Interpretable Stylistic Variation in Human and LLM Writing Across Genres, Models, and Decoding Strategies

Swati Rallapalli[1,*], **Shannon K. Gallagher**[2], **Ronald Yurko**[2], **Tyler Brooks**[1]
Chuck Loughin[1], **Michele Sezgin**[2], **Violet Turri**[1]
[1]Software Engineering Institute, AI Division, Carnegie Mellon University
[2]Department of Statistics & Data Science, Carnegie Mellon University
[*]Current affiliation: Independent Researcher
swatirallapalli@gmail.com

Abstract

Large Language Models (LLMs) are now capable of generating highly fluent, human-like text. They enable many applications, but also raise concerns such as large scale spam, phishing, or academic misuse. While much work has focused on detecting LLM-generated text, only limited work has gone into understanding the stylistic differences between human-written and machine-generated text. In this work, we perform a large scale analysis of stylistic variation across human-written text and outputs from 11 LLMs spanning 8 different genres and 4 decoding strategies using Douglas Biber's set of lexicogrammatical and functional features. Our findings reveal insights that can guide intentional LLM usage. First, key linguistic differentiators of LLM-generated text seem robust to generation conditions (e.g., prompt settings to nudge them to generate human-like text, or availability of human-written text to continue the style); second, genre exerts a stronger influence on stylistic features than the source itself; third, chat variants of the models generally appear to be clustered together in stylistic space, and finally, model has a larger effect on the style than decoding strategy, with some exceptions. These results highlight the relative importance of model and genre over prompting and decoding strategies in shaping the stylistic behavior of machine-generated text.

1 Introduction

As Large Language Models (LLMs) have seen significant advancement in recent times, they are demonstrating fluency and proficiency across a wide range of tasks (Brown et al., 2020; OpenAI, 2024a). As a result, the distinction between human-written text (HWT) and machine-generated text (MGT) is becoming increasingly blurred, making it challenging to differentiate between the two. While this can benefit applications such as conversational agents and assistive writing tools, it also raises concerns regarding traceability and misuse. For instance, misuse such as large scale spam, phishing attacks or academic plagiarism (Brown et al., 2020) is of concern. Furthermore, prior work (Spitale et al., 2023) shows that GPT-3-generated misinformation can be as hard or harder to recognize than human-generated misinformation.

Therefore, a lot of work has gone into classification of human-written text and machine-generated text. Existing approaches include supervised classifiers (Solaiman et al., 2019), methods based on language model probability signals (Gehrmann et al., 2019; Mitchell et al., 2023), and alternative techniques such as watermarking (Kirchenbauer et al., 2023). A comprehensive survey of these methods is provided by (Zanotto and Aroyehun, 2025). While it is interesting to understand whether we can successfully distinguish between LLM-generated and human-written text and the related nuances (e.g., how the results relate to model complexity), it is also important to understand the linguistic features that drive this distinction. This can provide deeper insights into how LLMs learn and generate language and the extent to which they adhere to or diverge from stylistic patterns observed in human language under different settings. These insights can, in turn, guide the development of better classifiers for distinguishing human-written from machine-generated text.

The analysis of stylistic patterns is also of growing interest to users. For instance, recent observations have highlighted characteristic stylistic tendencies in ChatGPT-generated text, such as the frequent use of em dashes, which has attracted considerable attention [1]. Some users seek to understand the underlying reasons for such patterns, while others aim to avoid them due to concerns about their association with machine-generated text. All in all

[1]https://www.nytimes.com/2025/09/18/magazine/chatgpt-dash-hyphen-writing-communication.html

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29243–29261
October 24-29, 2026 ©2026 Association for Computational Linguistics

this suggests the need for systematic understanding of stylistic features of machine-generated language. However, there is limited but growing body of work that explicitly characterizes the linguistic differences between human-written and machine-generated text.

One important paper in this direction is Reinhart et al. (2025). They construct two parallel corpora of human-written and LLM-generated texts from common prompts using several variants of Llama 3 and GPT-4o. With these corpora, they analyze Douglas Biber's set of lexical, grammatical, and rhetorical features drawn from several works by Biber and collaborators (Biber, 1988, 1995; Biber and Conrad, 2009) and identify that LLMs struggle to match human stylistic variation. They observe that systematic differences in the Biber features persist across model sizes, with larger differences for instruction-tuned models compared to base models.

Motivated by this line of work, we analyze the large-scale RAID dataset (Dugan et al., 2024), which contains human-written and LLM-generated text across 11 models, 8 genres and 4 decoding strategies using the Biber features. The goal is to understand LLM-generated text under different prompting strategies, across different genres and decoding methods – factors not explored in the study by Reinhart et al. Also, unlike the dataset used in that study, the RAID dataset does not provide models with access to initial human-written context or explicitly prompt them to mimic human writing styles. Therefore, our goal is to examine whether earlier findings generalize to these different settings at a larger scale, while also addressing questions that could not be explored due to dataset constraints.

Our findings for the LLMs analyzed are as follows: (i) key linguistic characteristics of machine-generated text remain robust across generation conditions and prompt settings; in particular, LLM outputs tend to be more information dense and noun-heavy, especially for instruction-tuned chat models, consistent with prior observations despite differences in data generation settings; (ii) the genre plays a stronger role than the source of text (i.e., human or a specific LLM), with texts from different sources but the same genre exhibiting more similar linguistic features than texts from the same source across genres, suggesting that LLMs adapt their style of writing for different genres (e.g., news versus reviews), even without explicit prompting; (iii) instruction-tuned models frequently cluster separately from base models, while the relationship of human-written text to these model groups varies across genres; and (iv) model choice has a larger effect on stylistic variation than decoding strategy (i.e., algorithm to pick next token from model output space), with some exceptions. These results highlight the relative importance of model and genre over prompting and decoding strategies in shaping the stylistic behavior of machine-generated text, and provide guidance to users when selecting among different LLM configurations. These findings can also inform the development of more robust human-versus-machine text classifiers by identifying stylistic patterns that generalize across generation settings, while also highlighting the need to account for genre and model specific variation.

2 Related Work

In this Section, we focus on the few other works that analyze linguistic features of machine-generated text.

A related study is (Zanotto and Aroyehun, 2025), which also analyzes human-written and machine-generated text using the RAID dataset. However, their analysis addresses complementary research questions using a selected set of linguistic features spanning morphology, syntax, and semantics. Our analysis on the other hand leverages Biber's established framework of 67 linguistic features. They find that human-written text tends to exhibit simpler syntactic structures and more diverse semantic content. They further examine variability across genres and models using these linguistic features as well as style embeddings (Patel et al., 2025), finding greater variability in human-written text. While this work examines linguistic differences across genres, models, and decoding settings, its cross-genre and cross-model analyses largely focus on linguistic variability. In contrast, our analysis investigates the relative influence of genre, model choice, and decoding strategy on the linguistic characteristics of machine-generated text. We further examine which Biber characteristics systematically distinguish human-written and machine-generated text and whether these patterns persist across models and generation settings.

Prior work by the same authors (Zanotto and Aroyehun, 2024) similarly reports greater variability in linguistic features for human-written text compared to LLM-generated text. Additionally, (Sandler et al., 2024) compares human and

ChatGPT-generated dialogues using Linguistic Inquiry and Word Count (LIWC) features, finding that human conversations exhibit greater variability and authenticity. These works emphasize variability in linguistic features, which is not the primary focus of our analysis.

Other recent work has examined linguistic differences between human-written and LLM-generated text, identifying systematic variations in stylistic and structural patterns (Zeleke et al., 2025; Muñoz-Ortiz et al., 2024) in specific settings. A recent survey (Terčon and Dobrovoljc, 2025) reviews existing work on the linguistic characteristics of machine-generated text and highlights the need for more large-scale and systematic analyses. In this work, we conduct a more comprehensive stylistic analysis of machine-generated text across diverse models, genres, and generation settings.

3 Approach

In this Section we overview the dataset and the features used in our analysis.

3.1 Dataset

We use the RAID dataset (Dugan et al., 2024), a large-scale benchmark generated for machine-generated text detection. RAID comprises over 6 million text samples spanning 12 sources, including human-written text and outputs from 11 large language models (LLMs): 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). The dataset covers 8 genres: Abstracts, Books, News, Poetry, Recipes, Reddit, Reviews, Wikipedia and includes 11 adversarial attacks. Text is generated under 4 decoding configurations, combining greedy decoding and random sampling, each with and without repetition penalties where applicable.

The dataset contains 13,371 human written documents, across the above 8 genres (roughly 2,000 documents for each category). The sources of the human-written documents are described in (Dugan et al., 2024) (Appendix E.1). The human-written documents were sampled from publicly available pre-2022 corpora to minimize contamination from AI-generated content. For each human-written document, machine-generated outputs were produced using 11 models, four decoding strategies, and 11 adversarial attacks. Generation was performed in a zero-shot manner using prompt templates based on the model type. "Chat" models fine-tuned on dialogue were prompted using chat-style instructions (e.g., `Write the abstract for the academic paper titled {title}`), while continuation-based "Non-Chat" models were prompted with prefix-style templates (e.g., `The following is the full text of the abstract for a research paper titled {title} from arxiv.org:`).

