Linguistic and Embedding-Based Profiling of Texts Generated by Humans and Large Language Models

Sergio E. Zanotto[1] **and Segun Aroyehun**[2]
[1]Department of Linguistics & Cluster of Excellence "The Politics of Inequality"
University of Konstanz
sergio.zanotto@uni-konstanz.de
[2]Department of Politics and Public Administration
University of Konstanz
segun.aroyehun@uni-konstanz.de

Abstract

The rapid advancements in large language models (LLMs) have significantly improved their ability to generate natural language, making texts generated by LLMs increasingly indistinguishable from human-written texts. While recent research has primarily focused on using LLMs to classify text as either human-written or machine-generated texts, our study focuses on characterizing these texts using a set of linguistic features across different linguistic levels such as morphology, syntax, and semantics. We select a dataset of human-written and machine-generated texts spanning 8 domains and produced by 11 different LLMs. We calculate different linguistic features such as dependency length and emotionality, and we use them for characterizing human-written and machine-generated texts along with different sampling strategies, repetition controls, and model release dates. Our statistical analysis reveals that human-written texts tend to exhibit simpler syntactic structures and more diverse semantic content. Furthermore, we calculate the variability of our set of features across models and domains. Both human- and machine-generated texts show stylistic diversity across domains, with human-written texts displaying greater variation in our features. Finally, we apply style embeddings to further test variability among human-written and machine-generated texts. Notably, newer models output text that is similarly variable, pointing to a homogenization of machine-generated texts.

1 Introduction

The rapid advancements in language models have significantly improved their ability to generate natural language, making machine-generated texts (MGT) increasingly indistinguishable from human-written texts (HWT). Indeed, recent studies indicate that disinformation produced by state-of-the-art large language models (LLMs) is often perceived as more credible than that created by humans (Spitale et al., 2023). This evolution has highlighted the importance of identifying MGT due to legitimate concerns about the potential for malicious actors to disseminate false information, as well as the broader need to uphold trust and authenticity across online platforms (Chakravarthi et al., 2025; Li et al., 2024; Sarvazyan et al., 2023). Best systems for this task, often referred to as Human/Machine Authorship Attribution, all imply the use of LLMs, and different studies show marginal gains of leveraging stylometric features combined with LLMs for successfully achieving this task (Alecakir et al., 2024; Wang et al., 2024a).

In this study, we utilize the RAID dataset (Dugan et al., 2024), a large-scale corpus designed for testing detection tools for machine-generated text and human-written text. This work focuses on analyzing different levels of linguistic analysis, such as morphology, syntax, and semantics, in order to characterize MGT from 11 different models with 4 different decoding strategies, along with the release date of the models. Thus, we analyze a set of representative linguistic features to distinguish human-written texts (HWT) from machine-generated texts (MGT), potentially uncovering distinctive linguistic patterns in a multi-domain setting (across 8 different domains). Moreover, we focus on linguistic variability over time to check whether there are patterns of linguistic homogenization in newer LLMs (Sourati et al., 2025; Padmakumar and He, 2023).

We test the following set of linguistic features per level of linguistic analysis: Textual level (Text Length, Sentence Length), Morphology (Morphological Complexity Index for Verbs and Nouns), Syntax (Dependency Tree Depth, Dependency Length), Lexical Level (Word Prevalence, Type-Token Ratio), Semantics (Semantic Similarity), Emotionality. We consider these features as representative of different levels of linguistic analysis of our interest, ensuring overlap with previous

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22853–22870

work on distinguishing between HWT and MGT (Zanotto and Aroyehun, 2024; Guo et al., 2023; Uchendu et al., 2020). We do not aim to test all possible linguistic features tested in other research on Human/Machine authorship attribution tasks (e.g., Simón et al., 2023; Uchendu et al., 2020).

We provide statistical results for Human/Machine authorship attribution for each of these features. Moreover, we train a binary logistic classifier using these features to distinguish between HWT and MGT. We use the logistic classifier to analyze feature importance to get an overall picture of the impact of these features in characterizing HWT and MGT.

To assess variability in linguistic representations across models and domains, we calculate the standard deviation of the Euclidean distance for our extracted features. Additionally, we map language models to their release dates to analyze temporal trends in model development. This framework highlights differences between models and reveals the impact of decoding strategies and domain linguistic constraints.

Furthermore, we apply a style embedding model (Patel et al., 2025) that has been used for authorship attribution to represent text styles, enabling further comparison of variability in HWT and MGT.

Our main contributions are: (i) We analyze linguistic differences between human-written texts (HWT) and machine-generated texts (MGT) with linguistic features across different linguistic levels such as morphology, syntax, and semantics to examine variability across different models and domains. We show how recent models tend to have similar linguistic variability, pointing to a risk of homogenization of texts. (ii) We employ a logistic classifier to identify different linguistic patterns between HWT and MGT by performing a feature importance analysis. (iii) We further use style embeddings to compare HWT and MGT showing that recent models tend to exhibit similar style variation within themselves, underlying how chat models output texts with more similar characteristics to HWT than their non-chat models.

2 Related Work

Scholars have explored various approaches to tackle the challenge of distinguishing between human-written and machine-generated texts. This task, often referred to as Human/Machine Authorship Attribution (Alecakir et al., 2024), involves detecting whether a text is produced by a human or a generative language model, or attributing authorship among different models.

The Human/Machine attribution of authorship to a text carries significant social relevance, especially in areas such as fake news detection (Kumarage et al., 2023; Jawahar et al., 2020). The need for explainability becomes particularly important when engaging with a broad audience of non-experts, who may not have the means to access or comprehend detection models (Gehrmann et al., 2019). As a result, numerous studies have focused on identifying human-explainable features that can differentiate between machine-generated (MGT) and human-written texts (HWT) (e.g., Dugan et al., 2023; Guo et al., 2023; Kumarage et al., 2023; Uchendu et al., 2020). To achieve this, researchers have employed diverse analytical approaches, including stylometric analysis (Kumarage et al., 2023; Ma et al., 2023), qualitative assessments (Guo et al., 2023; Gehrmann et al., 2019), and linguistic feature analysis (Wang et al., 2024b; Uchendu et al., 2020; Ferracane et al., 2017), to diverse corpora, contexts, and generation tasks. Notably, different studies highlight the difficulties encountered by humans in distinguishing machine-generated texts from human-written texts, but little has been done to address the real-life consequences of this issue. (e.g., Chakraborty et al., 2023; Jakesch et al., 2023; Fraser et al., 2024).

Moreover, classical machine learning algorithms, such as logistic regression, have been employed to train models on bag-of-words features to differentiate between HWT and MGT (Solaiman et al., 2019; Ippolito et al., 2020). Other traditional methods leverage linguistic features, including POS-tags (Ferracane et al., 2017), surface-level features such as readability indexes or punctuation marks (Doughman et al., 2024; Malviya et al., 2025), topic modeling (Seroussi et al., 2014), sentiment analysis (Hossen Rujeedawa et al., 2025), and LIWC (Linguistic Inquiry and Word Count) features to provide deeper insights into the characteristics of MGT (Uchendu et al., 2020; Li et al., 2014). HWT tend to be longer, express more sentiment polarity, especially negative sentiment, and show greater variability in discourse structures compared to MGT (e.g., Kim et al., 2024; Zanotto and Aroyehun, 2024).

With the advent of Large Language Models (LLMs), fine-tuned models such as RoBERTa have achieved state-of-the-art performance in many tasks (Crothers et al., 2023; Jawahar et al., 2020).

Thus, different shared tasks such as SemEval and IberLEF focus on the creation of the best models for tackling Human/Machine Authorship Attribution, with LLM-based systems always reaching the top positions (Wang et al., 2024a; Sarvazyan et al., 2023). Indeed, classifiers solely based on stylometric features reach poor accuracy in distinguishing Human/Machine authorship (Alecakir et al., 2024; Sharma and Mansuri, 2024), with marginal gains when combined with LLMs (Sharma and Mansuri, 2024). Models leveraging Transformers encoders with token-level probabilistic features offer state-of-the-art detection capabilities (Sarvazyan et al., 2024; Mitchell et al., 2023). Despite providing literature on classifiers for Human/Machine authorship attribution, the focus of our study is to examine possible characteristics that distinguish HWT and MGT on different levels of linguistic analysis. Few studies show that MGT tend to exhibit shorter texts on average, with lower emotional content and syntactically more complex structures compared to HWT (Zanotto and Aroyehun, 2024; Guo et al., 2023). However, they overlook distinctions across models, decoding strategies, or stylistic variability. In doing this, we test variability among models and provide results that possibly point to the tendencies of recent models to exhibit less variability in their output, possibly due to the recent practice of training on machine-generated data (Shumailov et al., 2023).

