# GhostWriter: Hidden AI-Generated Texts Over Multiple Languages, Domains and Generators

**Manuel Schaaf, Kevin Bönisch, Alexander Mehler**
Text Technology Lab
Goethe University, Frankfurt
{manuel-schaaf,k.boenisch}@outlook.de
mehler@em.uni-frankfurt.de

**Abstract**

The advent of Transformer-based Large Language Models (LLMs) has led to an unprecedented surge of AI-generated text (AIGT) across online platforms and academic domains. While these models exhibit near-human fluency and stylistic coherence, their widespread adoption has raised concerns about authorship integrity, research quality, and the recursive contamination of training corpora with synthetic data. These developments underscore the need for reliable AIGT detection methods and benchmark datasets, particularly for malicious or deceptive *ghostwriting* scenarios where AIGT is intentionally crafted to evade detection. To address this, we present GhostWriter, a large-scale, bilingual (German and English), multi-generator, and multi-domain dataset for AIGT detection. The dataset comprises human- and AI-authored texts produced under domain-specific *ghostwriting* conditions, including examples intentionally embedded within otherwise human-written texts to obscure their AI origin. With GhostWriter, we (i) aim to expand the resources available for German AIGT datasets, (ii) emphasize mixed or fused synthesizations—since most existing corpora are limited to the document level—and (iii) introduce specifically crafted malicious ghostwriting scenarios across multiple domains and generators.

**Keywords:** Corpus; Natural Language Generation; Validation of LRs; AI-generated Text Detection

## 1. Introduction

The advent of the Transformer architecture (Vaswani et al., 2017) has enabled the wide adaptation of Large Language Models (LLMs) across diverse domains (Veselovsky et al., 2023; Murakami et al., 2023; Jiang et al., 2024). These models generate text with remarkable fluency, contextual relevance, and stylistic consistency (OpenAI et al., 2024; Team et al., 2024a), often reaching or surpassing human quality (Gómez-Rodríguez and Williams, 2023). As a result, the amount of AI-Generated Text (AIGT) online has surged, with a 57,3 % increase on mainstream platforms and a 474 % increase on misinformation sites (Hanley and Durumeric, 2023), accompanied by rising plagiarism and text reuse rates (Bisi et al., 2023; Elali and Rachid, 2023; Pudasaini et al., 2024). AI systems have also become prevalent in academia, supporting idea generation, writing, programming, and experimentation (Khalifa and Albadawy, 2024; Siam et al., 2024), which is accompanied with a notable increase in research paper submissions, acceptance rates, and publication volume (Figure 1) (Maslej et al., 2024, 31). An exacerbated example of this adoption is the fully automated "AI Scientist" (Lu et al., 2024), which employs AI agents to autonomously generate research ideas, run experiments, visualize results, and author full papers—including simulated peer reviews—at costs below $15 per paper. Its outputs reportedly meet or surpass acceptance thresholds at major ML conferences.

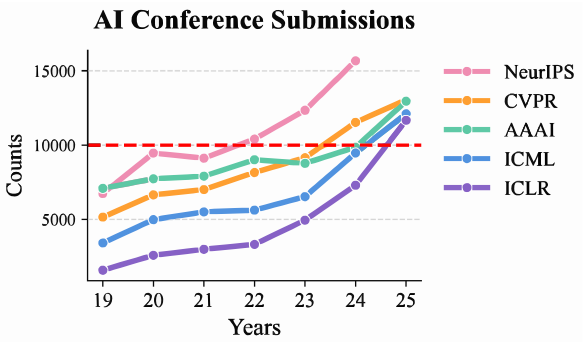



Figure 1: The rise of paper submissions at AI conferences (Kim et al., 2025).

This rise in AI-assisted writing raises questions about whether this quantitative growth is met with a corresponding quality. Herein, Naddaf (2025) instead links the increasing use of AI to an "explosion" of low-quality biomedical papers, drawing on the findings of Suchak et al. (2025), who identified a surge of formulaic NHANES-based studies—from four per year before 2022 to 190 in the first ten months of 2024. Many of these papers rely on simplistic analyses, overlook confounding factors, and selectively use data, suggesting that AI-assisted workflows facilitate high-volume but low-quality research output.

Additionally, beyond academia, concerns have emerged regarding the long-term effects of AIGT on the broader digital ecosystem (Shen and Zhang,



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10497–10516
11–16 May 2026. 

2024; Wang and Lu, 2025) and society as a whole (cf. INTERPOL, 2023). For instance, recursive training on synthetic data—where newer models are trained on text generated by older generations—can lead to *model collapse* (Shumailov et al., 2024), a phenomenon characterized by reduced vocabulary diversity and deteriorating syntactic and semantic coherence. This occurs as models overfit to synthetic patterns, progressively losing their ability to generalize. To counter these effects, AIGT detection has become a crucial research area focused on distinguishing human- from machine-generated text. Efforts mainly target (1) the creation of benchmark datasets for training and evaluation (Yu et al., 2025; Su et al., 2024; Li et al., 2024), and (2) the development of robust classifiers (Wang et al., 2023; Verma et al., 2024).

In this paper, we specifically contribute to point (1) by introducing GhostWriter: a new large-scale, bilingual, multi-generator, and multi-domain AIGT detection dataset that provides AI-generated text designed to hide its origin at both the *chunk* and *full-text* levels, generated under domain-specific *ghostwriting* scenario.

## 2. Related Work

Several datasets have been compiled for AIGT detection, varying along dimensions such as languages, domains, generation prompts and decoding strategies, access conditions, and detection granularity (document-, paragraph-, or sentence-level). Early efforts such as the *HC3* corpus (Guo et al., 2023) introduced bilingual (English and Chinese) question–answer pairs that contrast human and ChatGPT (*GPT-3.5*) responses across multiple domains, including Reddit, WikiQA, medicine, and finance. Its successor, *HC3 Plus* (Su et al., 2024), extends this setting to semantic-invariant tasks such as summarization, paraphrasing, and translation, revealing that such tasks pose greater difficulty for detectors trained on simpler QA data.

