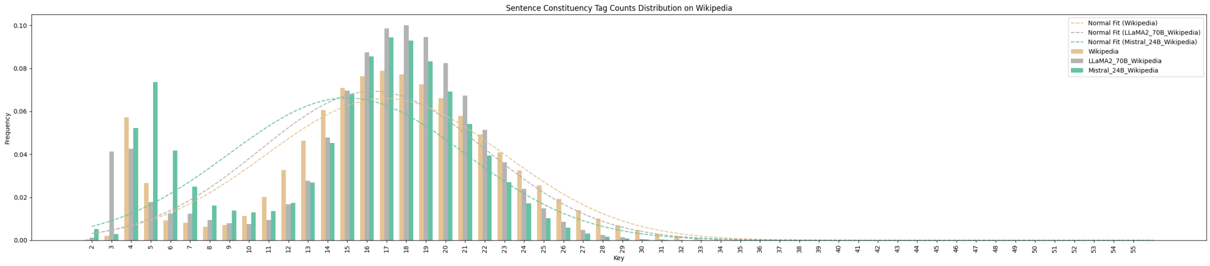

Figure 13: Proportion of sentences in Wikipedia (y-axis) that have a particular number of unique constituency tags (x-axis). Colors indicate whether the distribution belongs to the original source data, or the source domain as regenerated by Llama-V2 or Llama-V3. This comparison was made on data without any cleaning.

ELI5. For the ELI5 datasource, all model-regenerated distributions have a downward shifted mean, a narrower distribution and a strongly reduced right tail.

D.2 Additional Readability Scores

We also report the means, medians, and standard deviations of several readability scores, including Flesch-Kincaid grade level (see Table 5), for Wikipedia and CCNews domain data. We are using all 70B models for this analysis. We expect most of these metrics to correlate highly—as all of them are based on different combinations of words per sentence and syllables per word—but are including them in case they may be of interest to some readers. Across all readability metrics (means and medians), human-generated Wikipedia is deemed the simplest datasets, and Llama-V3-CCNews is deemed the most complex. Standard deviations are generally within the same range for all metrics, except for human-generated CCNews which has the most variation in readability. Second in highest standard deviations across scores is human-generated Wikipedia (Flesch-Kincaid Reading Ease, Linsear Write) or Llama-V3-generated CCNews (Gunning-Fog Index, Spache Readability Formula, Linsear Write). The fact that human-generated data has high standard deviations across the board (although occasionally in a tie with Llama-V3-CCNews) suggests the human distributions have more diversity in generations.

Flesch-Kincaid Reading Ease. We report the Flesch-Kincaid Reading Ease metric, which is similar to the Flesch-Kincaid grade level scoring in that it is calculated from number of words and syllables per sentence. A higher score indicates that the material is easier to read. Human-generated Wikipedia has the highest reading ease (50), and Llama-V3-generated CCNews has the lowest (39.5), but both fall into the range expected for college level texts. The Flesch-Kincaid Reading Ease are presented in Table 6.

Gunning-Fog Index. The Gunning-Fog index (Gunning, 1952) is another estimate of reading level, which is also based on the number of words per sentence and the number of syllables per word, but it generally has a lower value than the Flesh-Kincaid grade level and reading ease scores. The Gunning-Fog scores are presented in Table 7.

Linsear Write Scores. The Linsear Write Scores are something of a thresholded version of the other scores, where the words with more syllables are deemed "challenging" and words with fewer syllables are deemed "easy". The Linsear Write Scores are presented in Table 8.

Spache Readability Formula. The Spache Readability Formula (Spache, 1953) operates on a list of words that are expected to be familiar for children up until the fourth grade in the United States. The formula considers average sentence length and proportion of familiar words to determine its score. Of all the metrics reported, this score resolves the least differences between datasets. The Spache Readability Formula scores are presented in Table 9.

