A Systematic Analysis of Linguistic Features in AI-Generated English Text Detection Across Domains and Models

Yassir El Attar[♦1,2], **Esra Dönmez**[♦1,2], **Maximilian Maurer**[★3,4], **Agnieszka Falenska**[♦1,2]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart
[3]GESIS - Leibniz Institute for the Social Sciences [4]Heinrich-Heine University Düsseldorf
[♦]{yassir.el-attar, esra.doenmez, agnieszka.falenska}@ims.uni-stuttgart.de
[★]maximilian.maurer@gesis.org

Abstract

Interpretable linguistic features offer a promising approach for explaining *why* a given text appears machine-generated, particularly for non-expert users. However, existing findings on which features reliably indicate LLM-generated text remain fragmented across feature sets, models, and text domains. To address this gap, we conduct a large-scale empirical study assessing the robustness of linguistic signals for characterizing AI-generated text. Our analysis covers 284 interpretable linguistic features across outputs from 30 LLMs and 10 text domains under cross-model and cross-domain generalization settings. We show that classifiers based solely on linguistic features can reliably distinguish AI-generated from human-written text. However, many previously proposed indicators prove strongly context-dependent, with the exception of measures of lexical richness, which remain robust signals across model families and text domains. These results demonstrate which linguistic signals generalize across contexts and provide a foundation for more reliable, interpretable analyses of AI-generated language.

1 Introduction

Large Language Models (LLMs) are becoming increasingly adept at producing fluent and stylistically *human-like* text. The societal implications of this development are significant: the ease of producing large-scale synthetic text amplifies the risk of misinformation while simultaneously challenging cues used to assess authorship and factuality (Srivastava, 2025). This creates a growing need for *interpretable explanations* of why a text appears machine-generated, particularly in settings where unreliable or biased detectors can have significant consequences (Jiang et al., 2024).

Most existing approaches explain why a text appears machine-generated by interpreting the decisions of black-box classifiers (Zellers et al., 2019; Li et al., 2024a; Guo et al., 2025). However, such explanations are often difficult to interpret for non-experts (Miller, 2019; Kaur et al., 2020) and typically provide only local, example-based insights into the properties of generated text (Koike et al., 2026). In response, a growing line of research focuses directly on *interpretable linguistic features*, identifying stylometric and linguistic differences between human and AI-generated writing, such as lexical diversity and syntactic regularity (Muñoz-Ortiz et al., 2024; Opara; Reinhart et al., 2025). While these studies provide promising evidence that linguistic features can characterize AI-generated text, the findings remain fragmented across feature sets, LLMs, and text domains. Consequently, it remains unclear which signals reflect general properties of generated language and which arise from particular datasets or model settings (Kehkashan et al., 2025).

In this work, we take a systematic approach to clarify the fragmented evidence on linguistic signals of AI-generated text. To this end, we first update MAGE (Li et al., 2024a), a benchmark for AI-authorship detection across multiple LLMs and text domains, with newer models released since its construction. We then use the updated benchmark to study AI-authorship detection across 30 LLMs and 10 text domains. Our analysis uses 284 interpretable linguistic features, combining and extending features previously studied in stylometry, readability, and AI-text detection. We focus on cross-model and cross-domain generalization, where detectors are tested on unseen models or text domains, and ask three main questions:

RQ1 How robust are linguistic features for distinguishing LLM-generated from human-authored text across models and domains?

RQ2 Which linguistic features contribute most to distinguishing AI-generated from human-written text, and how does their contribution vary across

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 181–208
November 6-10, 2026 ©2026 Association for Computational Linguistics

model families?

RQ3 How does the contribution of linguistic features vary across text domains?

First, we show that classifiers based solely on linguistic features achieve strong performance across models and domains (§4), demonstrating that interpretable linguistic signals can reliably distinguish human- and AI-generated text. Second, feature-area ablations show that *lexical richness* is the most informative feature area overall and the only one whose removal consistently hurts performance on unseen model families, while the contribution of other areas differs across families (§5.2). Third, feature-area importance also varies across text domains: lexical richness remains the most consistent signal, whereas the usefulness of other areas depends on the target domain (§5.3). Taken together, this work provides systematic evidence that linguistic characteristics of AI-generated text exhibit both robust and context-dependent patterns: while some signals, such as lexical richness, generalize across models and domains, others depend on specific generation settings. This highlights the importance of evaluating linguistic explanations under realistic cross-model and cross-domain conditions.[1]

2 Related Work

Much of the literature on AI-generated text focuses on the task of **automatic detection** (Gehrmann et al., 2019; Hao et al., 2025; Zhang et al., 2024; Guo et al., 2024; Ji et al., 2024, inter alia), often relying on black-box systems (Zellers et al., 2019; Guo et al., 2025). Efforts to improve interpretability typically rely on post-hoc explanation methods, such as feature attribution (Ribeiro et al., 2016; Lundberg and Lee, 2017) or example-based explanations (Koike et al., 2026). In parallel, several large-scale benchmarks and shared tasks have been proposed to evaluate detection performance across datasets and models (Koike et al., 2024; Li et al., 2024a,b; He et al., 2024; Wang et al., 2024, inter alia). However, the primary focus of this strand of research remains improving detection accuracy rather than interpretability.

A smaller but growing line of research instead examines **interpretable linguistic signals** that distinguish AI-generated from human-written language. For example, Muñoz-Ortiz et al. (2024) identify morphosyntactic and semantic indicators that differentiate human- and AI-authored texts in the news domain, while Dönmez et al. (2025) analyze LLM-generated counterarguments. Reinhart et al. (2025) employ Biber's 67-feature tagset to explore linguistic patterns in generated text, Opara propose a stylometric detection approach based on handcrafted indicators, and Doughman et al. (2025) use a small set of linguistic features (POS/NER, readability, and lexical features) for post-hoc analysis of LLM-based detection. These studies demonstrate that interpretable linguistic features can reveal systematic differences between human and AI-generated texts. However, most analyses focus on a limited number of models, domains, or feature sets. Consequently, it remains unclear **how broadly such linguistic signals generalize across different models and text domains**.

Recent work has also highlighted the importance of **robustness and generalization** in AI-text detection. Large-scale evaluations show that many detectors and linguistic signals degrade under distribution shift, such as unseen models, domains, or generation settings (Li et al., 2024a; Wang et al., 2024; Dugan et al., 2024). Moreover, some apparent linguistic cues may arise from confounding factors such as decoding strategies, prompts, or dataset artifacts rather than properties of machine-generated language (Li et al., 2024a; Reinhart et al., 2025). This raises a broader question: **to what extent reported linguistic cues reflect genuine properties of AI-generated language**.

3 Methods

In this work, we train classifiers based on interpretable linguistic features to distinguish between human-written and LLM-generated text. Using these classifiers, we analyze which linguistic signals contribute most strongly to the prediction and evaluate their robustness across models and text domains. In this section, we describe the data used to train the classifiers, the linguistic features we extract, and the methodological steps taken to analyze linguistic signals of LLM-generated text.

3.1 Data

We build on the MAGE benchmark (Li et al., 2024a) for detecting English LLM-generated text. MAGE contains human-written texts from 10 **text domains**, covering seven writing tasks: opinion statements (CMV, Yelp), news articles (XSum,

[1] Our code for the experiments and analyses is available at https://github.com/YassirELATTAR/ling-feats-mgt.

TLDR), question answering (ELI5), story generation (WritingPrompts, ROCStories), commonsense reasoning (HellaSwag), knowledge illustration (SQuAD), and scientific writing (SciGen). Corresponding AI-generated texts are produced by 27 models from seven **model domains** (families): OpenAI, LLaMA, GLM, FLAN-T5, OPT, BigScience, and EleutherAI. MAGE includes generations obtained with three **prompt types**: continuation, topical, and specified prompts. In addition, it provides a separate **out-of-distribution (OOD)** test set with human-written texts from four additional domains – CNN/DailyMail, DialogSum, IMDb, and PubMed –and corresponding texts generated by GPT-4.

For our study, we adapt MAGE to control for prompt variation and to include newer models. First, we restrict the benchmark to generations obtained with *continuation prompts*, in which the model receives the first 30 words of a human-written text and continues it.[2] Second, we extend the main benchmark with three models released after MAGE was constructed – Llama-3.1-8B, Llama-4-Scout-17B-16E, and GLM-5.2 – bringing the total from 27 to 30 models. The remaining model families are unchanged, as no newer models from these families were available at the time of our experiments. Finally, we extend the OOD portion of the dataset with GPT-5.6-sol generations. A complete list of models and dataset statistics is provided in Appendix A.1.

3.2 Linguistic Features

We extract *interpretable linguistic features* with the `elfen` Python package (Maurer, 2026) from various groups: surface-level, syntactic, and morphological structures as well as information-theoretic, lexical richness, semantic, and named entity features, and features based on measurements of emotional and psycholinguistic grounding of tokens. For surface-level, lexical richness, and readability features, we retain only one measure from each set of theoretically equivalent features. For instance, both the type-token ratio (TTR) and the lemma-token ratio capture lexical richness, despite differing in formulation. Appendix A.2 provides a full overview of the features used in our experiments.

Except for raw counts of tokens, sentences, characters, lemmas, and syllables, we normalize count-based features (e.g., the number of named entities in a text) by token count to obtain relative frequencies. We remove uniform features (i.e., features with only one value over the dataset) and handle missing and infinite values through feature-specific imputation: binary features are imputed with the mode, integer features with the median, and continuous features with either the mean (for normally distributed features) or median (for skewed distributions), ensuring that the statistical properties of each feature are preserved. The final feature set consists of 284 features: emotion (36), information (2), psycholinguistic (77), dependency (47), lexical richness (3), POS (18), named entities (19), readability (2), surface (6), morphological (58), and semantic (16).[3]

3.3 Experiments

We run two sets of experiments: (1) LLM-authorship detection (**RQ1**), which evaluates the effectiveness and robustness of classifiers based on interpretable linguistic features, and (2) feature area ablation (**RQ2** and **RQ3**), which analyzes the contribution of different linguistic feature groups.

Classifiers We use a linear Support Vector Machine (SVM) with balanced class weights to account for the class imbalance between human- and LLM-generated texts in MAGE (details on model choice and parameters in Appendix A.3).

We compare our classifier with three baseline detection methods evaluated by Li et al. (2024a) when introducing MAGE: FastText, GLTR, and Longformer. We re-train all three baselines on the same continuation-only extended benchmark used in our experiments to ensure a direct comparison. We omit DetectGPT, which was also evaluated in MAGE, because it requires access to token-level likelihoods of the generating model, which are unavailable for the commercial models in our extended data. Implementation details are provided in Appendix A.6.

Metrics and Evaluation We evaluate detection performance using three metrics: **Macro F1**, which accounts for class imbalance in test sets; **AUROC** (Area Under the Receiver Operating Characteristic curve) which measures discrimination between classes across decision thresholds; and **AvgRec** (Average Recall, details in Appendix A.4).

[2] Potential prompt effects are discussed via a cross-dataset experiment in Appendix B.2

[3] Feature extraction and preprocessing steps are computed at the document level and fitted on the training data only, ensuring strict separation between train and evaluation splits.

Name	Setting	Description
TB1	Text domain & model specific	One classifier per text domain–model pair; trained and tested on data from the same domain–model combination.
TB2	Arbitrary text domains & model domain-specific	One classifier per model domain; trained and tested on all text domains, restricted to AI texts from a single model family.
TB3	Fixed text domain & arbitrary models	One classifier per text domain; trained and tested on all model domains, restricted to texts from a single text domain.
TB4	Arbitrary text domains & arbitrary models	A single classifier trained and tested on all available data across all text and model domains.
TB5	Arbitrary text domains & unseen model domain	Leave-one-model-domain-out evaluation: trained on all data except one model domain and tested on the held-out model domain.
TB6	Unseen text domains & arbitrary models	Leave-one-text-domain-out evaluation: trained on all data except one text domain and tested on the held-out text domain.
TB7	Unseen text domains & unseen models	Trained on all available data and evaluated on new text domains (CNN, DialogSum, IMDb, PubMed) and AI text generated by GPT-4 and GPT-5.6-sol for the respective domains.
TB8	Unseen text domain & unseen model domain	Leave-one text–model pair out: trained on all data except one text–model domain pair and tested on the held-out pair; similar to TB7 but using existing domains to control for text–model pair effects.

Table 1: Evaluation testbeds used to analyze classifier robustness across text domains and model domains.

Testbeds We evaluate detection performance on eight testbeds from Li et al. (2024a) that vary in generalization scenarios (see summary in Table 1 and Appendix A.5). While some testbeds operate at fine-grained model granularity (e.g., TB1 evaluates domain–model pairs across all 30 models), we present most analyses at the level of model families to reduce noise from individual model variation and improve interpretability.

LLM-Authorship detection Classifiers are evaluated under **in-domain (ID)** and **out-of-domain (OOD)** conditions across three evaluation scenarios: general classification, model domain effects, and text domain effects. General classification uses TB4 and TB1 in ID settings, and TB8 and TB7 in OOD settings. Model domain effects are evaluated with TB2 (ID) and TB5 (OOD), while text domain effects are evaluated with TB3 (ID) and TB6 (OOD).

Feature area ablations We conduct two types of ablation studies to investigate feature area importance: **(A)** Leave-one-out, where classifiers are trained on all eleven feature areas but one, and **(B)** Cumulative, where feature areas are removed sequentially according to their importance in (A), and classifiers are retrained with progressively fewer feature areas. We run (A) on all testbeds and (B) only for the domain-agnostic single-classifier settings (TB4 and TB7) to avoid the combinatorial explosion of ablation classifiers.

4 LLM-Authorship Detection

To address **RQ1**, we first examine general detection performance on TB1, TB4, TB7, and TB8, before using TB2/TB5 and TB3/TB6 to isolate model- and text-domain effects, respectively.

