Explaining Generalization of AI-Generated Text Detectors Through Linguistic Analysis

Anonymous ACL submission

Abstract

AI-text detectors achieve high accuracy on in-domain benchmarks, but often struggle to generalize across different generation conditions such as unseen prompts, model families, or domains. While prior work has reported these generalization gaps, there are limited insights about the underlying causes. In this work, we present a systematic study aimed at explaining generalization behavior through linguistic analysis. We construct a comprehensive benchmark that spans 6 prompting strategies, 7 large language models (LLMs), and 4 domain datasets, resulting in a diverse set of human- and AI-generated texts. Using this dataset, we fine-tune classification-based detectors on various generation settings and evaluate their cross-prompt, cross-model, and cross-dataset generalization. To explain the performance variance, we compute correlations between generalization accuracies and feature shifts of 80 linguistic features between training and test conditions. Our analysis reveals that generalization performance for specific detectors and evaluation conditions is significantly associated with linguistic features such as tense usage and pronoun frequency. [1]

1 Introduction

The ability to reliably detect AI-generated text is becoming increasingly critical as large language models (LLMs) are deployed in education, media, and content moderation (Guo et al., 2024; Hu et al., 2023). While recent detectors achieve near-perfect performance on standard benchmarks (Guo et al., 2023; Wang et al., 2024a), these evaluations typically assume that training and testing data are drawn from the same distribution. In real-world scenarios, however, AI-generated text varies widely across prompts, model families, and domains, raising serious concerns about how well detectors generalize under distribution shifts.

Prior studies have begun to examine generalization in the context of unseen prompts, models, or datasets (Xu et al., 2024; Liu et al., 2024). However, these efforts largely focus on reporting performance drops without probing the underlying causes. Meanwhile, recent benchmark datasets introduce diversity in generation settings (Wang et al., 2024a; Macko et al., 2023; Li et al., 2024), but offer limited interpretability regarding the features detectors rely on. A more systematic and interpretable approach is needed to explain **why** generalization succeeds or fails.

In this paper, we propose to understand generalization through the lens of linguistic analysis. We hypothesize that changes in surface-level linguistic features, such as verb tense, syntactic complexity, or pronoun usage, can partially account for generalization behavior. To test this hypothesis, we construct a comprehensive benchmark combining 7 LLMs (e.g., Deepseek (DeepSeek-AI, 2025), Mistral (Mistral AI, 2024)), 4 domains (abstracts, news, reviews, QA), and 6 prompting strategies (e.g., few-shot, chain-of-thought (CoT)), enabling evaluation across prompt, model, and dataset generalization. We train the AI-text detectors by fine-tuning on two state-of-the-art models (XLM-RoBERTa (Conneau et al., 2020) and DeBERTa-V3 (He et al., 2021)) for binary classification tasks. Each detector is trained with the texts generated by every possible condition (combination of prompt, model and dataset). The fine-tuned detectors are evaluated for cross-prompt, cross-model and cross-dataset generalization performance. Our results reveal substantial performance degradation under out-of-domain conditions, despite near-perfect in-domain accuracy.

To explain these generalization gaps, we perform a large-scale correlation analysis between detector generalization performance and score changes in 80 linguistic feature metrics across training and testing settings. We find that linguistic features have significant ($p<0.05$) correlations with generaliza-

[1] Code and data is available at: https://anonymous.4open.science/r/prompting-benchmark-EB7E

tion across different test settings. While some linguistic features (e.g., passive voice, short sentence ratio) are strongly correlated (correlation score > 0.7) with generalization behavior in specific configurations, there is no universal linguistic signal that explains all cases. Our findings suggest that linguistic features offer a useful, though partial, explanation of generalization, and that detectors may rely on different features depending on their training setup and the test conditions.

In summary, our main contributions are: (1) A new benchmark for evaluating AI-text detector generalization across 6 prompts, 4 domains, and 7 LLMs; (2) A comprehensive analysis linking linguistic feature shifts to detector generalization performance; (3) Insights into which linguistic features are most predictive of generalization behavior, helping to guide the development of more robust and interpretable detectors.

Ultimately, our work aims to move beyond raw performance scores toward a deeper understanding of generalization behavior in AI-text detection. While linguistic features alone do not capture the full complexity of generalization, they provide a valuable starting point for interpreting detector behavior in the wild.

2 Related Work

Datasets for AI Text Detection. Recent benchmarks have advanced AI-generated text detection by covering multiple languages (Wang et al., 2024a; Macko et al., 2024), domains (Li et al., 2024; Verma et al., 2024; Dugan et al., 2024), and generator models (Hu et al., 2023; Abassy et al., 2024; Tao et al., 2024). Several works also consider mixed-authorship settings (Yu et al., 2024; Zhang et al., 2024). However, few datasets jointly evaluate the impact of LLM type, domain, and prompt style in a systematic and controlled manner. Prompt engineering remains particularly underexplored, despite its known influence on generation behavior. To address these gaps, we introduce a new dataset that enables controlled experiments across 7 LLMs, 4 domains, and 6 prompting strategies.

AI Text Detection Models. Detection methods mainly fall into two broad categories: statistical detectors (e.g., GLTR, DetectGPT, Binoculars) and fine-tuned classifiers using pretrained LMs (e.g., RoBERTa, DeBERTa, XLM-R). While statistical detectors offer interpretability, classifier-based approaches consistently achieve stronger performance on benchmark datasets such as M4GT (Wang et al., 2024a), MULTITuDE (Macko et al., 2023), and MultiSocial (Macko et al., 2024). Our study builds on this foundation by fine-tuning two top-performing detectors, XLM-RoBERTa and DeBERTa-V3, across varied training settings to assess their generalization.

Generalization in Detection. Generalization is a core challenge in AI-text detection. Prior work has shown that detectors trained on one task or domain often fail when evaluated on others (Xu et al., 2024; Li et al., 2024; Bhattacharjee et al., 2024). However, most studies focus on reporting performance gaps without offering deeper explanations. Moreover, while some papers examine prompt-based variation, they typically limit prompting strategies or focus on handcrafted or adversarial prompts (Zhang et al., 2024). In contrast, our work evaluates generalization comprehensively across prompt styles, model architectures, and content domains, using naturalistic generation strategies (e.g., few-shot, CoT, zero-shot) common in practice.

Explaining Generalization Behavior. Several works have proposed high-level explanations for generalization variance. For example, Xu et al. (2024) attribute success to *prompt similarity* and *human–LLM alignment*, while Li et al. (2024) explore distributional differences using linguistic metrics like POS tags and named entity counts. However, these studies stop short of identifying which specific linguistic features correlate with generalization. Our work advances this line of inquiry through a detailed correlation analysis of 80 linguistic features, covering syntactic, stylistic, and discourse-level signals. We quantify feature shifts between training and testing data and link them to generalization performance, revealing interpretable signals, e.g., shifts in pronoun frequency or passive voice usage that influence detector robustness.

3 Dataset Creation

To provide a comprehensive study of generalization of AI-text detectors against prompt, LLM and dataset changes, we create our human-written and AI-generated text dataset by incorporating 6 prompting strategies, 7 LLMs with different parameter sizes from different model families, and 4 datasets from different domains.

3.1 Human-written text

We first randomly sample the human-written text of different domains from 4 datasets: (1) Scientific paper **abstracts** from the arXiv dataset (See et al., 2017); (2) Product **reviews** from the AmazonReviews2023 dataset (Hou et al., 2024); (3) **News** articles from the CNN/Daily Mail dataset (Clement et al., 2019); (4) Question and answers (**QA**) from the ASQA dataset (Stelmakh et al., 2023).

Data cleaning for human-written text. To remove obvious features for AI-text detectors, we use the following data cleaning steps for all the human-written texts: (1) Removing duplicates; (2) Normalizing punctuation; (3) Removing duplicated whitespace; (4) Removing URLs, e-mail addresses, and emojis; (5) Artifacts such as dates of article; (6) Filtering non-English text; (7) Filter too short text, as text length can impact the difficulty of the task (Wang et al., 2024b).

For abstracts and news articles, we use only texts that are at least 1,000 characters long. We set the minimum text length for reviews to 350 characters because the original text is short. For the QA dataset, we sample the longest texts considering the limited size of the data. For each of the datasets, we sample 3,000 examples and split them into training, validation and testing set with a split ratio of 50:17:33.

3.2 AI Text Generated with LLMs

The AI-text part of the dataset consists of texts generated by 7 LLMs using 6 diverse prompting strategies for each dataset. For each human-written text, we apply every LLM and prompting strategy to generate an AI-text counterpart under the same topic. Consequently, for each source dataset, the final data include 3,000 human-written texts and, for every model–prompt combination, 3,000 corresponding AI-generated texts. In total, the dataset comprises 516,000 texts: 12,000 human-written and 504,000 AI-text. Each AI text is matched with its human-written text counterpart for performing a binary classification.

LLMs. We employ LLMs with different parameter sizes and from 5 model families: (1) **Mistral 123B** (Mistral AI, 2024): Mistral-Large-Instruct-2411; (2) **Deepseek 70B** (DeepSeek-AI, 2025): DeepSeek-R1-Distill-Llama-70B; (3) **Llama 70B** (Meta AI, 2024): Llama-3.3-70B-Instruct; (4) **Qwen 72B, Qwen 32B, Qwen 14B** (Yang et al., 2024): Qwen2.5-72B/-32B/-14B-Instruct; (5) **Solar 22B** (Upstage, 2024): solar-pro-preview-instruct.