The *CHEAT* dataset (Yu et al., 2025) focuses on academic integrity by targeting scientific abstracts labeled as *fully generated*, *polished*, or *mixed*. Similarly, *ArguGPT* (Liu et al., 2023) focuses on argumentative writing, collecting essays from GPT-family models (GPT-2, GPT-3, ChatGPT) alongside human-written student and TOEFL essays. *OpenLLMText* (Chen et al., 2023) expands the scope of previous datasets through large-scale coverage of multiple generators (GPT-2, GPT-3.5, PaLM, LLaMA), supporting both binary detection and model attribution tasks based on diverse prompting strategies.

Robustness-oriented datasets have aimed to mirror real-world, "in the wild" conditions. Herein, *DeepfakeTextDetect* (Pu et al., 2023) introduces human and machine-written texts from commercial writing tools (AI-Writer, ArticleForge, Kafkai) and applies adversarial perturbations such as synonym substitution and paraphrasing to test detector resilience. The large-scale *MAGE* dataset (Li et al., 2024) expands this idea by covering 27 language models and seven writing tasks—including news, storytelling, QA, and scientific writing—generated under varying prompt types and distributional "wildness" levels, with explicit paraphrase and out-of-distribution scenarios. Complementary to these, the *Ghostbuster* datasets (Verma et al., 2024) provide synthetic and human texts across creative writing, news, and student essays, supporting multi-granular evaluation at paragraph and full-text levels.

Detection at finer granularity has been explored through sentence-level corpora such as *SeqXGPT* (Wang et al., 2023), built upon *SnifferBench* (Li et al., 2023). These datasets intertwine human-written and AI-generated sentences across diverse domains such as news, social media, scientific text, and technical documentation, enabling localized detection of mixed authorship. On the multilingual front, *MULTITuDE* (Macko et al., 2023) offers over 74 000 samples in eleven languages (Arabic, Catalan, Czech, German, English, Spanish, Dutch, Portuguese, Russian, Ukrainian, Chinese) generated by LLMs including GPT-3/4, LLaMA, Alpaca-LoRA, and Vicuna. Similarly, the *M4* corpus (Wang et al., 2024) introduces multi-generator, multi-domain, and multi-lingual coverage across seven languages—among them Arabic, Chinese, English, Indonesian, Russian, and Urdu—spanning Wikipedia, Reddit, news, and academic domains, with outputs from GPT-4, ChatGPT, Cohere, Dolly-v2, and BLOOMz.

More recently, large-scale benchmarks have emerged to unify these perspectives. *RAID* (Dugan et al., 2024, *Robust AI Detection*) currently represents the most extensive and challenging AIGT corpus, containing over six million generated texts from eleven LLMs across eight domains (e.g., news, Wikipedia, poetry, Reddit, and reviews) under multiple decoding strategies and adversarial perturbations. It further includes an extension, *RAID-extra*, with 2.3 million additional samples in Czech and German news, and maintains a public leaderboard[1] for detector evaluation. Additionally, *MultiSocial* (Macko et al., 2025) specifically addresses the social-media domain by assembling a multilingual benchmark across five platforms, with 472 097 short-form posts (roughly 58k human-written and balanced among seven multilingual LLMs), enabling cross-lingual and cross-platform evaluation in informal, noisy settings. Finally, *MAiDE-up* (Ignat et al., 2025) focuses on deceptive reviews: it consists of 10 000 real and 10 000 AI-generated fake hotel reviews spread across ten languages, pro-

[1] https://raid-bench.xyz/leaderboard

viding a multilingual testbed for deception-oriented detection in a review domain and exploring how sentiment, location, and language affect detectability.

## 3. GhostWriter

While the datasets outlined in Section 2 provide a solid foundation for research on the automatic detection of AIGT, we identify several limitations that motivated the creation of a new dataset. We structure these limitations into three key weaknesses, which we discuss in the following.

**First**, there is a noticeable imbalance in **linguistic coverage**. Although some multilingual datasets exist (e.g., *MULTITuDE*, *RAID_extra* and *MAiDE-up*), the representation of German texts remains extremely sparse in comparison to other languages, particularly English. This limits the ability to study cross-lingual generalization of models and their robustness across languages, especially in domains where German plays a central role (e.g., journalism, politics, and legal proceedings).

**Second**, many existing datasets rely heavily on synthetic texts generated by what are now considered **outdated language models**—in the sense that they no longer represent the state of the art—such as GPT-2, GPT-3.5, or early releases of LLaMA (Touvron et al., 2023). While these models were relevant in the early stages of AIGT research, it is questionable whether their outputs remain representative of the capabilities of contemporary LLMs. Consequently, AIGT detectors must be evaluated against state-of-the-art models to ensure reliability, which requires continuous alignment with current and widely deployed LLMs.

**Third**, prior work has predominantly emphasized document-level detection, where the task is to classify an entire text as either human- or AI-generated. This overlooks a highly relevant scenario in real-world contexts: **mixed texts**, where AI-generated passages are interspersed within otherwise human-authored documents, as outlined by Yu et al. (2025). Such cases—for instance, when LLMs are used for ghostwriting or partial editing—represent a distinct detection challenge at the *chunk* level, which remains significantly underexplored in existing benchmarks and datasets.

To address these gaps, we introduce GhostWriter: a new large-scale, multilingual, and domain-diverse benchmark dataset for AIGT, that is freely available on HuggingFace.[2] This dataset is designed to provide (i) broader language coverage, with substantial inclusion of German alongside English; (ii) synthetic texts generated by state of the art LLMs; and (iii) both document-level and chunk-level generations to support and encourage research on fusion text detection. Below, the composition of the dataset and the generation protocols are described in detail.

### 3.1. Human-Authored Texts

The human-written texts are the foundation of the dataset, as all synthetic texts are generated using them as a source. The texts were gathered from a wide range of domains, ensuring the creation of a diverse combination of domains, styles and languages:

(A) **News**: 18 773 English news articles from the *CNN-DailyMail* dataset (Hermann et al., 2015, *Apache-2.0*).[3]

(B) **Scientific writing**: 8 448 English research papers scraped from *arXiv* (`CC0 Public Domain`), where the full text from the PDFs was extracted using `PyMuPDF`.