3 Data

For our research, we use the RAID dataset (Dugan et al., 2024), a large-scale corpus comprising over 6.2 million text generations, developed to evaluate detection methods for outputs from language models. The training set of the dataset is constructed from 13,371 human-written documents sampled across eight diverse domains — Abstracts, Books, News, Poetry, Recipes, Reddit, Reviews, and Wikipedia — offering a broad spectrum of human-written and machine-generated English texts for linguistic analysis. For each document, a corresponding generation prompt is dynamically created using either "Chat" or "Non-Chat" templates, ensuring compatibility with the target language model. All the models use the same specific set of prompts (refer to Section B in Appendix for examples). "Chat" models are fine-tuned (often via supervised fine-tuning and reinforcement learning) on multi-turn interactions with humans. "Non-Chat" models are trained for next-token prediction on unstructured text. Texts are generated using eleven models, namely GPT-2 XL, GPT-3, GPT-4, ChatGPT, Mistral-7B, Mistral-7B (Chat), MPT-30B, MPT-30B (Chat), LLaMA 2 70B (Chat), Cohere, and Cohere (Chat), under four decoding strategies (i.e., greedy decoding and sampling decoding paired with the presence, where possible, or absence of repetition penalties). Greedy models are more deterministic and repetitive, while sampling models introduce randomness to make the response more diverse, whereas the presence of a penalty on repetition avoids excessive looping and redundancy. Note that we exclude texts generated using adversarial attacks from the original dataset. Hence, the selected subset of the dataset is designed to reflect realistic generation scenarios, providing a robust source for analyzing linguistic similarities and differences between machine-generated and human-written texts. In the end, we consider a total amount of 467,985 texts.

4 Methodology

We extract a range of linguistic features to analyze differences between human-written and machine-generated texts. These features include text length, sentence length, morphological complexity index, dependency length, dependency depth, word prevalence, type-token ratio, semantic similarity, and emotionality. For each feature, we compute the mean and standard error of the mean for human-written and machine-generated texts and assess similarity between human and model distributions using the Mann-Whitney U-test (McKnight and Najab, 2010). To further explore stylistic variability, we apply a style embedding model to extract a representation of the writing style of each text, enabling a comparison of style consistency across human-written and machine-generated texts. We train a logistic classifier on all extracted linguistic features to predict text authorship based on linguistic features. Additionally, we assess variability across models and domains by computing the standard deviation of the Euclidean distance for all linguistic features across HWT and MGT. We employ Principal Component Analysis (PCA) for dimensionality reduction using style embeddings to quantify variability within models and domains. All codes are available at `https://github.com/Sergio-E-Zanotto/LingAIHUman`.

4.1 Features Collection

We present here in detail the linguistic features that we calculate for characterizing human-written (HWT) and machine-generated texts (MGT).

Text Length: `Text Length` is calculated by counting the total number of alphanumerical tokens in a text using the tokenizer in SpaCy "en_core_web_sm". We average the length of all texts produced by humans or models with their own decoding and penalty settings. Text length provides shallow linguistic information on possible differences between humans and models and between different generation settings of models.

Sentence Length: `Sentence Length` is calculated by counting the total number of alphanumerical tokens per sentence, normalized by the total number of sentences produced by a specific model or human. This measure allows for a better fine-grained understanding of possible differences within text length, whereas two texts could have the same length, but a different number of sentences and therefore different sentence lengths.

Morphological Complexity Index: The `Morphological Complexity Index (MCI)` (Brezina and Pallotti, 2019) measures the diversity of word forms associated with the same lemma, reflecting the morphological richness of a text. It is calculated by extracting lemmas and their word forms using spaCy "en_core_web_sm", then randomly sampling subsets of five words to compute within-subset variety (how many unique word forms appear within a subset) and between-subset diversity (how different two subsets are). A higher MCI indicates greater morphological complexity, making it useful for analyzing linguistic richness and distinguishing between different registers and writing styles.

Dependency Tree Depth: `Dependency Tree Depth` is calculated using the dependency parser in SpaCy "en_core_web_sm". It represents hierarchical syntactic complexity. We calculate the maximum depth of the dependency tree for each sentence, from the syntactic head to the lowest leaf node. A deeper tree suggests more complex sentence constructions (e.g., multiple layers of subordinate clauses), whereas a shallower tree suggests a simpler structure.

Dependency Length: `Dependency Length` is calculated using the dependency parser in SpaCy "en_core_web_sm". It represents the distance in number of tokens between a word (dependent) and its syntactic head. We considered for this measure only the intervening words between the head and the last dependent word. This measure focuses on the linear arrangement of words. Longer dependency lengths can indicate that related words are spread farther apart, which reflects more complex syntactic structures.

Word Prevalence: `Word prevalence` refers to the proportion of people who recognize and understand a given word. It is a measure of lexical familiarity, capturing how widely known a word is within a population. To calculate word prevalence, we used the English Word Prevalence dataset (Brysbaert et al., 2019), which provides prevalence scores based on a large-scale crowdsourcing study involving over 220,000 participants. Each word's prevalence score represents the percentage of respondents who reported knowing the word. We computed the average prevalence per text by tokenizing the text using SpaCy "en_core_web_sm", extracting word forms and their lemmas, and matching them to their prevalence scores from the dataset. The prevalence scores were then averaged across all words in the text. This measure is useful for analyzing word difficulty, lexical accessibility, and text comprehensibility.

Type-Token Ratio: `Type_token_ratio` is calculated by counting the number of unique tokens and dividing it by the total number of tokens per text. We include this feature to verify if HWT and MGT tend to have similar lexical diversity in their texts.

Semantic Similarity: `Semantic Distance` is calculated using sentence embeddings derived from the Sentence Transformer model "paraphrase-MiniLM-L6-v2" (Reimers and Gurevych, 2019). We calculate the cosine similarity in pair-wise sentence comparisons and averaged the distance for each document. This feature describes the semantic content of a text in terms of consistency, as in Beaty and Johnson (2021).

Emotionality: `Emotionality` is calculated using the NRC Emotion Intensity Lexicon (Mohammad and Turney, 2013). The lexicon includes 8 different emotions: anger, disgust, fear, sadness, joy, anticipation, surprise, trust. We consider anger, fear and sadness as negative emotions, while we take joy as the only representative positive emotion in the lexicon following Aroyehun et al. (2023). We compute emotional load as the proportion of positive and negative emotion words in a given text. This feature helps to describe potential differences

in emotional content between HWT and MGT, as in Guo et al. (2023).

Style Embeddings: We compute style embeddings using the StyleDistance model (Patel et al., 2025). This model generates style embeddings with a dimension of 768 such that texts with similar stylistic features are closer in embedding space, independent of content. We use these embeddings to assess whether models and humans exhibit different stylistic variability in their generated texts. We selected StyleDistance (Patel et al., 2025) because it is a state-of-the-art model explicitly designed to generate content-independent style embeddings. The model has been shown to generalize well to real-world benchmarks and outperform leading style representation methods across multiple downstream tasks. Evaluations on style identification, style transfer, and authorship verification demonstrate that StyleDistance captures stylistic properties robustly, including under out-of-domain and out-of-distribution conditions. We use the resulting 768-dimensional embeddings as a latent representation of linguistic style, providing an additional analysis alongside our main feature-based approach. While we do not assume a one-to-one correspondence between embedding dimensions and linguistic features, these embeddings allow us to assess whether global stylistic patterns across models and domains align with those revealed through explicit linguistic features.

4.2 Models and Domain Variability

For each model and domain, we first compute the centroid by taking the mean of the values of our linguistic features. Then, for each individual observation in that model or domain, we calculate the Euclidean distance from its feature vector to the centroid. The overall variability for the models and domains is quantified as the standard deviation of these distances. These standard deviations reflect how spread out the observations are in the feature space relative to the centroid. Moreover, we apply dimensionality reduction using PCA to the style embeddings to match the linguistic feature count (10 dimensions). We compute variability within domains and models for the style embeddings by calculating their centroids, Euclidean distances, and standard deviations. Finally, we map models to their release dates to identify a possible trend in the evolution of models with respect to our set of linguistic features and style embeddings.

4.3 Feature Importance via a Logistic Classifier

We train a binary logistic regression classifier with a set of linguistic features to characterize via feature importance analysis human-written texts (label: 1) from machine-generated texts (label: 0). To ensure a balanced dataset, we downsample machine-generated texts, resulting in a training set of 21,376 documents and a test set of 5,344. Rather than comparing with existing methods, our focus is on assessing feature importance to get an overall picture of the impact of these features in characterizing HWT and MGT. We do not aim to build a state-of-the-art detection classifier nor to compare classification results with existing literature.