Testbed	Method	AvgRec	AUROC	Macro F1
TB4	FastText	.743	.776	.741
	GLTR	.556	.729	.458
	Longformer	.909	.985	.907
	SVM w/ Ling. Feats.	.838	.969	.832
TB7	FastText	.605	.634	.605
	GLTR	.569	.624	.498
	Longformer	.717	.892	.703
	SVM w/ Ling. Feats.	.812	.883	.811
TB1	SVM w/ Ling. Feats.	.954	.989	.787
TB8	SVM w/ Ling. Feats.	.808	.951	.602

Table 2: Classification results: in-domain (TB1 and TB4) and out-of-domain (TB7 and TB8).

4.1 General Detection Performance

Table 2 presents the classification results. Differences between Macro F1, AUROC, and AvgRec reflect how these metrics are computed. AUROC evaluates score ranking across all thresholds, whereas Macro F1 and AvgRec are computed from predicted labels at the classifier's decision boundary. Under stronger distribution shifts, scores can shift relative to this boundary, so Macro F1 may drop substantially even when ranking performance and balanced recall remain comparatively stable (Saito and Rehmsmeier, 2015; Lipton et al., 2014).

Aggregate testbeds We first consider TB4 and TB7, which provide an overall view of in-domain and out-of-domain detection, respectively. Both testbeds use a single classifier over all relevant text and model domains, making them the most direct settings for comparison with the MAGE baselines.

In the mixed in-domain setting (TB4), where a single classifier is trained and tested on all text and model domains, our linguistic-feature classifier achieves 83.2% Macro F1 and 0.969 AUROC. It substantially outperforms FastText and GLTR

Model Domain	ID (TB2)	OOD (TB5)	Difference (TB5 - TB2)
OPT	.913	.715	-.199
Eleuther	.881	.527	-.354
LLaMA	.786	.692	-.094
FLAN-T5	.766	.670	-.096
BigScience	.656	.594	-.063
GLM	.627	.562	-.065
OpenAI	.555	.610	+.055
Mean	.741	.624	-.117

Table 3: Macro F1 results of ID (TB2), OOD (TB5), and their difference per model domain. TB2 trains and tests on the same model domain data with arbitrary domains, while TB5 excludes the target model domain from training.

Task	Text Domain	ID (TB3)	OOD (TB6)	Difference (TB6 - TB3)
Story Generation	WP	.964	.878	-.086
	ROCT	.931	.698	-.233
Commonsense Reasoning	HSwag	.957	.745	-.212
News Articles	XSum	.945	.910	-.036
	TLDR	.916	.713	-.203
Scientific Writing	SciGen	.913	.925	+.012
Question Answering	ELI5	.876	.819	-.057
Knowledge Illustration	SQuAD	.803	.658	-.145
Opinion Statements	CMV	.925	.875	-.049
	Yelp	.803	.783	-.021
Mean		.903	.800	-.103

Table 4: Macro F1 results of ID (TB3), OOD (TB6) and their difference. TB3 trains and tests on the same domain with arbitrary models, while TB6 excludes the target domain from training.

and approaches the performance of Longformer in terms of AUROC (0.985). In the fully out-of-domain setting (TB7), which contains unseen text domains and unseen models, our classifier achieves 81.1% Macro F1 and 0.883 AUROC. Here, it outperforms all three baselines in Macro F1, while its AUROC remains close to that of Longformer (0.892). These results show that classifiers based only on interpretable linguistic features can remain competitive with the neural baselines.

Fine-grained testbeds We next turn to TB1 and TB8, which provide a more fine-grained view of detection performance across specific text–model combinations. In TB1, classifiers are trained and tested separately for each text domain–model pair. The average performance decreases to 78.7% Macro F1, indicating substantial variation across individual combinations that is hidden by the aggregate TB4 result. TB8 introduces a stronger cross-generalization setting in which both the target text domain and model domain are excluded from training. Performance decreases further to 60.2% Macro F1, showing that some text–model combinations are considerably harder to generalize to than the aggregate OOD setting in TB7 suggests. We examine these model- and text-domain effects in more detail in the next section.

4.2 Model Domain Effects

We compare model-domain performance in TB2 and TB5 (Table 3). When trained and tested on the data from the same model domain (ID, TB2), the classifiers perform differently for each model domain. For OPT and Eleuther, the classifiers perform above the domain-agnostic ID setting (cf. TB4 in Table 2) by 8.1 and 4.9 percentage points, respectively. In contrast, performance for BigScience, GLM, and OpenAI is lower by 17.6, 20.5, and 27.7 points, respectively. This indicates that certain model domains exhibit distinct linguistic signatures that alone provide sufficient signal for the model domain-specific classifiers.

In the OOD setting (TB5), performance drops by 19.9 and 35.4 percentage points for OPT and Eleuther, respectively, suggesting that model-domain-specific linguistic patterns in the training data contribute substantially to their detectability. Similarly, for the remaining model domains, we observe a moderate to low performance drop (by 9.6 to 6.3 percentage points). Surprisingly, in the OOD setting, the performance increases for the OpenAI models, suggesting that the models display linguistic markers that generalize over LLMs, which can be useful in OOD settings provided sufficient variety in LLM-generated training data.

Finally, we examine whether the model-domain effects observed in TB5 persist when the text domain is also unseen (TB8). Figure 1 (left) shows the distribution of Macro F1 scores across the ten held-out text domains for each model domain. The overall pattern is similar to TB5: performance is highest for OPT and LLaMA and lowest for EleutherAI and GLM. At the same time, the spread within model domains shows that detectability also varies substantially with the accompanying text domain.

4.3 Text Domain Effects

We observe text-domain effects in TB3 and TB6 (Table 4). In the ID setup (TB3), variation across text domains is less pronounced than in the model domain experiments, suggesting that text domain effects are not as detrimental to classifier performance as model domain for in-domain classification (cf. Table 3, ID to Table 4, ID). For most of the domains, the ID performance is above 90% F1, and the lowest performance is still relatively high, at 80.3%.

The OOD experiments reveal several noteworthy patterns: performance in TB6 drops most sharply for ROCT (ROCStories), HellaSwag, and TLDR, by 23.3, 21.2, and 20.3 percentage points, respectively; this is consistent with their larger distributional differences (see Appendix E.1). For SQuAD, the drop is more moderate, at 14.5 percentage points, while SciGen is the only domain for which performance improves, by 1.2 points. The remaining domains show smaller performance decreases. Overall, these results indicate that access to text-domain-specific training data can substantially affect detection performance.

Finally, we examine text-domain effects when the model domain is also unseen (TB8; Figure 1, right). Performance varies substantially across text domains, with average Macro F1 ranging from 43.3% for SQuAD to 81.1% for SciGen. This shows that the difficulty of generalizing to an unseen text–model combination depends strongly on the text domain. A similar heterogeneity is visible in TB7, where performance varies considerably across the four novel text domains (see Table 10 in Appendix B).

Answer to RQ1: Linguistic features provide robust and generalizable signals for distinguishing LLM-generated from human-authored text across models and text domains. Feature-driven classifiers achieve comparable results to detection methods that rely on neural inference (FastText, GLTR, Longformer) for either extraction or classification, while relying on simple and interpretable representations. Our results indicate the presence of generalizable linguistic markers of LLM-generated text; however, differences in ID and OOD performance highlight substantial influences of model and text domain.

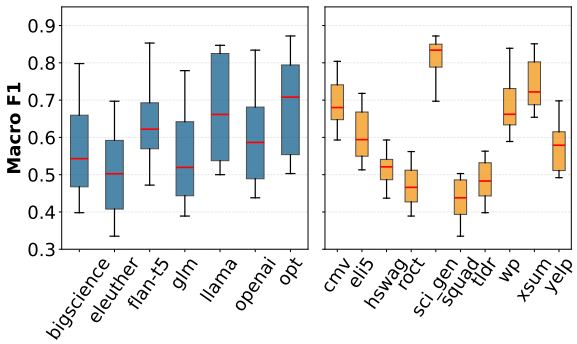


Figure 1: Distribution of Macro F1 scores across **TB8** (unseen text domain-model domain pairs), left pane for model domains and the right pane for text domains.

5 Features of AI-Authored Texts

We now zoom in on the effects of the types of linguistic features on the classification performance. We start by discussing the general classifiers and then focus on the model (**RQ2**) and text domains (**RQ3**).

5.1 Overall Feature-Area Importance

Table 5 displays the results from the leave-one-out feature-area ablation on the ID domain-agnostic classifier (TB4) and the OOD setting (TB7).

ID feature importance In the TB4 setting, we make two observations. First, dropping *lexical richness* features results in a stark performance drop of 14.3 percentage points, far larger than for any other feature area. The next two most influential areas are *information* and *POS*, with 1.5 and 1.1 drops, respectively. Second, the effects from the remaining eight feature areas are marginal, with the largest decrease of 0.3pp. To examine how performance changes as multiple feature areas are removed, we additionally conduct a cumulative ablation, removing feature areas sequentially according to their importance in the leave-one-out analysis. The resulting pattern is consistent with the leave-one-out results (left pane of Figure 2). Dropping the *readability* area initially improves performance, while removing *entities* leaves it essentially unchanged. As the remaining areas are removed one by one, the performance slowly degrades. However, even after removing ten areas, the classifier retains much of its performance, dropping only about 5 percentage points from the full-feature baseline (from 83.2% to 77.96%).

OOD feature importance In the OOD setting (TB7 in Table 5), the effect of removing *lexi-*

		Baseline	Surface	LexRich.	Emotion	Psycholing.	Read.	Morph.	POS	Depend.	Semantic	Entities	Inform.
F1 Macro	**TB4**	**.832**	.829	.689	.831	.829	.833	.831	.821	.830	.831	.833	.817
	TB7	**.811**	.820	.575	.816	.834	.810	.840	.799	.824	.811	.812	.828
# Feats. Removed		**0**	6	3	36	77	2	58	18	47	16	19	2

Table 5: Leave-one-out ablation results for TB4 (arbitrary text domains & arbitrary models) and TB7 (unseen text domains & model). Classifier performances when each feature area is dropped from the full feature set.

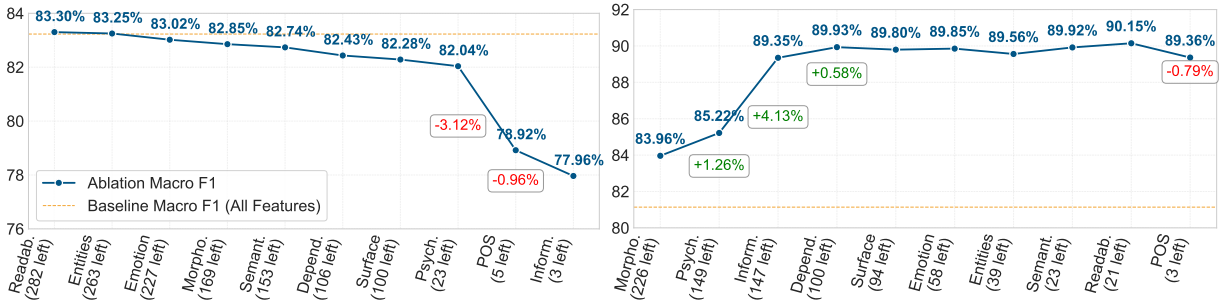


Figure 2: Macro F1 performance changes with cumulative feature area removal. The left pane shows results for TB4, and the right pane for TB7, in performance impact order from Table 5.

cal richness is similarly pronounced. Among the remaining feature areas, removing *POS* features decreases performance by 1.2 percentage points, while removing *readability* or *semantic* features has little effect. Removing several other feature areas improves performance, with the largest gain observed for *morphological* features (+2.9 points).

Unlike the ID setting, the cumulative ablation in TB7 shows that removing several feature areas can improve OOD performance (right pane of Figure 2). Removing the first four areas increases Macro F1 from 81.1% to 89.9% (+8.8 percentage points), after which performance remains relatively stable as further areas are removed. Notably, the classifier using only *lexical richness* features still achieves 89.36% F1, substantially above the full-feature baseline. This suggests that several feature areas capture signals that do not generalize well to the unseen domains, whereas lexical richness provides a particularly robust OOD signal.

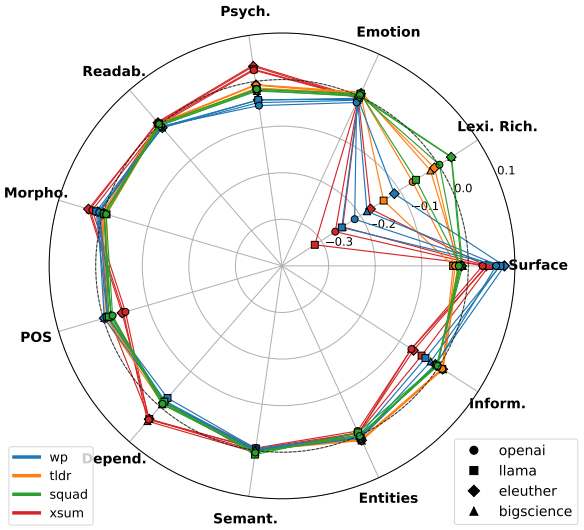


Figure 3: Ablation performance differences across 16 text domain-model domain after removing the corresponding feature area (selection based on Appendix C.3).

Joint text–model shift We next examine whether the observed patterns hold across individual unseen text–model pairs in TB8. A complete leave-one-out ablation over all 70 text-domain–model-domain pairs would require 770 classifiers. We therefore select four text domains and four model domains based on the performance analysis (see Appendix C.3 for details), yielding 16 representative pairs. Figure 3 shows two main patterns. First, removing *lexical richness* leads to the largest performance drops for most pairs, although its effect is much weaker for SQuAD. Second, the effects of the remaining feature areas are smaller (or even reversed) and vary across text–model pairs, with their removal sometimes improving and sometimes reducing performance. This indicates that, beyond lexical richness, the usefulness of linguistic features depends strongly on the particular text and model domain. Detailed pair-level results are provided in Appendix C.4.