Dataset	Mean	Median	STD	Sample Size
CCNews	12.6	11.8	8.2	561167
Llama-V2	12.7	12.8	2.1	708011
Llama-V3	13.8	13.8	3.2	702530
Wikipedia	10.8	10.7	2.7	3829535
Llama-V2	12.4	12.3	2.1	6601865

Table 5: Flesch-Kincaid Grade Level

Dataset	Mean	Median	STD	Sample Size
CCNews	48.2	50.4	24.9	561167
Llama-V2	43.4	43.2	11.3	708011
Llama-V3	39.5	39.4	12.8	702530
Wikipedia	50.0	51.1	12.8	3829535
Llama-V2	40.8	41.0	10.7	6601865

Table 6: Flesch-Kincaid Reading Ease

Dataset	Mean	Median	STD	Sample Size
CCNews	14.4	13.5	8.5	561167
Llama-V2	15.2	15.3	2.3	708011
Llama-V3	16.5	16.5	3.5	702530
Wikipedia	12.0	11.7	3.0	3829535
Llama-V2	14.2	14.2	2.4	6601865

Table 7: Gunning-Fog Index

Dataset	Mean	Median	STD	Sample Size
CCNews	16.6	15.1	14.1	561167
Llama-V2	15.9	15.9	2.6	708011
Llama-V3	17.6	17.5	4.3	702530
Wikipedia	12.9	12.7	4.2	3829535
Llama-V2	14.3	14.3	3.3	6601865

Table 8: Linsear Write Scores

Dataset	Mean	Median	STD	Sample Size
CCNews	8.1	7.8	3.2	561167
Llama-V2	7.7	7.7	0.7	708011
Llama-V3	8.0	8.0	1.2	702530
Wikipedia	7.6	7.6	1.0	3829535
Llama-V2	7.6	7.6	0.7	6601865

Table 9: Spache Readability Formula Scores

E Constituency Parse Depths Scores

Figure 16 shows the distribution of constituency parse depths. Depth scores for model-regenerated text have an upwardly shifted mean, and a more narrow distribution when compared to human-generated text.

F Full Yngve Score Distribution Plot

The full results for our Yngve metric on all models and domains can be found in Figure 17.

G Example regenerations from Llama-V2 on the Wikipedia Domain

We manually inspected several examples from Llama-V2 and Llama-V3 and noticed some trends. Several of these trends are also listed in the supplementary materials of Russell et al. (2025), which gives tips and tricks for detecting model generated content; these include stylized and formalized conclusions, and spelling normalizations.

Spelling Normalization. When the original article contained British spellings (e.g. *-ise* as opposed to *-ize*), Llama-2 did not uniformly preserve them, preferring American spellings for "organize", "emphasize", "decentralize".

Longer snipped of the Llama-regenerated Wikipedia article "A". *In addition to their use as indefinite articles, "a" and "an" can also be used as pronouns to refer to a noun or noun phrase that has been previously mentioned. For example: "I saw a cat in the park." (meaning a specific cat that was seen in the park) "I gave an apple to my friend." (meaning a specific apple that was given to a friend) In English, the letter "a" is also used as a prefix to indicate negation or absence, as in "a-" or "non-". For example: "Atheist" means someone who does not believe in God. "Amoral" means someone who does not have moral principles. "Asexual" means someone who does not experience sexual attraction. The letter "a" is also used as a suffix to indicate a* *noun that refers to a person or thing that performs a particular action or has a particular characteristic. For example: "Artist" means someone who creates art. "Athlete" means someone who is skilled in sports or physical activity. "Actor" means someone who performs in movies, television shows, or theater productions. The letter "a" is also used in many abbreviations and acronyms, such as "A.M." (meaning "ante meridiem" or "morning"), "P.M." (meaning "post meridiem" or "afternoon"), "A.S." (meaning "associate degree"), "B.A." (meaning "Bachelor of Arts"), "M.A." (meaning "Master of Arts"), and "Ph.D." (meaning "Doctor of Philosophy"). In addition to its use in English, the letter "a" is also used in many other languages, including Spanish, French, German, Italian, Portuguese, and many others. In these languages, "a" is used to represent a variety of vowel sounds, including the sounds /a/, /ɑ/, /ə/, and /ä/. In conclusion, the letter "a" is an important and versatile letter in the English language. It is used as an indefinite article, a pronoun, a prefix, a suffix, and in many abbreviations and acronyms. Its origins can be traced back to the Phoenician alphabet, where it represented a glottal stop, and it has evolved over time to represent a variety of vowel sounds in different languages.*