Prompts. We use 6 different prompting strategies based on existing research on prompt engineering. The prompts include: (1) **0-shot** prompts that only provide the metadata (e.g., title, text length) of each data sample; (2) **3-shot** prompts (Brown et al., 2020) that contains 3 human-written texts from the same dataset for in-context learning; (3) **Style** prompts (Zhang et al., 2024) which require LLMs to write in a style like the given human-written text example; (4) **0-shot CoT** prompts (Kojima et al., 2023) which consist of phrase "let's think step by step."; (5) **1-shot CoT** prompts (Wei et al., 2023) that contain manually written step-by-step instructions, the instruction is based on an example of a human-written text; and (6) **Self-refine** prompts (Madaan et al., 2023) that use the LLM itself to critique and improve its own responses. Self-refine prompts are multi-stage prompts that comprise 4 stages: firstly, the LLM is prompted to generate the AI text, then it is requested to provide feedback on how to make the generated AI text more human-like. Later, the LLM needs to incorporate the feedback to improve the initially generated AI text. The improved text is final if the LLM judges it sounds more human-written than the human-written text counterpart; otherwise, it goes back to the feedback step for at most 3 iterations. For each prompting strategy, we modified the prompt template used for each dataset to suit the task of text generation. To match the text length of the AI-generated texts to their human-written counterparts, we include information about character count in the prompts. A detailed discussion of the prompts can be found in the appendix A.1, along with the prompt templates used to create our dataset (Table 3).

Data Cleaning for AI-Generated Text. To prevent detectors from exploiting superficial artifacts rather than genuine linguistic characteristics, we extensively clean the LLM-generated texts by removing elements that could trivially reveal their artificial origin. Specifically, we remove formulaic AI responses (e.g., *"Certainly!"*, *"Sure!"*), structural markers such as section titles, bullet points, and numbered lists, placeholder tokens in square brackets (e.g., *[your name]*, *[insert e-mail address]*), extraneous metadata including review ratings, character-count information, and sentences beginning with *"Note:"* that describe the generation process, non-linguistic symbols such as asterisks (*), triple dashes (–), and hash symbols (#), as

well as model-specific tags such as `\think` and any preceding text in Deepseek reasoning outputs. This cleaning step ensures that the evaluation focuses on the linguistic properties of the generated text rather than on easily detectable formatting artifacts.

4 Generalization of AI-Text Detectors

We introduce three evaluation settings to assess the generalization ability of AI-text detectors across three dimensions: prompts, LLMs, and datasets.

4.1 Generalization Testing

Assume an AI-text detector M is fine-tuned on a training set $\mathcal{D}^{\text{train}}_{i,j,k}$, consisting of AI-generated texts produced with prompt p_i by LLM g_j for dataset d_k, along with human-written texts. We evaluate the cross-prompt, cross-model, and cross-dataset generalization of M under the following settings.

Cross-prompt (C-P) testing. This setting evaluates how well detectors generalize to texts generated with prompting strategies unseen during training. Each detector is evaluated on test sets produced by the same LLM and drawn from the same dataset, but generated with different prompts. This controlled setting ensures that the only changing factor is prompt strategies during evaluation. For example, a detector trained on the training split of QA texts generated by Llama 70B with 0-shot prompts is evaluated on the test split generated by Llama 70B with all 6 prompt types. Formally, the generalization accuracy of the detector, $\text{Acc}(M)$, from prompt p_i to all prompts is:

$$\Delta^{(p_i)}_{\text{gen}} = \left\{\text{Acc}\left(M\left(\mathcal{D}^{\text{test}}_{c,j,k} \mid \mathcal{D}^{\text{train}}_{i,j,k}\right)\right)\right\}_{c=1}^{6} \quad (1)$$

Where $\Delta^{(p_i)}_{\text{gen}}$ is a list of 6 accuracy values, with each value representing the generalization accuracy of a prompt. We carried out the test for each prompt and resulted in a 2-dimensional 6x6 vector (plot like left heatmaps in Figure 2), which presents the cross-prompt result of all prompts in one of the conditions, denoted as $\Delta^{(p)}_{\text{gen}}$.

Cross-model (C-M) testing. This setting evaluates generalization to texts generated by LLMs not seen during training. A detector fine-tuned on the training split from one LLM is evaluated on the test splits produced by other LLMs. For example, a detector trained on abstracts generated by Llama 70B is tested on abstracts generated by all 7 LLMs. We use 0-shot prompts (p_1) in this setting. Formally:

$$\Delta^{(g_j)}_{\text{gen}} = \left\{\text{Acc}\left(M\left(\mathcal{D}^{\text{test}}_{1,c,k} \mid \mathcal{D}^{\text{train}}_{1,j,k}\right)\right)\right\}_{c=1}^{7} \quad (2)$$

Similar to cross-prompt testing, we apply the formula to all LLMs, obtaining in a 7x7 vector as cross-model results $\Delta^{(g)}_{\text{gen}}$.

Cross-dataset (C-D) testing. This setting evaluates generalization across different dataset domains. A detector fine-tuned on the training split from one dataset is evaluated on test splits from other datasets generated by the same LLM. For example, a detector trained on abstracts generated by Llama 70B is tested on news, reviews, and QA data generated by the same LLM. We use 0-shot prompts (p_1) in this setting. Formally:

$$\Delta^{(d_k)}_{\text{gen}} = \left\{\text{Acc}\left(M\left(\mathcal{D}^{\text{test}}_{1,j,c} \mid \mathcal{D}^{\text{train}}_{1,j,k}\right)\right)\right\}_{c=1}^{4} \quad (3)$$

The corresponding cross-dataset result $\Delta^{(d)}_{\text{gen}}$ is a 4x4 vector.

4.2 Training setup

We fine-tune XLM-RoBERTa-base (Conneau et al., 2019) (referred to as RoBERTa) and DeBERTa-V3-small (He et al., 2021) (referred to as DeBERTa) for binary classification to distinguish between human-written and AI-generated text. These architectures have achieved state-of-the-art performance in prior work (Wang et al., 2024a). We train a separate detector for each combination of prompt type, AI-text generation model, and dataset type, resulting in 168 (i.e., 7x4x6) in-domain detectors when using, for example, RoBERTa for fine-tuning.

Model fine-tuning is performed using the following hyperparameters: `learning rate` = 2e-5, `num train epochs` = 3, `weight decay` = 0.01, `train batch size` = 16. The maximum sequence length is set to 512, corresponding to the maximum input length supported by XLM-RoBERTa. All our experiments are conducted on NVIDIA HGX H100, and approximately 400 GPU hours to replicate.

5 Explaining Generalization with Linguistic Analysis

To better understand what causes the variance of generalization performance, we perform a comprehensive linguistic feature analysis by measuring the correlation of 80 different feature metrics with the generation results. We first introduce the definition and metric of each linguistic feature and present the correlation evaluation method.

5.1 Linguistic Feature Definitions and Metrics

Our studied features can be categorized into Lexical diversity, Lexical density, Sentiment, Readabil-

ity, Part-of-Speech (POS), and Grammatical and Lexical analysis. We introduce selected features and metrics in the main paper and the complete set in Appendix A.4.

Readability. Readability refers to the ease of understanding a text. AI-generated text tends to be less readable than human-written text (Markowitz et al., 2024; Mathews et al., 2024). We use the **Gunning fog index** (Yadagiri et al., 2024) as one of the measures of readability, which is an estimated number of years needed to understand a given passage.

Part-of-Speech (POS). We use a selection of metrics from StyloMetrix (Okulska et al., 2023) to compare the frequency of **verbs, nouns, adjectives, numerals, etc**. The frequency of parts of speech is measured as the fraction of text covered by tokens representing a given part of speech. Previous research has discovered differences between human-written and AI-generated text in terms of frequency of certain POS (Georgiou, 2024). Therefore, POS analysis is relevant for gaining insights into the literary style of texts in our dataset.

Grammatical analysis. We use StyloMetrix (Okulska et al., 2023) metrics to compare texts in terms of grammatical categories related to verbs. Human-written texts have been shown to contain more passive voice than AI-generated texts (Georgiou, 2024). We measure the **incidence of passive and active voice** as the frequency of verbs in passive or active voice. We also compare the differences between the choice of tenses. The **frequency of past, present and future tenses** is measured as the fraction of the text covered by verbs in past, present and future tenses. For example, the incidence of past tenses is the number of verbs in past simple, past continuous, past perfect or past perfect continuous divided by the total number of words in the text.

Lexical analysis. As part of lexical analysis, we measure the **frequency of personal names**, and **adjectives in comparative and superlative degrees**. The pronoun-related metrics (Okulska et al., 2023) analyze the differences in the usage of pronouns in human-written and AI-generated text. We calculate the frequency of specific personal or reflexive pronouns and certain types of pronouns (e.g., "We", "It", "Our", "Yourself").

5.2 Correlation Between Generalization and Linguistic Features

To further investigate factors that influence generalization, we examine how changes in linguistic features correlate with generalization performance. For each linguistic feature f, we compute its shift between a given training configuration and the corresponding test configuration:

$$\Delta_f = f(\mathcal{D}^{\text{train}}) - f(\mathcal{D}^{\text{test}}) \quad (4)$$

where $\mathcal{D}^{\text{train}}$ denotes the set of AI-generated texts in the training configuration, and $\mathcal{D}^{\text{test}}$ denotes the texts in the corresponding test configuration. Corresponding to the generalization testing, we denote the cross-prompt, cross-model and cross-dataset feature shift as $\Delta_f^{(p)}$, $\Delta_f^{(g)}$, $\Delta_f^{(d)}$, which also contain the same size of values.

We then compute the Pearson correlation between **flattened** generalization accuracy and these feature shifts under the same conditions:

$$Corr(f) = |\text{Pearson}(\Delta_{\text{gen}}^{(n)}, \Delta_f^{(n)})| \quad (5)$$

where n indexes each cross-prompt, cross-model, or cross-dataset comparison. The resulting value $Corr(f)$ lies in $[0, 1]$, with $0.1 \leq Corr(f) < 0.3$ indicating a low correlation, $0.3 \leq Corr(f) < 0.5$, $0.5 \leq Corr(f) < 0.7$ and $Corr(f) \geq 0.7$ as moderate, high and strong correlations (DATAtab Team, 2025). This analysis allows us to identify which linguistic features are most strongly associated with robust generalization across different test settings.