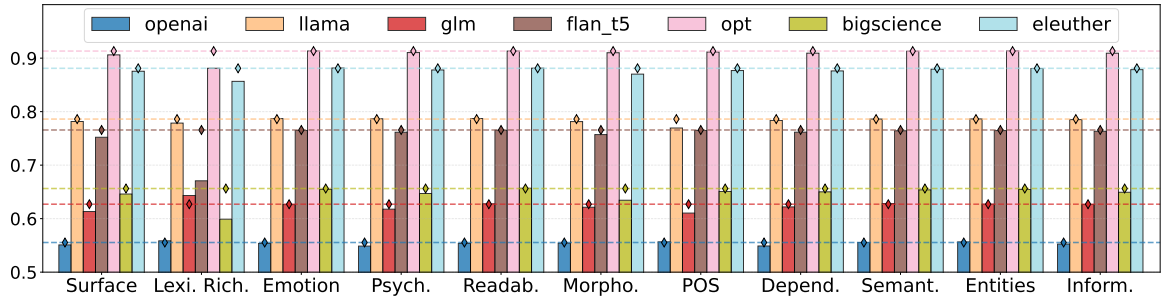


Figure 4: Ablation results on TB2. The horizontal dashed lines indicate the baseline results (all features) and the bars indicate the performance in Macro F1 after removing the corresponding feature area.

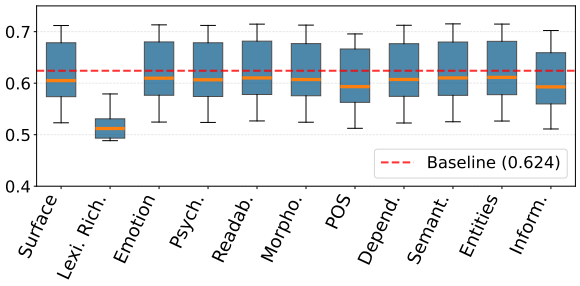


Figure 5: Ablation performance ranges on TB5 across the 11 feature areas after removing the corresponding feature area compared to the averaged baseline.

5.2 Feature-Area Importance Across Model Domains

Figure 4 shows the feature-area ablations for the ID model-domain setting (TB2). The effect of *lexical richness* varies substantially across model domains: removing it slightly improves Macro F1 for OpenAI (+0.3 points) and GLM (+1.6), but reduces performance for the remaining families, with the largest drops for FLAN-T5 (−9.5) and BigScience (−5.7). Removing most other feature areas has comparatively small effects. This suggests that the linguistic signals useful for in-domain detection differ across model families.

The pattern is more consistent in the OOD setting (TB5; Figure 5). Removing *lexical richness* reduces performance for every held-out model domain, by between 3 and 18 percentage points, with the largest drop for LLaMA. By comparison, removing *information* features decreases performance by only 1.2–2.4 points. This indicates that lexical richness provides a particularly robust signal for generalization across model domains (complete ablation results are reported in Table 16 in Appendix C). To provide qualitative intuition for these patterns, Appendix D presents representative human- and LLM-generated examples illustrating differences in the selected feature areas.

Answer to RQ2: *Lexical richness* has the largest overall impact on classifier performance, emerging as the strongest indicator of LLM-generated text. Feature ablations across model domains, however, reveal substantial variation between model families: the contribution of individual feature areas differs across families and between ID and OOD settings. In particular, while the importance of *lexical richness* varies across model domains in the ID setting, it provides a consistently useful signal when generalizing to unseen model domains. We further examine these model-family differences through an analysis of lexical-richness distributions in Appendix C.4.1.

5.3 Feature-Area Importance Across Text Domains

Figure 6 shows the feature-area ablation results for the ID and OOD text-domain settings. In the ID setting (TB3; upper pane), *lexical richness* has the largest and most consistent effect: removing it decreases Macro F1 across all ten text domains. The effect is negligible for SQuAD (-0.2 percentage points) and relatively small for HellaSwag (-2.6), but reaches 12.8 points for ELI5. For most other feature areas, the effects are smaller and vary across text domains, indicating that their usefulness is more domain-specific.

This pattern becomes stronger in the OOD setting (TB6; lower pane). Removing *lexical richness* decreases performance across all ten held-out text domains, with drops ranging from 2.0 percentage points for SQuAD to 27.8 points for XSum. The effects of the remaining feature areas are again mixed and depend on the target domain. For example, removing *surface* features improves performance for WritingPrompts and XSum, but decreases it for TLDR and ELI5. Overall, these results suggest that

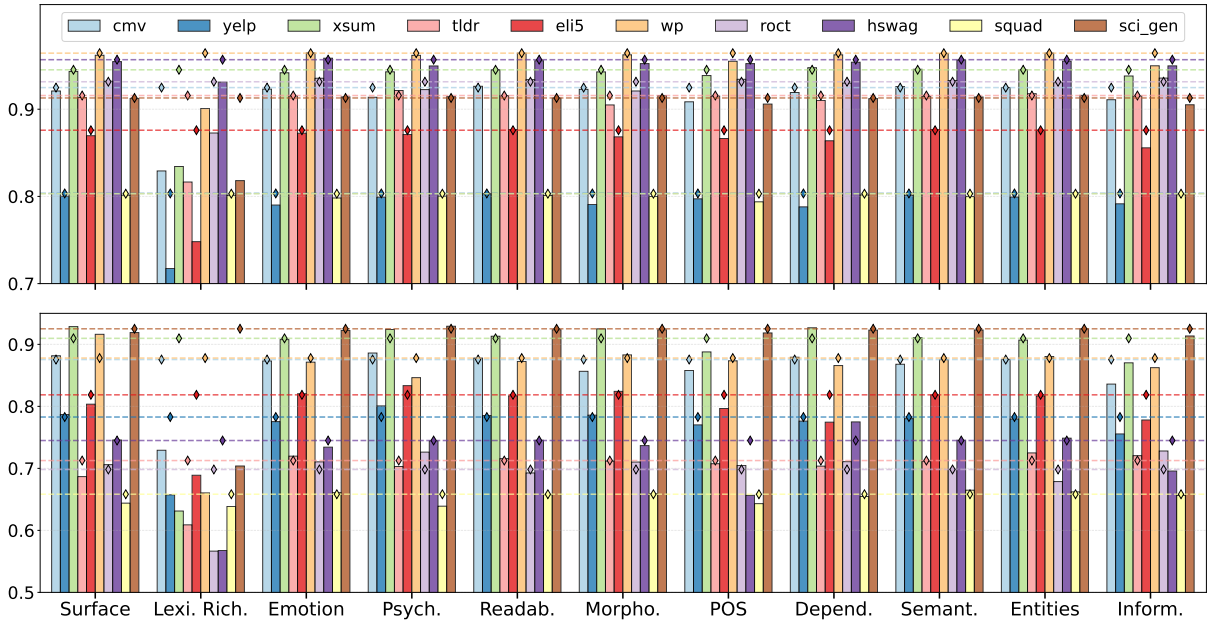


Figure 6: Macro F1 results of ablation study, top pane: TB3 and bottom pane: TB6. The horizontal dashed lines indicate baseline results for each model, and the bars the performance after dropping the corresponding feature area.

lexical richness provides the most consistent signal across text domains, whereas the contribution of other linguistic feature areas is more domain-dependent.

Answer to RQ3: Feature-area importance varies across text domains. *Lexical richness* is the most consistently influential feature area, while the importance of the remaining areas varies substantially across text domains (further analyses on this feature area in Appendix C.4). To further examine these differences, we compare feature distributions across text domains using Wasserstein distances (Appendix E.1). The resulting distributional differences provide additional context for the variation in feature informativeness and cross-domain detection performance.

6 Conclusions and Discussion

LLM-generated text increasingly populates both formal and informal discourse. Understanding the linguistic markers that signal such text is therefore essential, both to develop effective detection methods and to provide interpretable explanations to participants of such discourse.

Through extensive classification experiments, we showed that linguistic features (interpretable to both experts and non-experts) are highly informative of whether a text is generated by an LLM. At the same time, the linguistic markers of generated text are far from uniform. Many indicators previously proposed in the literature turn out to depend strongly on particular models or text domains. In contrast, measures of lexical richness consistently emerge as robust signals across contexts. Beyond individual features, systematic patterns also appear across groups of LLMs and types of text domains, suggesting that both model family and discourse context shape the characteristics of generated language.

These findings highlight both the promise and the limitations of linguistically grounded explanations of AI-generated text. While some signals generalize across contexts, others depend on specific models or domains. As a result, detection systems deployed in real-world settings, such as recognizing LLM-generated peer reviews or papers, must be evaluated under diverse conditions and updated as new models emerge. More broadly, our results emphasize the importance of systematically assessing the robustness of linguistic signals when using them to interpret or detect AI-generated language.

7 Limitations

Our setup and findings are limited in several ways. Firstly, while we find particular patterns in English data, this may not hold for other languages. In principle, our methods could be easily applied to any language that (a) has the necessary benchmarks

with a broad coverage available, and (b) external tooling to extract the features. The feature extraction tool we use covers a large number of languages, but has some blind spots, particularly regarding low-resource languages and/or languages from the African continent. Moreover, the applicability of the feature definitions, especially for lexical richness, depends on the particular language. Results in other languages may thus not necessarily be strictly comparable. Extending our analysis to typologically diverse and low-resource languages is therefore an important direction for future work.

Secondly, while we use, to the best of our knowledge, the broadest and most current available benchmark, it does not cover the newest models. Our results thus have limited informativeness on the behavior of these models. The stability of lexical richness in our results suggests it may capture a fundamental property of generated text that persists as models improve, but systematically evaluating this hypothesis on newer frontier models remains an important direction for future work, and underlines the need for continually updated benchmarks to assess whether findings on one set of models still hold.

Thirdly, our experiments cover two important dimensions that impact the detectability of machine-generated text: text domain and model (family). We do not consider prompt variants as a potential source of performance variation. Though outside of the scope of this paper and challenging due to a lack of availability of resources covering all three dimensions, we provide a small pilot experiment including prompt formulation as a factor of performance variation on one of the datasets (CMV) in Appendix B.2, where we find that model-domain alignment matters more than prompt formulation. A systematic study covering complex prompting strategies (personas, stylistic constraints, few-shot examples) is left for future work.

Finally, two further directions extend this work. First, performance variation across text domains – particularly for conversational text (DialogSum) where removing morphological and information-theoretic features improves detection – motivates future work on domain-specific feature selection rather than the development of entirely new feature sets. Second, the robustness of lexical richness against adversarial prompting (e.g., instructing models to use diverse vocabulary) is an open question, since classification relies on a combination of features, such interventions may also alter syntactic and semantic properties captured by other feature areas, which we leave for future investigation.

8 Ethical Considerations

While our work advances understanding of the features and quality of LLM-generated texts, it raises several important ethical considerations. First, insights from our findings raise potential dual-use concerns. On one hand, they could be misused by malicious actors to develop tools that better mimic human writing patterns. It is therefore essential not only to highlight these risks but also to support community efforts to identify, mitigate, and safeguard against such misuse. The extent to which specific linguistic features can be reliably controlled in generation remains an open question; existing work suggests that fine-grained control over individual linguistic properties is still challenging (Alhafni et al., 2024). On the other hand, understanding the linguistic signatures of AI-generated text has constructive applications. These insights can help guide generation toward more appropriate content, for example, by suggesting reformulations of user texts to match a desired stylistic profile. Additionally, classifiers based on fully interpretable linguistic features can serve as effective educational tools, helping the public recognize LLM-generated content. Finally, all human-authored texts used in our analyses were drawn from publicly available datasets and handled in accordance with established ethical research standards. No identifiable or private user data was used. Nonetheless, ongoing reflection on issues of consent, data provenance, and user agency remains vital when working with human discourse.

Acknowledgments

We would like to thank the reviewers for their helpful comments. We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. We thank the Institute for Natural Language Processing (IMS), University of Stuttgart, for providing the computational resources used in this work.

References

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Esra Dönmez and Agnieszka Falenska. 2025. "I understand your perspective": LLM persuasion through the lens of communicative action theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15312–15327, Vienna, Austria. Association for Computational Linguistics.

Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. 2025. AI argues differently: Distinct argumentative and linguistic patterns of LLMs in persuasive contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34595–34626, Suzhou, China. Association for Computational Linguistics.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Guangyu Shen, Kaiyuan Zhang, Shengwei An, Guanhong Tao, and Xiangyu Zhang. 2025. Profiler: Black-box AI-generated text origin detection via context-aware inference pattern analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24892–24912, Suzhou, China. Association for Computational Linguistics.

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. DeTeCtive: detecting AI-generated text via multi-level contrastive learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2025. Learning to rewrite: Generalized LLM-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6421–6434, Vienna, Austria. Association for Computational Linguistics.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, New York, NY, USA. Association for Computing Machinery.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting machine-generated texts: Not just "ai vs humans" and explainability is complicated. *Preprint*, arXiv:2406.18259.

Yang Jiang, Jiangang Hao, Michael Fauss, and Chen Li. 2024. Detecting chatgpt-generated essays in a large-scale writing assessment: Is there a bias against non-native english speakers? *Computers & Education*, 217:105070.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI 2020*. CHI 2020 Honorable Mention Award.

Tanzila Kehkashan, Raja Adil Riaz, Ahmad Sami Al-Shamayleh, Adnan Akhunzada, Noman Ali, Muhammad Hamza, and Faheem Akbar. 2025. Ai-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, 58:100793.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Institute for Simulation and Training.

Ryuto Koike, Masahiro Kaneko, Ayana Niwa, Preslav Nakov, and Naoaki Okazaki. 2026. ExaGPT: Example-based machine-generated text detection for human interpretability. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages

7702–7715, San Diego, California, United States. Association for Computational Linguistics.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(19), pages 21258–21266.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024a. 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.

Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024b. Spotting AI's touch: Identifying LLM-paraphrased spans in text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7088–7107, Bangkok, Thailand. Association for Computational Linguistics.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize f1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maximilian Maurer. 2026. elfen: A python package for efficient linguistic feature extraction for natural language datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–74, Rabat, Marocco. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad. 2018b. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Chidimma Opara. StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

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.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

Shashank Srivastava. 2025. Large language models threaten language's epistemic and communicative foundations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28662–28676, Suzhou, China. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ye Zhang, Qian Leng, Mengran Zhu, Rui Ding, Yue Wu, Jintong Song, and Yulu Gong. 2024. Enhancing text authenticity: A novel hybrid approach for ai-generated text detection. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 433–438.