The prompts were otherwise similar across settings, with the `title` field replaced with the title of the corresponding human-written document. Importantly, no explicit constraints were imposed on properties such as style or length, allowing the generations to more closely reflect realistic scenarios.

Most prior works use this dataset for detection of machine-generated text. We, on the other-hand, use it to study the lexical and stylistic properties of human-written versus machine-generated text. We use a subset of this dataset which excludes texts generated using adversarial attacks in our analysis. The goal of the selected subset is to reflect common generation settings. Therefore, in the end, our analysis dataset contains $467,985$ texts.

The dataset used for the analysis in (Reinhart et al., 2025) used a different set-up. First, it used a smaller-set of 6 models, namely, multiple versions of GPT-4o (OpenAI, 2024c,b) and Meta Llama3 (Meta AI, 2024). Second, the prompt settings are different. Instruction-tuned models are prompted with: "`In the same style, tone, and diction of the following text, complete the next 500 words, generate exactly 500 words, and note that the text does not necessarily end after the generated words:`", followed by an initial chunk of human-written text. The base (non-instruction-tuned) Llama 3 variants are provided only with the initial text chunk as a prefix. In either case, the models generate approximately 500 words of output, extending the given human-written text which was then compared against the corresponding second chunk of the original human-written document.

The differences in dataset construction lead to the following distinctions in our analysis. First, our setting captures a different range of stylistic variation: RAID generations are produced without explicit prompts to mimic human style, unlike (Reinhart et al., 2025), where models are guided to closely match human-written text and are provided with an initial text chunk. Second, we analyze genre level differences among the texts generated

by different sources. Third, we analyze the stylistic differences stemming from the different decoding styles available in the RAID dataset.

3.2 Features

As in (Reinhart et al., 2025) paper, we use 67 Biber features (Biber, 1985), extracted using the `pseudobibeR` (Brown, 2024) package. These features capture a broad range of grammatical and structural properties of text, including tense and aspect markers, clause types, nominalizations, and other syntactic constructions. Biber features are designed to quantify stylistic variation by characterizing how language is used, such as differences in information density or narrative style rather than what is being said (i.e., topic or semantic content). A detailed description of the feature set can be found in (Brown, 2024) and Appendix Table 2 of (Reinhart et al., 2025). The rate of occurrence for most features is counted per 1000 works of text, while some are derived measures. These values serve as features for our analysis.

Chunking and weighted averaging: Chunking was necessary due to the input-size limitations of the `pseudobibeR` package. For the classification experiments described in Section 3.2.1, we used a chunk size of $10,000$ tokens to obtain a sufficient number of training instances. For the remaining analyses, we used a chunk size of $900,000$ characters, below the package limit of 1 million characters, to process as much text as possible in each chunk. Feature values across chunks were aggregated using a token-count-weighted average.

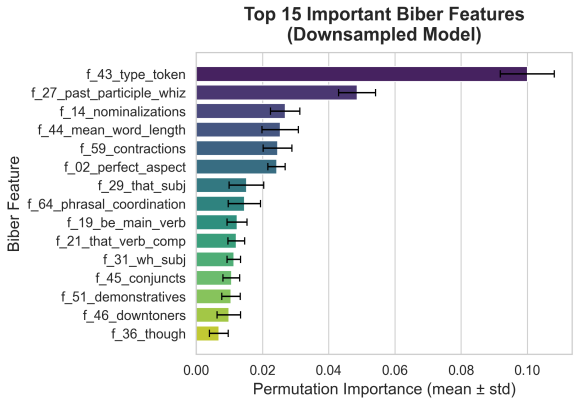


Figure 1: Top 15 most important Biber features for distinguishing human and LLM-generated text based on Random Forest classifier on down-sampled model.

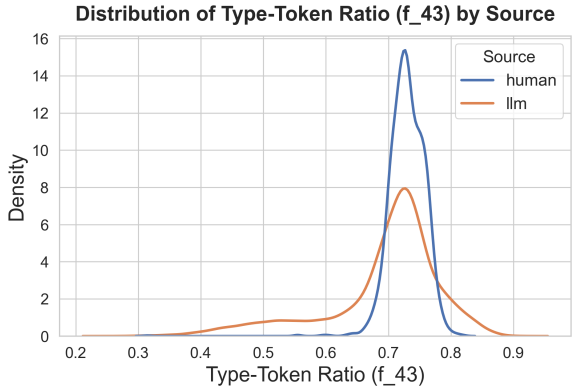


Figure 2: Feature Distribution for `f_43_type_token`

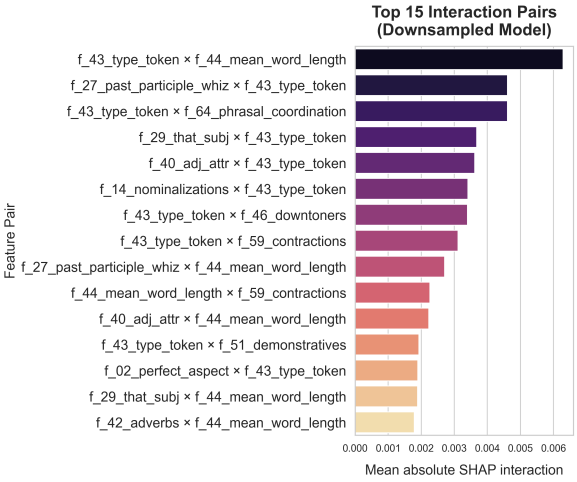


Figure 3: Top 15 SHAP interaction pairs based on Random Forest classifier on down-sampled model.

3.2.1 Important Biber features based on classification

While we look at 67 Biber features in our analysis, it is useful in some cases to zoom in on a few. Moreover, a dimension to understand is which of these features are actually helpful in discriminating between human-written text (HWT) and machine-generated text (MGT). The goal of this work is not the classification itself, but understanding the most discriminative features.

To this end, we train a Random Forest Classifier from `scikit-learn`[2] (with 200 trees) using Biber features on the RAID data. The dataset is inherently imbalanced, as each human-written document is paired with outputs from 11 LLMs. Therefore, we train two models: (i) a base model using the full imbalanced data, and (ii) a down-sampled model in which the majority class (MGT) is reduced. We use a $60\%/20\%/20\%$ train/validation/test split. For the down-sampled model, we tune the class ratio on the validation set using F1 score for the human

[2] RandomForestClassifier documentation scikit-learn

class. A ratio of 4:1 (MGT:HWT) yields the best performance. We then train both models on combined train and validation data and evaluate on the held-out test set.

On the test set, the down-sampled model achieved an F1 score of 0.67 for the human class and 0.99 for the LLM class, compared to 0.21 and 0.98, respectively, for the base model. However, the AUC is very close for both the models: 0.9775 for down-sampled model and 0.9755 for base-model, that is the discriminative ability of both models is comparable with a slight advantage with the down-sampled model. The corresponding ROC curves are shown in Appendix Figure 9. We use permutation importance from the down-sampled model to identify the top 15 most discriminative features, which are the focus of our analysis.

Figure 1 presents these top 15 features. We further examine them to understand the factors driving their discriminative power. Notably, the most important feature `f43_type_token` is a measure of lexical diversity or vocabulary richness, defined as the ratio of unique words (types) to the total number of words (tokens) in a given text. As we will see in Section 4, it is neither overused nor underused by LLMs. Then to better understand its importance, we analyze its distribution in Figure 2. This shows that for humans, it is narrow and highly predictable around the mean value. The distribution for all the features is in the Appendix Figures 10 and 11.

Also, in Section 4 we look at features that are overused or underused by LLMs, and we notice that 4 of the top 5 overused features are among the top 15 important features. There is also some representation from the underused features in the top 15 most discriminatory features. For other important features, discriminative power may arise from differences in their distributions, as illustrated by `f43_type_token`, or from their interactions with other features.

To investigate feature importance further, we look at the top interacting features using SHAP interaction values[3] in Figure 3 (computed on 500 random samples from the test set for computational efficiency). This allows us to better understand why certain features are important and to identify combinations of stylistic features that jointly contribute to distinguishing human-written and machine-generated text.

We observe that features involved in the strongest interaction pairs largely overlap with the most important features identified earlier via permutation importance. Note that `f_43_type_token` appears in a large fraction of the highest-ranked SHAP interaction pairs, suggesting that its contribution to classification is related not only to its individual predictive value but also to its interactions with multiple other linguistic features.

Comparing our important features with those identified in (Reinhart et al., 2025), 7 of the top 15 features overlap. The remaining differences are likely attributable to differences in the datasets and experimental settings. Additionally, 3 out of 4 features (`f_29_that_subj`, `f_14_nominalizations`, `f_64_phrasal_coordination`) identified in (Reinhart et al., 2025) as being most overused by instruction-tuned models appear in our top 15 most important features, providing further evidence of consistent stylistic trends. Our study complements the feature-importance analysis in (Reinhart et al., 2025) by additionally examining interactions among Biber features and their joint contribution to distinguishing human-written and machine-generated text.

4 Results

4.1 Feature Overuse/Underuse by LLMs

We analyze overuse and underuse of Biber features in LLM-generated text relative to human-written text, both in aggregate across models and at the level of individual LLMs.