5 Results

In this section, we present the linguistic features used to analyze human-written texts (HWT) and machine-generated texts (MGT). These features capture differences in text structure, morphological complexity, syntactic complexity, lexical diversity, and emotionality. We show the results and feature importance from the classifier based on the calculated linguistic features. Figure 1 illustrates the variation in linguistic features across domains, showing that MGT exhibit comparatively low variability. Section E in the Appendix presents further visualizations and discussions of observed patterns across domains. These plots highlight the influence of genre-specific constraints on different domains. In more creative domains such as poetry and books, human-written texts tend to be longer and semantically more diverse. Across all domains, human-written texts consistently show simpler syntactic structures, while maintaining a relatively stable use of common words, as measured by Word Prevalence. Notably, this measure varies among HWT between Reddit and news: human-written texts on Reddit show the highest use of common words, whereas News articles show the lowest.

Table 1 shows the statistical analysis of our selected set of features for HWT and MGT in the entire dataset. Bold entries are those models that are statistically similar to humans for that feature. Different models and decoding strategies show different behaviours per feature in comparison to humans. We notice a tendency for chat models and for sampling strategies to output texts that are more similar to humans. Thus, we rely on the feature importance of a logistic classifier to extract further ten-

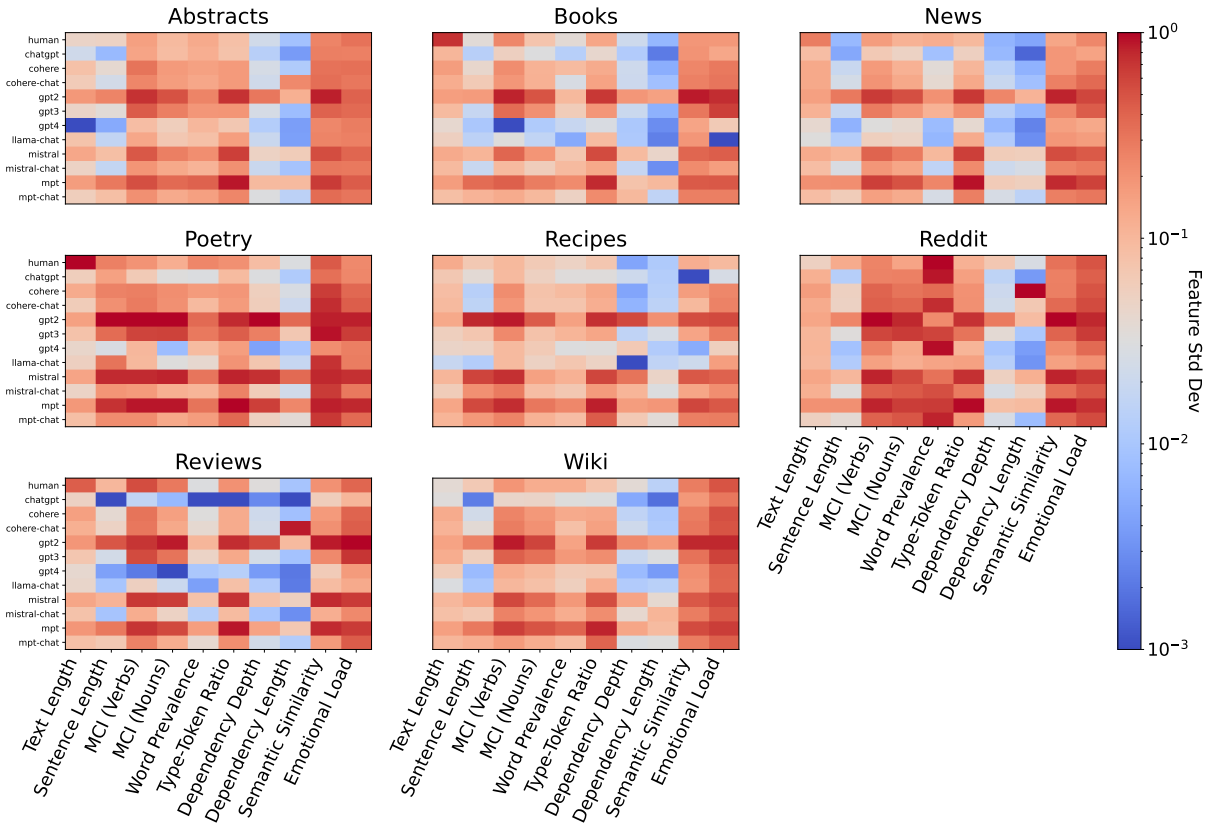

Figure 1: Multi-panel heatmap displaying log-normalized feature standard deviation across different text domains and models

dencies. Figure 2 illustrates the feature importance for the overall dataset. The classifier primarily relies on simpler syntactic structures (e.g., shorter dependency length and depth) and semantic properties (e.g., lower word prevalence, reduced semantic similarity, and a lower type-token ratio) for distinguishing between human-written and machine-generated texts. We present an analysis of each feature in the next section.

5.1 How do HWT and MGT differ in terms of linguistic features?

Text Length: Table 1 shows the average length of HWT and MGT in the entire dataset. Previous studies found HWT to exhibit, on average, longer texts than MGT (Guo et al., 2023). In our study, this tendency is confirmed with a lower margin, since some models with the no repetition penalty setting generate longer texts on average than humans (e.g., gpt2). However, Section E in the Appendix presents the feature analysis across domains. Notably, in the Poetry and Book domains, HWT are longer and show greater variability than MGT, likely reflecting the effect of more "creative" domains with less rigid genre-specific constraints. This may enable human authors to express individual creativity more distinctly than models.

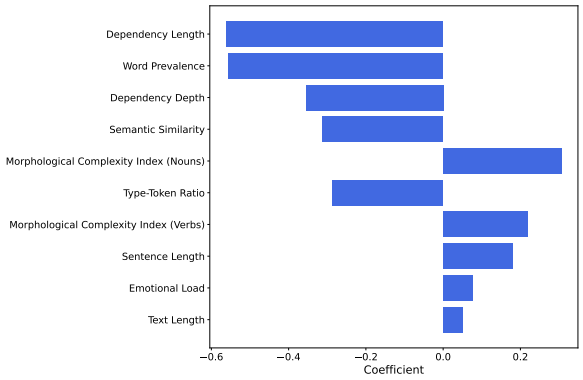

Figure 2: Feature importance for the logistic classifier across all domains. A positive coefficient means that as a feature value increases, the model is more likely to predict human authorship, while a negative coefficient indicates that higher values make the model less likely to predict human authorship.

Sentence Length: Table 1 shows the average sentence length of HWT and MGT in the entire dataset. We do not notice a clear pattern that distinguishes humans from different models with decoding strategies or penalty control. Figure 2 shows