Conclusion summaries and value judgments. The regenerated article on "Albedo" concluded with *By understanding albedo, scientists and policymakers can make informed decisions about environmental management, urban planning, agriculture, and other fields that affect human life and the environment.*

"Abraham Lincoln": *In conclusion, Abraham Lincoln was a strong leader who fought to preserve the Union and end slavery. His eloquence, political skills, and commitment to democracy and freedom have made him a beloved figure in American history. His legacy endures, and he remains an icon of American values and ideals.*

"Alabama": *From its Native American roots to its current status as a thriving Southern state, Alabama has contributed to the nation's cultural, economic, and political landscape. Its people, events, and landmarks continue to inspire and captivate the nation's attention, and its legacy will undoubtedly endure for generations to come.*

"Achilles": *Achilles' legacy continues to live on, and his story will continue to inspire and captivate audiences for generations to come.*

"Aristotle": *Aristotle's legacy is a testament to his profound impact on human thought and culture, and his ideas will continue to shape our understanding of the world for generations to come.*

"Academy Awards": *Whether it's the red carpet glamour, the thrill of winning an Oscar, or the memories of unforgettable moments, the Academy Awards continue to capture the hearts and imaginations of people around the world.*

H Example regenerations from Llama-V2 on the CCNews Domain

Conclusion summaries and sales pitches. The Llama-V2 regenerated article on "China launches 1st X-ray Space Telescope" concludes with *In conclusion, the launch of China's first X-ray space telescope, 'Insight', marks a significant milestone in the country's space exploration history. The telescope is expected to provide valuable insights into some of the most complex and fascinating phenomena in the universe and push the boundaries of detection technology. The launch is a testament to China's commitment to advancing the field of astronomy and its growing influence in the global space community.*

An article on Nestlé being optimistic that ingredient prices will fall concludes with *In conclusion, Nestle's optimism about its ability to offset the impact of rising ingredient prices and meet its sales growth target is a positive sign for the company's future prospects. The company's diversified portfolio, geographic presence, and commitment to sustainability will help to mitigate any negative impact from economic uncertainty and drive long-term growth. As the company continues to invest in premium and healthier products, as well as sustainable sourcing, it is well-positioned to meet the changing needs of consumers and deliver on its growth targets.*

"Magical Show Mary Poppins to Wrap Albuquerque Little Theatre's 87th Season": *The ALT production of Mary Poppins runs until June 12th, with performances on Fridays, Saturdays, and Sundays. Tickets are available through the ALT website or at the box office. Don't miss your chance to experience this magical show and be a part of the ALT's 87th season.*

"Rutherford Co. deputies searching for missing woman - FOX Carolina 21": *The disappearance of Patricia Murray has sent shockwaves through the community, and the Rutherford County Sheriff's Office is committed to doing everything they can to* *bring her home safely. As the search continues, the community remains hopeful that she will be found soon, and that she will be reunited with her loved ones.*

"Southfields Residential Care: A Comprehensive Guide": *In conclusion, at Southfields Residential Care, we understand the importance of providing a safe, supportive, and welcoming environment for our residents. We strive to create a place where people can live comfortably and safely, while still maintaining their independence and individuality. If you're looking for a care home that truly cares, look no further than Southfields Residential Care.*

I Comparison of original and regenerated CCNews

We show a direct comparison of human-generated, Llama-V2-generated, and Llama-V3-generated CCNews data in Table 10. Anecdotally, the human-generations appear less similar to the Llama ones than the Llama ones appear to each other.