4.1.1 Aggregate over all LLMs

We first analyze aggregate feature usage across LLMs relative to human-written text. For each Biber feature, the input consists of a token-count-weighted average value for each of the 11 LLMs and a corresponding weighted-average value for human-written text. We compute the mean and standard error of the feature values across the 11 LLMs and normalize these by the corresponding mean feature usage in human-written text. We then focus on the: (i) top 5 overused and underused features, and (ii) visualize the ratios for all features.

Top overused/underused features: We identify the five most overused and underused Biber features in LLM-generated text, considering all features (not limited to those in Figure 1). We do not define overuse and underuse using fixed thresholds. Instead, we compute the average ratio of feature usage in LLM-generated text to that in human-written

[3] SHAP TreeExplainer documentation

text, rank all features by this ratio, and report the five highest and lowest ranked features as the most overused and underused, respectively.
(1) Top overused features in LLM-generated text:

1. `f_29_that_subj`: *that* relative clauses on subject position (e.g., the dog *that bit me*)
2. `f_59_contractions`: Contractions (e.g., *can't*, *won't*)
3. `f_27_past_participle_whiz`: Past participial postnominal clauses (e.g., The solution *produced by this process*)
4. `f_14_nominalizations`: Nominalizations (noun forms derived from verbs or adjectives)
5. `f_34_sentence_relatives`: Sentence relatives (e.g., Bob likes fried mangoes, *which is the most disgusting thing I've ever heard of.*)

Some of these features contribute to a more information dense, noun-heavy style, particularly nominalizations and past participal postnominal clauses while others reflect more elaborative or conversational nature of LLM-generated text.
(2) Top underused features in LLM-generated text:

1. `f_32_wh_obj`: *Wh*-relative clauses functioning as objects (e.g., the man *who Sally likes*)
2. `f_33_pied_piping`: Pied-piping relative clauses: relative clauses moved in the sentence by 'wh-' questions (e.g., the manner *in which he was told*)
3. `f_36_though`: Concessive subordinators (e.g., *though*, *although*)
4. `f_66_neg_synthetic`: Synthetic negation (e.g., *No answer is good enough*)
5. `f_50_discourse_particles`: Discourse particles in sentence-initial position (e.g., *well*, *now*, *anyway*)

These features generally reflect complex style of writing and are relatively infrequent even in human text (they have a mean human frequency between $0.2 - 1.2$, whereas for the top overused features it is between $1 - 14$), suggesting that LLMs further suppress already rare constructions.
(3) Observations: Note that 4 of the top 5 overused features are among the important features for classification (Figure 1). Only one of the top 5 underused features is among top 15 important features. Since the underused features generally occur at low frequencies in both human-written and machine-generated text, their discriminatory ability maybe limited.

Visualize ratios: We look at the ratio of mean feature usage of LLMs to Humans for top 15 most discriminatory features in Figure 4 (and all the features in Appendix Figure 12). Note the log scale for plotting (ratio is not log, just the visualization is in log scale). The error bars represent the SEM across the 11 LLMs, normalized by the human-written feature value.

The patterns in Figure 4 are broadly consistent with findings in (Reinhart et al., 2025). Even though the datasets used for the analysis are very different and models analyzed have only some overlaps, (Reinhart et al., 2025) also found that instruction-tuned models favor present participial clauses (`f_25_present_participle` in Appendix Figure 12), what they describe as 'that' clauses as subjects (`f_29_that_subj`), nominalizations (`f_14_nominalizations`), and phrasal co-ordination (`f_64_phrasal_coordination`), which are generally markers of more information dense, noun-heavy style of writing. These features are among the top overused features even in our aggregate analysis (in Section 4.1.2 we will also see that overuse is more pronounced in the instruction-tuned "Chat" models). This is interesting because of the differences in the experimental setup: in (Reinhart et al., 2025), models were conditioned on the first 500 tokens of human-written text and explicitly prompted to mimic human style, whereas neither condition holds for the RAID dataset used in our analysis.

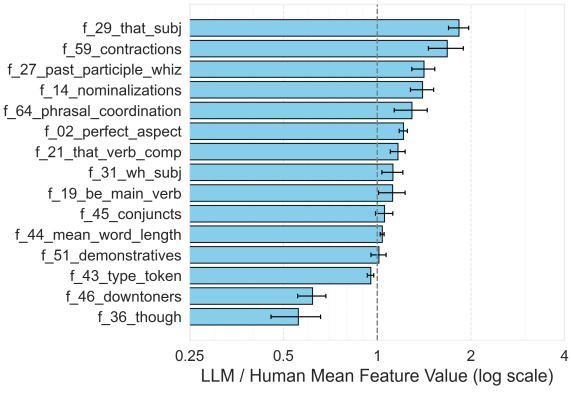


Figure 4: Ratio of Feature Usage LLM Vs. Human for Selected Features.

As an additional robustness analysis, Appendix Figure 13 reports 95% bootstrap confidence inter-

vals for the mean LLM-to-human feature ratios.

4.1.2 Individual LLMs

We look at the ratio of feature usage of individual LLMs to humans for top 15 selected features in Figure 5. We plot the log of ratio for visual clarity. This reveals some interesting trends. First, many features seem to be showing similar trends across LLM models. For instance `f_29_that_subject` is overused significantly by all models, `f_46_downtoners` is underused by all. Some features show a mixed trend for e.g., `f_31_wh_subj`, `f_19_be_main_verb`. This is also the trend we see in the corresponding figure for all features is in the Appendix 14. Second, to re-iterate what we mentioned in Section 4.1.1, despite differences in experiment settings, we observe trends similar to (Reinhart et al., 2025). Specifically, they found that instruction-tuned models tend to favor present participial clauses (please refer to Appendix 14 for this feature `f_25_present_participle` as it was not selected among the top important features), 'that' clauses as subjects, nominalization, and phrasal coordination, which are characteristics of more information dense, noun-heavy style of writing. These features are among the top overused features even in our previous aggregate analysis. Here, additionally, we would like to highlight that the overuse is more pronounced in the instruction-tuned "Chat" models. Finally, we notice certain models are closer to human writing than others, for e.g., the Cohere model exhibits feature usage ratios that are consistently closer to 1.

4.2 Visualization of Features Across Sources and Genres

In this analysis, we look at the writing style across genres and sources (human or a specific LLM) using the Biber features. For each genre–source pair, we compute the average of each feature, yielding a 67-dimensional feature vector (one dimension per Biber feature). Each feature value represents the weighted average over all text chunks within that genre and source. This results in a matrix $\mathbf{X} \in \mathbb{R}^{96\times67}$, corresponding to 8 genres and 12 sources (i.e., $8 \times 12 = 96$). We perform z-score normalization on this matrix for each feature to have zero mean and unit standard deviation. We then perform Principal Component Analysis (PCA) and visualize the first two principal components in Figure 6. Throughout the paper, we use the `PCA`[4] and `StandardScaler`[5] implementations from `scikit-learn` for Principal Component Analysis and z-score normalization, respectively. We further plot confidence ellipses for each category, centered at the mean of the genre and axis lengths are proportional (twice in our implementation) to the standard deviation along those axes. This leads to an interesting observation – the genre plays a more prominent role in shaping the writing style than the source, i.e., texts from same genre tend to cluster together irrespective of the source. Further, the confidence ellipses are elongated for `Reviews` and `Poetry`, indicating a larger variance along the first principal component, which is likely due to creative differences in `Poetry` and subjectivity in `Reviews`. Finally, while `Poetry`, `Abstracts` cluster away from the center showing more specialized or niche writing styles, genres like `Books`, `News` and `Wikipedia` cluster closer to the center, indicating a more general writing style.

4.3 Hierarchical Clustering of Sources of Text

We study the similarities between text sources by clustering the 11 LLM-generated text documents together with human-written text. As genre is a dominant factor, we perform clustering separately for each genre to avoid mixing cross-genre and cross-source patterns. For each genre–source pair, we construct a feature vector using the weighted average of Biber features across all text chunks. We then z-score normalize each feature across sources. To reduce redundancy among correlated features, we apply PCA and retain components that explain 95% of the variance, typically reducing the dimensionality from 67 to 6–7 components. We perform hierarchical clustering on these PCA-reduced representations using Ward's method[6] and visualize the resulting dendrograms in Appendix Figure 15.

To summarize clustering patterns across genres, we construct a consensus tree (Figure 7)[7] using a relaxed threshold ($p = 0.4$), retaining groups that occur in at least 40% of the genre-specific trees. This relaxation is to allow more compact and interpretable summary of recurring clustering patterns. The full set of dendrograms (Appendix Figure 15) shows that human-written text groups with the base models in five of the eight genres.