model	Text Length	Sentence Length	MCI (VERBs)	MCI (NOUNs)	Word Prevalence	Type-Token Ratio	Dependency Depth	Dependency Length	Semantic Similarity	Emotionality
human	300.61 ± 2.766	23.53 ± 0.134	8.73 ± 0.011	8.73 ± 0.006	2.27 ± 0.001	0.64 ± 0.001	6.83 ± 0.017	2.41 ± 0.006	0.30 ± 0.001	0.38 ± 0.001
chatgpt_greedy_no	**260.35 ± 0.950**	19.29 ± 0.065	8.60 ± 0.009	8.71 ± 0.006	2.28 ± 0.001	0.62 ± 0.001	**6.70 ± 0.012**	2.27 ± 0.004	0.38 ± 0.001	0.37 ± 0.001
chatgpt_sampling_no	272.73 ± 0.960	20.00 ± 0.077	**8.70 ± 0.008**	8.80 ± 0.006	2.28 ± 0.001	0.63 ± 0.001	**6.75 ± 0.011**	**2.32 ± 0.004**	0.37 ± 0.001	**0.38 ± 0.001**
cohere-chat_greedy_no	189.48 ± 0.743	21.47 ± 0.106	7.91 ± 0.012	8.29 ± 0.008	2.28 ± 0.001	0.63 ± 0.001	6.72 ± 0.013	**2.44 ± 0.038**	0.33 ± 0.001	0.34 ± 0.001
cohere-chat_sampling_no	190.51 ± 0.772	21.58 ± 0.105	7.92 ± 0.012	8.31 ± 0.008	2.28 ± 0.001	**0.64 ± 0.001**	6.73 ± 0.014	2.58 ± 0.104	0.33 ± 0.001	0.34 ± 0.001
cohere_greedy_no	234.41 ± 0.911	21.29 ± 0.106	8.30 ± 0.012	8.50 ± 0.007	2.28 ± 0.000	0.61 ± 0.001	6.58 ± 0.013	2.80 ± 0.144	0.32 ± 0.001	0.36 ± 0.001
cohere_sampling_no	239.54 ± 0.934	22.11 ± 0.132	8.33 ± 0.012	8.54 ± 0.007	2.27 ± 0.001	0.63 ± 0.001	6.64 ± 0.014	2.61 ± 0.076	0.31 ± 0.001	0.37 ± 0.001
gpt2_greedy_no	364.78 ± 0.678	79.62 ± 0.939	2.98 ± 0.023	5.02 ± 0.022	2.26 ± 0.001	0.10 ± 0.001	16.59 ± 0.242	5.15 ± 0.107	0.59 ± 0.003	0.17 ± 0.001
gpt2_greedy_yes	243.15 ± 1.059	25.84 ± 0.407	8.91 ± 0.018	8.38 ± 0.012	2.29 ± 0.000	0.67 ± 0.002	7.67 ± 0.090	2.78 ± 0.054	0.29 ± 0.001	0.35 ± 0.001
gpt2_sampling_no	314.53 ± 0.904	**22.45 ± 0.107**	8.98 ± 0.014	8.79 ± 0.008	2.29 ± 0.000	0.58 ± 0.001	6.76 ± 0.018	2.56 ± 0.006	0.29 ± 0.001	0.37 ± 0.001
gpt2_sampling_yes	297.14 ± 1.062	26.90 ± 0.085	9.38 ± 0.015	8.93 ± 0.010	2.30 ± 0.000	0.78 ± 0.001	7.37 ± 0.014	2.55 ± 0.005	0.25 ± 0.001	0.39 ± 0.001
gpt3_greedy_no	136.86 ± 0.659	19.95 ± 0.182	6.61 ± 0.017	7.34 ± 0.015	2.27 ± 0.001	0.63 ± 0.001	6.54 ± 0.035	2.38 ± 0.012	0.33 ± 0.001	0.29 ± 0.001
gpt3_sampling_no	139.58 ± 0.672	19.42 ± 0.112	6.84 ± 0.017	7.50 ± 0.014	**2.27 ± 0.001**	0.66 ± 0.001	6.44 ± 0.023	2.35 ± 0.008	0.32 ± 0.001	0.30 ± 0.001
gpt4_greedy_no	267.44 ± 0.782	18.21 ± 0.048	8.51 ± 0.009	8.81 ± 0.004	2.28 ± 0.001	0.60 ± 0.001	6.54 ± 0.015	2.37 ± 0.005	0.35 ± 0.001	0.37 ± 0.001
gpt4_sampling_no	289.58 ± 0.816	19.14 ± 0.038	8.69 ± 0.008	8.94 ± 0.004	2.27 ± 0.001	0.65 ± 0.001	6.56 ± 0.011	2.44 ± 0.004	0.35 ± 0.001	0.39 ± 0.001
llama-chat_greedy_no	281.95 ± 0.547	23.19 ± 0.143	8.84 ± 0.007	8.85 ± 0.004	2.29 ± 0.000	0.57 ± 0.001	7.08 ± 0.014	2.48 ± 0.005	0.34 ± 0.001	0.38 ± 0.001
llama-chat_greedy_yes	268.22 ± 0.572	23.62 ± 0.151	9.00 ± 0.007	8.93 ± 0.004	2.29 ± 0.000	0.66 ± 0.001	7.07 ± 0.014	2.41 ± 0.004	0.31 ± 0.001	0.38 ± 0.001
llama-chat_sampling_no	281.03 ± 0.547	23.12 ± 0.140	8.83 ± 0.007	8.85 ± 0.004	2.28 ± 0.000	0.58 ± 0.001	7.05 ± 0.014	2.48 ± 0.005	0.34 ± 0.001	0.38 ± 0.001
llama-chat_sampling_yes	265.09 ± 0.585	23.61 ± 0.138	8.97 ± 0.007	8.96 ± 0.004	2.29 ± 0.000	0.69 ± 0.001	7.09 ± 0.014	**2.40 ± 0.005**	0.30 ± 0.001	0.39 ± 0.001
mistral-chat_greedy_no	205.28 ± 0.634	23.17 ± 0.176	8.03 ± 0.012	8.41 ± 0.008	2.29 ± 0.001	0.59 ± 0.001	7.17 ± 0.028	**2.52 ± 0.021**	0.38 ± 0.001	0.34 ± 0.001
mistral-chat_greedy_yes	200.14 ± 0.608	21.79 ± 0.101	8.53 ± 0.009	8.65 ± 0.007	2.30 ± 0.001	0.72 ± 0.001	7.04 ± 0.013	2.36 ± 0.005	0.33 ± 0.001	0.36 ± 0.001
mistral-chat_sampling_no	210.89 ± 0.663	22.07 ± 0.099	8.26 ± 0.010	8.55 ± 0.007	2.29 ± 0.001	0.62 ± 0.001	7.02 ± 0.014	2.49 ± 0.022	0.37 ± 0.001	0.36 ± 0.001
mistral-chat_sampling_yes	199.95 ± 0.698	23.51 ± 0.097	8.43 ± 0.011	**8.67 ± 0.008**	2.30 ± 0.001	0.77 ± 0.001	7.31 ± 0.014	2.39 ± 0.005	0.32 ± 0.001	0.37 ± 0.001
mistral_greedy_no	316.24 ± 0.651	50.66 ± 0.718	5.20 ± 0.027	6.73 ± 0.021	2.26 ± 0.001	0.24 ± 0.001	10.94 ± 0.158	3.77 ± 0.088	0.44 ± 0.003	0.25 ± 0.001
mistral_greedy_yes	208.68 ± 0.823	20.60 ± 0.150	8.86 ± 0.013	8.56 ± 0.008	2.29 ± 0.001	0.78 ± 0.001	6.60 ± 0.031	2.44 ± 0.030	0.28 ± 0.001	0.36 ± 0.001
mistral_sampling_no	286.23 ± 0.743	21.29 ± 0.092	8.90 ± 0.011	8.81 ± 0.006	2.28 ± 0.001	0.60 ± 0.001	6.52 ± 0.012	2.54 ± 0.005	0.28 ± 0.001	0.37 ± 0.001
mistral_sampling_yes	236.98 ± 0.869	31.47 ± 0.140	8.97 ± 0.011	8.80 ± 0.006	2.30 ± 0.000	0.85 ± 0.000	7.85 ± 0.018	2.68 ± 0.006	0.26 ± 0.001	**0.38 ± 0.001**
mpt-chat_greedy_no	157.84 ± 0.623	22.06 ± 0.113	7.58 ± 0.011	8.12 ± 0.009	2.29 ± 0.001	0.66 ± 0.001	7.09 ± 0.021	2.43 ± 0.007	0.38 ± 0.001	0.33 ± 0.001
mpt-chat_greedy_yes	165.66 ± 0.766	26.20 ± 0.125	8.01 ± 0.012	8.46 ± 0.009	2.29 ± 0.001	0.87 ± 0.001	7.44 ± 0.019	2.47 ± 0.011	0.33 ± 0.001	0.36 ± 0.001
mpt-chat_sampling_no	161.26 ± 0.648	22.42 ± 0.112	7.64 ± 0.011	8.18 ± 0.009	2.29 ± 0.001	0.67 ± 0.001	7.08 ± 0.018	2.42 ± 0.005	0.37 ± 0.001	0.34 ± 0.001
mpt-chat_sampling_yes	180.93 ± 0.911	31.72 ± 0.194	8.01 ± 0.013	8.47 ± 0.010	2.28 ± 0.001	0.92 ± 0.001	7.66 ± 0.018	2.75 ± 0.013	0.31 ± 0.001	0.37 ± 0.001
mpt_greedy_no	361.78 ± 0.569	43.87 ± 0.660	5.27 ± 0.026	6.73 ± 0.020	2.26 ± 0.001	0.20 ± 0.001	9.46 ± 0.133	3.52 ± 0.068	0.50 ± 0.002	0.25 ± 0.001
mpt_greedy_yes	192.30 ± 0.945	48.82 ± 0.314	8.33 ± 0.016	8.33 ± 0.013	2.29 ± 0.001	1.00 ± 0.000	8.89 ± 0.025	3.55 ± 0.042	0.25 ± 0.001	0.36 ± 0.001
mpt_sampling_no	349.96 ± 0.613	21.02 ± 0.088	9.23 ± 0.011	8.98 ± 0.006	2.28 ± 0.001	0.57 ± 0.001	6.53 ± 0.012	2.48 ± 0.008	0.27 ± 0.001	**0.38 ± 0.001**
mpt_sampling_yes	245.68 ± 1.194	47.13 ± 0.388	8.49 ± 0.021	8.36 ± 0.016	2.28 ± 0.001	1.00 ± 0.000	7.79 ± 0.025	3.53 ± 0.025	0.22 ± 0.001	0.36 ± 0.001

Table 1: Average values ± standard errors (SE) of calculated linguistic features per model in the dataset. The pattern of each model name is: model + decoding strategy + yes/no penalty on repetitions. Models corresponding to *non-bold* entries are statistically different from human writing in the respective feature. *Bold* entries are **not** statistically different based on the Mann-Whitney U-test ($p \geq 0.05$), indicating that the feature distribution in the machine-generated texts is not distinguishable from human-written texts.

that the classifier uses sentence length as a distinguishing feature, with a tendency for HWT to be longer than those of MGT.