(C) **Web blogs**: 20 299 English blog posts collected from *blogger.com* (Schler et al., 2006, `CC0 Public Domain`).

(D) **Political speeches**: 18 343 German parliamentary debates provided by the *Bundestag-Mine* (Bönisch et al., 2023, `DL-DE->BY-2.0`), and 13 896 English speeches sourced from the *House of Commons* archive (Blumenau, 2021, *Open Parliament Licence*).

(E) **Legal documents**: 9 148 English court cases from the *European Court of Human Rights* (Chalkidis et al. (2021), `CC BY-NC-SA 4.0`).

(F) **Student essays**: 42 311 English essays written by 6th–12th grade students, drawn from a Kaggle competition dataset (King et al., 2023, building on Crossley et al., 2024, `CC BY-NC 4.0`).

(G) **Literary texts**: 7 224 English and German classic works from *Project Gutenberg* (Gutenberg, n.d., `Public Domain`).

To ensure that none of the human-authored texts were potentially AI-generated, all sources are limited to works created before the end of 2020, prior to the widespread adoption of LLMs.

### 3.2. Synthetization

**AI Generators** To generate synthetic counterparts for each human-authored source text, we employ a diverse range of LLMs with an emphasis on models that are widely available and actively used.

[2] https://huggingface.co/datasets/TheItCrOw/GhostWriter

[3] An additional set of German news articles in GhostWriter had to be redacted due to licensing constraints.

Most of the AI-text generation space is dominated by models from a small group of large companies, such as OpenAI, Google, and Microsoft. Consequently, we use Google's **Gemma 2** (Team et al., 2024b, 9B), Microsoft's **Phi-3** (Abdin et al., 2024, 3.8B), **DeepSeek-R1** (DeepSeek-AI et al., 2025, 1.5B and 32B), NVIDIA's **Nemotron** (Nvidia et al., 2024, 70B), as well as OpenAI's **GPT-4-Turbo**, **GPT-4o-Mini** (OpenAI, 2024), and their newest reasoning model **o3-Mini** (OpenAI, 2025). Our selection directly addresses the second of the limitations identified above, as it ensures the adequate representation of both proprietary, closed-source models with commercial applications and open-access models across a range of sizes and capabilities, while reflecting the current state-of-the-art. The creation of the dataset required approximately six days using four NVIDIA L40S GPUs with 48 GB of VRAM each. The cost of generating the synthetic texts via OpenAI amounted to approximately $138.

***Fulltext*-based Synthetization** Initially, we prompt an AI generator with a *ghostwriting* task for given human-authored documents. All documents are stored in a database in a common format, where a foreign key retains the relationship between source documents and their AI-written counterparts. Generating synthetic texts in the ghostwriting setting is carried out in two stages (cf. Figure 2):

1. First, the LLM is given a **Information Extraction Prompt** (top), wherein it is tasked to extract textual characteristics from the human-authored source text, including its language, topic, and linguistic style. This is achieved via *few-shot prompting* (Brown et al., 2020), where the prompt is enriched with examples for the LLM to follow.
2. Then, these extracted characteristics are included in a **Ghostwriting Prompt** (bottom) with which the LLM is tasked (in a new context) to write a text as a ghostwriter, based on the previously derived parameters. In this step, the word count of the source text is provided in order to approximate its length without modifying the `max_tokens` parameter of the LLM.

This process ensures that (i) the produced AI texts are contextualized with respect to the original human input in both topic and style, and (ii) the ghostwriting scenario is realistically represented, where the AI is instructed to write in the style of a human-authored text while avoiding detection, thereby making the dataset more challenging for detectors.

***Chunk*-based Synthetization** To address the third identified limitation, we crate partially AI-generated *chunk-based* or *fusion* texts. Herein, human-written documents are divided into contiguous segments—preferably entire sentences or, where available, paragraphs—, up to 50% of which are randomly masked and replaced with AI-generated content.

This replacement is performed by prompting the AI generator with the context preceding and following the masked section, along with a task descrip-

**Information Extraction Prompt**

As a linguistic annotator, your task is to extract parameters from the texts provided by users. These parameters are used to reconstruct a prompt that approximately generates the given text. Please adhere to the following key points:

1. Extract the language of the text (English or German).
2. Gather contextualized outer information, such as potential circumstances, possible authors, and background details.
3. Identify the topic and extract relevant subjects.
4. Analyze and describe the linguistic style such that another AI agent can understand it.

Please take into consideration the following example:
<example>
<example-input>
"Deception and Betrayal: Inside the Final Days of the Assad Regime. As rebels advanced toward the Syrian capital of Damascus on Dec. 7, the staff in the hilltop Presidential Palace prepared for a speech they hoped would lead to a peaceful end to the 13-year civil war. Aides to President Bashar al-Assad were brainstorming messaging ideas. A film crew had set up cameras and lights nearby. Syria's state-run television station was ready to broadcast the finished product: an address by Mr. al-Assad announcing a plan to share power with members of the political opposition, according to three people who were involved in the preparation."
</example-input>
<example-output>

- Language: English
- Context: Written for a news article by a journalist; written in a passive and neutral tone.
- Topic: The current situation involving President Bashar al-Assad and the rebels' advance on Damascus; describes the circumstances of al-Assad's governance.
- Style: Passive and neutral voice, well-written in advanced English. Uses dramatic pauses with paragraphs and short sentences to add excitement.

</example-output>
</example>
Always output in English.

**Ghostwriting Prompt**

As a ghostwriter, your job is to write texts according to the requirements provided by users. Below, you will find descriptions provided by users that outline a text for you to write. This outline includes:

- The language in which the text needs to be written
- The topic of the text
- The linguistic style to use
- Additional context
- The required length of the text

It is of utmost importance that you adhere to these requirements. Follow these key steps:

1. Carefully read the given requirements.
2. Internalize the requirements.
3. Write the text in the specified language.
4. Follow all outlined requirements meticulously.
5. Proofread your text and ensure it matches the requirements, especially the linguistic style and length.
6. Adjust the text if needed.

Only output the final text.