[4] PCA documentation

[5] StandardScaler documentation

[6] `https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html`

[7] `https://rdrr.io/cran/ape/man/consensus.html`

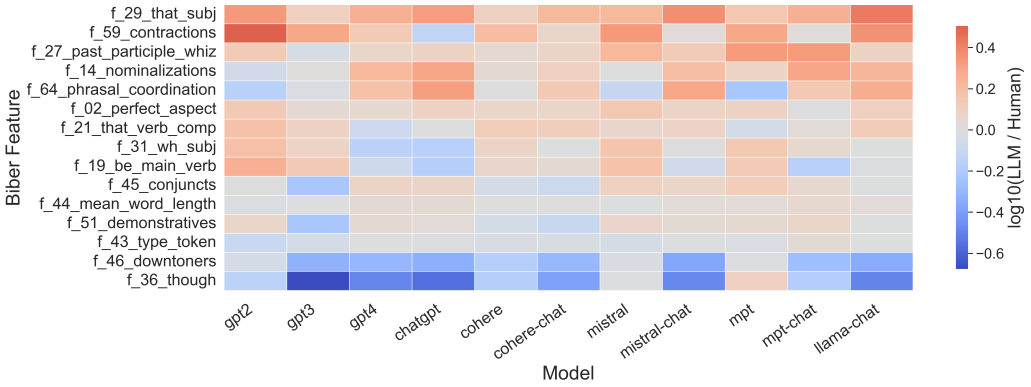


Figure 5: Heat map of Log10 Biber Feature Ratios (LLM / Human) for Selected Important Features

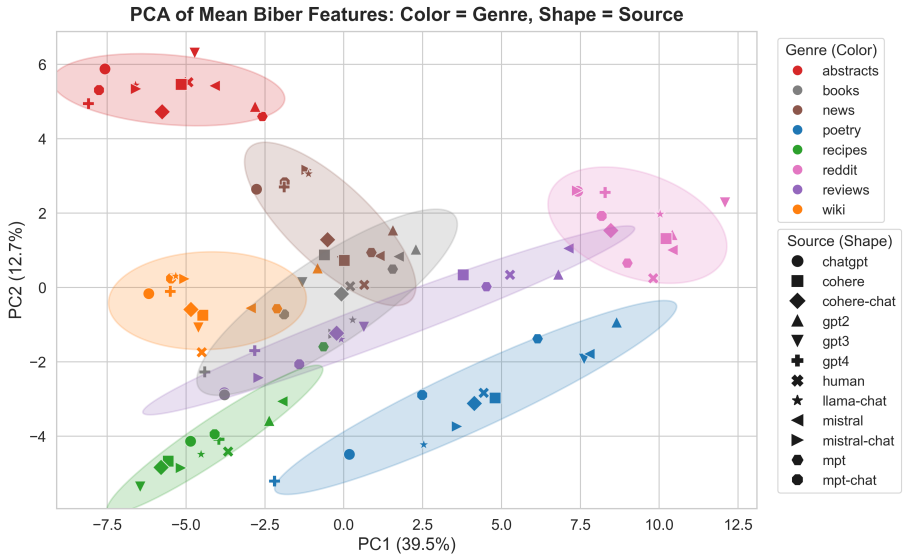


Figure 6: Scatterplot of first two PCA dimensions of mean Biber features.

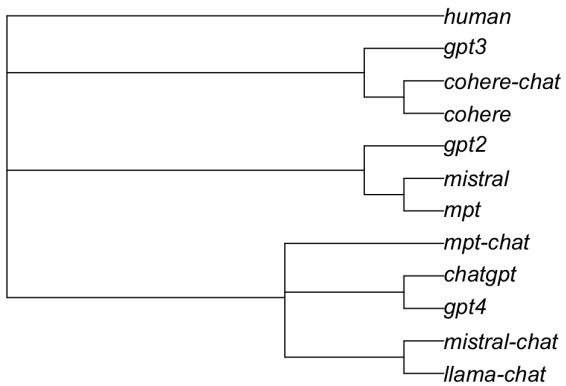


Figure 7: Relaxed consensus tree summarizing hierarchical clustering results over all genres.

This could be related to findings in (Reinhart et al., 2025), which suggest that instruction tuning can shift chat-based models further away from human-written text. We further observe that *Mistral*, *GPT2*, and *MPT* tend to cluster together. Also, *GPT4* and *ChatGPT* form a group, and *Cohere* and *Cohere-Chat* cluster closely. These patterns show the similarity between base models and their chat variants in some of the model families.

As a robustness check, we also performed hierarchical clustering directly on all 67 standardized Biber features, without PCA. The resulting clustering patterns were similar, indicating that the core clustering result is not an artifact of dimensionality reduction. For brevity, we report only the results based on the more compact PCA-reduced representation.

4.4 Writing Styles per LLM Decoding Strategy

RAID dataset contains outputs from multiple LLMs under different decoding strategies. Specifically, greedy decoding and random sampling with and without repetition penalties (where applicable). Decoding strategy determines how tokens are selected from a language model's probability distribution. Greedy decoding picks the most likely token at every time step. Random sampling draws

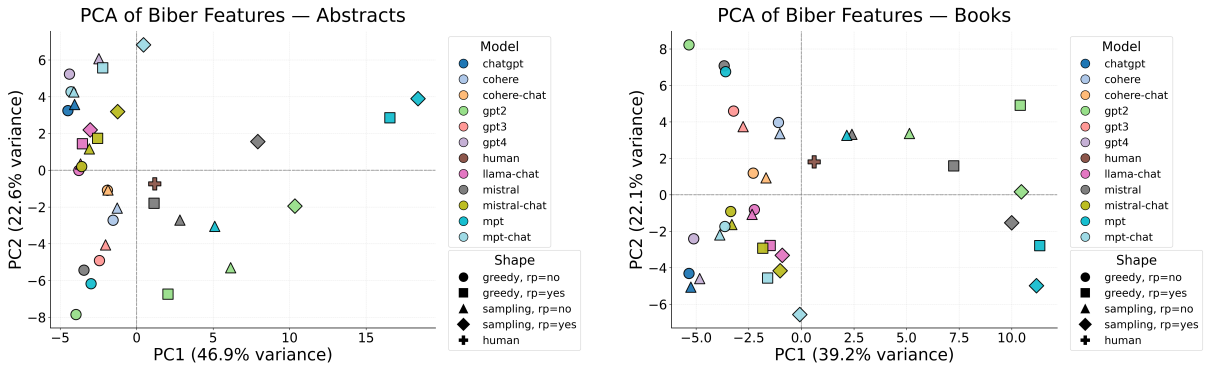


Figure 8: Visualizations of Biber features across selected RAID genres. Colors indicate model identity, while marker shapes indicate decoding and repetition-penalty settings. Human text is shown with a distinct marker.

from the language model output distribution. When repetition penalty is applied, the probability of previously generated tokens is reduced. In this section, we analyze how Biber feature usage varies across decoding strategies relative to human-written text. As prior analyses indicate that genre is a dominant factor, we conduct this analysis separately for each of the eight genres to isolate the effect of decoding. For each genre, we visualize the first two principal components of z-score normalized Biber features. For brevity we present two genres (one more specialized- `Abstracts` and one more generic `Books`) in Figure 8. All the genres are presented in the Appendix for reference (Figure 16). We make the following observations. First, across all genres, human-written text tends to fall approximately in the center. Second, model generally has a stronger effect than decoding strategy: points corresponding to the same model (color) tend to cluster together, with some exceptions (e.g., points from *GPT-2*, *MPT*, *Mistral* appear dispersed). For these models decoding strategy can have a noticeable effect in some cases. For example, sampling without repetition penalty tends to produce similar patterns, while greedy decoding without repetition penalty also groups together across these models. This may be because these models tend to cluster together irrespective of the decoding strategy (Section 4.3).

5 Conclusion

In this work, we analyze the RAID dataset, a large-scale benchmark for machine-generated text detection, using Biber features to characterize the stylistic variation between machine-generated and human-written text across multiple models, genres and decoding strategies. For the LLMs analyzed, our findings can be summarized as follows. First, the dominant stylistic patterns in LLM-generated text appear robust to variations in prompt settings. Second, genre plays a stronger role in shaping stylistic features than the source of text itself, showing that LLMs have effectively learned to mimic the genre-specific variation in human writing. Third, chat-based and base models tend to form distinct clusters, suggesting systematic stylistic differences that may be associated with instruction tuning and other post-training procedures. Finally, within each genre, model generally has a stronger influence on stylistic features than decoding strategy, with some exceptions.

Understanding these stylistic differences is important for several reasons. It can inform more effective use of LLMs, for example by guiding prompting strategies to achieve desired writing styles. Moreover, it can also help with identifying characteristic signatures of machine-generated text in order to aid detection. For instance, training or conditioning first on genre may especially be important for attaining accurate detection as genre is a major source of stylistic variation. This aligns with prior work that has shown that detector performance can degrade when generalizing to unseen text domains (Li et al., 2024; Kuznetsov et al., 2024). Our finding that model choice produces greater stylistic variation than decoding strategy also suggests that diversity across generator models may be particularly important when training detectors, although decoding strategy can still affect detector performance (Dugan et al., 2024).

Finally, as LLMs continue to evolve, ongoing analysis will be necessary to track how their stylistic characteristics change over time.

6 Limitations

Our analysis leverages Douglas Biber's set of linguistic features to analyze the models, genres, and decoding strategies covered in the RAID dataset. Therefore, our findings are inherently limited by the patterns represented in the dataset and by those that can be captured through Biber features. In addition, our study only considers English-language text.