Morphological Complexity Index (MCI): Table 1 shows the MCI for nouns and the MCI for verbs across models in the dataset. Again, we notice that human-written texts tend to be somewhat in the middle between models with different decoding strategies and the presence of penalties (e.g., mpt). However, HWT rather tend to have a higher inflectional diversity for both verbs and nouns than MGT (See Figure 2).

Dependency Tree Depth: Table 1 shows the average syntactical dependency depth per model across the entire dataset. We notice that HWT tend to score lower in dependency depth than older models, while recent models such as GPT-4 show a more human-like amount of depth in the syntactic tree. This evolution is clear when compared to other studies in which the syntactic depth of HWT was lower than all the other models (Zanotto and Aroyehun, 2024).

Dependency Length: To further investigate the syntactic characteristics of human-written texts and machine-generated texts, Table 1 shows the average dependency length per model across the entire dataset and complements the dependency depth results. Indeed, HWT score lower than older models like GPT-2 in dependency length and depth, while newer models produce similarly short or even shorter dependencies.

Word Prevalence: Table 1 shows the average word prevalence per model across the dataset. While HWT score slightly lower than most models, the differences are minimal. Both HWT and MGT exceed a two z-score threshold, indicating that 98% of the population is familiar with the words used.

Type-Token Ratio: Table 1 depicts the average type-token ratio per model across the dataset. Human-written texts maintain a balanced ratio of around 0.6, indicating a balanced mix of unique words and repetitions. Models with a repetition penalty show higher type-token ratios than those without (e.g., mistral-chat).

Semantic Similarity: Table 1 shows that human-written texts (HWT) exhibit a lower mean similarity than most models, indicating greater semantic diversity in human writing (See Figure 14 in Appendix). Repetition penalties help reduce redundancy, while sampling generates more diverse content than greedy decoding, as demonstrated with GPT-2.

Emotionality: In Table 1, we can notice how HWT score high in average emotionality in the dataset, but it does not score the highest, unlike what previous studies found (Guo et al., 2023).

Indeed, differences between models seem to be driven by the decoding strategy and the presence of the penalty, where the majority of models with a sampling decoding strategy and a penalty on repetition score higher than the others (see Figure 14 in the Appendix). Looking at possible differences between emotional load for positive and negative emotions, Figure 10 in the Appendix shows how there are some exceptions where HWT do not consistently score higher on negative emotions than MGT, contrary to previous findings (Guo et al., 2023).

5.2 Domain Variability

Figure 3 shows the variability of linguistic features across domains. The results highlight variations in the linguistic feature space, reflecting the well-established influence of genre-specific constraints on linguistic variability (Biber and Conrad, 2019). Indeed, more creative domains like poetry show more variability than less creative domains such as abstracts. Figure 11 in Appendix shows the variability based on style embeddings across domains. Style embeddings capture less variability in poetry than in other domains. To better understand the linguistic features potentially driving these differences, an avenue for future work will be to extract the set of individual features used to train the style embeddings model and identify which ones are responsible for the different patterns. This alternative approach is beyond the scope of the current study.

Moreover, we disentangle humans and models to understand if they behave differently in terms of variability for both our linguistic features and for the style embeddings. Figure 4 shows the domain variability of HWT in the linguistic feature space, while Figure 5 shows the domain variability of MGT in the linguistic feature space. HWT exhibit more variability than MGT across domains, arguably as they are more sensible to the well-known effect of different genre-specific constraints (Biber and Conrad, 2019). Figure 4 shows how poetry is the most variable domain, arguably as the most "creative" domain of the dataset. However, the domain variability of HWT and MGT appears more similar when using the style embeddings. Figure 12 in Appendix shows how Wiki is the most variable domain, in contrast to the linguistic features analysis. However, the variation in terms of standard deviation is comparable with the results based on the linguistic features. Figure 13 in Appendix shows poetry to be the least variable domain, and exhibits marginally lower variability across several domains, for example Reddit or Abstracts. We argue that these differences stem from the different features that are considered for the analyses: explicit linguistic features vs. latent style embedding representations. Further investigation would be needed to understand which dimensions can be attributed to the different observations.

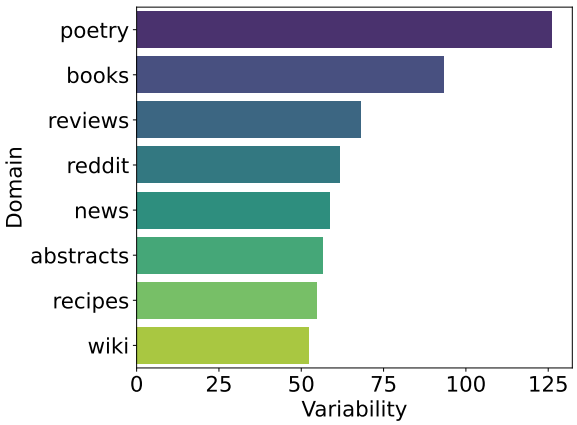

Figure 3: Domain variability with linguistic features in the entire dataset

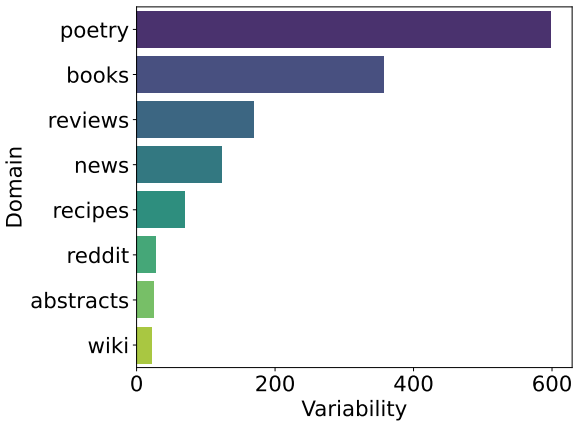

Figure 4: Domain variability with linguistic features in the entire dataset for humans

5.3 Homogenization of Model Outputs

Figure 6 presents the variability of linguistic features across different models and their release dates. The results indicate a trend where MGT exhibit lower linguistic variability compared to HWT and score increasingly similarly to one another, suggesting a homogenization of linguistic styles of models for our set of features. Despite differences in the overall magnitude of variability, Figure 7 shows that machine-generated texts exhibit similar

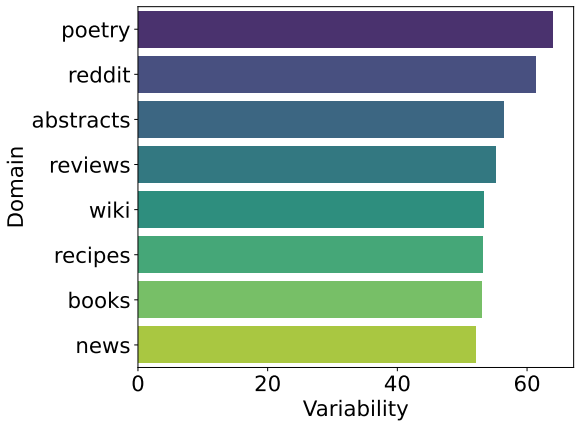

Figure 5: Domain variability with linguistic features in the entire dataset for non-humans

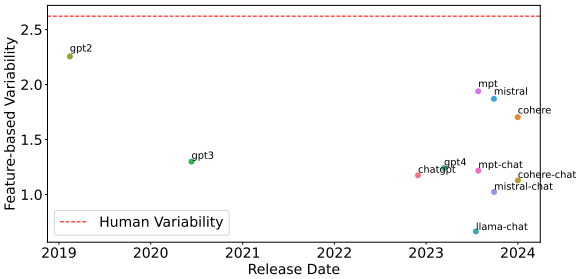

Figure 6: Model Variability with linguistic features in the entire dataset across different release years

variability in style embeddings both among themselves and in comparison to human-written texts. However, the chat models tend to exhibit variability comparable to HWT and higher than their non-chat versions, as expected given the different design of the models.