A Methodology

A.1 Dataset Details

Additional dataset statistics for the MAGE data used in our experiments are reported in Table Table 6. The 30 models in the main benchmark

belong to seven model families: OpenAI (GPT-3.5-turbo, text-davinci-002/003; parameter counts undisclosed), LLaMA (7B, 13B, 30B, 65B, Llama-3.1-8B, and Llama-4-Scout-17B-16E, with 17B activated and 109B total parameters), GLM (GLM-130B and GLM-5.2, with approximately 753B total parameters), FLAN-T5 (small: 77M, base: 248M, large: 0.8B, XL: 3B, XXL: 11B), OPT (125M, 350M, 1.3B, 2.7B, 6.7B, 13B, and 30B, including 1.3B and 30B IML variants), BigScience (BLOOM-7B1: 7.1B, T0-3B: 3B, T0-11B: 11B), and EleutherAI (GPT-J-6B: 6B, GPT-NeoX-20B: 20B).

For the newly added models, we follow the MAGE generation procedure (Li et al., 2024a): each model receives the first 30 words of a human-written text and generates a continuation. We use the same 10 text domains and preserve the original train/validation/test split assignments. Three generations for ROCStories were identical to their human-written source texts and were discarded.

For the OOD portion of the dataset, we additionally generate texts with GPT-5.6-sol for the same four domains and human-written instances used in the original GPT-4 set: CNN/DailyMail, DialogSum, IMDb, and PubMed. This allows both models to be compared on the same human-written texts.

A.2 Full List of Linguistic Features

In the following, we provide an overview of the full set of features used in our experiments. We report features per feature area and rationales for the selection of features wherever we do not use all the available features.

Surface level We use the number of tokens, characters, sentences, lemmas, and words over 5 characters, and the raw sequence length in characters (including whitespaces and special characters). We do not consider all additional surface-level features, as we consider them theoretically equivalent to the ones we select.

We do not include the number of types (unique tokens), as it is expected to carry similar information as the number of tokens and the number of lemmas. We do not include sentence-normalized measures, as all information is captured by the combination of the number of sentences and the respective measure to be sentence-normalized. We do not include the average word length, as similar but more interpretable information is captured by the number of long words.

Parts of speech We use the number of tokens per universal dependencies POS tag (de Marneffe et al., 2021), and the POS variability (i.e., the number of different POS tags divided by the number of tokens).

Readability We use the total number of syllables in a text, and the Flesch reading ease (Kincaid et al., 1975) as measures of readability. We do not consider all other readability features, as we consider them theoretically equivalent to the Flesch reading ease or the number of syllables. More specifically, they are all designed to measure the reading level, either measured in school grade levels, age, or a more abstract index of how easy to read a given text is.

Lexical richness We use the type-token ratio (TTR), the number of global token hapax legomena (i.e., the number of tokens per text that only occur once across the respective train/evaluation split, computed independently for each split), and the lexical density (i.e., the percentage of adjectives, nouns, verbs, and adverbs per text). We do not consider all other lexical richness features, as we consider them theoretically equivalent to at least one of the selected ones. More specifically, they all measure how many different *kinds* of words relative to the overall number of words are present in a text. Our choice of TTR gives us a general measure, the global token hapax legomena provides information about rare words, and the lexical density about content words.

Information-theoretic We use both information-theoretic features, the compressibility and Shannon entropy of a text.

Named entities We use the number of named entities per entity type and the number of entities overall in a text.

Semantic We use the number of hedges in a text, the average number of synsets overall and per nouns, verbs, adjectives, and adverbs, and the number of tokens with a high and with a low number of synsets overall and per nouns, verbs, adjectives, and adverbs, respectively.

Morphological We use all available universal dependencies morphological features that are not uniform in our data.

Syntactic dependencies We use the number of universal dependency relations per relation type,

Split	CMV	Yelp	XSum	TLDR	ELI5
Train	4,223H / 22,788AI	31,827H / 22,929AI	4,708H / 28,451AI	2,826H / 22,211AI	16,706H / 27,948AI
Val	2,436H / 2,797AI	2,657H / 2,864AI	3,274H / 3,555AI	2,526H / 2,732AI	3,146H / 3,464AI
Test	2,403H / 2,814AI	2,652H / 2,847AI	3,283H / 3,554AI	2,535H / 2,742AI	3,156H / 3,495AI
Split	**WP**	**ROCT**	**HSwag**	**SQuAD**	**SciGen**
Train	6,356H / 27,203AI	3,287H / 27,910AI	3,129H / 26,882AI	15,820H / 22,340AI	4,436H / 21,091AI
Val	3,133H / 3,381AI	3,284H / 3,483AI	3,288H / 3,330AI	2,524H / 2,775AI	2,531H / 2,612AI
Test	3,099H / 3,437AI	3,275H / 3,487AI	3,292H / 3,355AI	2,508H / 2,796AI	2,538H / 2,551AI
Overall Total: Train (93,318H / 249,753AI), Val (28,799H / 30,993AI), Test (28,741H / 31,078AI)					

Table 6: MAGE dataset statistics using continuation prompts only. Training includes all available AI generations from 30 models across 7 families (model domains); validation and test sets are approximately balanced (H = Human, AI = AI-generated).

and the dependency tree width, tree depth, tree branching factor, the ramification factor, and the number of noun chunks.

Emotion and sentiment We use all emotion and sentiment features available in `elfen`. This includes the number of positive and negative sentiment tokens per text, the number of tokens with a high and a low valence, arousal, and dominance, the average valence, arousal, and dominance, the average intensity, and the number of high and low intensity tokens per Plutchik emotion.

Psycholinguistic We use all available psycholinguistic features. This includes the number of tokens with high and low concreteness, age of acquisition, prevalence, socialness, iconicity, and sensorimotor association, and the average rating across tokens per text for these dimensions, respecively.

A.3 Classifier

We chose a linear Support Vector Machine (SVM) for its effectiveness in high-dimensional feature spaces (284 features), computational efficiency for training hundreds of classifiers across testbeds, and interpretability through feature coefficients.

We employ a linear SVM using scikit-learn's `SVC` (v1.7.0) implementation[4] with the following configuration: linear kernel, regularization parameter ($C = 1.0$; default value), balanced class weighting to handle class imbalance, random seed 42. The training was accelerated using Intel Extension for Scikit-learn (`sklearnex`).

Features were standardized using z-score normalization (zero mean, unit variance) via `StandardScaler`, fitted on the training set and applied to validation and test sets.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

A.4 Evaluation Metrics

We compute Average Recall (AvgRec) as the arithmetic mean of recall for human-written and AI-generated texts:

$$\text{AvgRec} = \frac{\text{Recall}_{\text{human}} + \text{Recall}_{\text{AI}}}{2}.$$

Macro F1 is computed by averaging the F1 scores of the human and AI classes. We additionally report accuracy and per-class precision, recall, and F1 where relevant.

A.5 Testbeds

We build on the settings designed by Li et al. (2024a) (summarized in Table 7) and depart from their setup in the following ways.

TB1 (Fixed-domain & Model-specific) trained and tested classifiers on individual domain-model pairs, establishing baseline performance for 10 text domains and 30 AI models (300 classifiers).

TB1-1 extended this setting by adding a family-level analysis: One classifier per text domain-model domain pair (70 classifiers), trained on all generations of a single model family/domain.

TB2 (Arbitrary-text-domains & Model-domain-specific) trained on all text domains for each model family/domain separately.

TB3 (Fixed-text-domain & Arbitrary-model-domains) evaluated cross-model generalization within individual text domains.

TB4 (Arbitrary-text-domains & Arbitrary-model-domains) similar to Li et al. (2024a): combining all available domains and models to assess overall detection capability.

TB5 (Unseen model domains) evaluated a scenario of leave-one-out by leaving data generated by a specific model family (model domain) out of the training and then evaluated on the excluded model domain.

TB6 (Unseen text domains) similarly assessed text domain generalization by training on all text domains except one and testing on the excluded one.

TB7 (Unseen text-domains & Unseen-models) Li et al. (2024a) evaluate on four new text domains and a new model (GPT-4). However, we add GPT-5.6-sol as an unseen model generating data for the same domains and adding a substantially more recent model.

TB7-1 extended this analysis by evaluating each unseen text domain individually, for each model, to examine text domain-specific performance variations (8 classifiers).

TB8 (Unseen text domain–model domain pairs) We add a setting excluding both the target text domain and the target model domain from training, giving 70 held-out pairs (10×7). It isolates the joint shift of TB5 and TB6.

A.6 MAGE Baselines Replication

Li et al. (2024a) evaluate three baseline detectors that we use for comparison: FastText, GLTR, and Longformer. Longformer is trained with their released code and hyperparameters; for FastText and GLTR, no code is released, so we reconstruct the setup from the description in their paper and use library defaults wherever a setting is unspecified. We run each baseline twice: once on the original benchmark to verify our replication against the published numbers, and once on our continuation-only extended data, so that the comparison in Table 2 reflects differences in the data rather than in the implementation.

FastText Word bi-grams as the only text features, 100 epochs, all other settings at their defaults. On the original benchmark our replication is within 0.03 AvgRec of the published numbers.

GLTR GPT-2-XL (Radford et al., 2019) gives the number of tokens per text whose rank falls in the Top-10, Top-100, and Top-1000; these are passed to a logistic regression whose solver, penalty and regularization strength are grid-searched under 5-fold cross-validation.

Longformer (Beltagy et al., 2020) Fine-tuned from `longformer-base-4096` with MAGE's hyperparameters: sequence length 2048, effective batch size 16, learning rate 3×10^{-5}, 5 epochs, mixed precision, seed 42629309. The continuation-only run is distributed over four GPUs with the effective batch size unchanged.

MAGE refines the Longformer decision boundary for the out-of-distribution testbeds; we report the default decision rule unless stated otherwise.

A.7 Reproducibility Resources

All experiments were conducted using Python 3.10.18 with the following key dependencies: scikit-learn 1.7.0[5] for SVM implementation and evaluation metrics, Intel Extension for Scikit-learn (sklearnex) 2025.9.0[6] for accelerated training, pandas 2.3.0[7] for data manipulation, numpy 2.2.6[8] for numerical operations, matplotlib 3.10.3[9] and seaborn 0.13.2[10] for visualization, spaCy 3.8.7[11] with model en_core_web_lg 3.8.0, and ELFEN 1.1.9[12] for linguistic feature extraction. More details and required packages can be found in the requirements file (*requirements.txt*) on the repository. All experiments used consistent random seeds (seed=42) for train-test splits and model initialization.

A.8 Artifact Licensing

Our code and extensions to the MAGE benchmark are released under the Apache License 2.0, consistent with the license of the original MAGE benchmark. Third-party datasets, model resources, and linguistic resources used in this work remain subject to their respective original licenses and terms of use and are not relicensed by us. We use these artifacts solely for research purposes consistent with their stated access conditions.

B Classification Results

B.1 Detailed Testbeds Results

Summary results for all testbeds are reported in Table 8; the main findings are discussed in Section 4.

[5] https://scikit-learn.org
[6] https://uxlfoundation.github.io/scikit-learn-intelex/
[7] https://pandas.pydata.org
[8] https://numpy.org
[9] https://matplotlib.org
[10] https://seaborn.pydata.org
[11] https://spacy.io
[12] https://github.com/mmmaurer/elfen

TB#	Name	Training	Testing	# Classifiers
TB1	Fixed-domain & Model-specific	Single text domain + Single model	Same text domain + Same model	300
TB1-1	Fixed-text-domain & Model-domain-specific	Single text domain + Model domain	Same domain + Same model domain	70
TB2	Arbitrary-text-domains & Model-domain-specific	All text domains + Model domain	All text domains + Same model domain	7
TB3	Fixed-text-domain & Arbitrary-model-domains	Single text domain + All models	Same text domain + All models	10
TB4	Arbitrary-text-domains & Arbitrary-model-domains	All text domains + All models	All text domains + All models	1
TB5	Unseen model domains	All text domains + 6 model domains	All domains + 1 held-out model domain	7
TB6	Unseen text domains	9 text domains + All models	1 held-out text domain + All models	10
TB7	Unseen-text & -model domains	All text domains + All models	4 new text domains + GPT-4 & GPT-5.6-sol	1
TB7-1	Unseen-text-domain & Unseen-model	All text domains + All models	1 of 4 new text domains + 1 of 2 models (GPT4 or GPT-5.6-sol)	8
TB8	Unseen text-model-domain pairs	9 text domains + 6 model domains	1 held-out text domain + 1 held-out model domain	70

Table 7: Overview of experimental testbeds for AI-generated text detection. Each testbed evaluates different generalization scenarios, ranging from fixed conditions (TB1) to completely unseen domain-model combinations (TB7, TB8).

Table 9 reports Macro F1 for TB1.1, reports TB7 results on the four unseen text domains combined and on each domain separately, and Table 11 reports Macro F1 for all 70 held-out text domain–model domain pairs (TB8).

B.2 Cross-Dataset Experiment

To assess prompt generalization, we conducted cross-dataset experiments using the Change-MyView (CMV) dataset from Dönmez and Falenska (2025), which employs direct response prompts rather than continuation prompts. We evaluated: (1) training on MAGE-CMV and testing on their CMV dataset, and (2) training on their dataset and testing on MAGE-CMV. This setting tests whether models can generalize across different prompting strategies for the same domain. Statistics of the final CMV dataset are in Table 12. Table 13 reports in-distribution results on CMV.

Cross-dataset results are reported in Tables 14 and 15, respectively. In both scenarios, **model domain alignment between training and test data has a large impact on detection performance, while prompt formulation appears to have a smaller effect**.

C Ablation Results

In this section, we discuss further the ablation studies conducted on all testbeds (ablation study **A**) and on domain-agnostic settings (ablation study **B** -cumulative).

C.1 Model and Text Domain Effects Additional Analysis

Table 16 reports the complete ablation results for TB2, TB3, TB5, and TB6. Figure 7 visualizes the TB5 results, and Figure 8 shows the distribution of TB3 and TB6 results across feature-area ablations.