The analysis could be expanded by incorporating additional linguistic analysis frameworks such as NeuroBiber (Alkiek et al., 2025), as well as semantic or content-based analyses. Comparing these approaches with Biber feature analysis could help better understand their similarities and differences. Another limitation is the applicability of the findings to newer models or models not included in the RAID dataset. As language models continue to evolve rapidly, our understanding of their stylistic characteristics must evolve as well.

7 Ethical Considerations

While our goal is to promote transparency and the responsible deployment of LLMs, this work carries potential risks common to stylistic analyses of machine-generated text. For instance, the insights presented here could be used by individuals to generate text that more closely resembles human writing. Additionally, the conclusions may be overused or misapplied in automated moderation systems or authorship judgments. We should remain vigilant against unintended misuse of such stylistic analysis methods.

We acknowledge the use of AI assistants for coding support and language editing during the preparation of this manuscript.

Acknowledgments and Copyright

Copyright 2026 Association of Computational Linguistics. This material is based upon work supported by the Department of War under Air Force Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. References herein to any specific entity, product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity. This work is licensed under a Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/).
Copyright © 2026 Association for Computational Linguistics (ACL). All Rights Reserved. This work product was created in part using generative AI. DM26-0872.

References

Kenan Alkiek, Anna Wegmann, Jian Zhu, and David Jurgens. 2025. Neurobiber: Fast and interpretable stylistic feature extraction. *Preprint*, arXiv:2502.18590.

Douglas Biber. 1985. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, 23(2):337–360.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

David W. Brown. 2024. pseudobiber: Aggregate counts of linguistic features. `https://github.com/browndw/pseudobibeR`. R package version 1.1.

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

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.

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.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Robust AI-generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17036–17055, Miami, Florida, USA. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed: 2025-09-21.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

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

OpenAI. 2024a. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-09-21.

OpenAI. 2024c. Hello gpt-4o. Accessed: 2025-09-21.

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. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. *Preprint*, arXiv:2401.16587.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

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

Luka Terčon and Kaja Dobrovoljc. 2025. Linguistic characteristics of ai-generated text: A survey. *Preprint*, arXiv:2510.05136.

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

Sergio E. Zanotto and Segun Aroyehun. 2025. Linguistic and embedding-based profiling of texts generated by humans and large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22841–22858, Suzhou, China. Association for Computational Linguistics.

Brook Zeleke, Amish Soni, and Lydia Manikonda. 2025. Human or genai? characterizing the linguistic differences between human-written and llm-generated text. In *Companion Publication of the 17th ACM Web Science Conference 2025*, Websci Companion '25, page 34–37, New York, NY, USA. Association for Computing Machinery.

A Appendix

This section provides additional visualizations that support and extend the results presented in the main text. Figure 9 shows the ROC curve comparing the base model and the downsampled model explained in Section 3.2.1. The distribution of all the features referred to in the same Section are in Figures 10 and 11. Figure 12 shows the ratio of aggregate feature usage across all LLMs to that in human-written text, over all features, as explained in Section 4.1.1 (paragraph on visualization of usage ratios).

As an additional robustness analysis, we estimate 95% bootstrap confidence intervals for

the mean LLM-to-human feature ratios using the chunk-level feature values prior to aggregation. For this analysis, we use the original chunks (up to 900,000 characters) from which the aggregate feature values were computed. For each of 10,000 bootstrap iterations, chunks are resampled with replacement separately for human-written text and for each of the 11 LLMs, and token-count-weighted feature averages are recomputed. We then compute the ratio between each LLM and human-written text and average these ratios across the 11 LLMs. The 2.5th and 97.5th percentiles of the resulting bootstrap distribution are reported as the 95% confidence interval in Figure 13. Note that the error bars in this figure represent bootstrap confidence intervals, whereas those in Figure 12 represent the standard error of the mean (SEM) across the 11 LLMs, normalized by the human-written feature value. We retain the relatively large chunk sizes used in the original aggregate analysis so that the bootstrap analysis directly corresponds to the estimates reported in Figure 12. As a sensitivity check, we also repeated the bootstrap analysis using smaller chunks and obtained tighter confidence intervals with qualitatively similar results.

Figure 14 plots the ratio of individual LLM feature usage relative to humans, across all features, as explained in Section 4.1.2. Figure 15 visualizes the clustering of text sources for each individual genre as explained in Section 4.3. Figure 16 visualizes the effect of decoding strategies over all the genres, as explained in Section 4.4.

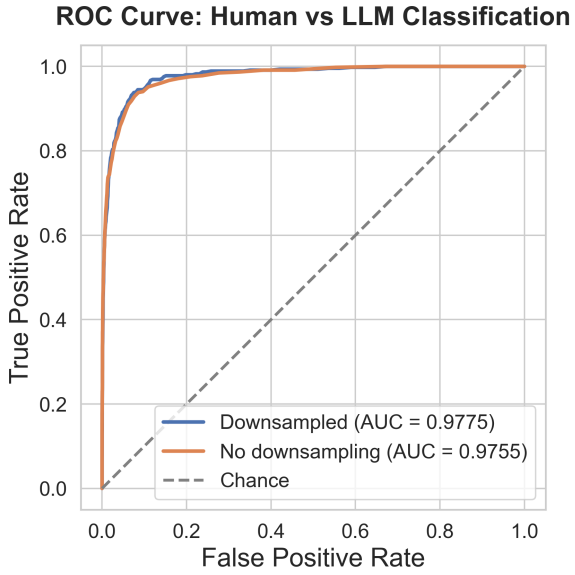


Figure 9: ROC curve comparing base model and downsampled model.

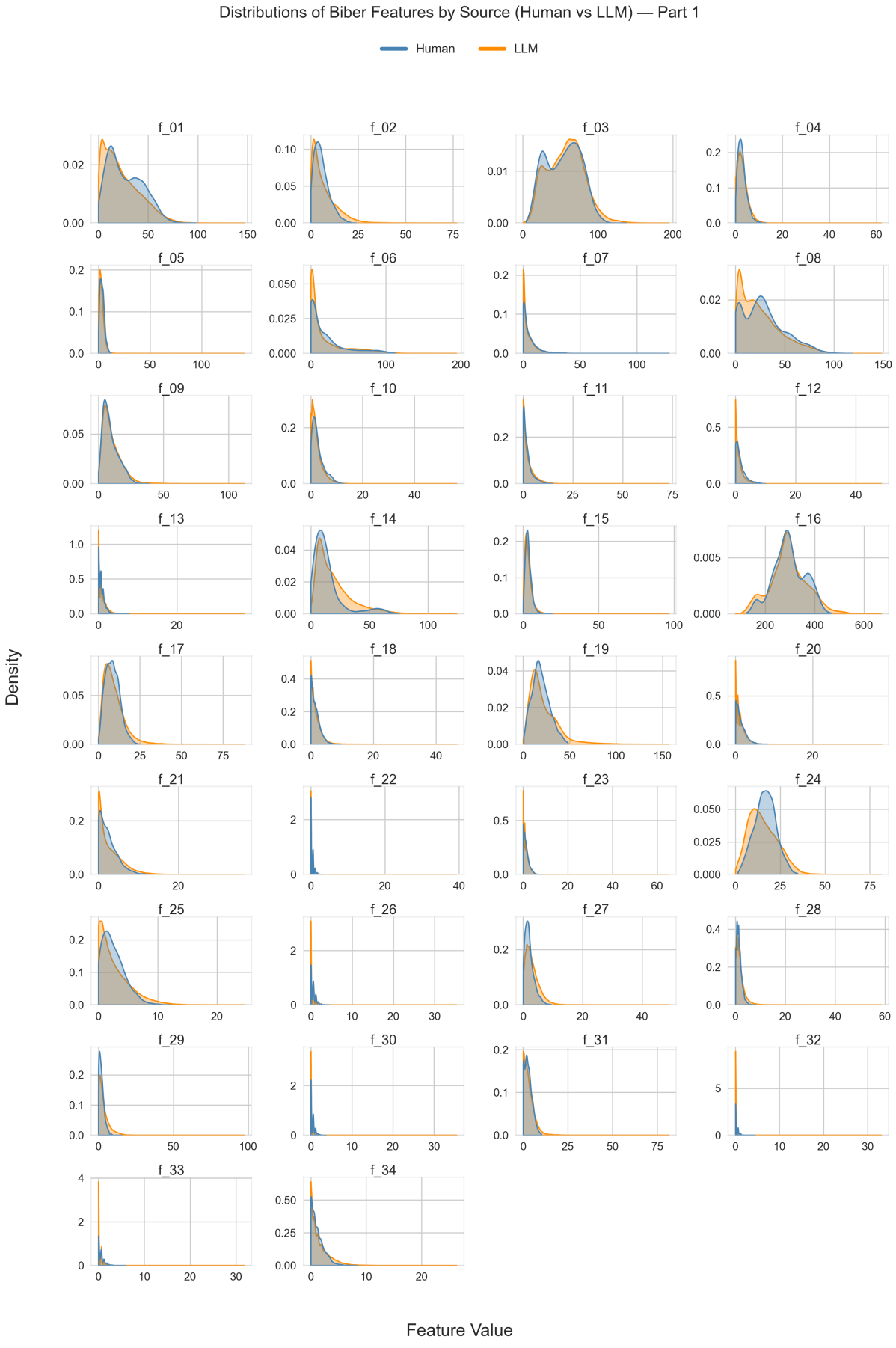


Figure 10: Distribution of Biber features (Part 1).