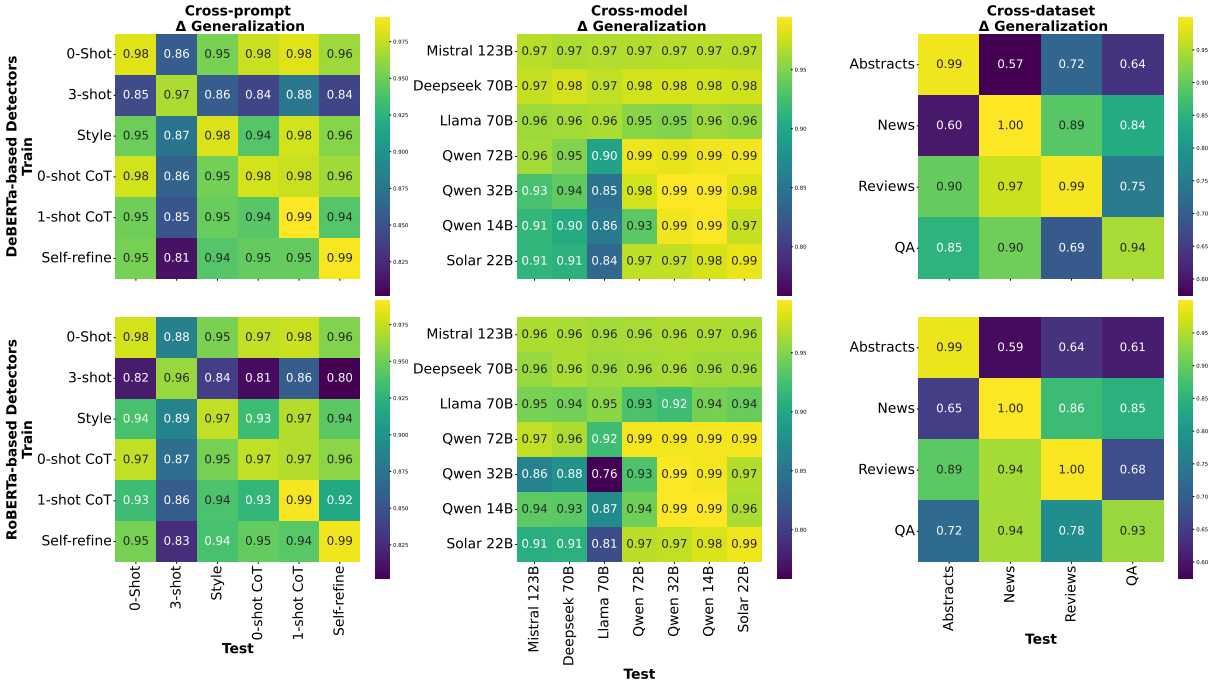
Setting-specific correlation (Table 2) is measured under a specific testing combination. For example, the setting-specific correlation of cross-prompt generalization and linguistic feature shifts is only measured on texts from Llama 70B and the Abstract dataset.

Overall correlation (Table 1) is measured under all combinations. For example, the overall correlation of cross-prompt generalization and linguistic feature shifts is measured on the texts of every combination of LLMs and datasets.

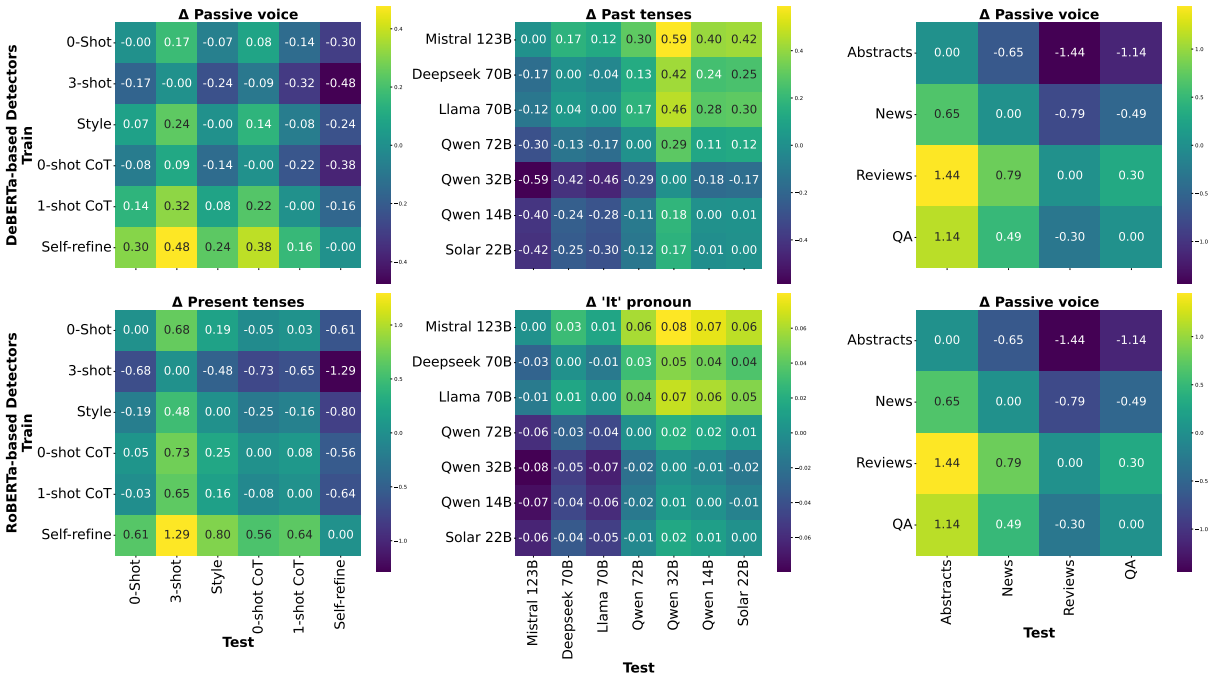
Aggregation of Results. To provide a high-level summary of cross-prompt, cross-model, and cross-dataset generalization and feature shifts (as Figure 1a and 1b), we report accuracy and shift values averaged over the dimensions that are not the focus of the evaluation. For instance, when presenting overall cross-prompt results, we average accuracy scores across all 7 LLMs and 4 datasets.

6 Results and Analysis

We analyze the results in Table 1, Table 2, and Figure 1 to understand how generalization performance is shaped by shifts in linguistic features.



(a) The aggregated generalization performance.



(b) The aggregated feature shifts between training and test configurations.

Figure 1: Comparison of aggregated generalization performance and aggregated feature shifts across all evaluation settings. Similar patterns in the two heatmaps indicate that certain feature shifts are correlated with reduced generalization accuracy.

Finding 1: Linguistic features have significant correlations with generalization results. Table 1 shows that several features exhibit significant correlations (bold values) with detector generalization. For example, overall *cross-model generalization* that averaged across datasets is moderately correlated (0.416) with the proportion of past-tense verbs, indicating that stylistic verb usage in training data influences transfer. In contrast, cross-prompt generalization averaged across datasets and LLMs shows only weak correlation with passive voice. These results highlight that some features play a more critical role than others in determining generalization success.

Linsuitic Feature	Feature Metric	DeBERTa-based C-P	DeBERTa-based C-M	DeBERTa-based C-D	RoBERTa-based C-P	RoBERTa-based C-M	RoBERTa-based C-D
Readability	Gunning fog	0.056	**0.248**	**0.231**	0.043	**0.308**	***0.261***
Part-of-Speech	Numerals	***0.108***	***0.363***	0.086	**0.076**	**0.157**	0.031
Grammatical	Passive voice	**0.109**	**0.281**	**0.296**	0.054	**0.221**	**0.287**
	Active voice	0.010	**0.194**	0.066	0.061	***0.373***	0.014
	Present tenses	0.046	**0.147**	0.144	**0.116**	**0.328**	0.173
	Past tenses	0.018	**0.416**	0.031	0.006	**0.324**	0.049
	Future tenses	**0.062**	0.066	0.159	**0.069**	**0.172**	0.111
Lexical	Personal names	**0.076**	***0.381***	0.030	0.046	**0.182**	0.071
	Adjectives in comparative degree	0.027	0.123	***0.267***	0.023	**0.317**	***0.251***
	Adjectives in superlative degree	***0.095***	**0.239**	0.034	0.047	0.123	0.055
	"We" pronoun	0.014	0.041	0.015	***0.108***	0.030	0.037
	"It" pronoun	0.029	**0.236**	0.184	**0.077**	**0.385**	0.120
	"Our" possessive pronoun	0.004	0.007	***0.261***	***0.113***	0.063	**0.246**
	"Yourself" pronoun	0.030	**0.183**	0.157	0.061	***0.366***	0.111

Table 1: The **overall** correlation between generalization performance with different linguistic features. This table only presents the features that fall into the top 3 correlated features in one of the settings. The **significant (p<0.05) correlation is bolded**. We underline the strongest correlation for each setting, and *italicize the other scores within the top 3 correlated features*. The results of other features that are less correlated are shown in the Appendix.

	DeBERTa-based Abstracts	DeBERTa-based News	DeBERTa-based Reviews	DeBERTa-based QA	DeBERTa-based ALL	RoBERTa-based Abstracts	RoBERTa-based News	RoBERTa-based Reviews	RoBERTa-based QA	RoBERTa-based ALL
Mistral 123B	**0.421**	**0.636**	**0.577**	**0.709**	**0.395**	0.155	**0.703**	**0.573**	**0.630**	**0.219**
Deepseek 70B	**0.342**	**0.605**	**0.415**	**0.735**	**0.276**	**0.528**	**0.506**	**0.459**	**0.728**	**0.400**
Llama 70B	**0.412**	0.209	**0.736**	**0.584**	**0.346**	0.245	0.248	**0.758**	**0.572**	**0.380**
Qwen 72B	0.072	**0.590**	0.317	0.301	0.098	0.128	**0.519**	**0.549**	**0.357**	**0.212**
Qwen 32B	0.214	**0.684**	**0.527**	**0.678**	**0.183**	**0.377**	**0.672**	**0.617**	**0.492**	**0.308**
Qwen 14B	0.275	**0.476**	**0.553**	**0.560**	**0.218**	0.264	**0.457**	**0.605**	**0.580**	**0.204**
Solar 22B	**0.563**	**0.482**	**0.517**	**0.614**	**0.303**	**0.703**	**0.688**	0.320	**0.582**	**0.426**
ALL	**0.196**	**0.231**	**0.437**	**0.255**	**0.109**	**0.246**	**0.155**	**0.486**	**0.224**	**0.116**

Table 2: The cross-prompt correlation between generation performance and the most correlated linguistic feature when evaluated on different datasets and models. The **significant correlation is bolded**. We underline the strongest correlation for each dataset. Similar cross-model and cross-dataset results are in Appendix.