Model domain Effects (TB2 & TB5) The upper part of Table 16 reports results for TB2 (ID) and TB5 (OOD).

Text Domain Effects (TB3 & TB6) The lower part of Table 16 reports results for TB3 (ID) and TB6 (OOD), and Figure 8 shows their distributions.

C.2 Zoom-in TB7: Unseen Domains and Unseen Model

For the setting of completely unseen text and model domains, we tested two configurations: combining all four unseen domains (CNN, PubMed, IMDb,

Setting	Method	Acc	AUROC	F1-Macro
In-distribution Detection				
1. Fixed-Domain & Model (300 classifiers)	SVM w/ Ling. Feats.	0.9410	0.9893	0.7866
1.1. Fixed-Domain & Model-domain (70 classifiers)	SVM w/ Ling. Feats.	0.9142	0.9884	0.8445
2. Arbitrary-Domains & Fixed-Model (7 classifiers)	SVM w/ Ling. Feats.	0.8334	0.9853	0.7407
3. Fixed-Domain & Arbitrary-Models (10 classifiers)	SVM w/ Ling. Feats.	0.9055	0.9861	0.9034
4. Arbitrary-Domains & Arbitrary-Models (1 classifier)	SVM w/ Ling. Feats.	0.8348	0.9692	0.8323
Out-of-distribution Detection				
5. Unseen Models (7 classifiers)	SVM w/ Ling. Feats.	0.7294	0.9573	0.6241
6. Unseen Domains (10 classifiers)	SVM w/ Ling. Feats.	0.8091	0.9640	0.8004
7. Unseen Domains & Unseen Model (1 classifier)	SVM w/ Ling. Feats.	0.8114	0.8829	0.8114
7.1. Unseen Domains & Unseen Models (per domain) (8 classifiers)	SVM w/ Ling. Feats.	0.8083	0.9126	0.7892
8. Unseen Domain-Model Pairs (70 classifiers)	SVM w/ Ling. Feats.	0.6891	0.9508	0.6024

Table 8: Summary results for all testbeds (TB1-8). For testbeds with multiple classifiers, metrics represent the mean performance averaged across all individual classifiers within that testbed. Results are grouped into *in-distribution* (TB1-4) and *out-of-distribution* (TB5-8) detection scenarios.

Model domain	CMV	ELI5	HSwag	ROCT	SciGen	SQuAD	TLDR	WP	XSum	Yelp	Avg.
BigScience	0.844	0.750	0.784	0.828	0.867	0.701	0.759	0.949	0.892	0.752	**0.813**
Eleuther	0.935	0.934	0.870	0.959	0.945	0.956	0.871	0.971	0.951	0.851	**0.924**
FLAN-T5	0.926	0.900	0.820	0.835	0.951	0.855	0.881	0.986	0.967	0.793	**0.891**
GLM	0.792	0.659	0.913	0.970	0.698	0.540	0.755	0.874	0.793	0.561	**0.756**
LLaMA	0.916	0.833	0.956	0.986	0.799	0.710	0.904	0.947	0.870	0.760	**0.868**
OpenAI	0.742	0.632	0.919	0.733	0.742	0.563	0.656	0.839	0.725	0.608	**0.716**
OPT	0.960	0.911	0.951	0.964	0.967	0.936	0.945	0.978	0.966	0.858	**0.944**
Avg.	**0.879**	**0.801**	**0.875**	**0.875**	**0.852**	**0.785**	**0.806**	**0.939**	**0.867**	**0.737**	**0.842**

Table 9: TB1.1 results: F1-Macro performance for fixed domain and model domain combinations (70 classifiers total). Each cell represents a separate classifier trained and tested on data from one domain and one model domain. Row averages show model domain detectability; column averages show domain-specific performance.

Dataset	GPT-4		GPT-5.6-sol	
	AUROC	F1 Macro	AUROC	F1 Macro
CNN	0.9986	0.9793	1.0000	0.9925
DialogSum	0.8275	0.4882	0.7577	0.4431
IMDb	0.9853	0.9048	0.8741	0.8150
PubMed	0.9838	0.8845	0.8735	0.8059
ALL FOUR	0.9226	0.8278	0.8432	0.7796

Table 10: Results for TB7: unseen-domains and unseen-models, using the full set of features. The unseen models are GPT-4 and GPT-5.6-sol.

DialogSum) into a single test set, and testing each domain separately and per model.

Combined Unseen Domains Results for the four unseen text domains combined are reported in Table 5 and discussed in Section 5.

Separate Unseen Domains Figure 9 reports the ablation results for each unseen text domain separately, for GPT-4 (top) and GPT-5.6-sol (bottom).

C.3 TB8 Ablation: Selection of Domain-Model Pairs

Running ablation on all TB8 combinations (10 text domains × 7 model domains × 11 feature areas = 770 classifiers) is computationally expensive. We selected 16 representative pairs based on sensitivity analysis.

Model and Domain Selection We measure how much the feature-area ablation effects vary (standard deviation across the 11 ablations) for each text and model domain in the ID testbeds (TB2, TB3) and the OOD testbeds (TB5, TB6) (Figure 11), and compute the change from ID to OOD (Figure 10).

Selected 4 text domains: XSum and WritingPrompts (the largest increases in standard deviation from ID to OOD), and TLDR and SQuAD (almost no change).

Selected 4 model domains: LLaMA and OpenAI (the largest increases in standard deviation from ID to OOD), and EleutherAI and BigScience (almost no change).

This yields 16 pairs covering the full spec-

Model domain	CMV	ELI5	HSwag	ROCT	SciGen	SQuAD	TLDR	WP	XSum	Yelp	Avg.
BigScience	0.652	0.565	0.521	0.451	0.798	0.398	0.450	0.662	0.722	0.517	**0.574**
Eleuther	0.593	0.513	0.437	0.389	0.697	0.335	0.398	0.589	0.653	0.492	**0.510**
FLAN-T5	0.680	0.651	0.593	0.562	0.853	0.472	0.508	0.697	0.765	0.592	**0.637**
GLM	0.642	0.534	0.466	0.403	0.779	0.389	0.436	0.639	0.669	0.505	**0.546**
LLaMA	0.783	0.685	0.529	0.510	0.847	0.500	0.563	0.839	0.840	0.638	**0.673**
OpenAI	0.699	0.594	0.507	0.466	0.834	0.438	0.483	0.627	0.706	0.579	**0.593**
OPT	0.804	0.718	0.553	0.514	0.872	0.503	0.556	0.765	0.851	0.698	**0.683**
Avg.	**0.693**	**0.608**	**0.515**	**0.471**	**0.811**	**0.433**	**0.485**	**0.688**	**0.744**	**0.574**	**0.602**

Table 11: Results of TB8; Macro F1 performance on unseen domain-model-domain pairs. Training data excludes both the target domain and target model domain, creating 70 unique domain-model combinations. Each cell shows the F1-Macro score when testing on a domain-model pair that was completely excluded from training. The **Avg** row shows the average performance across all text domains for each model domain, while the **Avg** column shows the average performance across all model domains for each domain. Overall statistics: Mean = 0.602, Std = 0.138, Range = [0.335, 0.872].

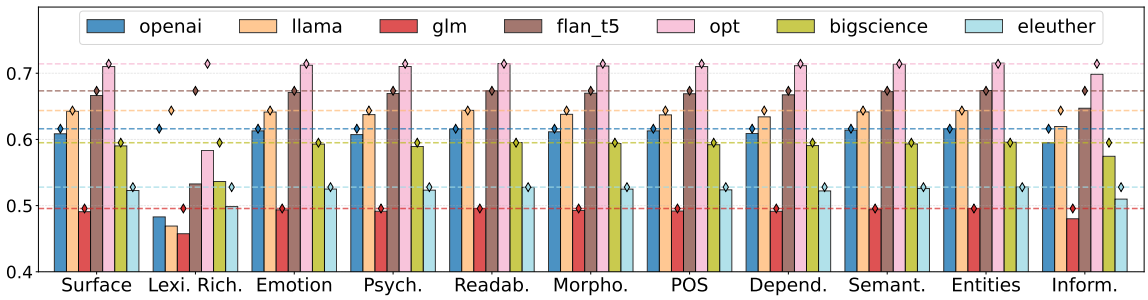


Figure 7: Visualization of results of ablation study on **TB5** (unseen models). The horizontal dashed lines indicate the original baseline results for each model, and the bars indicate the results of dropping the corresponding feature area.

Source	Number of Samples
Human	157,880 comments (13,498 posts)
GPT	13,498 (1 comment per post)
LLaMA	13,489 (1 comment per post)
Mistral	13,498 (1 comment per post)

Table 12: Final dataset sizes for the ChangeMyView (CMV) corpus used in training and evaluation.

AI Text	Accuracy	AUROC	F1 Macro
GPT	0.9911	0.9994	0.9911
Llama	0.9852	0.9948	0.9852
Mistral	0.9920	0.9986	0.9920
AI-combined	0.9884	0.9973	0.9884

Table 13: Results for **CMV dataset** for the different settings using the complete features set (284 features). *We also report accuracy since the dataset is balanced.*

trum from stable (SQuAD-EleutherAI) to volatile (XSum-LLaMA).

Results Figure 12 shows the distribution of ablation effects across the 16 pairs, Figure 13 reports all effects of at least 2 percentage points per pair, and Figure 14 summarizes the patterns per feature area.

Overall, we observe that the text domain largely determines which feature areas matter, while the model domain modulates the size, and in some cases the direction, of these effects.

C.4 Lexical Richness Results in Unseen Domain-Model Pairs (TB8)

Our ablation studies identified lexical richness as the most influential feature area. In the cumulative ablation on TB7, a classifier using *only* lexical richness features outperformed the full-feature baseline (Figure 2). This motivated us to test whether the same holds for individual unseen domain-model pairs in TB8.

We train classifiers using only the three lexical richness features for all 70 text domain–model domain pairs in TB1.1 (ID) and TB8 (OOD), and compare them to the corresponding full-feature

Mage Test data	Accuracy	AUROC	F1 Macro	AI-Recall	CMV-Training size	MAGE-CMV-test size
i- All models	0.7769	0.7862	0.7687	0.5973	28,339 H, 28,340 AI	2,403H, 2,339 AI
ii-OpenAI-Llama	0.9083	0.9385	0.8725	0.7947	18,890 H, 18,891 AI	2,403 H, 750 AI

Table 14: Results of two settings for cross-CMV-MAGE setup: (i) classifier is trained on training split from ChangeMyView dataset with all AI-texts (GPT, LLaMA, & Mistral) and tested on MAGE-CMV data with AI-texts from all 30 models. (ii) classifier trained on training split from ChangeMyView but with AI-texts from only GPT & LLaMA, then tested on MAGE-CMV with AI-texts from OpenAI & LLaMA models.

Mage Test data	Accuracy	AUROC	F1 Macro	AI-Recall	MAGE-CMV-Training size	CMV-test size
i- All models	0.4796	0.5529	0.4528	0.2582	4,223 H, 18,956 AI	8,097 H, 8,097 AI
ii-OpenAI-Llama	0.8496	0.9121	0.8492	0.9016	4,223 H, 6,029 AI	5,398 H, 5,397 AI

Table 15: Results of two settings for cross-CMV-MAGE setup: (i) classifier is trained on training split from MAGE-CMV data with AI-texts from all 30 models and tested on ChangeMyView dataset with all AI-texts (GPT, LLaMA, & Mistral). (ii) classifier trained on training split from MAGE-CMV with AI-texts from OpenAI & LLaMA models, then tested on ChangeMyView but with AI-texts from only GPT & LLaMA.

baselines. Figure 15 reports the resulting differences in Macro F1.

C.4.1 Feature Distributions Analysis

Given the substantial impact of lexical richness, we analyze the distributions of its three features: TTR (n_types/n_tokens), the number of global token hapax legomena (tokens of a text that occur only once in the respective data split; see Section A.2), and lexical density ($n_lexical_tokens/n_tokens$).

Figures 17 and 18 show the distributions of the three features by label (row a), text domain (row b), and model domain (row c).

D Qualitative Examples

To complement our quantitative analysis, we present qualitative examples that illustrate how the extracted linguistic features manifest in human-written versus AI-generated text. We select two feature areas–Emotion and Lexical Richness– based on their performance and clear distinction in both texts, and choose the domain-model pairs that best highlight the contrast identified in our experiments.

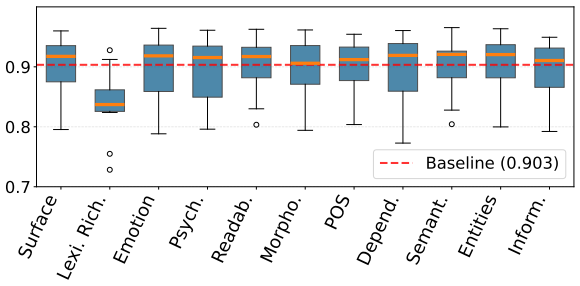
D.1 Emotion Features — Domain: Yelp, Model: OpenAI text-davinci-003

We select the Yelp domain paired with OpenAI's `text-davinci-003` for the emotion feature visualization (Figure 19), as Yelp reviews are inherently opinion-driven and exhibit a wide range of emotional expression among the text domains. For each sample, we display two emotion intensity dimensions—`anger` (red) and `disgust` (blue)—where darker shading indicates higher intensity scores from the NRC Affect Intensity Lexicon (Mohammad, 2018b). Valence is encoded background highlighting: green indicates high valence (> 0.66) and pink indicates low valence (< 0.33) based on the NRC VAD Lexicon (Mohammad, 2018a). Aggregate statistics (`avg_intensity`, `avg_valence`, `n_low/high_valence`) are shown above each sample.