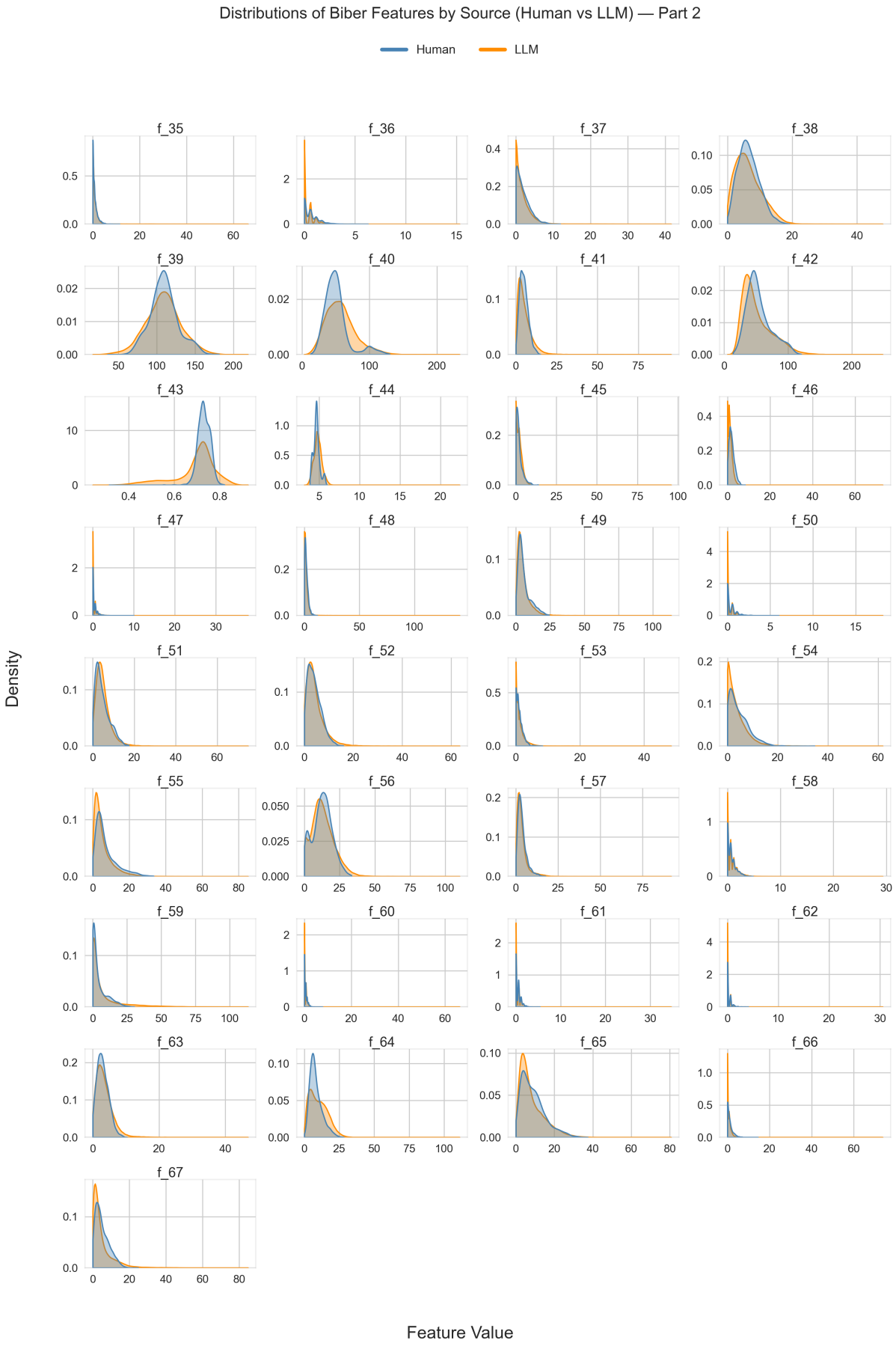


Figure 11: Distribution of Biber features (Part 2).

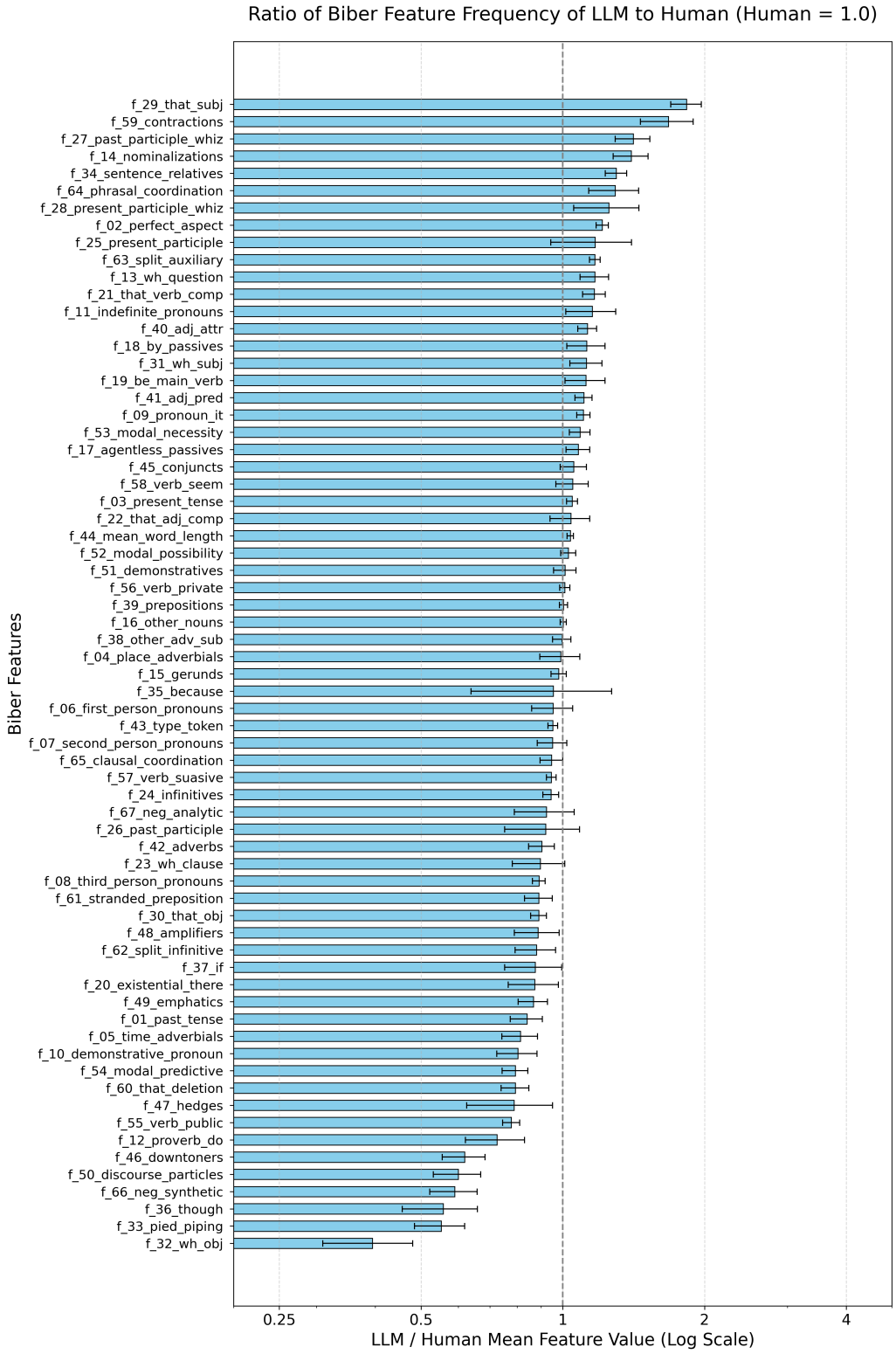


Figure 12: Ratio of Feature Usage LLM Vs. Human for All Features.