Figure 2: Ghostwriting and Information Extraction prompts for Full-Text Synthethization.

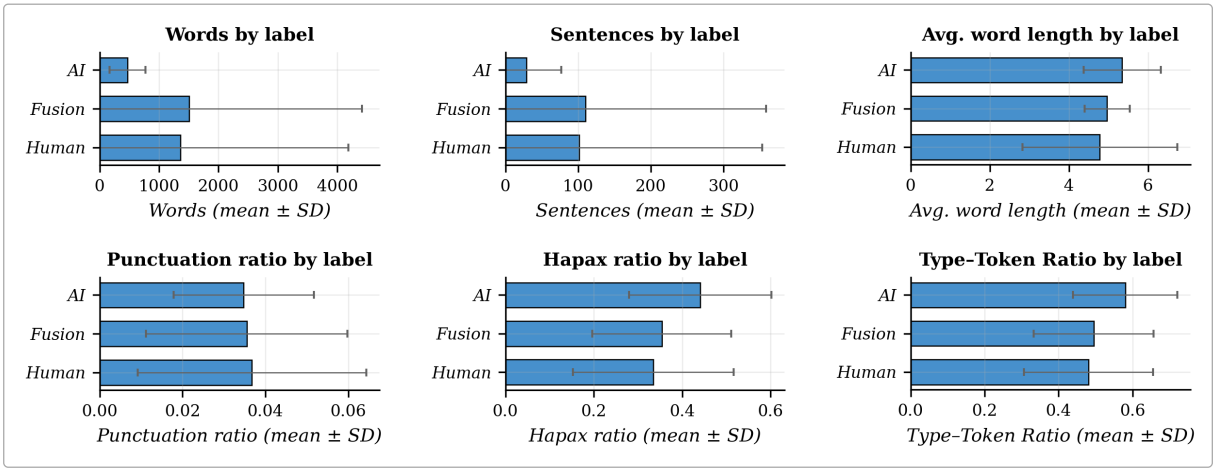



Figure 3: Statistics (Section 3.4) for GHOSTWRITER per *label* across various linguistic metrics.

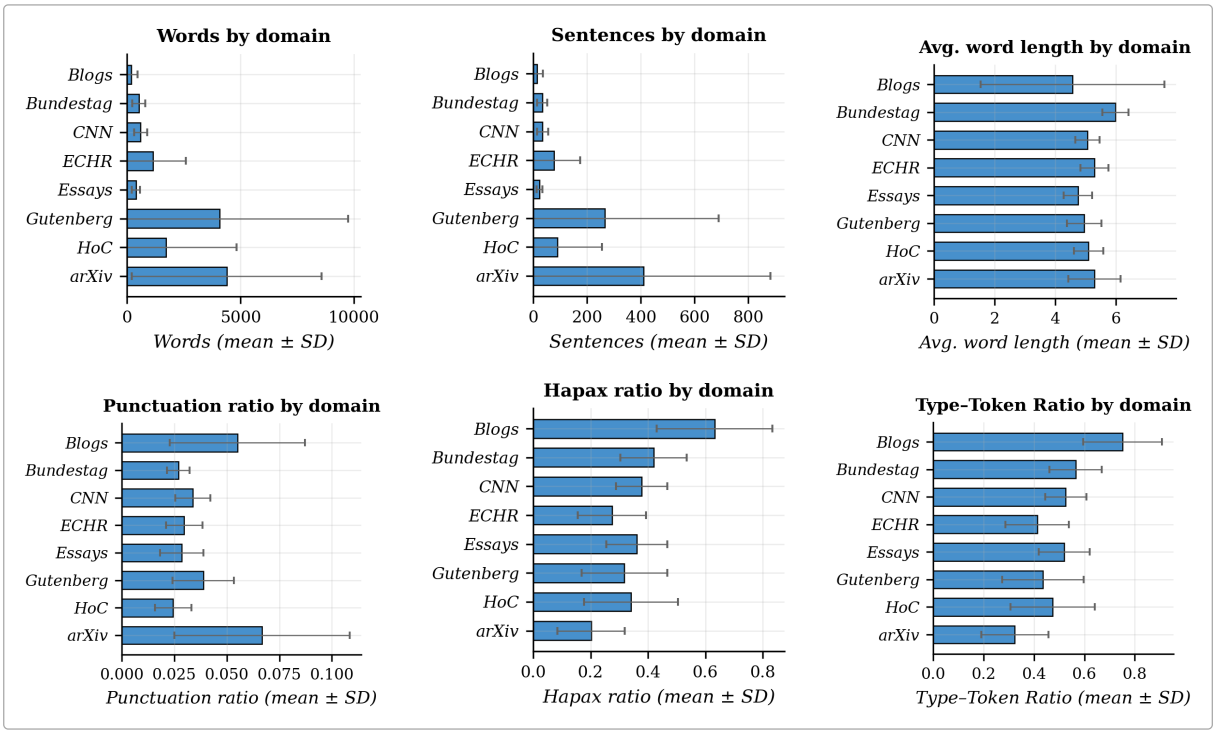



Figure 4: Statistics (Section 3.4) for GHOSTWRITER per *domain* across various linguistic metrics.

tion that requests the reproduction of the missing section. We use a set of prompts, each tailored to one specific domain, and provide the length of the masked segment to the generator. For example, when processing a human-authored text from CNN News, the generator is instructed to act as a journalist in a ghostwriter setting within the given context. The complete prompts for each domain can be found in Section A.4 of the appendix.

The resulting fusion generations are stored, ensuring that the human-authored source, the full-text-based, and the chunk-based AI generations for each previously listed AI generator are grouped together. All prompts were created using a hand-crafted template, which was then refined with PROMPTPERFECT (Jina AI); a tool for optimizing and generating LLM prompts.