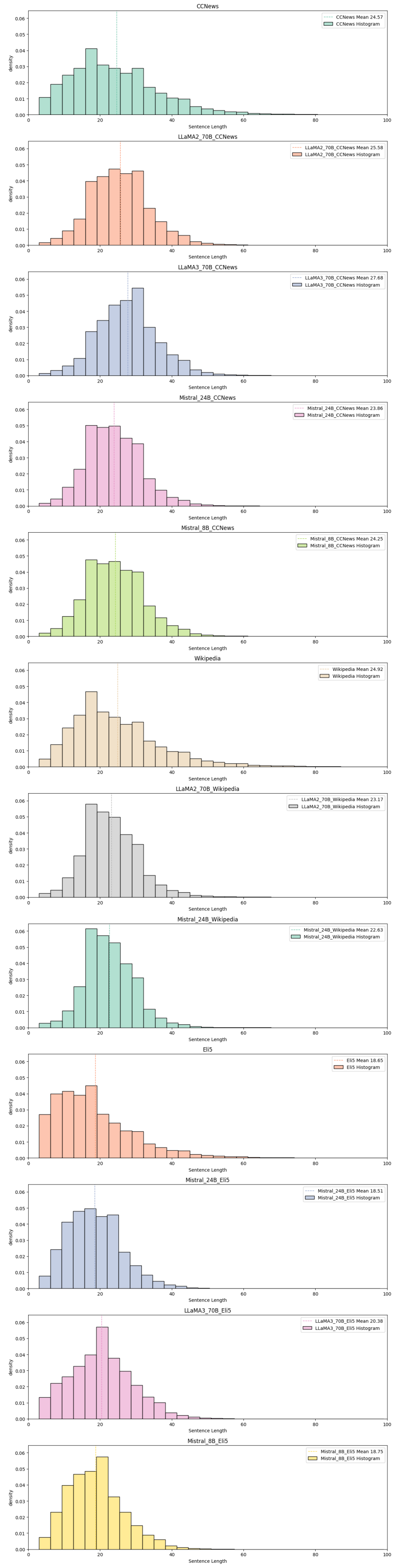

Figure 14: The full plot of sentence length for all the models.

Dataset	Prompt Title	Text
Human	ARKit 1.5 Demos Showing Up Online	*iOS 11.3 developer beta has been out for a few days and we’re seeing some ARKit 1.5 examples showing up on YouTube and Twitter. The demos are rudimentary right now, but will likely improve as developers get more time with Apple’s updates augmented reality technology. One of the big new features in ARKit 1.5 is vertical plane, or wall, detection. That’s what the early demos are focusing on.*
Llama-V2	ARKit 1.5 Demos Showing Up Online	*With its improved wall detection, lighting estimation, and other features, ARKit 1.5 is set to revolutionize the way we interact with the world around us. As developers continue to explore the potential of this technology, we can expect to see even more innovative and practical applications of AR in the months and years to come. Whether you’re a tech enthusiast, a developer, or simply someone interested in the future of technology, ARKit 1.5 demos are certainly worth keeping an eye on.*
Llama-V3	ARKit 1.5 Demos Showing Up Online	*With its introduction of vertical plane detection, ARKit 1.5 is set to revolutionize the world of augmented reality, enabling developers to create more immersive, interactive, and engaging experiences. As the technology continues to evolve, it will be exciting to see how developers and companies harness its potential to transform industries and revolutionize the way we interact with the world around us. With the release of iOS 11.3 and ARKit 1.5, Apple is once again demonstrating its commitment to innovation and its desire to push the boundaries of what is possible with technology.*
Human	Magical Show Mary Poppins to Wrap Albuquerque Little Theatre’s 87th Season	*The jack-of-all trades, Bert, introduces us to England in 1910 and the troubled Banks family. Young Jane and Michael have sent many a nanny packing before Mary Poppins arrives on their doorstep. Start the conversation, or Read more at BroadwayWorld.com.*
Llama-V2	Magical Show Mary Poppins to Wrap Albuquerque Little Theatre’s 87th Season	*Mary Poppins, the beloved musical that the hearts of audiences for decades, is set to wrap up the 87th season of the Albuquerque Little Theatre (ALT) in New Mexico. The show, which opened on May 13th, has been delighting audiences with its magical blend of song, dance, and storytelling. Based on the popular Disney film of the same name, Mary Poppins tells the story of the Banks family, who are in search of a new nanny to care for their children, Jane and Michael. The family’s fortunes take a turn for the better when Mary Poppins, a practically perfect nanny, arrives on their doorstep. With her enchanting voice, graceful movements, and a sprinkle of magic, Mary wins over the hearts of the Banks family and their friends, bringing joy and order to their lives*
Llama-V3	Magical Show Mary Poppins to Wrap Albuquerque Little Theatre’s 87th Season	*The beloved musical, Mary Poppins, is set to conclude the Albuquerque Little Theatre’s (ALT) 87th season, bringing a spoonful of sugar and a dash of magic to the city’s theater scene. The show, directed by Henry Avery, promises to be a spectacular finale to a season that has seen a wide range of productions, from classic dramas to contemporary comedies. The story of Mary Poppins has been enchanting audiences for generations, and the ALT production is expected to be no exception. The musical, based on the books by P.L. Travers, tells the tale of the Banks family, who are struggling to find joy and discipline in their lives. The arrival of Mary Poppins, played by actress Katie Calahan, brings a much-needed injection of fun and discipline into the household, as she teaches the children valuable lessons about the importance of imagination, hard work, and family.*