6 Discussion

Our analysis of linguistic features highlights different tendencies between human-written (HWT) and machine-generated texts (MGT) in the RAID dataset. Our feature importance analysis shows that human-written texts tend to have less complex syntactic structures and more varied lexical and semantic content. The tendency to produce less complex syntactical structures is in line with previous studies (Muñoz-Ortiz et al., 2024). Our variability analysis shows that models tend to produce outputs that are very similar between themselves and are less varied than humans. We attribute this result to the importance of individual style in human-written texts. Nevertheless, models that differ in decoding strategies and repetition penalties produce outputs that reflect their intended design. Moreover, the

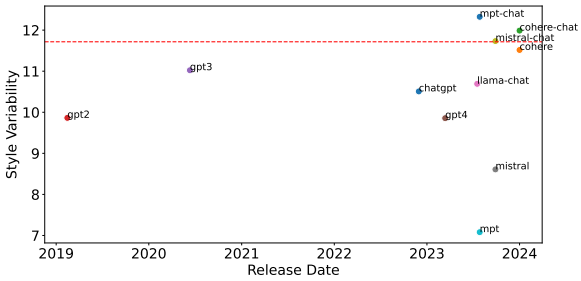

Figure 7: Model Variability with style embeddings in the entire dataset across different release years

domain variability analysis suggests an influence of linguistic constraints in less flexible and less creative domains such as abstracts (Biber and Conrad, 2019). Notably, newer models tend to be similar in terms of variability between themselves for our linguistic features, possibly pointing to the phenomenon of "model collapse" (Shumailov et al., 2023), where models tend to exhibit low variance due to training on generated data. Analyses based on style embeddings reveal differing levels of variability between chat models and non-chat models, with the former exhibiting variability comparable to that of humans, while the latter show lower variability. These results can be attributed to the effect of reinforcement learning from human feedback, which is used to train chat models.

7 Conclusion and Future Work

In our analysis of linguistic features of human-written (HWT) and machine-generated texts (MGT), we show different linguistic tendencies of human-written texts and how recent models tend to generate texts with linguistic characteristics similar to human-written texts, but exhibit a lower variability within themselves based on our set of features. Future work should explore diverse corpora with varying characteristics to verify these differences across different domains and languages. Expanding the range of linguistic features, especially those related to content such as the use of metaphors or figurative language, could provide deeper insights. Moreover, our setting of Human/Machine authorship attribution should be expanded from a binary human/non-human setting to a multi-class classification, where detection models have to attribute authorship to humans and to different LLMs.

8 Limitations

One limitation of our study lies in its generalizability. We rely on a dataset covering eight domains in the English language. Thus, the applicability of our findings is limited when considering broader linguistic variations across domains and languages. The eight domains covered in our dataset provide valuable insights, but they may not be representative of all possible linguistic contexts characteristic of all textual domains. Furthermore, texts in various languages may have unique linguistic features that limit the relevance of our results to non-English contexts.

While this study identifies and measures a range of linguistic features relevant for comparing human- and machine-generated texts, it does not examine how interactions between these features might provide additional insights. This represents a notable limitation, as individual features may not only operate independently but also in combination. However, analyzing such interactions poses methodological challenges due to the large number of possible feature combinations, especially in the absence of prior hypotheses or a theoretical framework indicating which interactions are likely to be salient.

Our analysis focuses on a set of LLMs that may already have been superseded by more advanced versions due to the rapid advancements in the field. This poses a challenge for the temporal validity of our findings as future LLMs could exhibit different linguistic patterns.

The RAID dataset does not include metadata identifying different human authors of the sampled texts. As a result, we are unable to analyze potential stylistic variations among human writers.

Extending this study to multiple languages, datasets, and model versions could potentially enhance the applicability of our findings across broader contexts and evolving technologies. However, such an extension would require significant data curation efforts and depend on the availability of multilingual linguistic pipelines capable of characterizing texts at scale. Nevertheless, this study and our findings can serve as a foundation for further exploration of LLM outputs.

9 Ethical Considerations

In developing our approach, we acknowledge the potential for unintended bias, particularly against non-native English speakers. Some of the linguistic features we analyze may capture characteristics in texts produced by English language learners. This overlap raises important ethical questions. It is crucial to emphasize that our primary objective is to advance the theoretical understanding of language patterns in texts generated by humans and LLMs, rather than to create tools for real-world applications. The features and techniques described in this paper are intended for research purposes and should not be directly applied in practical systems without careful consideration of their broader implications. Any potential real-world application would require extensive additional research and safeguards. We strongly caution against using these features or similar approaches in high-stakes decision-making processes or in any context where they could disadvantage individuals based on their language competency.

Acknowledgments

S.E.Z. is supported by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379. S.A. acknowledges the support of ERC Advanced Grant 101020961 PRODEMINFO.

References

Huseyin Alecakir, Puja Chakraborty, Pontus Henningsson, Matthijs Van Hofslot, and Alon Scheuer. 2024. Groningen team a at SemEval-2024 task 8: Human/machine authorship attribution using a combination of probabilistic and linguistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1926–1932, Mexico City, Mexico. Association for Computational Linguistics.

Segun Taofeek Aroyehun, Lukas Malik, Hannah Metzler, Nikolas Haimerl, Anna Di Natale, and David Garcia. 2023. Leia: Linguistic embeddings for the identification of affect. *EPJ Data Science*, 12(1):52.

Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior research methods*, 53(2):757–780.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written l2 texts. *Second language research*, 35(1):99–119.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word

prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51:467–479.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sajeetha Thavareesan, Elizabeth Sherly, Saranya Rajiakodi, Balasubramanian Palani, Malliga Subramanian, Subalalitha Cn, Dhivya Chinnappa, et al. 2025. Proceedings of the fifth workshop on speech, vision, and language technologies for dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2024. Exploring the limitations of detecting machine-generated text. *arXiv preprint arXiv:2406.11073*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Elisa Ferracane, Su Wang, and Raymond Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 584–593, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *arXiv preprint arXiv:2406.15583*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Muhammad Irfaan Hossen Rujeedawa, Sameerchand Pudaruth, and Vusumuzi Malele. 2025. Unmasking ai-generated texts using linguistic and stylistic features. *International Journal of Advanced Computer Science & Applications*, 16(3).

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, Baltimore, Maryland. Association for Computational Linguistics.

Kevin Li, Kenan Hasanaliyev, Sally Zhu, George Altshuler, Alden Eberts, Eric Chen, Kate Wang, Emily Xia, Eli Browne, and Ian Chen. 2024. Team MLab at SemEval-2024 task 8: Analyzing encoder embeddings for detecting LLM-generated text. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1463–1467,

Mexico City, Mexico. Association for Computational Linguistics.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human–differentiation analysis of scientific content generation. *arXiv preprint arXiv:2301.10416*.

Shrikant Malviya, Pablo Arnau-González, Miguel Arevalillo-Herráez, and Stamos Katsigiannis. 2025. Skdu at de-factify 4.0: Natural language features for ai-generated text-detection. *arXiv preprint arXiv:2503.22338*.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

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

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. *Preprint*, arXiv:2410.12757.

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*. Association for Computational Linguistics.

Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-salvador. 2024. Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Surbhi Sharma and Irfan Mansuri. 2024. Team innovative at SemEval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1172–1176, Mexico City, Mexico. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Lara Alonso Simón, José Antonio Gonzalo Gimeno, Ana María Fernández-Pampillón Cesteros, Marianela Fernández Trinidad, and María Victoria Escandell Vidal. 2023. Using linguistic knowledge for automated text identification. In *IberLEF@ SEPLN*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The shrinking landscape of linguistic diversity in the age of large language models. *arXiv preprint arXiv:2502.11266*.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv preprint arXiv:2412.03025*.

A Classifier results

This section presents the results from the logistic classifier and the support vector machine (SVM) trained on a set of linguistic features. We also provide a qualitative error analysis for the logistic classifier, illustrating examples of texts that were misclassified.

A.1 Linguistic features Classifier

We use a set of linguistic features we calculated to build a logistic classifier for assessing feature importance in distinguishing HWT and MGT. Table 2 provides the details of the overall results. The classifier achieves an accuracy of 0.61 across all domains. While we do not aim to compare our results against specific baselines or prior studies, it is worth noting that our accuracy scores are consistent with findings from related work on human/machine authorship attribution using classifiers based solely on linguistic features, which typically report performance in the 0.5–0.6 range (Alecakir et al., 2024; Sharma and Mansuri, 2024).

Metric	Non-Human	Human
Accuracy	0.61	
Precision	0.62	0.60
Recall	0.56	0.65
F1-score	0.59	0.62

Table 2: Classification performance of the logistic regression classifier based on linguistic features

In order to demonstrate the robustness of these results, we build a second classifier using an SVM on our set of linguistic features. Table 3 shows the results from the SVM classifier in distinguishing HWT and MGT. The classifier achieves an accuracy of 0.60 across all domains, in line with the results from the logistic classifier. Moreover, Figure 8 shows the results from the feature importance analysis of the SVM model. The pattern of the feature importance analysis for the SVM classifier is very similar to that of the logistic classifier (see Figure 2 in Section 5). Particularly, all features maintain the same direction, while the coefficients are slightly lower for the SVM classifier.