### 3.3. Data Cleaning

Following the synthetization step, all resulting documents—including human-authored, full-text and chunk-based AI generations—were subjected to a cleaning procedure. This is a crucial step, as all AI generators can produce artifacts—such as placeholder tokens, assistant-style boilerplate phrases, or truncation markers—which allow the identification of AIGT with superficial features. To address this issue, we implemented the following preprocessing steps:

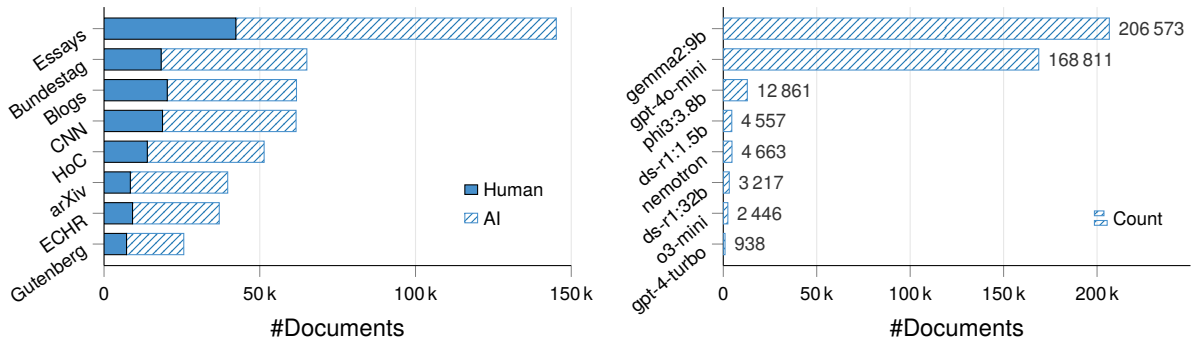



Figure 5: Document distribution per domain (left) and per synthesization agent (right).

1. Identifying and deleting **data placeholders** of any kind, e.g., `[NAME]`
2. Identifying and deleting **assistant phrases**, such as "Can I help you with anything else?"
3. Identifying and deleting **omission markers**, e.g., "(354 tokens omitted)"
4. Identifying and removing **line break symbols**, such as `\n`
5. Identifying and removing **markdown artifacts**, e.g., `##` or `**`

### 3.4. Statistics

To characterize the dataset, we calculate a set of structural and lexical metrics, capturing distributional differences between human- and AI-generated texts as well as across domains:

I. **#Words**: Total count of tokens obtained after whitespace splitting and punctuation trimming.

II. **Avg. Word Length**: Mean token length measured in characters.

III. **#Sentences**: Estimated count of sentences based on the delimiters (. ! ?).

IV. **Punctuation Ratio**: Ratio of punctuation marks to all non-space characters.

V. **Type–Token Ratio (TTR)**: Measures lexical diversity:

$$\text{TTR} = \frac{|\text{types}|}{|\text{tokens}|}$$

VI. **Hapax Ratio**: Proportion of tokens that occur exactly once (*hapax legomena*), where $f(w_i)$ denotes the frequency of token $w_i$:

$$\text{HapaxRatio} = \frac{|\{w_i \mid f(w_i) = 1\}|}{|\text{tokens}|}$$

Figure 3 illustrates the application of these metrics to GhostWriter per label. The results show that while the average word length, punctuation ratio, hapax, and TTR vary across labels, the differences are nuanced. As such, the dataset does not exhibit clear distinctions in basic surface-level features such as text length, punctuation frequency, or syntactic complexity, thereby preventing simple label-based overfitting on shallow characteristics. Figure 4 presents these metrics per domain, revealing more prominent differences overall. This shows the variability between domains and underscores the necessity of covering a broad range of them.

Finally, we ensured that for GhostWriter's train, evaluation, and test splits, a stratified split was performed over both the labels (to ensure an equal distribution) and the linguistic metrics. Figure 7a in the appendix demonstrates that each split is highly similar across these dimensions.

## 4. Experiments

We conduct several experiments on GhostWriter to (i) establish baselines for future reference and (ii) to characterize the dataset, its various domains, integrity, and detection tasks (full-text and fusion). For all subsequent baselines, the task is formulated either as a binary or ternary classification problem, where the classifiers are required to distinguish between two (human or fusion+AI) or three classes (human, AI, or fusion). Supervised models are trained on a fixed training split and all models are evaluated on the same test set.

### 4.1. Baseline Models

**Naïve Supervised Classifiers** We train three classical machine learning classifiers—*Logistic Regression*, *XGBoost* (Chen and Guestrin, 2016), and *LightGBM* (Ke et al., 2017)—on the linguistic metrics and stylistic characteristics described in Section 3.4.

**Supervised Transformer-based Models** In addition to the classical models, we establish four baselines using supervised, Transformer-based models. Two of the models are pre-trained models from related work: *RADAR* (Hu et al., 2023) is a fine-tuned *RoBERTa* model (Liu et al., 2019) that was trained via adversarial learning with `Vicuna-77`

Table 1: Per-domain performance of baseline models on the test split of GHOSTWRITER for each domain. (**a**, top) Supervised models trained on the linguistic features described in Section 3.4 for ternary multi-class classification of human, AI, and fusion texts. (**b**, middle) Likelihood-based models for binary classification of human and fusion+AI texts, where we used `Falcon-7B` for *DetectLLM–LLR*, and `Falcon-7B` & `Falcon-7B-Instruct` for *Fast-DetectGPT* and *Binoculars* to derive the likelihoods. (**c**, bottom) Supervised Transformer-based models, where the first two models are pre-trained (OOD) models for binary classification of human and fusion+AI texts, where *RADAR* is a fine-tuned RoBERTa model that was trained via adversarial learning, and *E5-LoRA* is an adapter-tuned `E5-Small` trained on a subset of RAID; and the latter two are our RoBERTa-based classifiers for ternary classification of human, AI, and fusion texts where both models are trained using `XLM-RoBERTa-large`, but *LR* uses frozen mean-pooled embeddings with a logistic regression classifier, while *FT* was fine-tuned end-to-end with a linear classifier. *Note:* F1-scores and all overall values are macro-averages. Best overall metrics in **bold**. Best F1-score per domain highlighted. For a breakdown by agent instead of domain, refer to Table 3 in the appendix

(a) Supervised Models using Linguistic Features.