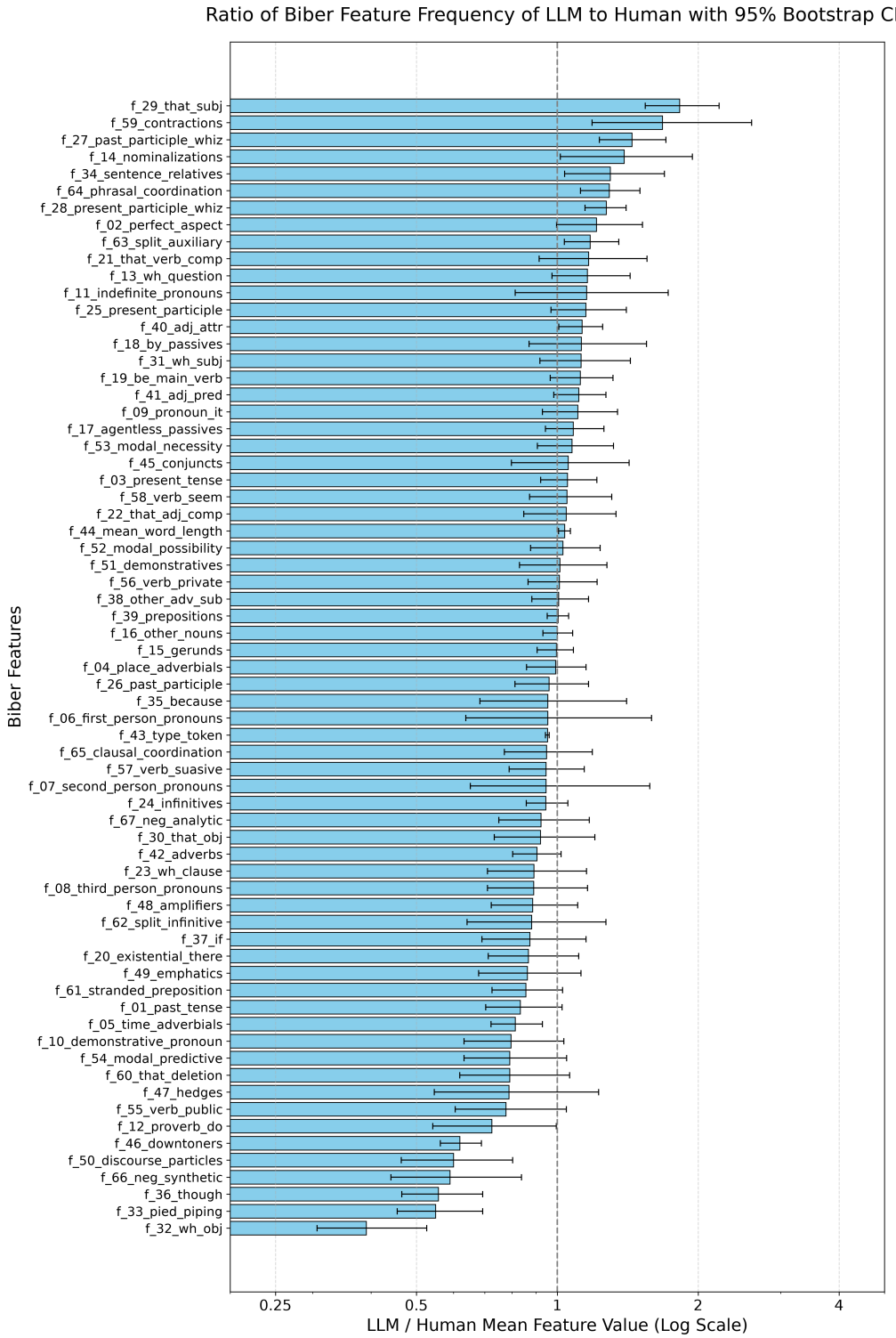


Figure 13: Ratio of Feature Usage LLM Vs. Human for All Features with 95% bootstrap Confidence Intervals.

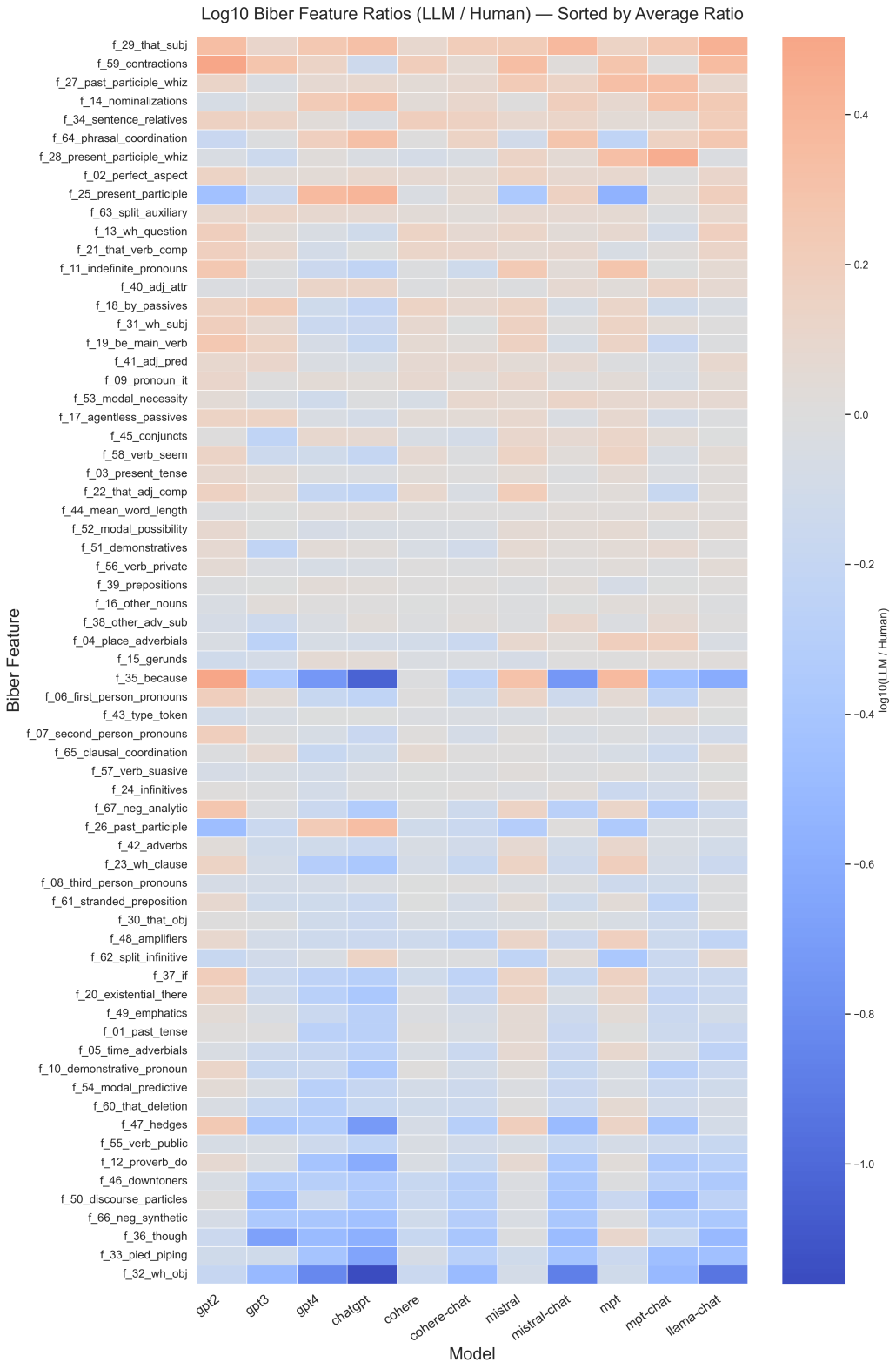
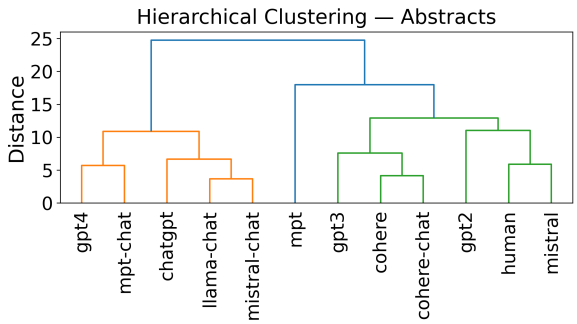
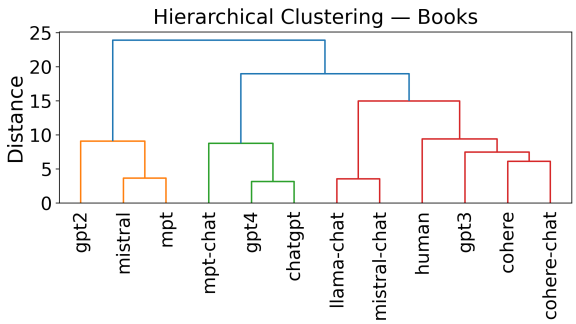


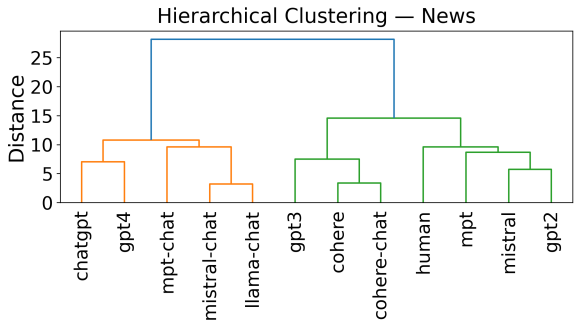
Figure 14: Heat map of Log10 Biber Feature Ratios (LLM / Human) for All Features