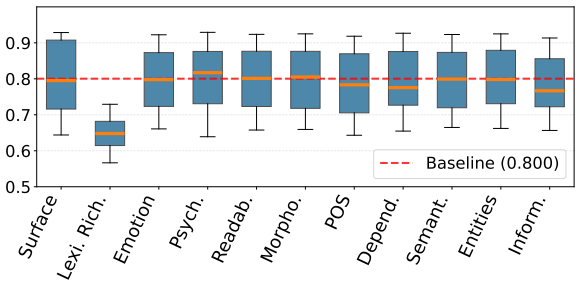
D.2 Lexical Richness — Domain: XSum, Models: LLaMA & OpenAI

We select XSum paired with LLaMA (-7B,-13B) and OpenAI (`text-davinci-003`) for the lexical richness visualization (Figure 20), as these pairs (combinations of domain and models) exhibited one of the largest performance drops in TB8 when lexical richness features were removed (cf. Figure 13), and the spider plot analysis (Figure 14) confirmed that lexical richness is the dominant feature area for these pairs. Furthermore, XSum (news summarization) produces structurally diverse texts where lexical patterns differ markedly between human and AI writing.

Each token is annotated along three dimensions: *italic* marks first occurrences (unique types contributing to TTR), **bold** marks lexical tokens (nouns, verbs, adjectives, and adverbs contributing to lexical density), and underline marks hapax legomena (tokens appearing only once in the text). Aggregate statistics (`ttr`, `lexical_density`, `n_hapax_legomena`, `n_tokens`, `n_types`) are displayed above each sample.



(a) Visualization of results distribution of ablation study on **TB3** (domain-fixed—arbitrary-model-domains) across the 11 feature areas.



(b) Visualization of results distribution of ablation study on **TB6** (unseen domains) across the 11 feature areas.

Figure 8: Distribution of F1-Macro scores across feature area ablations for in-distribution (TB3) and out-of-distribution (TB6) testbeds focusing on text domains.

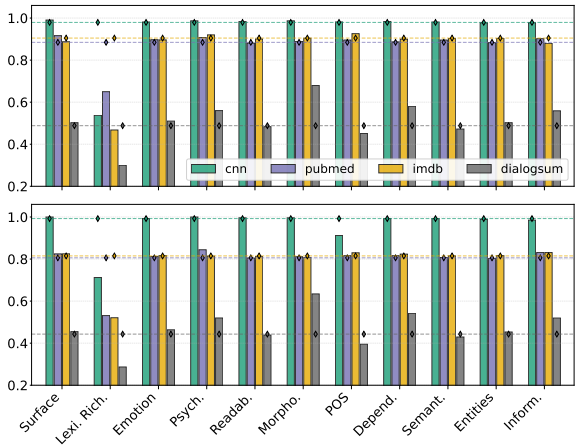


Figure 9: Visualization of results of ablation study on **TB7** using separate unseen domains (cnn, pubmed, imdb, & dialogsum) and two unseen models, shown in separate panels: GPT-4 (top) and GPT-5.6-sol (bottom). The panels are separated so that the effect of each ablated feature area can be compared across a model seen in the original MAGE data and a substantially more recent one. The horizontal dashed lines indicate the original baseline results for each domain, and the bars indicate the results of dropping the corresponding feature area.

E Distributional Similarity of Text and Model Domains

E.1 Wasserstein Distance of Text Domains

Figure 21 reports pairwise Wasserstein distances between human-written text domains for the ten most discriminative features. The domain pairs form two groups. Pairs involving ROCStories, HellaSwag, WritingPrompts, and TLDR show higher divergence, mainly in surface-level features (e.g., `raw_sequence_length`, `n_tokens`, `n_syllables`, `n_lemmas`); three of these domains show the largest performance drops in TB6 (Table 4). Pairs involving CMV, XSum, ELI5, SQuAD, and Yelp show lower divergence, in line with their smaller OOD drops.

`entropy` and `n_PRON_Number_Sing` follow a different pattern, with higher divergence for pairs involving Yelp and SQuAD. This suggests that, while surface features drive most between-domain differences, other feature areas capture complementary domain-specific signals.

E.2 Pearson's r Calculations of Model Domains

Building on the distributional analysis in Section C.4.1, we compute pairwise Pearson's r between the TTR distributions of each pair of model families within each text domain, yielding ten r values per pair. We group the pairs into three categories following the ablation results (Section 5): *Smaller–Smaller* (e.g., FLAN-T5 vs. OPT), *Smaller–LLaMA/OpenAI* (e.g., EleutherAI vs. LLaMA), and *LLaMA–OpenAI*.

Figure 22 reports the distributions of r for all 21 model family pairs. The correlations are close to zero for all pairs, with no clear differences between the three groups, indicating that the TTR patterns of model families do not co-vary across text domains. This is consistent with the model-dependent effects of lexical richness in the ablation studies (Section 5).

Figure 16 reports the per-domain values for selected pairs that reflect contrasts identified in the ablation analysis: GLM paired with smaller models (left) and OpenAI paired with models of different scale (right). Correlations remain close to zero across all text domains, without systematic domain-specific deviations.

Domain	Baseline	Surface	Lexi.Rich.	Emotion	Psych.	Readab.	Morpho.	POS	Depend.	Semant.	Entities	Inform.
Testbed 2												
OpenAI	.555	.551 (-.005)	.558 (+.003)	.554 (-.002)	.549 (-.007)	.554 (-.002)	.554 (-.001)	.557 (+.001)	.549 (-.006)	.556 (+.000)	.556 (+.001)	.552 (-.003)
LLaMA	.786	.782 (-.005)	.779 (-.008)	.787 (+.001)	.787 (+.001)	.787 (+.001)	.781 (-.005)	.769 (-.017)	.784 (-.003)	.786 (+.000)	.786 (+.000)	.785 (-.001)
GLM	.627	.613 (-.014)	.643 (+.016)	.626 (-.001)	.618 (-.009)	.628 (+.001)	.621 (-.006)	.610 (-.017)	.622 (-.005)	.628 (+.001)	.627 (-.000)	.626 (-.001)
FLAN-T5	.766	.752 (-.014)	.671 (-.095)	.765 (-.001)	.761 (-.004)	.765 (-.001)	.757 (-.009)	.765 (-.001)	.762 (-.004)	.764 (-.002)	.765 (-.001)	.763 (-.002)
OPT	.913	.906 (-.007)	.881 (-.032)	.913 (+.000)	.910 (-.003)	.913 (+.000)	.910 (-.003)	.911 (-.002)	.909 (-.004)	.913 (-.000)	.913 (+.000)	.909 (-.004)
BigScience	.656	.646 (-.010)	.599 (-.057)	.654 (-.002)	.647 (-.009)	.657 (+.001)	.634 (-.022)	.651 (-.005)	.650 (-.006)	.654 (-.002)	.655 (-.002)	.649 (-.007)
Eleuther	.881	.875 (-.006)	.857 (-.024)	.882 (+.001)	.878 (-.003)	.881 (+.000)	.870 (-.011)	.877 (-.004)	.876 (-.005)	.879 (-.002)	.881 (-.000)	.878 (-.003)
Testbed 5												
OpenAI	.610	.605 (-.005)	.489 (-.121)	.610 (-.000)	.607 (-.004)	.610 (+.000)	.607 (-.003)	.593 (-.017)	.607 (-.003)	.610 (+.000)	.611 (+.001)	.593 (-.017)
LLaMA	.692	.691 (-.001)	.512 (-.180)	.690 (-.002)	.689 (-.004)	.693 (+.001)	.688 (-.005)	.682 (-.010)	.687 (-.005)	.691 (-.001)	.692 (+.000)	.673 (-.019)
GLM	.562	.558 (-.004)	.488 (-.073)	.561 (-.001)	.558 (-.004)	.562 (+.000)	.559 (-.003)	.548 (-.014)	.558 (-.003)	.560 (-.002)	.562 (+.000)	.544 (-.017)
FLAN-T5	.670	.666 (-.004)	.525 (-.144)	.670 (-.000)	.668 (-.001)	.670 (+.000)	.666 (-.004)	.651 (-.019)	.666 (-.004)	.669 (-.001)	.670 (+.000)	.645 (-.024)
OPT	.715	.712 (-.003)	.579 (-.136)	.713 (-.001)	.712 (-.003)	.715 (+.000)	.713 (-.002)	.696 (-.019)	.712 (-.002)	.715 (+.000)	.715 (+.000)	.702 (-.012)
BigScience	.594	.590 (-.004)	.536 (-.057)	.592 (-.001)	.591 (-.003)	.594 (+.000)	.592 (-.002)	.577 (-.016)	.591 (-.003)	.593 (-.001)	.593 (-.000)	.575 (-.019)
Eleuther	.527	.523 (-.003)	.497 (-.029)	.524 (-.002)	.524 (-.003)	.527 (+.000)	.524 (-.002)	.512 (-.014)	.523 (-.004)	.525 (-.001)	.526 (-.000)	.511 (-.015)
Testbed 3												
CMV	.925	.921 (-.004)	.829 (-.096)	.923 (-.002)	.914 (-.011)	.926 (+.002)	.923 (-.002)	.908 (-.016)	.919 (-.006)	.926 (+.001)	.925 (+.000)	.911 (-.014)
Yelp	.803	.800 (-.003)	.717 (-.086)	.790 (-.013)	.799 (-.004)	.803 (-.000)	.791 (-.013)	.797 (-.006)	.788 (-.015)	.801 (-.002)	.799 (-.004)	.791 (-.012)
XSum	.945	.943 (-.002)	.834 (-.111)	.942 (-.003)	.943 (-.002)	.945 (+.000)	.943 (-.002)	.939 (-.007)	.948 (+.002)	.946 (+.001)	.945 (+.000)	.938 (-.007)
TLDR	.916	.913 (-.002)	.817 (-.099)	.915 (-.001)	.921 (+.006)	.915 (-.000)	.905 (-.011)	.915 (-.001)	.910 (-.006)	.915 (-.001)	.918 (+.002)	.915 (-.001)
ELI5	.876	.870 (-.006)	.748 (-.128)	.873 (-.003)	.871 (-.005)	.877 (+.001)	.868 (-.008)	.866 (-.010)	.864 (-.012)	.877 (+.001)	.876 (+.000)	.856 (-.020)
WP	.964	.962 (-.002)	.901 (-.063)	.965 (+.000)	.962 (-.003)	.964 (+.000)	.962 (-.002)	.955 (-.009)	.962 (-.002)	.964 (-.000)	.964 (+.000)	.950 (-.015)
ROCT	.931	.932 (+.001)	.873 (-.059)	.935 (+.004)	.922 (-.009)	.934 (+.002)	.921 (-.011)	.934 (+.003)	.930 (-.001)	.933 (+.001)	.930 (-.001)	.936 (+.004)
HSwag	.957	.955 (-.002)	.931 (-.026)	.959 (+.002)	.950 (-.007)	.956 (-.001)	.952 (-.004)	.952 (-.004)	.954 (-.003)	.956 (-.000)	.955 (-.002)	.950 (-.007)
SQuAD	.803	.801 (-.002)	.801 (-.002)	.798 (-.005)	.800 (-.003)	.801 (-.002)	.800 (-.003)	.794 (-.009)	.802 (-.001)	.800 (-.003)	.804 (+.001)	.802 (-.001)
SciGen	.913	.912 (-.001)	.818 (-.095)	.914 (+.001)	.914 (+.001)	.913 (-.000)	.916 (+.003)	.906 (-.007)	.912 (-.001)	.914 (+.001)	.916 (+.003)	.905 (-.008)
Testbed 6												
CMV	.875	.882 (+.006)	.729 (-.146)	.873 (-.002)	.886 (+.011)	.878 (+.002)	.857 (-.019)	.858 (-.018)	.879 (+.004)	.868 (-.008)	.876 (+.000)	.836 (-.040)
Yelp	.783	.787 (+.004)	.657 (-.126)	.775 (-.007)	.801 (+.018)	.785 (+.003)	.786 (+.003)	.770 (-.013)	.776 (-.007)	.780 (-.003)	.780 (-.003)	.755 (-.027)
XSum	.910	.929 (+.019)	.631 (-.278)	.908 (-.002)	.924 (+.014)	.913 (+.003)	.925 (+.015)	.888 (-.022)	.927 (+.017)	.911 (+.001)	.907 (-.003)	.870 (-.039)
TLDR	.713	.687 (-.026)	.609 (-.104)	.720 (+.007)	.703 (-.010)	.716 (+.003)	.712 (-.001)	.708 (-.005)	.704 (-.009)	.711 (-.002)	.725 (+.012)	.721 (+.008)
ELI5	.819	.803 (-.015)	.689 (-.130)	.820 (+.002)	.833 (+.015)	.817 (-.002)	.824 (+.006)	.797 (-.022)	.774 (-.044)	.818 (-.000)	.817 (-.002)	.778 (-.041)
WP	.878	.916 (+.038)	.660 (-.218)	.871 (-.007)	.846 (-.032)	.873 (-.005)	.883 (+.005)	.874 (-.004)	.866 (-.012)	.875 (-.003)	.880 (+.002)	.862 (-.015)
ROCT	.698	.706 (+.008)	.567 (-.132)	.711 (+.012)	.726 (+.028)	.693 (-.006)	.710 (+.012)	.705 (+.007)	.711 (+.013)	.699 (+.000)	.679 (-.020)	.728 (+.030)
HSwag	.745	.746 (+.001)	.568 (-.177)	.734 (-.011)	.744 (-.001)	.745 (+.000)	.737 (-.008)	.656 (-.088)	.775 (+.030)	.745 (+.000)	.749 (+.004)	.696 (-.049)
SQuAD	.658	.644 (-.014)	.639 (-.020)	.661 (+.003)	.639 (-.019)	.658 (-.001)	.659 (+.001)	.643 (-.015)	.655 (-.004)	.665 (+.006)	.662 (+.004)	.656 (-.002)
SciGen	.925	.919 (-.006)	.704 (-.221)	.922 (-.003)	.929 (+.004)	.924 (-.002)	.923 (-.002)	.918 (-.007)	.923 (-.002)	.923 (-.002)	.925 (-.000)	.913 (-.012)

Table 16: Complete results of the ablation study (A) where we drop a feature group at each training cycle for four different testbeds (**TB2**, **TB3**, **TB5**, & **TB6**) and the difference to baseline where the classifier is trained on the full set of features (284). The values in red indicate a decrease in performance, while green values indicate an improvement compared to the baseline result.