Finding 2: Certain dataset–model combinations reveal very strong feature dependencies. Although overall cross-prompt correlations appear weak in Table 1, Table 2 reveals that in specific configurations the effect is dramatic. For instance, on Llama-70B outputs for the Reviews dataset, cross-prompt generalization is strongly correlated (>0.7) with the number of short sentences. Fig 2a shows that changes in this feature align directly with sharp drops in performance when generalizing from 1-shot CoT to other prompting strategies. This demonstrates that some detectors are highly sensitive to prompt-induced shifts in linguistic structure.

Finding 3: Different detectors rely on different linguistic features. RoBERTa-based detectors and DeBERTa-based detectors do not exploit the same linguistic signals. For example, the cross-model generalization of RoBERTa models is moderately correlated (0.385) with the frequency of "It" pronouns, whereas for DeBERTa the correlation is only 0.236. This suggests that detectors may learn fundamentally different features for distinguishing human and AI text even when trained on the same data.

6.1 Linguistic Analysis of Generalization Results

Figure 4 summarizes average performance across the three generalization settings. As expected, **in-domain testing achieves near-perfect accuracy**, as shown in the diagonal cells of Figure 1a, confirming the strong baseline capabilities of our detectors. Detailed in-domain results are reported in Table 4 in the Appendix.

6.1.1 Cross-prompt Generalization

The most striking pattern is that the **3-shot prompt is consistently the hardest to generalize to and from**, with accuracy dropping to 80–89%. Other

prompting strategies show relatively minor effects.

Explanation. Figure 1b shows averaged feature shifts, but strong effects can be masked by aggregation. For example, Figure 2b highlights a clear pattern: AI texts that use the "We" pronoun in similar contexts are more difficult to generalize to for the detectors. This confirms Findings 2 and 3: when we zoom in on specific dataset–model pairs, clear linguistic drivers of generalization emerge.

6.1.2 Cross-model Generalization

A major finding is that **detectors trained on Qwen or Solar outputs perform poorly on Llama-generated text**, whereas generalization across other LLMs is more stable.

Explanation. Figures 1a and 1b reveal that cross-model generalization is moderately influenced by shifts in past-tense usage and "It" pronoun frequency. Qwen and Solar outputs share similar linguistic profiles, which diverge from Llama's, explaining this degradation.

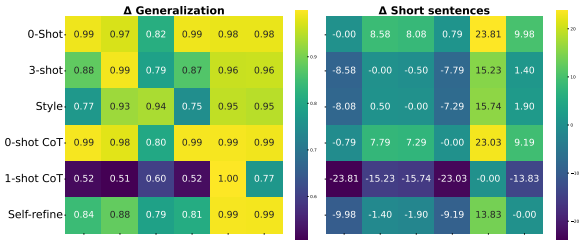
6.1.3 Cross-dataset Generalization

The most pronounced performance gap appears here: **detectors trained on abstracts generalize poorly to other datasets, achieving as low as 57% accuracy when tested on news articles**. Conversely, detectors trained on reviews or QA data transfer more successfully to news.

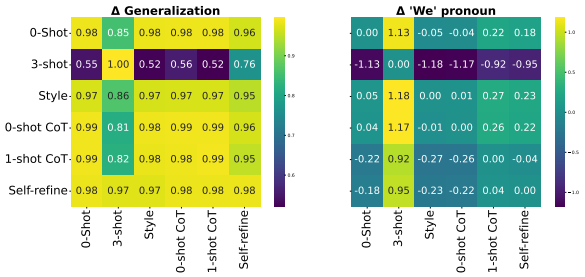
Explanation. Across all detectors, cross-dataset generalization shows moderate correlation with passive voice usage. Abstracts exhibit a higher rate of passive constructions, which likely makes them a poor source domain for training detectors that must generalize broadly.

6.1.4 Discussion

While our analysis highlights the role of linguistic features in explaining the generalization behavior of AI-text detectors, we acknowledge that these features represent only one facet of a more complex landscape. Generalization performance is likely influenced by a broader set of factors, including semantic coherence, discourse structure, and detector-specific inductive biases. Our findings should therefore be interpreted as offering a linguistic perspective rather than a comprehensive account of generalization. Nonetheless, by systematically correlating linguistic feature shifts with detection performance, our study contributes valuable insights into how stylistic and grammatical signals may impact detector robustness across prompts, models, and domains.



(a) Llama 70B, Reviews



(b) Deepseek 70B, Abstracts

Figure 2: Cross-prompt generalization and feature shifts when evaluating on a specific model and dataset. A clearer correlation is observed than the overall cross-prompt correlation in Figure 1.

7 Conclusion

This work presents an interpretable investigation into the generalization behavior of AI-text detectors. While prior studies primarily report detection performance, we go further by examining why generalization succeeds or fails through a linguistic lens. Across a large-scale benchmark incorporating diverse prompts, LLMs, and domains, we show that state-of-the-art detectors, despite near-perfect in-domain accuracy, often struggle in cross-prompt, cross-model, and cross-dataset scenarios. To explain these generalization behaviors, we quantify shifts in 80 linguistic features between training and testing distributions and uncover statistically significant correlations between these shifts and generalization performance.

Our analysis reveals that features such as pronoun usage, verb tense, and passive voice are predictive of generalization gaps, but their influence varies across detectors and settings. This suggests that detectors latch onto different linguistic signals depending on their training context, impacting their robustness in deployment.

These findings underscore two key points: (1) evaluation must extend beyond in-domain testing to realistically assess detector reliability, and (2) linguistic analysis provides a principled and interpretable path toward diagnosing and improving generalization.

Limitations

While our work offers new insights into the generalization behavior of AI-text detectors through linguistic analysis, several limitations remain.

First, our study focuses on English-language text and detectors trained on English corpora. Although our methodology can be extended to multilingual settings, the linguistic features and generalization patterns may differ significantly across languages due to variations in grammar and stylistic conventions.

Second, we rely on surface-level linguistic features (e.g., POS tags, sentence length, voice, pronouns) that can be extracted using standard NLP tools. While these features provide interpretable signals, they may not capture deeper semantic or discourse-level properties that also influence detector decisions.

Third, the detectors we evaluate are based on fine-tuned encoder-only transformer models. Other architectures, such as generative or retrieval-augmented models, may exhibit different generalization behaviors and rely on alternative linguistic features.

Fourth, our correlation-based analysis reveals associations but does not establish causal relationships between feature shifts and performance drops. Further research using controlled interventions or counterfactual examples would be needed to verify causality.

Lastly, our dataset covers a wide but still limited set of domains, models, and prompting strategies. As the landscape of LLMs and prompting methods continues to evolve, future work should assess whether our findings hold for more recent or unseen generation techniques.

Despite these limitations, our study provides a strong foundation for more principled and interpretable evaluations of generalization in AI-text detection.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-DetectAIve: a Tool for Fine-Grained Machine-Generated Text Detection. arXiv preprint. ArXiv:2408.04284 [cs].

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Eagle: A domain generalization framework for ai-generated text detection. Preprint, arXiv:2403.15690.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. Preprint, arXiv:2005.14165.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. Preprint, arXiv:1905.00075.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. CoRR, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of Quantitative Linguistics, 17(2):94–100.

DATAtab Team. 2025. Pearson correlation: A beginner's guide. https://datatab.net/tutorial/pearson-correlation. DATAtab: Online Statistics Calculator. DATAtab e.U., Graz, Austria.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David Hua. 2023.

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.

Georgios P. Georgiou. 2024. Differentiating between human-written and ai-generated texts using linguistic features automatically extracted from an online computational tool. Preprint, arXiv:2407.03646.

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

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. Advances in Neural Information Processing Systems, 37:88320–88347.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. Preprint, arXiv:2403.03952.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. Advances in neural information processing systems, 36:15077–15095.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. Preprint, arXiv:2205.11916.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. arXiv preprint. ArXiv:2305.13242 [cs].

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. On the detectability of chatgpt content: Benchmarking, methodology, and evaluation through the lens of academic writing. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24, page 2236–2250, New York, NY, USA. Association for Computing Machinery.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2024. MultiSocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts. arXiv preprint. ArXiv:2406.12549 [cs].

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9960–9987. ArXiv:2310.13606 [cs].

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. Preprint, arXiv:2303.17651.

Anastasia Margolina and Anastasia Kolmogorova. 2023. Exploring Evaluation Techniques in Controlled Text Generation: A Comparative Study of Semantics and Sentiment in ruGPT3large-Generated and Human-Written Movie Reviews. In COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES". RSUH.

David M. Markowitz, Jeffrey T. Hancock, and Jeremy N. Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. Journal of Language and Social Psychology, 43(1):63–82.

Daniel Mathews, Justin P Varghese, and Libin Chacko Samuel. 2024. Classifying ai-generated summaries and human summaries based on statistical features. In 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, pages 1–5.

Akshay Mendhakar and Darshan H S. 2023. Parts-of-speech (pos) analysis and classification of various text genres. Corpus-based Studies across Humanities, 1(1):99–131.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2025-05-01.

Mistral AI. 2024. Mistral large instruct 2411. Accessed: 2025-05-01.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting Linguistic Patterns in Human and LLM-Generated Text. arXiv preprint. ArXiv:2308.09067 [cs].

Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. Stylometrix: An open-source multilingual tool for representing stylometric vectors. Preprint, arXiv:2309.12810.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. Preprint, arXiv:2405.10129.

Jacques Savoy. 2020. Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling. Springer International Publishing, Cham.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. Asqa: Factoid questions meet long-form answers. Preprint, arXiv:2204.06092.

Zhen Tao, Zhiyu Li, Dinghao Xi, and Wei Xu. 2024. CUDRT: Benchmarking the Detection of Human vs. Large Language Models Generated Texts. arXiv preprint. ArXiv:2406.09056 [cs].

Upstage. 2024. Solar pro preview instruct. https://huggingface.co/upstage/solar-pro-preview-instruct. Instruction-tuned 22B-parameter LLM optimized for single-GPU performance. Preview released September 2024.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. arXiv preprint. ArXiv:2305.15047 [cs].