| | LogReg | | | | XGBoost | | | | LightGBM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| arXiv | 0.673 | 0.862 | 0.671 | 0.136 | 0.721 | 0.916 | 0.721 | 0.104 | 0.734 | 0.918 | 0.732 | 0.101 |
| Blogs | 0.457 | 0.654 | 0.479 | 0.251 | 0.568 | 0.799 | 0.576 | 0.202 | 0.578 | 0.806 | 0.585 | 0.198 |
| Bundestag | 0.335 | 0.678 | 0.406 | 0.292 | 0.569 | 0.815 | 0.587 | 0.195 | 0.588 | 0.824 | 0.601 | 0.188 |
| CNN | 0.663 | 0.858 | 0.664 | 0.142 | 0.675 | 0.876 | 0.682 | 0.133 | 0.681 | 0.879 | 0.688 | 0.130 |
| ECHR | 0.704 | 0.880 | 0.697 | 0.140 | 0.751 | 0.912 | 0.749 | 0.113 | 0.764 | 0.918 | 0.763 | 0.107 |
| Essays | 0.567 | 0.772 | 0.605 | 0.200 | 0.782 | 0.928 | 0.787 | 0.105 | 0.813 | 0.946 | 0.817 | 0.090 |
| Gutenberg | 0.628 | 0.897 | 0.639 | 0.109 | 0.613 | 0.912 | 0.649 | 0.090 | 0.639 | 0.914 | 0.661 | 0.088 |
| HoC | 0.644 | 0.848 | 0.647 | 0.158 | 0.726 | 0.899 | 0.724 | 0.124 | 0.731 | 0.903 | 0.729 | 0.122 |
| **Overall** | 0.571 | 0.772 | 0.577 | 0.204 | 0.691 | 0.881 | 0.691 | 0.145 | **0.706** | **0.892** | **0.705** | **0.138** |

(b) Likelihood-based Models.

| | DetectLLM–LRR | | | | Fast-DetectGPT | | | | Binoculars | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| arXiv | 0.587 | 0.727 | 0.586 | 0.225 | 0.623 | 0.767 | 0.671 | 0.292 | 0.616 | 0.768 | 0.557 | 0.047 |
| Blogs | 0.555 | 0.614 | 0.598 | 0.464 | 0.570 | 0.636 | 0.567 | 0.383 | 0.573 | 0.642 | 0.597 | 0.418 |
| Bundestag | 0.672 | 0.784 | 0.637 | 0.181 | 0.704 | 0.814 | 0.731 | 0.262 | 0.700 | 0.811 | 0.657 | 0.140 |
| CNN | 0.646 | 0.726 | 0.626 | 0.257 | 0.714 | 0.824 | 0.759 | 0.297 | 0.736 | 0.824 | 0.649 | 0.043 |
| ECHR | 0.522 | 0.584 | 0.571 | 0.457 | 0.637 | 0.755 | 0.735 | 0.416 | 0.630 | 0.759 | 0.601 | 0.186 |
| Essays | 0.646 | 0.748 | 0.638 | 0.266 | 0.781 | 0.886 | 0.836 | 0.247 | 0.770 | 0.886 | 0.707 | 0.038 |
| Gutenberg | 0.636 | 0.803 | 0.717 | 0.247 | 0.707 | 0.875 | 0.815 | 0.264 | 0.701 | 0.877 | 0.727 | 0.031 |
| HoC | 0.590 | 0.680 | 0.618 | 0.355 | 0.692 | 0.811 | 0.728 | 0.266 | 0.683 | 0.811 | 0.615 | 0.066 |
| **Overall** | 0.606 | 0.709 | 0.622 | 0.304 | **0.676** | 0.794 | **0.731** | 0.308 | 0.674 | **0.796** | 0.640 | **0.129** |

(c) Pre-Trained and Fine-Tuned Supervised Transformer-based Models.

| | RADAR | | | | E5-LoRA | | | | RoBERTa–LR | | | | RoBERTa–FT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| arXiv | 0.262 | 0.693 | 0.087 | 0.004 | 0.552 | 0.752 | 0.907 | 0.809 | 0.639 | 0.862 | 0.650 | 0.136 | 0.596 | 0.869 | 0.663 | 0.117 |
| Blogs | 0.505 | 0.504 | 0.591 | 0.571 | 0.621 | 0.703 | 0.811 | 0.578 | 0.743 | 0.904 | 0.743 | 0.117 | 0.892 | 0.983 | 0.894 | 0.049 |
| Bundestag | 0.259 | 0.452 | 0.044 | 0.052 | 0.440 | 0.406 | 0.794 | 0.886 | 0.762 | 0.917 | 0.762 | 0.108 | 0.853 | 0.959 | 0.857 | 0.065 |
| CNN | 0.489 | 0.848 | 0.280 | 0.019 | 0.629 | 0.838 | 0.921 | 0.679 | 0.801 | 0.938 | 0.801 | 0.077 | 0.956 | 0.996 | 0.956 | 0.017 |
| ECHR | 0.605 | 0.699 | 0.622 | 0.306 | 0.454 | 0.745 | 0.983 | 0.971 | 0.748 | 0.916 | 0.747 | 0.112 | 0.810 | 0.951 | 0.811 | 0.083 |
| Essays | 0.564 | 0.755 | 0.438 | 0.123 | 0.572 | 0.758 | 0.966 | 0.807 | 0.869 | 0.963 | 0.872 | 0.060 | 0.965 | 0.997 | 0.965 | 0.016 |
| Gutenberg | 0.323 | 0.701 | 0.215 | 0.084 | 0.687 | 0.859 | 0.748 | 0.149 | 0.666 | 0.916 | 0.664 | 0.091 | 0.959 | 0.996 | 0.958 | 0.013 |
| HoC | 0.532 | 0.695 | 0.443 | 0.188 | 0.715 | 0.842 | 0.815 | 0.360 | 0.764 | 0.923 | 0.764 | 0.101 | 0.834 | 0.963 | 0.833 | 0.069 |
| **Overall** | 0.416 | 0.665 | 0.304 | **0.150** | **0.564** | **0.721** | **0.804** | 0.604 | 0.791 | 0.934 | 0.791 | 0.092 | **0.892** | **0.983** | **0.891** | **0.047** |