F Use Of AI Assistants

In this work, GitHub Copilot[13] (version as of November 2025) was used as a code completion/suggestion tool. Additionally, AI-assisted writing tools like Grammarly[14] (accessed November 2025 - January 2026) have been used for spelling checks and grammar corrections.

[13] https://github.com/features/copilot

[14] https://app.grammarly.com/

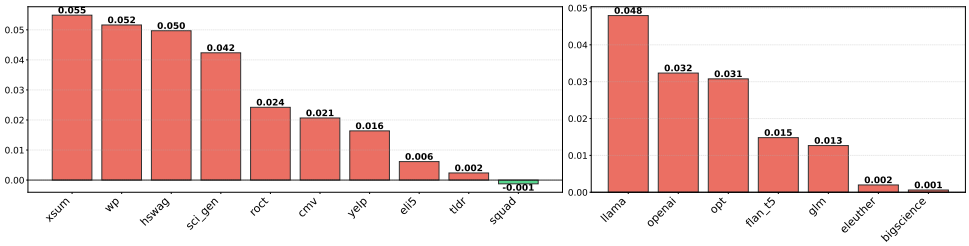


Figure 10: Change in standard deviation between unseen (**TB5** & **TB6**) and fixed (**TB2** & **TB3**) scenarios for text domains (left) and model domains (right). Red bars indicate increased sensitivity when unseen, green bars indicate decreased sensitivity. Higher absolute values represent larger shifts in model behavior across different experimental setups.

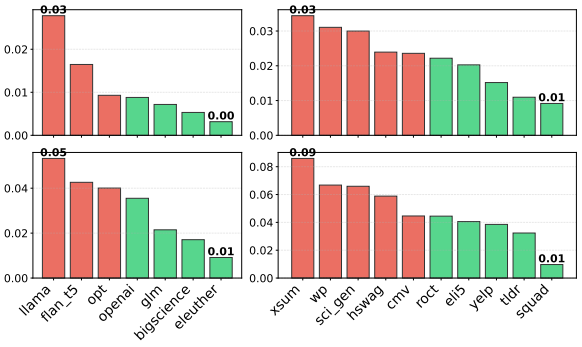


Figure 11: Results of the standard deviation between ablation study across the different testbeds and the ablation study using feature areas. Red bars indicate increased sensitivity compared to the mean, green bars indicate decreased sensitivity compared to the mean (horizontal line).

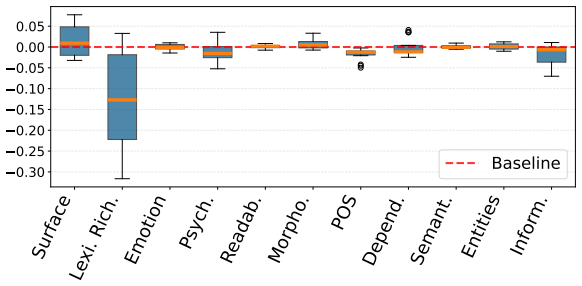


Figure 12: Leave-one-out features-area ablation performance variance across 16 text domain-model domain

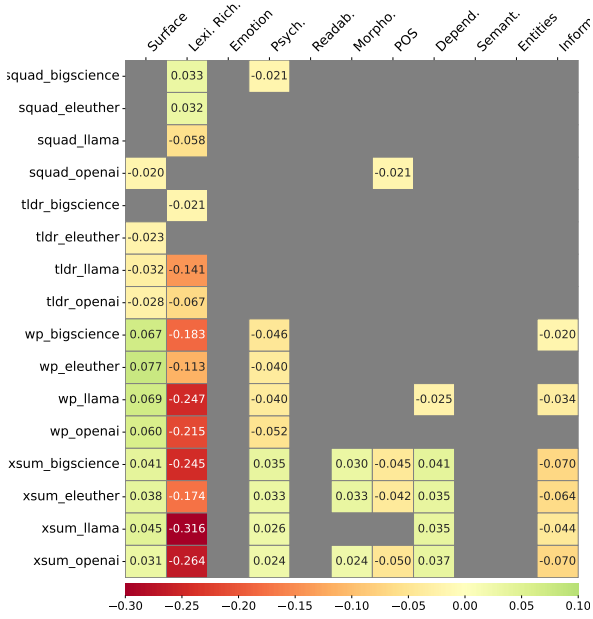


Figure 13: The difference in the performance for the 16 selected text domain-model-domain pairs from the previous study. We keep only the deltas ($\Delta = ablation_results - baseline$) of $\pm 2\%$ effect.

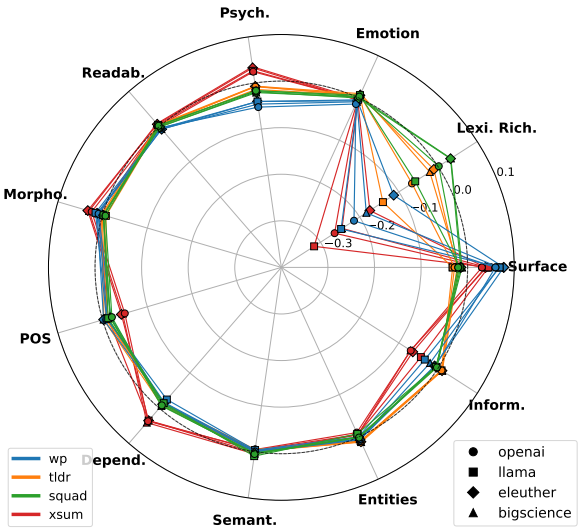
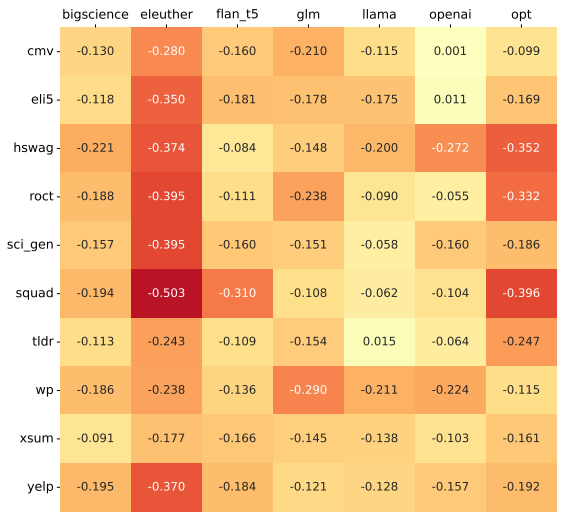
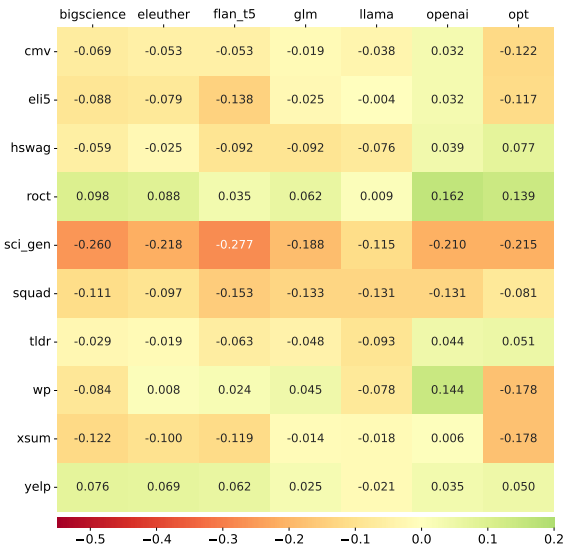


Figure 14: The difference in the performance for the 16 selected text-domain-model-domain pairs from the previous study using the 16 text-domain-model-domain pairs. The text domains separated by colors and and the model domains by markers/ shapes.



(a) Results of **TB1.1** domain-model pairs. Positive values (green) indicate lexical richness alone outperforms all 284 features; negative values (red) indicate the full feature set is superior.



(b) Results of **TB8** domain-model pairs. Positive values (green) indicate lexical richness alone outperforms all 284 features; negative values (red) indicate the full feature set is superior.

Figure 15: Performance difference (F1-Macro) between lexical-richness-only and full-feature classifiers.

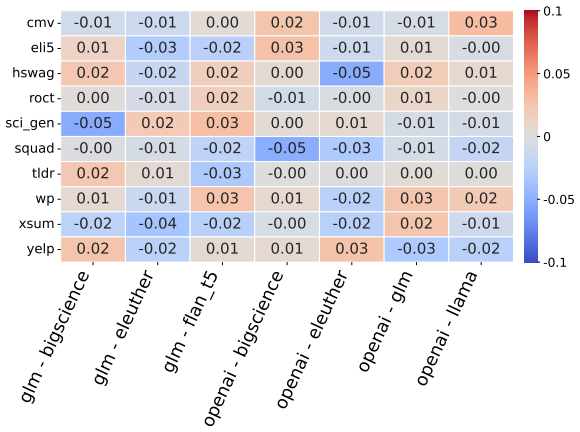


Figure 16: Heatmap of pairwise Pearson's r of TTR distributions between selected model family pairs across text domains. The left columns show GLM paired with smaller models, while the right columns show OpenAI paired with models of varying scale. Cell values indicate the Pearson's r between the TTR distributions of the two model families within each text domain.

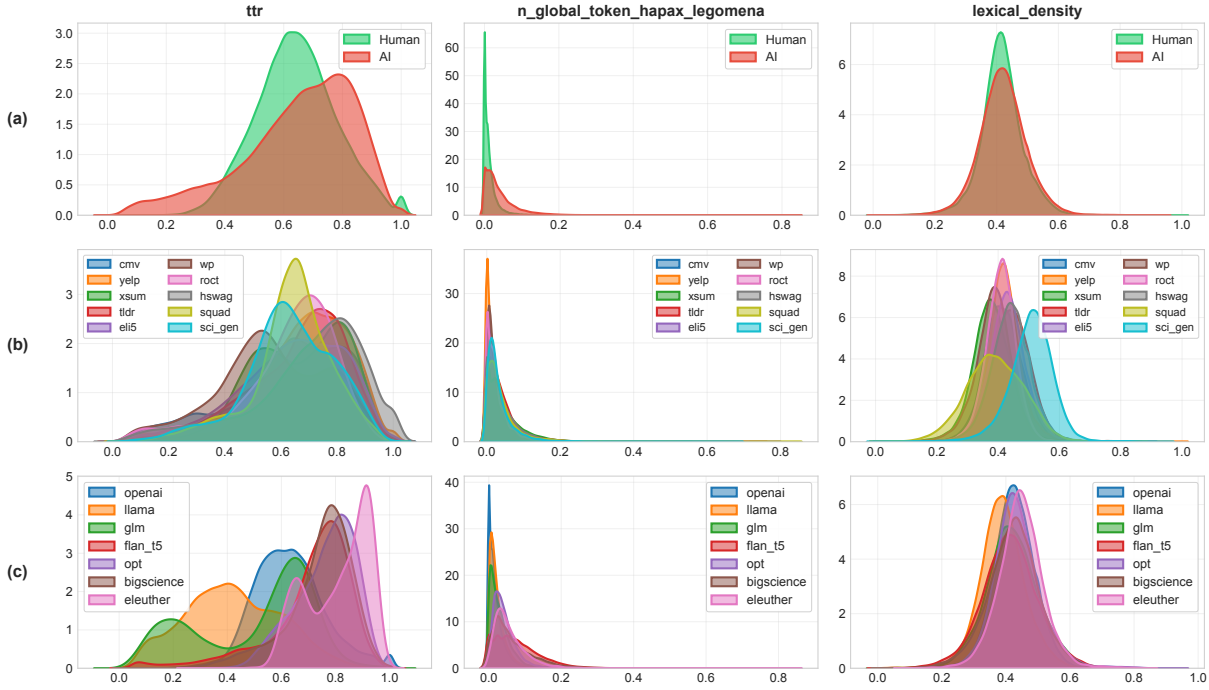


Figure 17: Density distributions of lexical richness features (TTR, hapax legomena, lexical density) across (a) Human vs AI labels, (b) text domains, and (c) AI model domains. Overlapping histograms show the discriminative power of these features across different groupings.

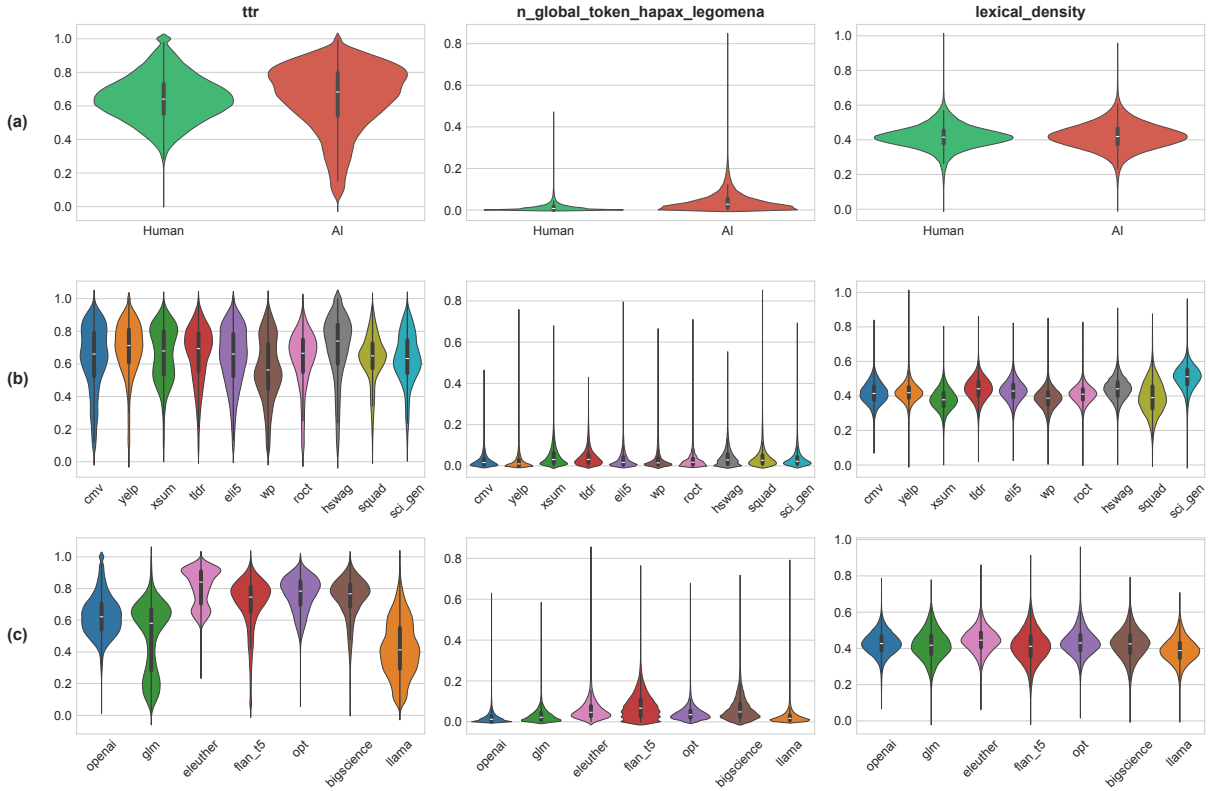


Figure 18: Violin plots for the three lexical richness features: TTR (left), hapax legomena (middle), and lexical density (right). Each row displays distributions across Human vs AI, text domains, and model domains, revealing both central tendency and distributional shape.