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. 2024a. 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.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. Preprint, arXiv:2201.11903.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2024. On the generalization of training-based ChatGPT detection methods. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7223–7243, Miami, Florida, USA. Association for Computational Linguistics.

Annepaka Yadagiri, Lavanya Shree, Suraiya Parween, Anushka Raj, Shreya Maurya, and Partha Pakray. 2024. Detecting AI-generated text with pre-trained models using linguistic features. In Proceedings of the 21st International Conference on Natural Language Processing (ICON), pages 188–196, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Hatice Yildiz Durak, Figen Eğin, and Aytuğ Onan. 2025. A Comparison of Human-Written Versus AI-Generated Text in Discussions at Educational Settings: Investigating Features for ChatGPT, Gemini and BingAI. European Journal of Education, 60(1):e70014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejed.70014.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2024. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. Preprint, arXiv:2304.12008.

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

Yuehan Zhang, Yongqiang Ma, Jiawei Liu, Xiaozhong Liu, Xiaofeng Wang, and Wei Lu. 2024. Detection vs. anti-detection: Is text generated by ai detectable? In Wisdom, Well-Being, Win-Win, pages 209–222, Cham. Springer Nature Switzerland.

A Appendix

A.1 Overview of prompting techniques

To test diverse prompting techniques, we choose six different prompt types. The prompt types are:

0-shot: Our zero-shot prompts are simple instructions for the language models that include a basic piece of information about the text and desired length of the text. For example, the template for the prompt to generate abstracts is *Write an abstract for an article titled "{title}". The abstract should be around {length} characters long.*

Chain-of-Thought (0-shot CoT): *A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output* (Wei et al., 2023). Chain-of-thought prompting is a strategy that aims to improve reasoning by dividing the task into smaller steps (Wei et al., 2023). A chain-of-thought prompt usually relies on exemplars, containing a prompt and a correct response. The example response is expressed as a series to steps which lead to a final output. This prompts the model to reason step-by-step.

For the purpose of our dataset, we adapt a **1-shot CoT** prompting strategy to the task of text generation: this prompt subtype consists of a human-written step-by-step instruction for the model. The instruction is based on an example of a human-written text from the dataset.

We also use **0-shot CoT** prompts (Kojima et al., 2023), which consist of adding *let's think step by step* to the baseline prompt to cause step-by-step reasoning.

Few-shot In-context learning (3-shot) (Brown et al., 2020): The few-shot prompt contains several examples of input-output pairs for a given task. Our few-shot prompts include three human-written examples from the dataset.

Style examples (Style): In Zhang et al. (2024), prompts with style guidelines have been effective in prompting LLMs to generate output that evades detection methods. In this study, the models have been prompted to simulate the styles of several famous writers. We adapt one of the prompts of this study for our task. We use the style example prompt, but use an example from the human-written part of our dataset as an example. Even though using a text written by a famous author (e.g. Shakespeare) as a style example has been successful in preventing detection (Zhang et al., 2024), this prompt modification has been done to create a more realistic setting.

Self-refine (Madaan et al., 2023): Self-refinement prompts are prompts that use the LLM itself to critique and improve its own responses. The prompts of this kind are comprised of three stages: generating the output, generating feedback for the output and applying the feedback to the output. The second and third steps are repeated until a stopping condition is met. In this case, we base the stopping condition on the ability of the model itself to distinguish its own outputs from human-written text. To achieve this, we use evaluation prompts based on the GPT-4 evaluation prompt used in (Madaan et al., 2023). For the purpose of our task, we ask the model to decide which text sounds more human-written. The resulting set of prompts is made of the following stages:

- Initialization prompt: the same as our baseline zero-shot prompt
- Feedback prompt: prompt used for generating feedback on how to make the text seem more human-written
- Iterate prompt: used to get the next iteration
- Evaluate prompt: used to check whether the stopping condition (the model classifying the text as human-written) has been met

A.2 Prompt templates

Prompt	Dataset	Prompt template
0-shot	Abstracts	"Write an abstract for an article in {category} with a title: \"{title}\". The abstract should be around {length} characters long."
	News	"Write a news article based on the following highlights:\n\"{highlights}\"\nYour article should be around {length} characters long."
	Reviews	"Write an Amazon review for the item \"{item_name}\" with a title \"{title}\" and a rating of {rating}. The review should be around {length} characters long."
	QA	"{question}\n Your answer should be around {length} characters long."
0-shot CoT	Abstracts	"Write an abstract for an article in {category} with a title: \"{title}\". Let's think step by step. Your answer should only include the abstract. The abstract should be around {length} characters long."
	News	"Write a news article based on the following highlights:\n\"{highlights}\"\nLet's think step by step. Your answer should only include the article. The article should be around {length} characters long."
	Reviews	"Write an Amazon review for the item \"{item_name}\" with a title \"{title}\" and a rating of {rating}. Let's think step by step. Your answer should only include the review. The review should be around {length} characters long."
	QA	"{question}\nLet's think step by step. Your answer should only include the answer to the question. The answer should be around {length} characters long."
1-shot CoT	Abstracts	"I want to write an abstract for an article in computer science. The article is titled \"ConQRet: Benchmarking Fine-Grained Evaluation of Retrieval Augmented Argumentation with LLM Judges\". \n 1. First, I introduce the context of the research and explain the motivation:\n\"Computational argumentation, which involves generating answers or summaries for controversial topics like abortion bans and vaccination, has become increasingly important in today's polarized environment. Sophisticated LLM capabilities offer the potential to provide nuanced, evidence-based answers to such questions through Retrieval-Augmented Argumentation (RAArg), leveraging real-world evidence for high-quality, grounded arguments. However, evaluating RAArg remains challenging, as human evaluation is costly and difficult for complex, lengthy answers on complicated topics. At the same time, re-using existing argumentation datasets is no longer sufficient, as they lack long, complex arguments and realistic evidence from potentially misleading sources, limiting holistic evaluation of retrieval effectiveness and argument quality.\"\nThen, I describe how I addressed the gaps in current research and give a detailed description of my methodology:\n\"To address these gaps, we investigate automated evaluation methods using multiple fine-grained LLM judges, providing better and more interpretable assessments than traditional single-score metrics and even previously reported human crowdsourcing. To validate the proposed techniques, we introduce ConQRet, a new benchmark featuring long and complex human-authored arguments on debated topics, grounded in real-world websites, allowing an exhaustive evaluation across retrieval effectiveness, argument quality, and groundedness. We validate our LLM Judges on a prior dataset and the new ConQRet benchmark.\"\nFinally, I describe the results and their implications for the research on this topic:\n\"Our proposed LLM Judges and the ConQRet benchmark can enable rapid progress in computational argumentation and can be naturally extended to other complex retrieval-augmented generation tasks.\"\nBased on the provided step-by-step instruction, write an abstract for an article in {category} titled \"{title}\"."
	News	"I want to write a news article about the following events:\nDarsh Patel, 22, was hiking with friends in the Apshawa Preserve in West Milford on Sunday when a bear started following them.\nThe group fled in different directions and when the four other hikers could not find Patel, they called police.\nPatel's body was found two hours later.\nThe 300-pound bear was circling the body and could not be scared away.\nIt was shot dead in accordance with Division of Fish and Wildlife guidelines.\nOn Saturday locals splitting wood filmed a bear rifling through their garbage.\n\n\nFirst, I introduce the event and its content to the readers:\nLocals in northern New Jersey believe they filmed a black bear hunting for food hours before a 22-year-old hiker was mauled to death in nearby woods at the weekend. Two men splitting wood on Saturday captured a video of a bear going through garbage just a few feet from where they were working, before scampering off into the woods, according to CNN. On Sunday, Darsh Patel, a senior majoring in information technology and informatics at Rutgers University, was found dead in Apshawa Preserve - about 45 miles northwest of New York City - with a 300-pound bear guarding his body. Officials say the attack was the first fatal bear-human encounter on record in New Jersey. Just a day after the footage was shot, a black bear mauled a 22-year-old student to death in the woods nearby.\nThen, I provide more details related to the event:\nPatel had been hiking with four friends in the 526-acre woods. The five friends noticed the bear beginning to follow them and ran, splitting up as they did. When they couldn't find Patel, they called police, who found his body about two hours later. The bear was about 30 yards from the body and circling, Department of Environmental Protection spokesman Larry Ragonese said, and wouldn't leave even after officers tried to scare it away by making loud noises and throwing sticks and stones. The male bear was killed with two rifle blasts and is being examined at a state lab for more clues as to why it may have pursued the group of five hikers.\nThen, I provide opinions on the event, quoting officials, experts or witnesses of the event:\nKelcey Burguess, principal biologist and leader of the state Division of Fish and Wildlife's black bear project, said the bear could have been predisposed to attack but more likely was looking for food. State and local officials stressed that bear attacks are rare even in a region of the state that may have as many as 2,400 bruins in its dense forests. \"This is a rare occurrence,\" West Milford police Chief Timothy Storbeck said, noting that his department receives six to 12 calls per week regarding bears, usually involving them breaking into trash cans. Locals: Residents in northern New Jersey often spot bears in and around their yards. There are as many 2,400 bruins in the area's dense forests, but until now had never been a fatal human-bear attack. Wildlife officials believe there is a current shortage of the acorns and berries that bears eat. The hikers had granola bars and water with them, Storbeck said. Officials don't believe the hikers provoked the bear but they may have showed their inexperience when they decided to run. The safest way to handle a bear encounter is to move slowly and not look the bear in the eye, DEP spokesman Larry Ragonese said. New Jersey Division of Fish and Wildlife guidelines direct law enforcement to euthanize \"Category I\" bears, which are deemed an \"immediate threat to human safety.\" NJ Advance Media reports that the New Jersey State Medical Examiner, the Fish and Wildlife Division of the state Department of Environmental Protection and the West Milford Police Department are looking into the circumstances of Patel's death.\nFinally, I conclude the report by highlighting the relevance of the event:\n\"Bear sightings are not unusual by any stretch in New Jersey,\" said Bob Considine, spokesperson for the Department of Environmental Protection. \"They have been seen in all 21 counties, although they're obviously most common in the northwest part of the state.\" Black bears rarely pose a threat to humans and often retreat when confronted. In 2006, a tabby cat scared a black bear up a tree in West Milford. The bear only climbed down and left after the cat's owner had called it back into the house.\n\nNow, I want to write an article on a different topic:\n{highlights}\nWrite the article, following the steps described above. Your answer should only include the article. The article should be around {length} characters long.",
	Reviews	"I want to write an Amazon review for an item called \"Haier RDG350AW 6.5 Cubic Foot Front Load Gas Dryer, White\". The rating will be 5.0.\nFirst, I choose a title for my review that describes my opinion and experience well. The title of my review will be:\n\"Very Affordable Dryer\".\nThen, I would state my initial experience with the product:\n\"I was a little worried about buying this cause it had some bad reviews, but it's a really great deal.\"\nThen I would describe the pros and cons of the product, expressing the reasons for my rating:\n\"It didn't touch my gas bill period. Yes larger loads take a while to dry, maybe up to 3 to 4 hours. It's just really energy efficent. Like I said my gas bill didn't budge with this being hooked up.\"\nNow, I want to write a review of an item called \"{item_name}\" and give it a rating of {rating}.\nFollow the directions described above to come up with the review. Your answer should only include the review and the title. The review should be one paragraph."
	QA	"If I was asked the question: \"When do the English state schools finish summer term and holiday begins?\",\nFirst, I would give a general overview of the answer:\"In the English school system, state schools run from early September to mid or late July of the following year.\"\nThen, I would go into more detail:\n\"The summer term (also known as the third term) runs from late April and finishes mid to late July with a week-long half term break in between. The summer holiday begins in late July and usually runs about six weeks long, ending in September.\"\nFinally, I would add nuance or additional context to my answer:\n\"The schools on the Trinity terms end their school year and begin summer holidays a few weeks earlier, at the end of June.\"\nFollowing the steps described above, answer the question: \"{question}\" in one paragraph."