Table 10: Examples of how Llama-regenerated CCNews text differs from the original. The original is presented in its entirety, the regenerate examples came from the end of the article in the case of ARKit 1.5, and from the middle in the case of Mary Poppins. Red denotes a near exact text match between Llama-V2 and Llama-V2 CCNews, Orange denotes less exact match, potentially with position shifts across the sentence.

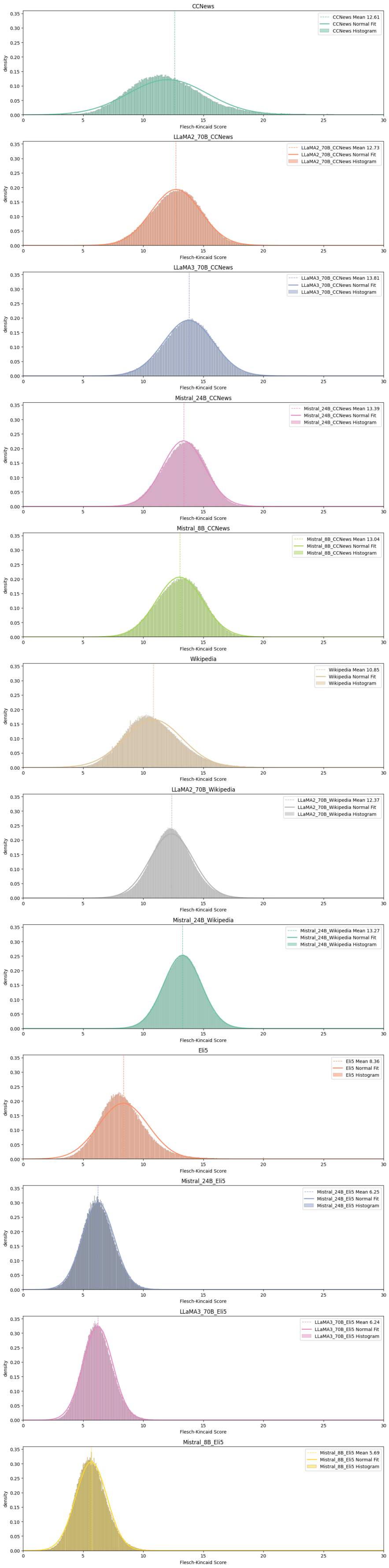

Figure 15: The full plot of Flesch-Kincaid Readability for all the models.