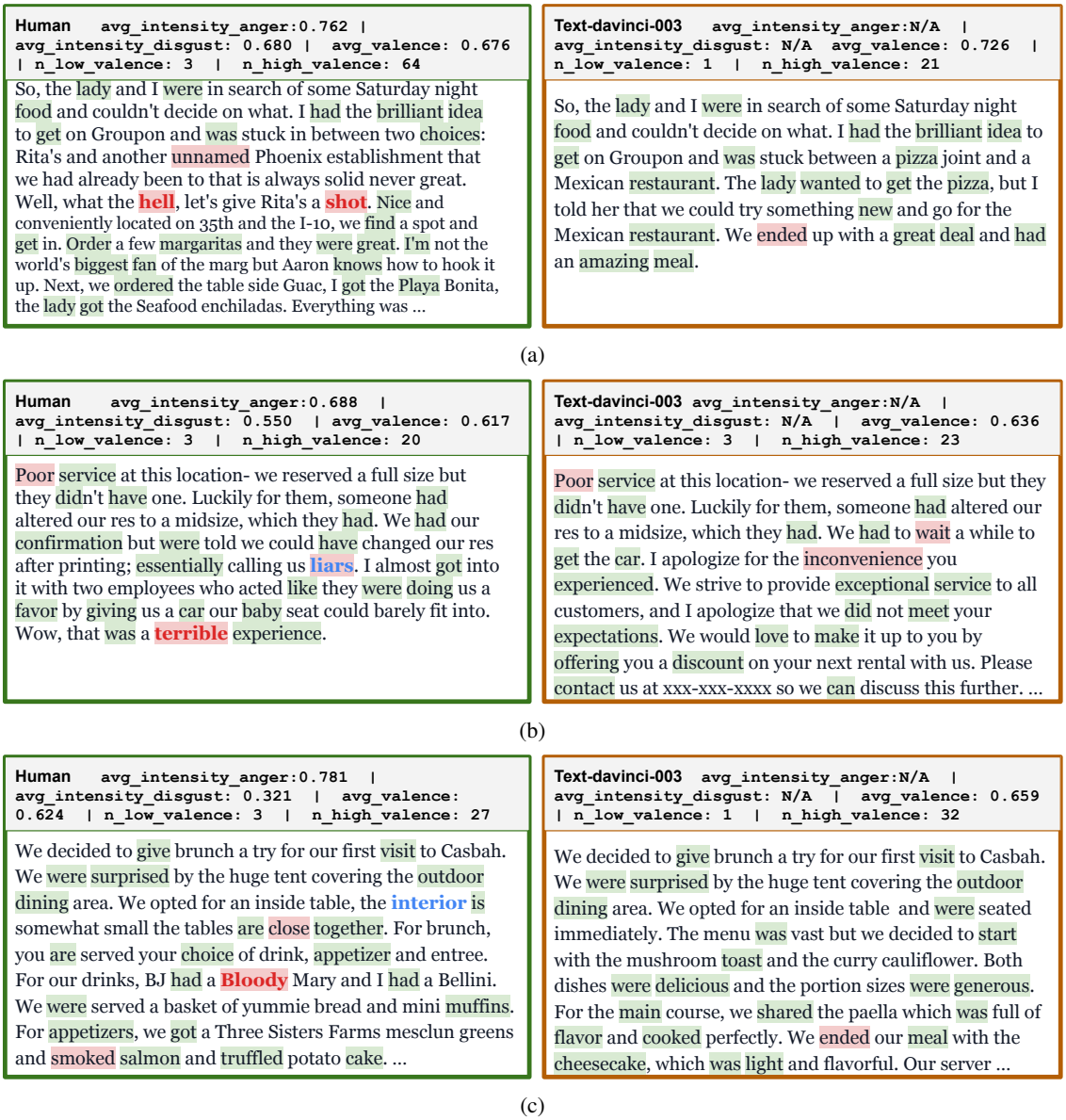


Figure 19: Parallel human–AI text pairs from the Yelp domain (AI model: Text-davinci-003) illustrating differences in a few selected features from emotion area. Visual markers: **red bold text** = anger intensity, **blue bold text** = disgust intensity, green highlight = high valence (> 0.66) and pink highlight = low valence (< 0.33). feature values (average anger and disgust intensity, average valence, and high/low valence word counts) are shown above each text. Emotion intensity is derived from the NRC Affect Intensity Lexicon and valence from the NRC VAD lexicon, following the `elfen` toolkit's extraction procedure. Some texts are truncated for space; ellipses (...) indicate omitted content.

Human ttr: 0.271 | lexical_density: 0.298 | n_hapax: 114 | n_tokens: 822 | n_types: 223

Media playback is not supported on this device After an even first 30 minutes, Danny Wright tapped in from James Dayton's free-kick to put the Robins in front. But Tubbs levelled after the restart, latching onto Maxime Biamou's header and slotting in from 10 yards out. Robert Dickie headed against the bar for the Robins **late on**, *before Deacon fired in to win the game in added time. Sutton have won just one of their past eight games in the National League - including a 6-0 home defeat by Dover on Tuesday - but* Deacon's ...

Llama-7B ttr: 0.694 | lexical_density: 0.426 | n_hapax: 62 | n_tokens: 108 | n_types: 75

Media playback is not supported on this device Our mission is to inspire, empower and motivate our online community in the pursuit of feeling BodyPositive. And we'll be with you every step of the way, because who says diet is a dirty word. Feel free to browse our extensive library of articles on a variety of topics. This is the first time that the FDA has approved a drug to treat obesity, and it is a big deal. In the 1990s, pharmaceutical companies tried to develop drugs to **treat obesity**, *but the* **drugs** *had...*

(a) Human (Left) vs. LLaMA-7B (right).

Human ttr: 0.276 | lexical_density: 0.314 | n_hapax: 137 | n_tokens: 978 | n_types: 270

Andreu was denied from the spot by a brilliant save from *Raith keeper Kevin Cuthbert 15 minutes* from *time. United were reduced to ten men when former Rovers defender Lewis Toshney* was *sent off in 55* **minutes** *after collecting a second yellow card. Rudi Skacel struck the bar for* Raith *but neither side could find the net. Dundee* United *manager Ray McKinnon is desperate* to *add more of a goal threat* to *his team and Dutch trialist striker Felitciano Zschusschen watched* from the *stands. The Tangerines struggled ...*

Llama-13B ttr: 0.675 | lexical_density: 0.400 | n_hapax: 82 | n_tokens: 160 | n_types: 108

Andreu was denied from the spot by a **brilliant save** from *Raith keeper Kevin Cuthbert 15 minutes* from *time. United were reduced to ten men when former Rovers defender Lewis Toshney* was *dismissed for a second bookable offence. Manager Jackie McNamara admitted he* was *delighted* to *see his side's recent revival continue with another impressive performance at a venue where* United *have a mixed record in* **recent years**. *"I think the pitch is always a little bit dodgy here but the guys are gelling, that's...*

(b) Human (left) vs. LLaMA-13B (right).

Human ttr: 0.292 | lexical_density: 0.330 | n_hapax: 211 | n_tokens: 1191 | n_types: 348

The Latics are seven points from safety with six games remaining including tough trips to Brighton and Reading. David McGoldrick put Ipswich ahead midway through the first half, slotting in Myles Kenlock's pass, before Freddie Sears raced clear to *double* the *lead. Wigan improved after* the *break, Bartosz Bialkowski saving brilliantly* from *Ryan Colclough* and *Alex Gilbey, but* Sears *pounced late on* to *make it 3-0.* The *win relieved some of* the *pressure* on *Town boss Mick McCarthy, following* seven...

Text-davinci-003 ttr: 0.558 | lexical_density: 0.289 | n_hapax: 68 | n_tokens: 190 | n_types: 106

The Latics are seven points from safety with six games remaining including tough trips to Brighton and Reading. David McGoldrick put Ipswich ahead midway through the first half, slotting in from *close range after good work* from David *Healy*. The Latics *had a great chance* to *equalise on* the *stroke of* **half-time**, *but Grant Holt fired over* the *bar* from *12 yards out*. Ipswich *doubled their lead* in the *second* **half** *when Luke Chambers nodded* in a *corner* from *Jay Emmanuel-Thomas* to *make it 2-0. Substitute Jon ...*

(c) Human (left) vs. Text-davinci-OO3 (right).

Figure 20: Parallel human-AI text pairs from the XSum domain illustrating differences in lexical richness features. Each pair shares the same opening words, with the AI text generated as a continuation. Visual markers: *italic*= first occurrence of a token (unique type, contributing to TTR), **bold** = lexical token (noun, verb, adjective, or adverb, contributing to lexical density, and underline = hapax legomenon (token appearing only once in the text). Feature values (TTR, lexical density, hapax count, token count, and type count) are shown above each text. Texts are truncated for space; ellipses (. . .) indicate omitted content.

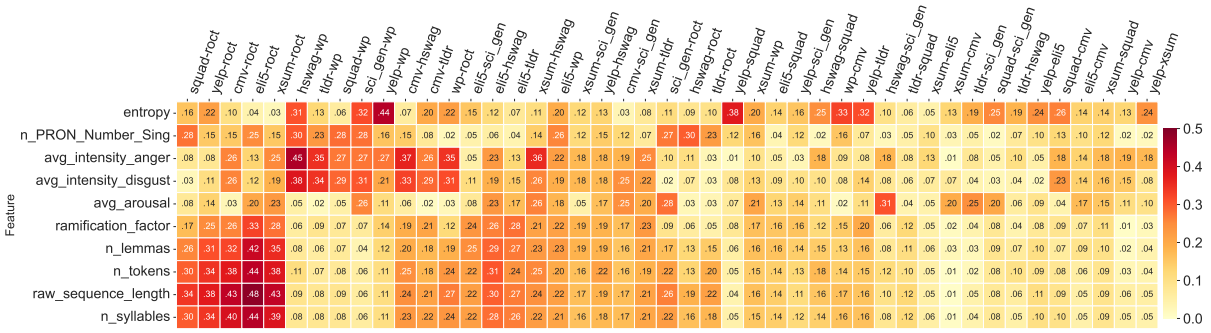


Figure 21: Clustered heatmap of pairwise Wasserstein distances for the top 10 most discriminative linguistic features across all human text domain pairs. Rows represent features that most consistently appear in the top 10 most discriminative features across domain pairs, ranked by their frequency of occurrence. Columns represent each pair of domains. Higher values (darker cells) indicate greater distributional divergence between the two domains for that feature, suggesting that the feature captures domain-specific linguistic patterns. Lower values (lighter cells) indicate similar distributions across domains, suggesting domain-invariant characteristics

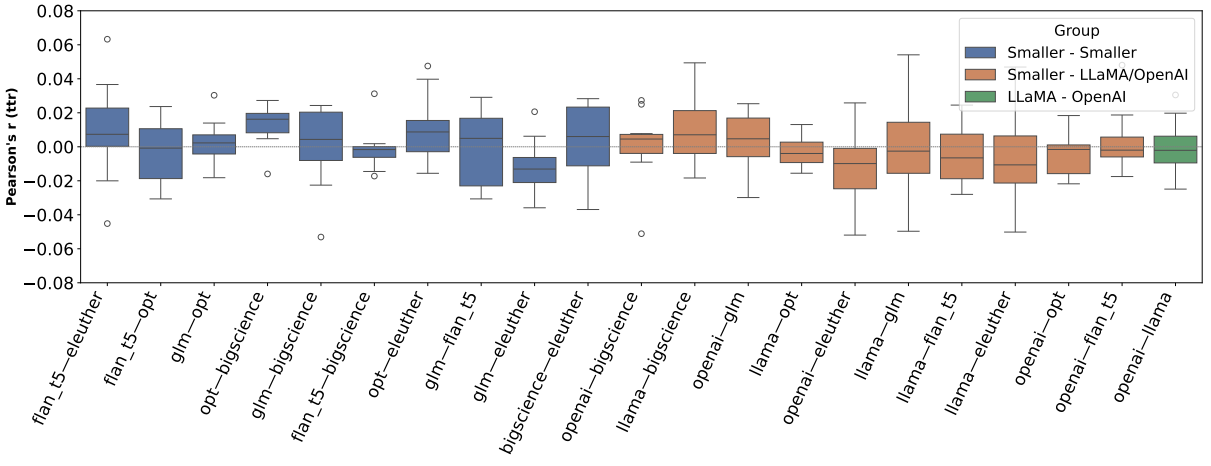


Figure 22: Pairwise Pearson's r of TTR feature distributions across model family pairs, with distributions computed over the ten text domains. Pairs are grouped into three categories: *Smaller–Smaller* (blue; e.g., FLAN-T5 vs. OPT), *Smaller–LLaMA/OpenAI* (orange; e.g., Eleuther vs. LLaMA), and *LLaMA–OpenAI* (green). TTR was selected as the focal feature given its pronounced human–AI distributional separation observed in Section C.4.1.