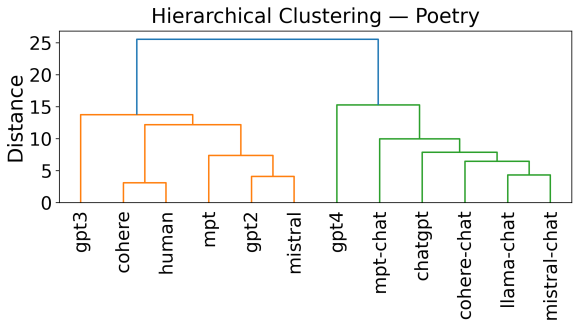
(a) Abstracts



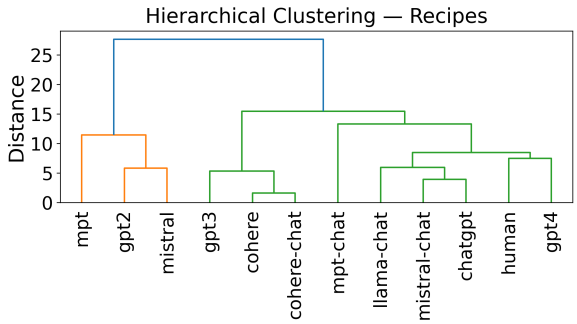
(b) Books



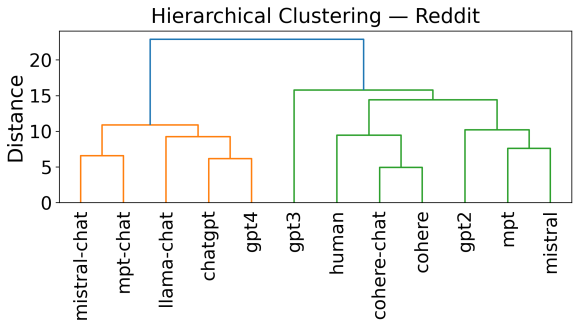
(c) News



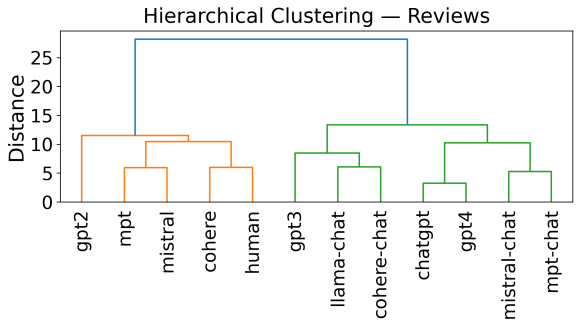
(d) Poetry



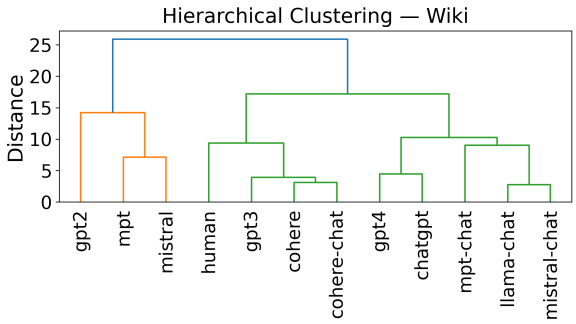
(e) Recipes



(f) Reddit

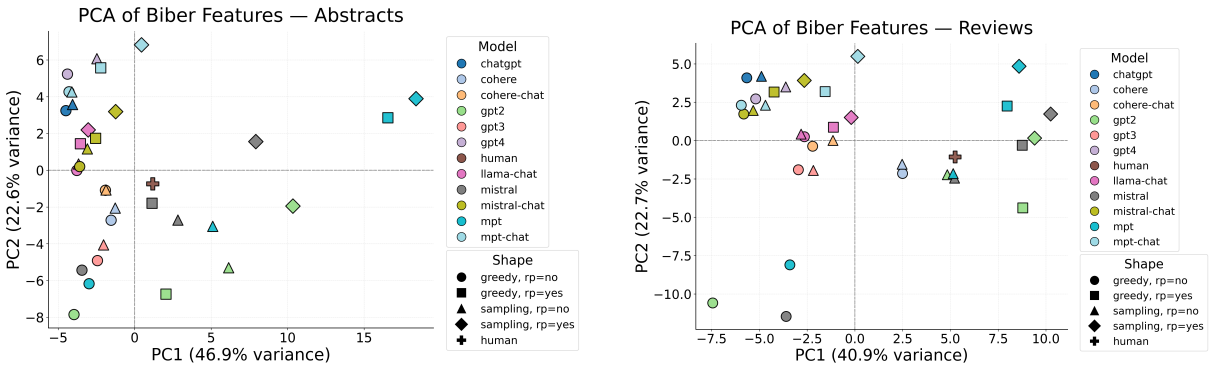


(g) Reviews



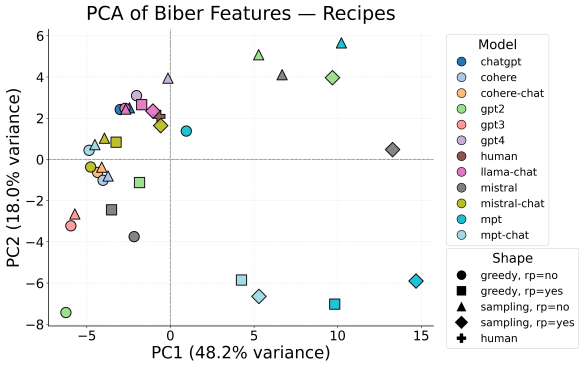
(h) Wikipedia

Figure 15: Hierarchical clustering dendrograms using PCA with components capturing 95% variance on normalized biber features.

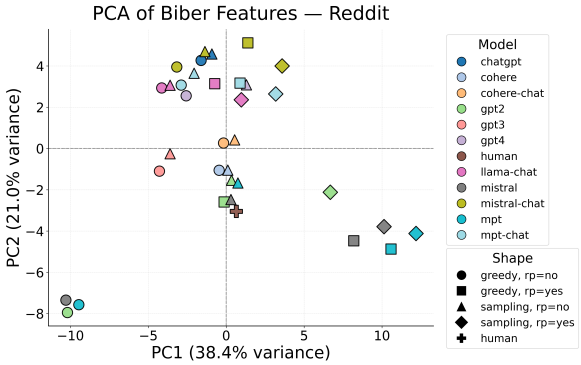


(a) Abstracts

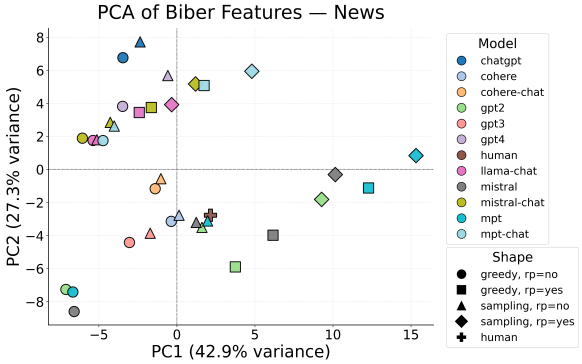
(b) Reviews



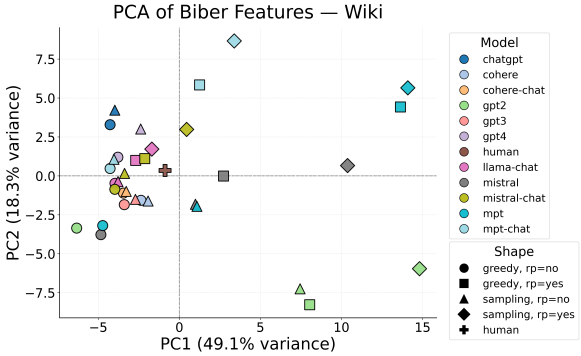
(c) Recipes



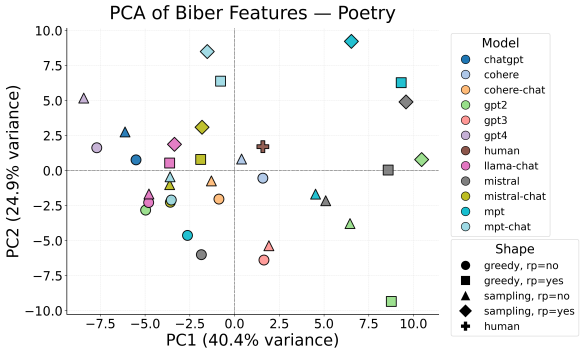
(d) Reddit



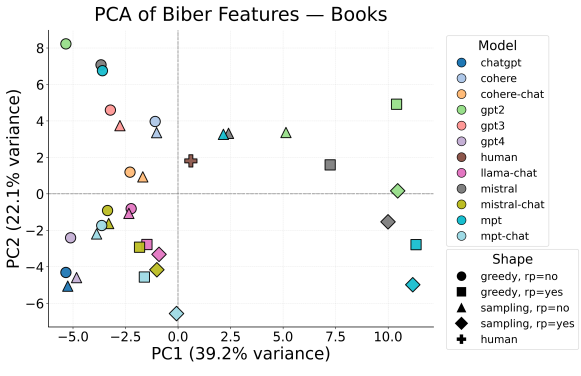
(e) News



(f) Wiki



(g) Poetry



(h) Books

Figure 16: PCA visualizations of Biber features across eight RAID categories. Colors indicate model identity, while marker shapes indicate decoding and repetition-penalty settings. Human text is shown with a distinct marker.