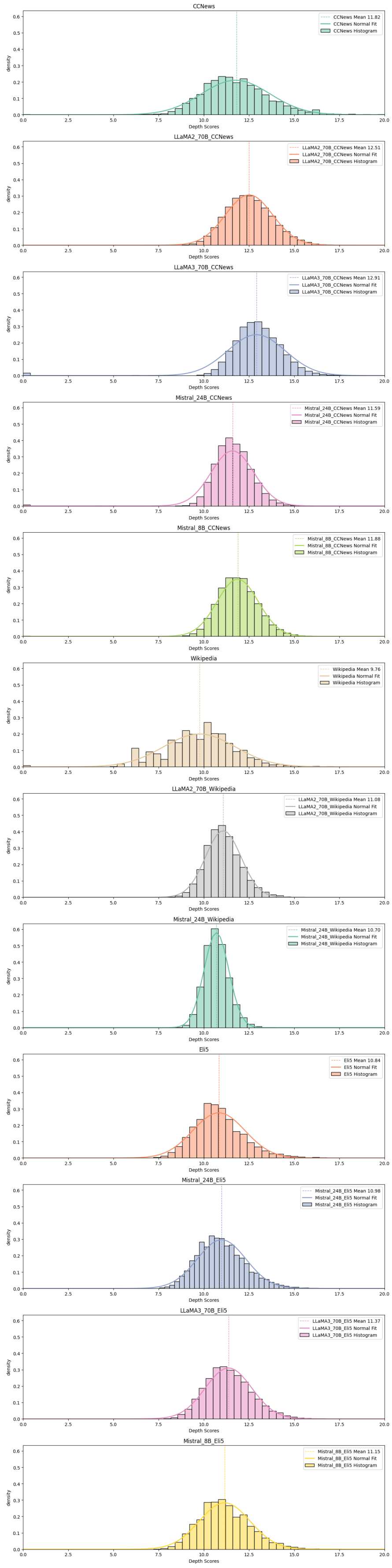

Figure 16: The distribution of constituency parse depth. Llamadatasets have a narrower distribution.

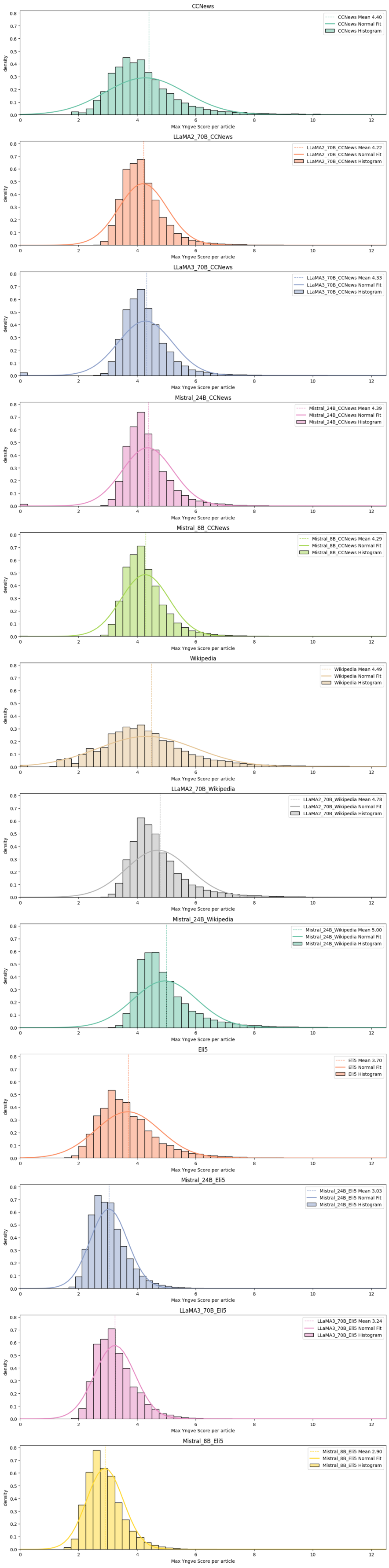

Figure 17: The full plot of Yngve score for all the models.