Metric	Non-Human	Human
Accuracy	0.60	
Precision	0.62	0.59
Recall	0.55	0.66
F1-score	0.58	0.62

Table 3: Classification performance of the Support Vector Machine classifier using linguistic features

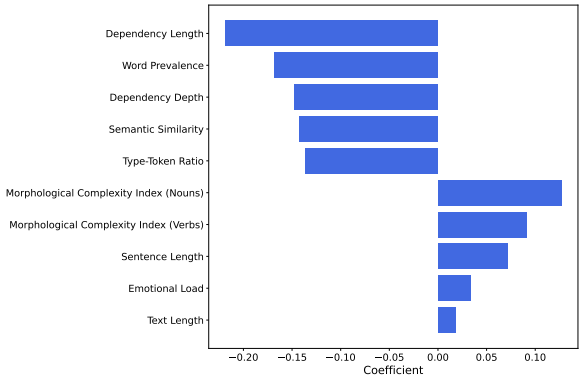

Figure 8: Feature importance for the SVM classifier across all domains. A positive coefficient means that as the feature value increases, the model is more likely to predict human authorship, while a negative coefficient indicates that higher values make the model less likely to predict human authorship.

A.2 Qualitative Error Analysis

We present a qualitative error analysis of the logistic classifier by examining misclassified texts. Figure 9 displays the confusion matrix, while Table 6 provides three examples of false positives (machine-generated texts predicted as human-written texts) and false negatives (HWT predicted as MGT). To explore whether some specific features systematically drive these errors, we use Cohen's d (Cohen, 2013) to measure the standardized mean difference between misclassified and correctly classified cases. We focus on false positives (FP) versus true negatives (TN) and false negatives (FN) versus true positives (TP). This allows us to identify the features that likely drive misclassification.

Table 4 shows that false positives are primarily associated with texts that are longer and exhibit

greater morphological complexity (both verbs and nouns) compared to true negatives. Thus, when machine-generated texts are long and structurally rich, the classifier tends to mistake them for HWT.

In contrast, Table 5 shows that false negatives often involve human-written texts that are shorter and morphologically simpler than true positives. Additionally, misclassified HWT tend to show higher semantic similarity, as human-written texts tend to show greater semantic diversity than MGT, according to our feature importance analysis of the logistic regression classifier.

Future work could investigate whether such cases are also challenging for human annotators, potentially leading to similar misclassification.

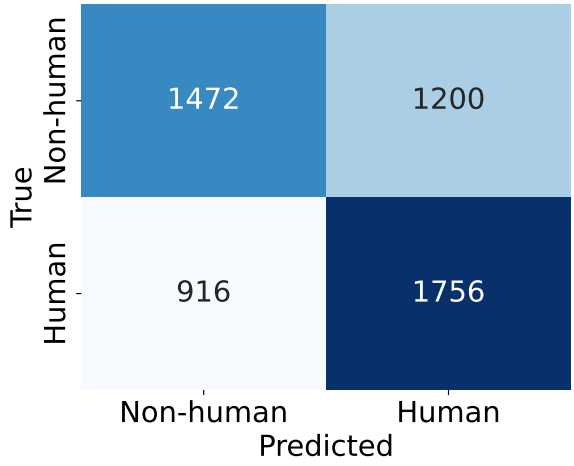

Figure 9: Confusion matrix showing classification performance of the logistic regresson classifier for human vs. machine-generated texts

Feature	**Cohen's d**
Text Length	1.20
Morph. Complexity (Verbs)	0.88
Morph. Complexity (Nouns)	0.81
Semantic Similarity	-0.66
Emotional Load	0.52
Type–Token Ratio	-0.45
Dependency Depth	-0.22
Sentence Length	-0.20
Dependency Length	-0.18
Word Prevalence	-0.16

Table 4: Cohen's d for linguistic features comparing false positives and true negatives. Positive values indicate features that are more salient in machine-generated texts (MGT) that were misclassified as human-written (HWT).

Feature	**Cohen's d**
Morph. Complexity (Verbs)	-1.07
Semantic Similarity	0.92
Morph. Complexity (Nouns)	-0.75
Emotional Load	-0.61
Text Length	-0.58
Type–Token Ratio	0.57
Dependency Length	0.46
Dependency Depth	0.44
Sentence Length	0.28
Word Prevalence	0.15

Table 5: Cohen's d for linguistic features comparing false negatives and true positives. Positive values indicate features that are more salient in HWT misclassified as MGT.

False Positives (machine-generated predicted as human-written texts)
Hey fellow Redditors, I'm on day 35 of my CT (camel trail) and I'm starting to feel a little stuck. I've been following the same routine for weeks now, and I can't help but wonder if I'm clinging to my nightmare because it's all I know at this point. [...].
In this paper, we explore the connection between Brunet-Derrida particle systems, free boundary problems, and Wiener-Hopf equations. We first introduce the concept of Brunet-Derrida particle systems, which are stochastic systems that describe the evolution of a collection of particles that interact with each other through a non-local competition mechanism. [...].
The trees are all so naked, and shivering with cold! They clatter over one another like drunken men when there's wind; and the snow lies hard on their arms, as if it were trying to get them into bed—trying too hard, for they don't move an inch in reply.... [...].
False Negatives (human-written predicted as machine-generated texts)
David Innes and his captive, a member of the reptilian Mahar master race of the interior world of Pellucidar, return from the surface world in the Iron Mole invented by his friend and companion in adventure Abner Perry. Emerging in Pellucidar at an unknown location, David frees his captive. [...].
The Children of Zion, published in January 1998, is considered as a documentary that was based on a collection of fragments of records compiled in Palestine in 1943 by the Eastern Center for Information, a Polish government group. [...].
Newspaper columnist Mitch Albom recounts time spent with his 78-year-old sociology professor, Morrie Schwartz, at Brandeis University, who was dying from Lou Gehrig's disease (ALS). Albom, a former student of Schwartz, had not corresponded with him since attending his college classes 16 years earlier. [...].

Table 6: Examples of misclassified texts by the logistic regression classifier

B Prompt Examples

In this section, we present examples of prompts from the RAID dataset used to generate MGT. All models rely on the same specific set of prompts. For instance:

- `Write the abstract for the academic paper titled "Model Theory for a Compact Cardinal."`
- `Write a recipe for "Anise Toasts Recipe."`

C Positive and Negative Emotions

In Figure 10, we show the differences in positive and negative emotion intensities of different models with their decoding strategy and the presence or not of a penalty on repetition. Indeed, HWT score high on negative emotion intensity, while more recent models tend to use fewer negative emotion words, arguably due to their alignment with human and/or machine feedback.

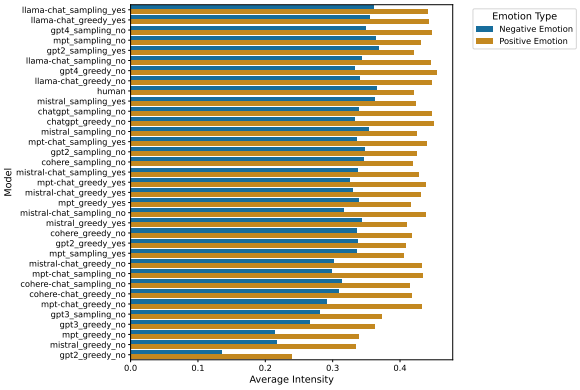

Figure 10: Negative and Positive Emotion Intensity per Model. This figure presents the emotional intensity across humans and different models.

D Domain variability with style embeddings

Figure 11 shows the variability within domains for HWT and MGT based on style embeddings for the entire datasets. Moreover, Figure 12 and Figure 13 disentangle the variability respectively of HWT and MGT. We argue that differences in variability between linguistic features and style embeddings stem from the different representations that are considered for the analyses. Future research should explore potential explanations for the differing variability rankings across text domains.