Continued on next page

Prompt	Dataset	Prompt template
3-shot	Abstracts	"role" : "user", "content" : {title_1} "role" : "assistant", "content" : {human-written_abstract_1} "role" : "user", "content" : {title_2} "role" : "assistant", "content" : {human-written_abstract_2} "role" : "user", "content" : {title_3} "role" : "assistant", "content" : {human-written_abstract_3} "role" : "user", "content" : {title}
	News	"role" : "user", "content" : {highlights_1} "role" : "assistant", "content" : {human-written_article_1} "role" : "user", "content" : {highlights_2} "role" : "assistant", "content" : {human-written_article_2} "role" : "user", "content" : {highlights_3} "role" : "assistant", "content" : {human-written_article_3} "role" : "user", "content" : {highlights}
	Reviews	"role" : "user", "content" : {product_name_1} "role" : "assistant", "content" : {human-written_review_1} "role" : "user", "content" : {product_name_2} "role" : "assistant", "content" : {human-written_review_2} "role" : "user", "content" : {product_3} "role" : "assistant", "content" : {human-written_review_3} "role" : "user", "content" : {product}
	QA	"role" : "user", "content" : {question_1} "role" : "assistant", "content" : {human-written_answer_1} "role" : "user", "content" : {question_2} "role" : "assistant", "content" : {human-written_answer_2} "role" : "user", "content" : {question_3} "role" : "assistant", "content" : {human-written_answer_3} "role" : "user", "content" : {question}
Style	Abstracts	"As an academic paper writer, your task is to write an abstract of a research paper in a specific writing style. Write in the writing style of an example but ignore the content and topic of the example. You will be provided with the style example. You will be provided with the title for your abstract.\nStyle example: {example}\nTitle: {title}"
	News	"As a news article writer, your task is to write a news article in a specific writing style. Write in the writing style of an example but ignore the content and topic of the example. You will be provided with the style example. You will be provided with the summary of the topic for your article.\nStyle example: {example}\nSummary: {highlights}"
	Reviews	"As an Amazon review writer, your task is to write a review for an item in a specific writing style. Write in the writing style of an example but ignore the content and topic of the example. You will be provided with the style example. You will be provided with the name of the item you have to review.\nStyle example: {example}\nItem: {item_name}"
	QA	"As a highly intelligent question answering bot, your task is to answer questions in specific writing styles. Write in the writing style of an example but ignore the content and topic of the example. You will be provided with the style example. You will be provided with the question.\nStyle example: {example}\nQuestion: {question}"
Self-refine	Abstracts	1. "Write an abstract for an article in {category} with a title: \"{title}\". The abstract should be around {length} characters long." 2. "You will see an abstract for a scientific article. Your task is to provide feedback on how to make the text seem more human-like. Consider sentence length, sentence structure, vocabulary and readability.\nAbstract: {text}\nFeedback: " 3. "Based on the feedback, improve the text below:\nText: {text}\nFeedback: {feedback}." 4. "Which text sounds more human-written?\nText A: {text_a}\nText B: {text_b}\n\nPick your answer from [\"Text A\", \"Text B\", \"both\", \"neither\"]. Generate a short explanation for your choice first. Then, generate \"Text A seems more human-written\" or \"Text B seems more human-written\" or \"Both texts seem human-written\" or \"Neither of the texts sounds human-written\"
	News	1. "Write a news article based on the following highlights:\n\"{highlights}\"\nYour article should be around {length} characters long." 2. "You will see a news article. Your task is to provide feedback on how to make the text seem more human-like. Consider sentence length, sentence structure, vocabulary and readability.\nArticle: {text}\nFeedback: " 3. "Based on the feedback, improve the text below:\nText: {text}\nFeedback: {feedback}." 4. "Which text sounds more human-written?\nText A: {text_a}\nText B: {text_b}\n\nPick your answer from [\"Text A\", \"Text B\", \"both\", \"neither\"]. Generate a short explanation for your choice first. Then, generate \"Text A seems more human-written\" or \"Text B seems more human-written\" or \"Both texts seem human-written\" or \"Neither of the texts sounds human-written\"

Continued on next page

Prompt	Dataset	Prompt template
	Reviews	1. "Write an Amazon review for the item \"{item_name}\" with a title \"{title}\" and a rating of {rating}. The review should be around {length} characters long." 2. "You will see an Amazon review. Your task is to provide feedback on how to make the text seem more human-like. Consider sentence length, sentence structure, vocabulary and readability.\nReview: {text}\nFeedback: " 3. "Based on the feedback, improve the text below:\nText: {text}\nFeedback: {feedback}." 4. "Which text sounds more human-written?\nText A: {text_a}\nText B: {text_b}\n\nPick your answer from [\"Text A\", \"Text B\", \"both\", \"neither\"]. Generate a short explanation for your choice first. Then, generate \"Text A seems more human-written\" or \"Text B seems more human-written\" or \"Both texts seem human-written\" or \"Neither of the texts sounds human-written\"
	QA	1. "{question}\n Your answer should be around {length} characters long." 2. "You will see an answer to a question. Your task is to provide feedback on how to make the text seem more human-like. Consider sentence length, sentence structure, vocabulary and readability.\nAnswer: {text}\nFeedback: " 3. "Based on the feedback, improve the text below:\nText: {text}\nFeedback: {feedback}." 4. "Which text sounds more human-written?\nText A: {text_a}\nText B: {text_b}\n\nPick your answer from [\"Text A\", \"Text B\", \"both\", \"neither\"]. Generate a short explanation for your choice first. Then, generate \"Text A seems more human-written\" or \"Text B seems more human-written\" or \"Both texts seem human-written\" or \"Neither of the texts sounds human-written\"

Table 3: Prompt details.

A.3 Detailed classification results

The results of the detailed in-domain accuracy are shown in Table 4.

A.4 Linguistic analysis

A.4.1 Methods used for linguistic analysis

Lexical diversity: We choose Moving Average Type-Token Ratio to measure the lexical diversity of the texts, as it is independent of the text length (Covington and McFall, 2010). Additionally, we compare the texts in terms of the **number of unique words**, as this feature has been shown to be relevant in the previous research (Opara, 2024; Yildiz Durak et al., 2025).

Lexical density: Lexical density is the percentage of content words in the text. Machine-generated text is said to achieve higher lexical density (Savoy, 2020) than human-written text, which means that it contains fewer function words. The content words are adjectives, adverbs, nouns, and verbs.

Sentiment: AI-generated text has been shown to differ from human-written text in terms of sentiment: some authors have written of *positivity bias* present in AI-generated texts (Markowitz et al., 2024; Muñoz-Ortiz et al., 2023; Margolina and Kolmogorova, 2023), which means that text generated by large language models tends to contain more positive emotions compared to human-written texts. Other research suggests that human-written texts are more varied in terms of the richness of emotional content (Zanotto and Aroyehun, 2024). We use the TextBlob library to calculate the **polarity** and **subjectivity** scores for the texts in our dataset. **Polarity** is a score between -1 and 1, where -1 denotes a negative sentiment, while 1 denotes a positive sentiment. **Subjectivity** relates to the amount of personal opinion included in the text.

Readability: Readability refers to the ease of understanding a text. AI-generated text tends to be less readable than human-written text (Yadagiri et al., 2024; Markowitz et al., 2024; Mathews et al., 2024). We use the **Gunning fog index** and the **Flesch reading ease test** as measures of readability. The **Gunning fog index** is an estimated number of years needed to understand a given passage. The **Flesch reading ease test** is a metric of readability, in which texts that are easier to read receive a higher score. In the analysis of readability, we also include the **text length** in characters, as well as sentence length statistics. Machine-generated text tends to be less varied than human-written text in terms of sentence length (Desaire et al., 2023; Muñoz-Ortiz et al., 2023). Following (Desaire et al., 2023), we calculate the **average sentence length**, the **standard deviation from the average sentence length**, and the **number of very long** (35 words or more) and **very short** (10 words or less) **sentences** in each text.