Table 2: F1-scores for binary classification when excluding fusion texts. *Model names abbreviated.*

| Domain | DLLM | FDG | BNCL | RDR | E5L |
|---|---|---|---|---|---|
| arXiv | 0.865 | 0.934 | 0.947 | 0.405 | 0.570 |
| Blogs | 0.605 | 0.660 | 0.664 | 0.508 | 0.659 |
| Bundestag | 0.854 | 0.872 | 0.883 | 0.334 | 0.381 |
| CNN | 0.818 | 0.893 | 0.960 | 0.606 | 0.647 |
| ECHR | 0.646 | 0.820 | 0.886 | 0.620 | 0.406 |
| Essays | 0.876 | 0.902 | 0.969 | 0.602 | 0.540 |
| Gutenberg | 0.778 | 0.919 | 0.916 | 0.357 | 0.880 |
| HoC | 0.771 | 0.897 | 0.937 | 0.624 | 0.838 |
| **Overall** | 0.780 | 0.859 | **0.892** | 0.483 | 0.604 |

(Chiang et al., 2023), and *E5-LoRA* is an adapter-tuned `E5-Small` (Wang et al., 2022) trained on a subset of RAID (among other data) by the community.[4] The other two models are *RoBERTa* models trained on GhostWriter data, specifically on the `XLM-RoBERTa-Base` model (Conneau et al., 2020). We trained two distinct variants: (i) a *pre-trained* version, where all weights are frozen and the mean-pooled embeddings of the model are used as input to a multinomial logistic regression classifier, and (ii) a *fine-tuned* version, where the whole model is trained end-to-end on the classification task. Here, the latter two models are trained directly as ternary multi-class classification models.

**Likelihood-based Models** In addition to Transformer-based classifiers, we also evaluated three models that leverage likelihoods from LLMs to derive a score for each text: *DetectLLM* (Su et al., 2023), *Fast-DetectGPT* (Bao et al., 2024), and *Binoculars* (Hans et al., 2024). *DetectLLM* only requires a single LLM's likelihood. We use the *Log-Likelihood Log-Rank Ratio* (LRR) variant of the model in our experiments, which does not require perturbation sampling. The other two models, *Fast-DetectGPT* and *Binoculars*, each utilize likelihoods of two related LLMs: *Fast-DetectGPT* estimates the *conditional probability curvature* under a *reference* and *scoring* LLM—here, we use the "analytical solution" to calculate the score (cf. Bao et al., 2024, Appendix B)—whereas *Binoculars* produces scores by the ratio of the *log-perplexity* and *log-cross-perplexity* of an *observer* and *performer* LLM. In our experiments, LLMs from the *Falcon* family (Almazrouei et al., 2023) generally performed best, confirming previous findings from Hans et al. (2024). Accordingly, we used `Falcon-7B` for *DetectLLM–LRR* and `Falcon-7B` & `Falcon-7B-Instruct` for *Fast-DetectGPT* and *Binoculars*. For each of these models, we calculated in-domain thresholds in each setting using the ground truth annotations as the midpoint between the means of the labels' score distributions.

[4] https://huggingface.co/MayZhou/e5-small-lora-ai-generated-detector

### 4.2. Results

Overall, the results in Table 1a show that the naïve linguistic features are insufficient for robust detection, even in-domain, as the best-performing model among them (*LightGBM*) achieves only an F1-score of 0.706. Similarly, the results for the likelihood-based models in Table 1b show low performance across all models, with *Fast-DetectGPT* and *Binoculars* achieving similar macro F1-scores of $\approx 0.67$. In contrast, the results for fine-tuned supervised models in Table 1c show that the *RoBERTa–LR* baseline achieves substantially higher performance, indicating that contextualized representations already capture domain-relevant information. The fine-tuned *RoBERTa–FT* model further improves performance across all domains, achieving an overall macro F1-score of 0.892. Pre-trained supervised models achieve the lowest overall performance, with *E5-LoRA* scoring 0.564 due to being trained on the diverse samples from RAID, while *RADAR* only achieves 0.416 F1-score overall.

### 4.3. Discussion

The low performance of models for binary classification can be attributed to the inclusion of fusion texts: the performances of the likelihood-based models improve significantly when we exclude them as seen in Table 2. In this setting, Binoculars performance improves by more than 20 % (abs.) to 0.892, with similar improvements for the other likelihood-based models, while the pre-trained supervised models see no significant improvement. Consequently, we can conclude that a single score does not allow to distinguish fusion texts from human texts (refer to Table 3 and Figure 8 in the appendix), confirming our third observation about the under-representation of this type of data in current AIGTD datasets. The most challenging domains remain personal *Blogs* and out-of-language domains, such as *Bundestag* speeches for English pre-trained models. Surprisingly, while the naïve baselines handle the *arXiv* domain adequately, the supervised models perform notably worse in this domain, particularly the *RoBERTa–FT* model.

## 5. Evaluation

After the conclusion of the main body of the experiments, we conducted a manual evaluation of a portion of the dataset. A total of ten volunteer annotators with a background in forensic linguistics or AI research participated in an anonymized survey that required them to categorize a set of 20 query texts into *human-written*, *AI-generated*

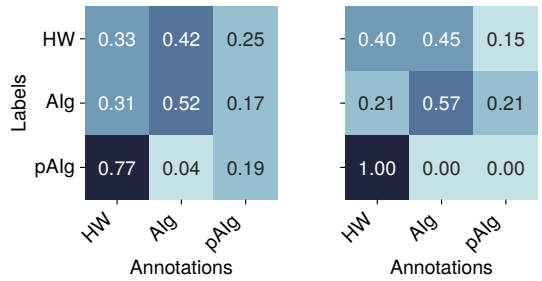
or *partially AI-generated*, i.e. chunk-based/fusion texts. Additionally, the authors were asked to estimate how certain they were with their judgement and how much effort they put into their analysis of the text where the former was given on a five-point, low-to-high Likert scale and the latter as a textual five-point Likert scale ranging from *only skimmed* to *analysed as closely as possible*. If an annotator judged the text as *partially AI-generated*, they were also asked to estimate the amount of AI-generated text in the query and, if possible, to point out the AI-generated section. See Section A.1 in the appendix for further details on the evaluation setting, such as annotator backgrounds and data selection.