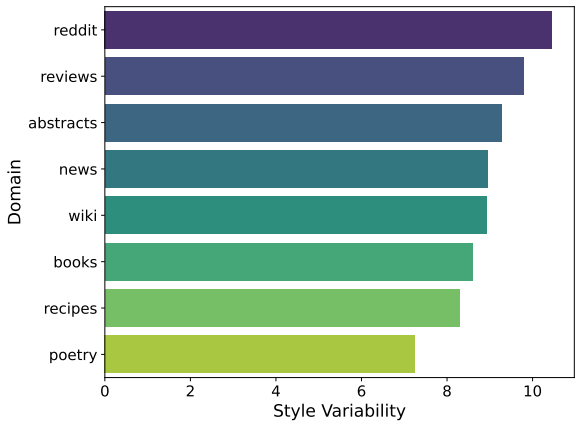

Figure 11: Domain variability with style embeddings in the entire dataset

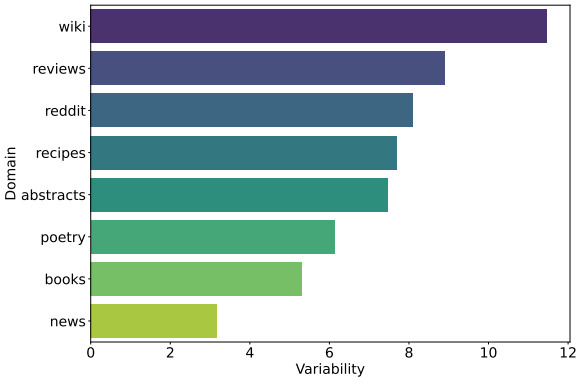

Figure 12: Domain variability with style embeddings in the entire dataset for humans

E Linguistic Features Visualization

Figure 14 shows the mean values of the linguistic features (standard deviation in Figure 1) across domains for each model in the dataset. The following linguistic features are considered: Text Length, Sentence Length, Morphological Complexity Index (MCI) for nouns, Morphological Complexity Index (MCI) for verbs, dependency depth, dependency length, word prevalence, type-token ratio, semantic similarity and emotionality. Notably, we observe a difference in variability between domains, especially when dealing with more "creative" domains such as Poetry and Books (See Figures 14 and 1). Especially, human-written texts show greater differences in average text length within these domains.

Poetry Domain In the poetry domain, on aggregate, HWT score highest in average text length and emotionality, while maintaining fairly simpler syntactic structures than most models. We can also notice that human-written texts tend to score higher in morphological complexity for verbs and nouns.

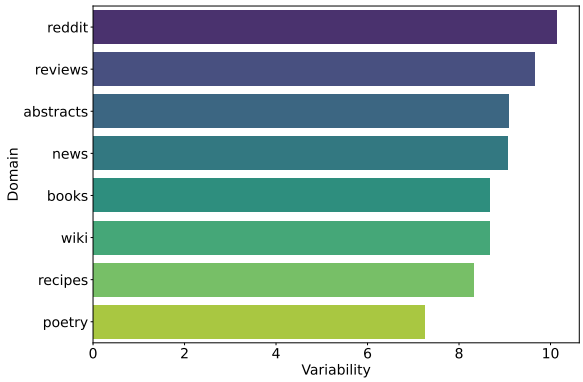

Figure 13: Domain variability with style embeddings in the entire dataset for non-humans

Books Domain In the books domain, human-written texts score highest in average text length. However, HWT score lower on emotionality than some models, and they employ more complex syntactic structures than in Poetry. Nevertheless, we notice that HWT tend to score high in morphological complexity for verbs and nouns in this domain as well.

Abstracts Domain For the abstracts domain, the patterns are different in comparison with poetry and books domains. Human-written texts score lower on average text length and emotionality than the other two domains. We can notice how recent models (e.g., GPT4) behave extremely more consistently and more similar to HWT than their older models (e.g., GPT2) in text length, sentence length, syntactic depth and length, and also in semantic similarity.

Recipes Domain In our set of linguistic features, the recipes domain shows that human-written texts score lowest in dependency length, underlying the tendency to employ simpler syntactic structures than MGT. This tendency involves semantic similarity as well, where HWT appear extremely more consistent in the semantic content than in previously discussed domains.

Reddit Domain In the Reddit domain, HWT score lower on morphological complexity than in previously discussed domains. This could point to differences in language register (e.g., less formal vs more formal) between a user-based domain with no genre-specific constraints and more strict domains (Biber and Conrad, 2019). Again, newer models show more similar patterns to human-written texts than their older versions, especially in semantic similarity (e.g., GPT).

Reviews Domain The reviews domain shows that human-written texts employ a slightly higher use of frequently known words than in the other discussed domains, as we can see in Figure 14.

Wikipedia Domain In the Wikipedia domain, HWT score very low in average text length, as well as in morphological complexity and emotionality, in comparison with the other discussed domains. Clearly, we can observe the effect of very rigid genre-specific constraints and how humans have to strictly abide by them.

News Domain The news domain shows that HWT score high in average text length, morphological complexity, and fairly high in emotionality in comparison to most models. Unlike in other domains, human-written texts also score lowest in word prevalence, indicating the use of less common words, and lower in semantic similarity, suggesting a more varied semantic content.

F Experimental Setup Details

All experiments were carried out with an NVIDIA A100 GPU (40GB memory) and standard CPU resources. We did not train or fine-tune any models; instead, we relied on pre-trained encoders purely for inference:

- **StyleDistance**, based on RoBERTa-base with approximately 125M parameters.
- **paraphrase-MiniLM-L6-v2**, a MiniLM encoder with approximately 22M parameters.

The calculation of linguistic features (e.g., syntactic depth, sentence length, morphological complexity, emotion intensity) was performed using `spaCy` (`en_core_web_sm`) and other lexicon-based methods on CPU. Only the transformer-based components (style embeddings and semantic similarity) were executed on the GPU, requiring about 3 GPU hours. The overall computational budget was therefore modest, with CPU-bound feature extraction accounting for most of the runtime.

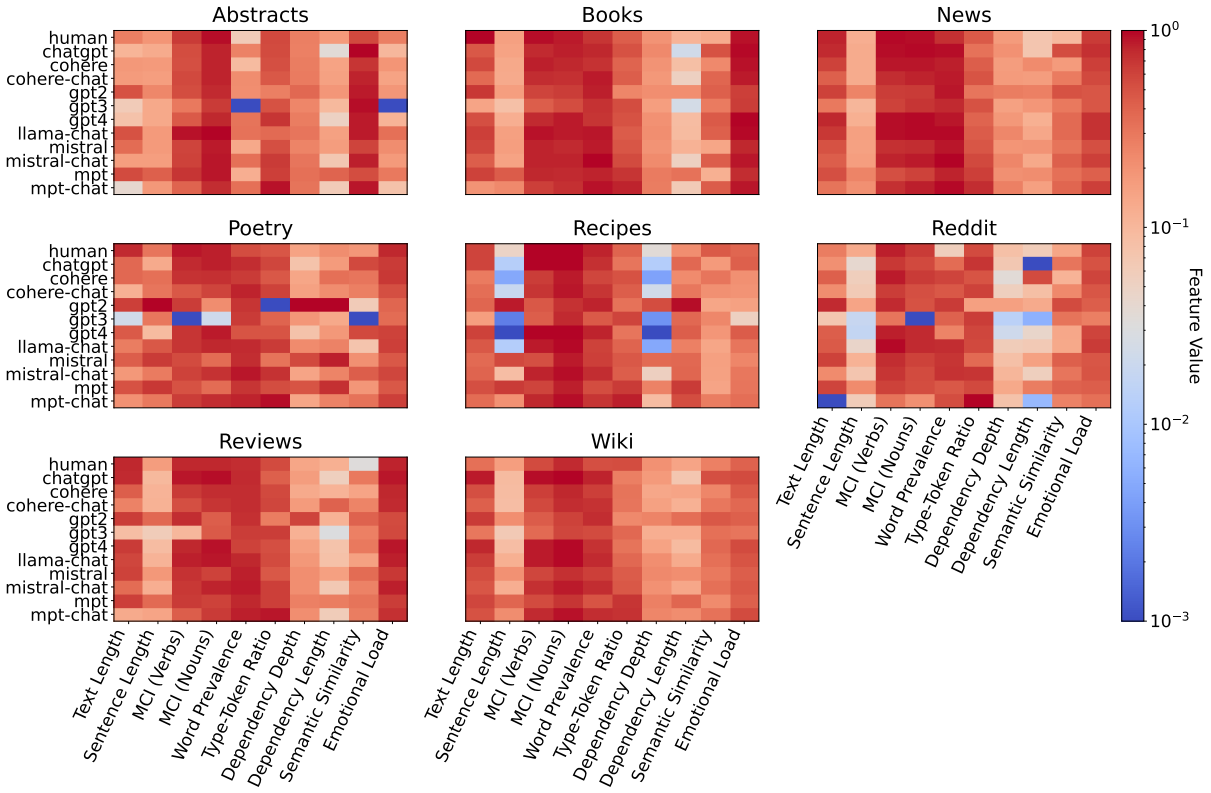

Figure 14: Multi-panel heatmap displaying log-normalized feature values across different text domains and models