Part-of-Speech (POS): We use a selection of metrics from StyloMetrix (Okulska et al., 2023) to compare the frequency of verbs, nouns, adjectives, pronouns, determiners, conjunctions and numerals across the different dataset types, models and prompts. The frequency of parts of speech is measured as the fraction of text covered by tokens representing a given part of speech. The frequency of POS has been shown to be different across different text genres, with non-fiction texts typically achieving a higher frequency of nouns than fiction texts (Mendhakar and S, 2023). A high frequency of verbs can be associated with more narrative texts, while a high frequency of adjectives is common for more descriptive texts. Additionally, previous research has discovered differences between human-written and AI-generated text in terms of frequency of certain POS (Georgiou, 2024). Therefore, POS analysis is relevant for gaining insights into the literary style of texts in our dataset.

Grammatical analysis: We use StyloMetrix (Okulska et al., 2023) metrics to compare texts in terms of grammatical categories related to verbs. In previous research, human-written texts have been shown to contain more passive voice than AI-generated texts (Georgiou, 2024). We measure the **incidence of passive and active voice** as the frequency of verbs in passive or active voice. We also compare the differences between the choice of tenses. The **frequency of past, present and future tenses** is measured as the fraction of the text covered by verbs in past, present and future tenses. For example, the incidence of past tenses is the number of verbs in past simple, past continuous, past perfect or past perfect continuous divided by the total number

Detector	Model	Dataset	Prompt type					
			0-Shot	3-Shot CoT	1-Shot CoT	Style	3-Shot	Self-refine
DeBERTa	Llama3.3 70b	Abstracts	0.988	0.9875	0.993	0.9795	0.9945	0.9865
		News	0.999	0.999	0.9985	0.9775	0.7935	1.0
		Reviews	0.984	0.988	0.9995	0.9615	0.9805	0.9875
		QA	0.8866	0.8966	0.9989	0.9604	0.8233	0.9509
	Qwen 14b	Abstracts	0.998	0.991	0.994	0.998	0.999	0.999
		News	1.0	0.9995	0.9995	1.0	1.0	0.999
		Reviews	0.9985	0.996	1.0	1.0	0.9965	0.9985
		QA	0.98	0.9662	0.9916	0.9926	0.9668	0.9736
	Qwen 32b	Abstracts	0.9955	0.9915	0.994	0.9965	1.0	0.999
		News	1.0	0.999	0.9995	1.0	1.0	1.0
		Reviews	0.999	0.9955	1.0	0.998	0.9985	0.9955
		QA	0.9763	0.981	0.9852	0.9942	0.9626	0.9736
	Qwen 72b	Abstracts	0.988	0.9845	0.9845	0.991	0.9955	0.991
		News	0.999	0.9995	0.9995	1.0	0.998	0.999
		Reviews	0.995	0.9975	0.9995	0.9945	0.9965	0.989
		QA	0.9699	0.9583	0.9889	0.942	0.8903	0.9662
	Solar 22b	Abstracts	0.9855	0.9915	0.992	0.989	0.9985	0.9975
		News	1.0	0.9995	0.999	1.0	0.9994	0.9995
		Reviews	0.998	0.998	0.9995	0.9975	0.9965	0.999
		QA	0.9626	0.9852	0.9947	0.9773	0.9705	0.9926
	Mistral 123b	Abstracts	0.986	0.989	0.9895	0.984	0.9895	0.9965
		News	0.9985	0.998	0.9995	0.9975	0.962	1.0
		Reviews	0.9905	0.9895	0.9945	0.993	0.9765	0.994
		QA	0.9014	0.8333	0.9299	0.8819	0.9167	0.9905
	Deepseek 70b	Abstracts	0.9885	0.989	0.992	0.9825	0.9965	0.9885
		News	0.998	0.9985	0.998	0.9985	0.999	1.0
		Reviews	0.9945	0.994	0.9945	0.99	0.9955	0.9955
		QA	0.9341	0.9019	0.9483	0.8903	0.9151	0.954
RoBERTa	Llama3.3 70b	Abstracts	0.9865	0.9765	0.9925	0.9295	0.9795	0.9835
		News	0.99	0.996	0.988	0.901	0.782	0.9965
		Reviews	0.987	0.9925	0.9995	0.938	0.99	0.9915
		QA	0.8481	0.8586	1.0	0.9578	0.8291	0.9451
	Qwen 14b	Abstracts	0.9945	0.993	0.995	0.998	0.999	0.9965
		News	0.999	1.0	0.9995	1.0	0.9985	1.0
		Reviews	0.9995	0.999	0.9995	1.0	0.998	1.0
		QA	0.9826	0.9652	0.9784	0.9852	0.9541	0.9662
	Qwen 32b	Abstracts	0.993	0.989	0.9935	0.993	0.999	0.9985
		News	1.0	1.0	1.0	0.999	0.9985	0.9975
		Reviews	0.998	0.9995	0.9965	0.9985	0.9975	0.9955
		QA	0.9789	0.9789	0.981	0.9873	0.9209	0.9789
	Qwen 72b	Abstracts	0.986	0.9825	0.9875	0.979	0.9945	0.982
		News	0.998	1.0	0.999	0.9995	0.999	0.999
		Reviews	0.996	0.9955	1.0	0.9945	0.998	0.9915
		QA	0.981	0.9399	0.9831	0.9504	0.8623	0.9699
	Solar 22b	Abstracts	0.9825	0.987	0.9795	0.9825	0.995	0.9995
		News	0.9975	0.995	0.994	0.9995	1.0	0.998
		Reviews	0.9975	0.9975	1.0	0.997	0.9935	1.0
		QA	0.9742	0.9831	0.9963	0.9821	0.9747	0.9937
	Mistral 123b	Abstracts	0.9835	0.982	0.982	0.9665	0.987	0.9925
		News	0.993	0.996	0.998	0.992	0.9825	0.999
		Reviews	0.9955	0.981	0.9915	0.991	0.9805	0.9935
		QA	0.8877	0.8318	0.9383	0.9024	0.8866	0.9826
	Deepseek 70b	Abstracts	0.9835	0.986	0.988	0.9655	0.996	0.9825
		News	0.997	0.986	0.9995	0.998	0.9985	0.998
		Reviews	0.994	0.995	0.9925	0.9925	0.9965	0.9965
		QA	0.8813	0.8734	0.9314	0.8645	0.8513	0.9185

Table 4: Accuracy of the fine-tuned detectors on in-domain data.

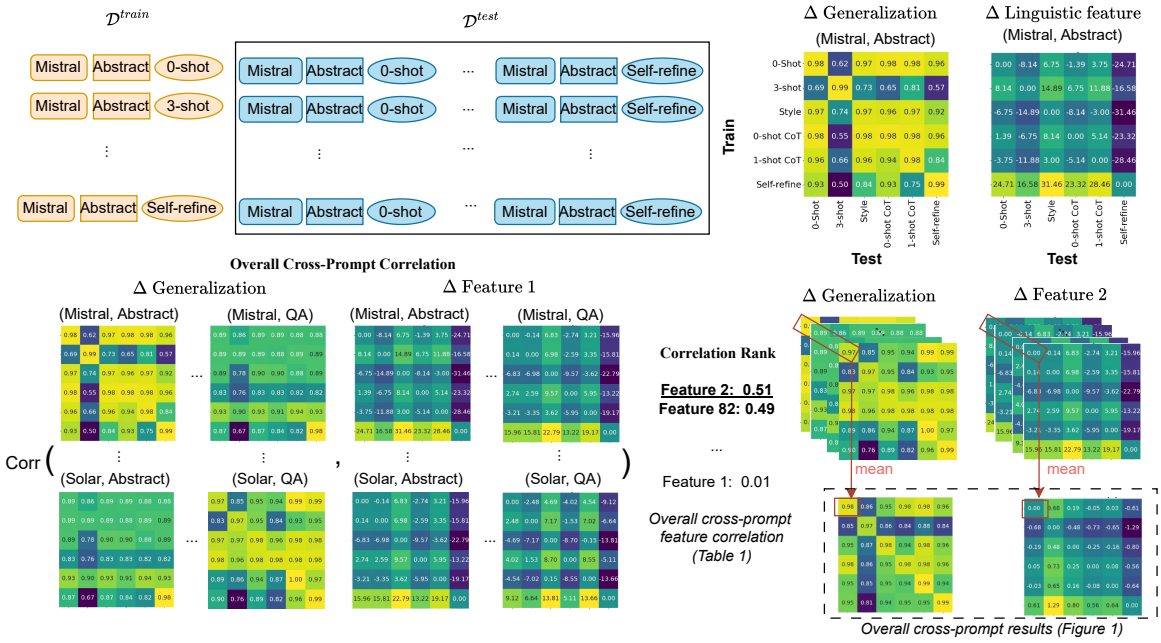


Figure 3: The workflow to get the overall cross-prompt generalization and feature shift results.

of words in the text.

Lexical analysis: As part of lexical analysis, we measure the **frequency of proper and personal names**, as well as the **frequency of adjectives and adverbs in positive, comparative and superlative degrees** and the **frequency of nouns in possessive case**. We use the StyloMetrix (Okulska et al., 2023) pronoun-related metrics to analyze the differences in the usage of pronouns in human-written and AI-generated text, as well as across the different prompt types, dataset types and models. We calculate not only the frequency of pronouns (POS_PRO), but also the frequency of specific personal or reflexive pronouns and general frequency of certain types of pronouns (for example, the frequency of all first person singular pronouns).

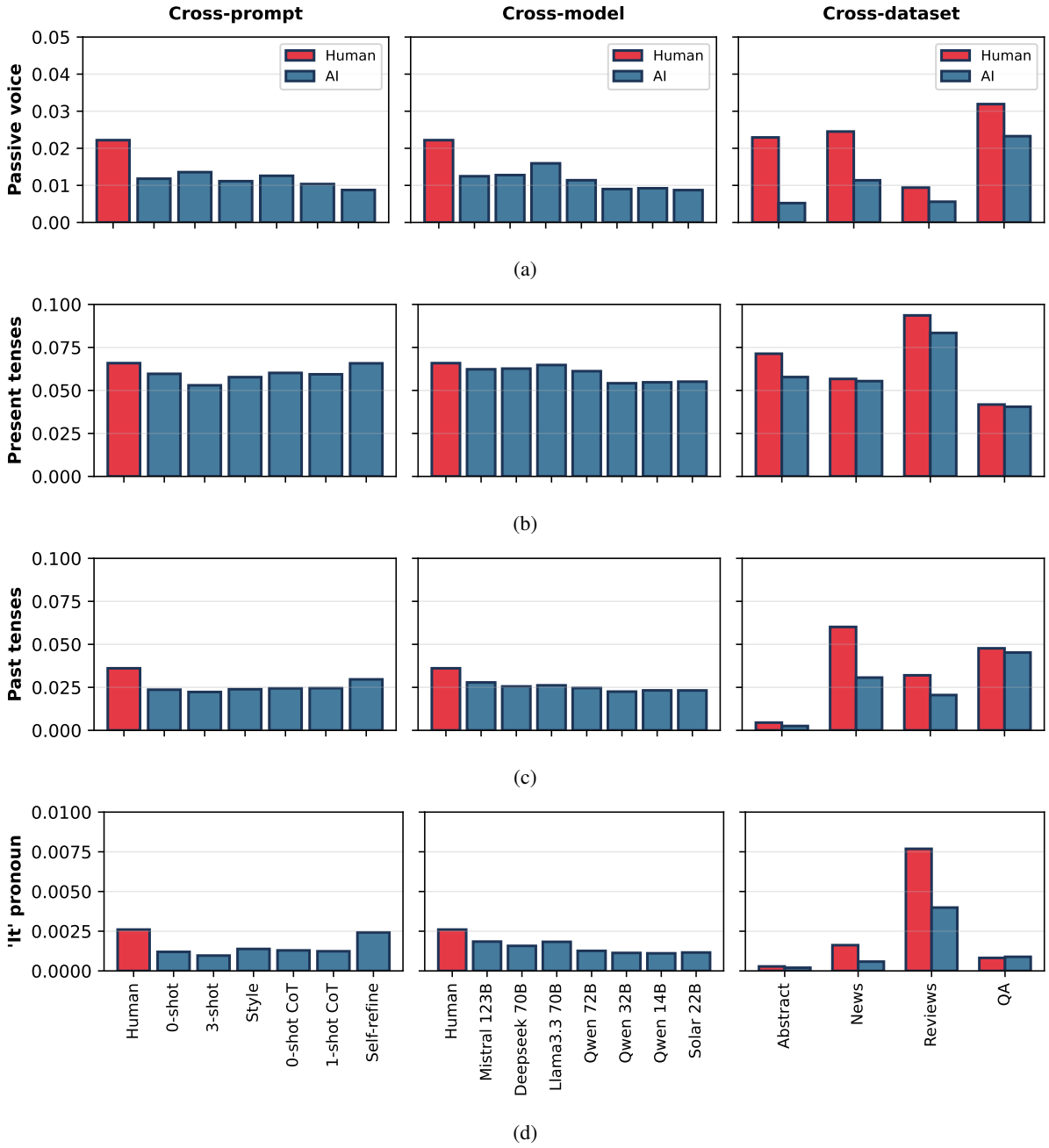


Figure 4: The more detailed comparison of different linguistic features across different configurations as well as the differences between human and AI text. We present the features that have the strongest correlations for different dimensions (underlined features in Table 4 in the main paper).

Linsuitic Feature	Feature Metric	DeBERTa-based C-P	DeBERTa-based C-M	DeBERTa-based C-D	RoBERTa-based C-P	RoBERTa-based C-M	RoBERTa-based C-D
Lexical diversity	MATTR	0.035	**0.312**	0.097	0.031	**0.326**	0.100
	L MATTR	0.038	**0.313**	0.087	0.021	**0.349**	0.079
	Unique words	0.004	0.118	0.116	0.044	**0.163**	0.074
Lexical density	Number of function words	0.008	**0.148**	0.104	0.042	**0.282**	0.184
Readability	FLESCH	0.006	**0.254**	**0.198**	0.006	**0.330**	**0.251**
	Sentence length	**0.069**	**0.242**	0.093	**0.073**	**0.288**	0.122
	Long sentences	0.020	**0.161**	0.099	0.061	**0.196**	0.066
	Short sentences	**0.072**	**0.223**	0.040	0.041	**0.290**	0.043
	Sentence length std	0.034	0.104	0.004	0.058	**0.210**	0.064
	Length in characters	0.002	0.117	0.083	0.045	**0.191**	0.012
Sentiment	Polarity	0.025	0.125	**0.224**	**0.106**	**0.261**	**0.187**
	Subjectivity	0.019	0.064	0.001	0.058	0.134	0.054
POS	Verbs	0.050	0.080	0.005	**0.072**	0.042	0.041
	Nouns	0.047	**0.307**	0.164	**0.066**	**0.322**	0.180
	Adjectives	0.017	**0.147**	0.038	0.037	0.014	0.092
	Adverbs	0.005	**0.201**	**0.217**	0.016	**0.213**	**0.195**
	Determiners	0.010	**0.215**	**0.218**	0.044	**0.286**	**0.241**
	Interjections	0.031	0.111	0.157	0.032	**0.169**	0.096
	Conjunctions	0.042	0.061	0.158	0.020	**0.190**	0.098
	Particles	0.027	0.012	0.026	0.006	**0.212**	0.047
	Numerals	0.008	**0.270**	0.163	0.003	**0.265**	**0.202**
	Pronouns	0.013	0.064	0.057	0.051	**0.159**	0.007
Lexical	Content words	0.023	0.128	0.064	0.039	**0.278**	0.148
	Function words	0.005	0.045	0.086	0.035	**0.251**	0.164
	Content words types	0.041	0.100	**0.195**	0.001	**0.239**	**0.238**
	Function words types	0.012	0.037	0.137	0.048	0.006	**0.194**
	Proper names	**0.070**	**0.360**	0.013	0.022	0.134	0.053
	Nouns in possessive case	0.049	0.062	0.102	0.002	0.081	0.167
	Adjectives in positive degree	0.012	**0.147**	0.039	0.043	0.018	0.097
	Adverbs in positive degree	0.003	**0.188**	**0.191**	0.025	**0.188**	0.184
	Adverbs in comparative degree	0.006	**0.191**	**0.225**	0.016	**0.206**	**0.198**
	Adverbs in superlative degree	0	**0.191**	**0.226**	0.022	**0.204**	**0.197**
	'I' pronoun	0	**0.202**	0.134	0.033	**0.327**	0.085
	'He' pronoun	0.045	**0.173**	0.045	0.031	**0.191**	0.059
	'She' pronoun	**0.086**	**0.176**	0.004	**0.080**	**0.316**	0.020
	'It' pronoun	**0.067**	0.092	0.167	**0.067**	**0.203**	**0.199**
	'You' pronoun	0.015	**0.160**	0.085	0.011	0.113	0.028
	'They' pronoun	0.015	0.072	0.101	0.018	0.128	0.128
	'Me' pronoun	0.044	0.096	0.041	0.025	0.077	0.032
	'You' object pronoun	0.016	0.077	0.153	0.050	**0.161**	0.100
	'Him' object pronoun	0.041	**0.161**	0.069	0.042	0.034	0.097
	'Her' object pronoun	0.040	0.080	0.072	**0.072**	**0.203**	0.031
	'Us' pronoun	0.041	**0.215**	0.101	0.012	0.086	0.137
	'Them' pronoun	0.052	0.030	0.073	**0.080**	0.080	0.012
	'My' pronoun	0.021	**0.214**	0.127	0.053	**0.341**	0.109
	'Your' pronoun	0.007	**0.250**	0.085	0.009	**0.254**	0.096
	'His' pronoun	0.005	0.078	0.086	0.023	0.059	0.152
	'Her' possessive pronoun	0.034	0.004	0.053	**0.063**	0.102	0.129
	'Its' possessive pronoun	**0.089**	**0.293**	**0.245**	**0.079**	**0.323**	**0.244**
	'Their' possessive pronoun	**0.065**	**0.237**	0.121	0.015	**0.228**	0.132
	'Yours' pronoun	0.026	0.001	**0.221**	0.009	0.103	0.163
	'Theirs' pronoun	0.004	0.011	**0.255**	0.008	0.084	**0.237**
	'Hers' pronoun	0.061	0.006	0.067	0.031	0.130	0.021
	'Ours' possessive pronoun	0.003	0.106	0.021	0.010	**0.186**	0.005
	'Myself' pronoun	0.023	**0.157**	0.088	0.017	**0.297**	0.073
	'Himself' pronoun	0.012	**0.332**	0.052	0.027	**0.232**	0.043
	'Herself' pronoun	0.036	**0.248**	0.025	0.032	**0.227**	0.001
	'Itself' pronoun	0.032	0.132	**0.221**	0.055	0.027	**0.204**
	'Ourselves' pronoun	0	0.103	0.073	0.018	**0.269**	0.068
	'Yourselves' pronoun	0.018	**0.177**	0.108	0.005	**0.194**	0.083
	'Themselves' pronoun	**0.064**	**0.230**	0.019	**0.100**	**0.278**	0.015
	First person singular pronouns	0	**0.202**	0.134	0.033	**0.327**	0.085
	Second person pronouns	0.014	**0.204**	0.056	0.011	**0.192**	0.005
	Third person singular pronouns	0.047	0.014	0.093	0.031	**0.198**	0.140
	Third person plural pronouns	0.050	**0.182**	0.117	0.039	0.129	0.157
General	Incidence of verbs in infinitive	0.037	0.050	0.020	**0.095**	**0.153**	0.085

Table 5: The correlation between generalization performance with different linguistic features. The **significant correlation is bolded**.

	DeBERTa-based					RoBERTa-based				
	Abstracts	News	Reviews	QA	ALL	Abstracts	News	Reviews	QA	ALL
0-shot	0.587	0.569	0.663	0.591	0.416	0.399	0.677	0.655	0.355	0.385

Table 6: The cross-model correlation between generation performance and the most correlated linguis- tic feature when evaluated on different datasets.

	DeBERTa-based 0-shot	RoBERTa-based 0-shot
Mistral 123B	0.310	0.350
Deepseek 70B	0.418	0.468
Llama 70B	0.389	0.374
Qwen 72B	0.425	0.308
Qwen 32B	0.426	0.418
Qwen 14B	0.487	0.545
Solar 22B	0.269	0.188
ALL	0.296	0.287

Table 7: The cross-dataset correlation between generation performance and the most correlated linguis- tic feature when evaluated on different models.