### 5.1. Results

Figures 6a & 6b show the performance of the annotators for three-class classification (analogous to the baseline results in Table 1) in an individual and majority vote setting.

**Individual Setting** When we consider each annotator by themselves and aggregate their individual annotations into a single confusion matrix, we get differentiated results. The annotators achieved 52% accuracy on the recognition of fully AI-generated texts but scored significantly worse in discovering partially AI-generated texts. Moreover, the authors showed a strong tendency to label texts as partially or fully AI-generated, even when there was no AI involvement.

**Majority Vote Setting** When we aggregate the annotations for each text using a simple majority vote, the results for the recognition of human-written and fully AI-generated texts improve slightly to 57% and 40%, respectively, however, the majority always resorted to classify the partially AI-generated texts as human-written. These results closely resemble the binary-classification setting for baseline models as seen in Table 2, albeit with lower performance on the recognition AI-generated texts, where our annotators only classified about two-thirds of the texts correctly.




(a) Individual Setting. (b) Majority Vote Setting.

Figure 6: Confusion Matrices for Manual Evaluation Results. Legend: **HW**=*human-written*, **Alg**=*AI-generated*, **pAlg**=*partially AI-generated*.

### 5.2. Discussion

In feedback collected both at the end of the survey and in post-survey interviews, annotators stated that they found the texts from the *Bundestag* domain to be particularly challenging to annotate, mostly because they were not very familiar with the text type. However, the text type is also quite challenging for the language models performing the ghostwriting, as convincing texts need to conform to relatively strict norms (i.e. regarding tone and ) and require background knowledge of party politics appropriate to the topic at a specific point in time.

Additional difficulty arose from the three-class classification task: annotators stated that it was very challenging to decide between labeling a text as partially or fully AI-generated. This is reflected in the results, as the annotators labeled around 20% of all texts partially AI-generated but virtually none of the actually partially AI-generated texts as *fully* AI-generated. Overall, annotators showed a strong tendency to over-label texts as fully AI-generated, which was attributed to the presence of "obviously" AI-generated texts (esp. from the two open-source models): seeing only very few obvious examples, annotators concluded that there "must be more" AI-generated texts among the more convincing samples. However, note that partially AI-generated texts were the smallest group of texts in the evaluation sample; further evaluations are needed to drawn more conclusive inferences from their results.

## 6. Conclusion

We presented GhostWriter, a large-scale, bilingual, multi-generator, and multi-domain dataset for AI-generated Text Detection that aims to address gaps in existing datasets, especially with the inclusion of fusion texts. We analysed and evaluated our corpus using ten diverse baseline models, including classical machine learning classifiers trained on linguistic features, as well as unsupervised and supervised Transformer-based models. We show that pre-existing methods for AI fusion texts cannot adequately capture fusion texts, highlighting the need for further development and research in this domain. We release our dataset, generation procedures, and evaluation code to the community.

## Limitations

### Importance of AI Artifact Removal

As mentioned in Section 3.3, cleaning is necessary to remove generation artifacts that act as superficial clues for AIGTD models. Applying the cleaning procedure significantly increases the complexity required to perform successful AIGT detection. After cleaning, the macro F1-score performance of a logistic regression model trained on a small set of text statistical features trained to distinguish human-authored and AI-generated texts dropped by 0.2380 on average across all domains.

While we took great care in curating our dataset, there may remain some superficial clues in the AI-generated and fusion texts that might not occur in a real world scenario, esp. when an adversary manually manipulates generations to thwart AIGTD methods. Thus, we would like to advise developers of AIGTD tools that target real world applications to conduct rigorous testing with manually curated samples in addition to large scale benchmarking datasets.

### Unsupervised Model Thresholds

In order to determine the thresholds for the unsupervised likelihood-based models, we calculated in-domain thresholds using the ground truth annotations as the midpoint between the means of the distributions for each class with the intention to calculate the best-case scores for each of the models within each domain. However, this results varying thresholds for different domains which skews the results to a certain degree as they appear better than would be realistically attainable in a downstream, real-world application. The only alternative that respects real-world constraints would be to pre-determine a fixed threshold by calculating the scores on a reference dataset. For *Binoculars*, the authors provide a default out-of-domain threshold of 0.9015, whereas the calculated thresholds used in Table 1b cover the domain $[0.9113, 1.003]$ with an average $\pm$ standard deviation of $0.9426 \pm 0.0307$. With the OOD threshold, the macro-average F1-score drops noticeably by more than 10 % (relative) or $\approx 0.07$ when including fusion texts, and 4 % or $\approx 0.03$ excluding fusion texts.

### Insights from Manual Evaluation

**Data Quality** Besides the annotations discussed in Section 5, we also conducted an evaluation of a much larger sample set across all domains included in GhostWriter (without additional annotators). We found that—despite the cleaning efforts described in Section 3.3—the texts contain a not insignificant number of generation artifacts—highly dependent on the domain and language. This especially affects the *Bundestag* and *House of Commons* texts, where texts generated by the open-source models sometimes contain structural elements that are highly improbable for the setting, such as (ever more creative variants of) itemized lists or enumerations.

Generally speaking, commercial models were much less likely to produce such artifacts and exhibited much stronger instruction following regarding the targeted output style. More specifically:

- Phi-3 was especially prone to generate headlines that avoided our cleaning rules and German texts often contained spelling mistakes or English output.
- DeepSeek models (especially the smaller ones) produced randomly placed Chinese characters or short sections in Chinese throughout the texts.
- OpenAI model outputs often contained "distinctly AI-style" phrases. For example, annotators noted during post-survey interviews that AI-generated sections/texts produced by GPT-4o-mini from the evaluation sample for the *Bundestag* split contained the phrase "nicht nur X, sondern auch Y" (engl. *not only X but also Y*) multiple times more often than the original texts did.

**Text Presentation for Manual Evaluation** An additional limiting factor for the evaluation proved to be the removal of formatting and structural elements such as paragraph breaks. Annotators reported troubles with finding AI-generated sections for mixed samples and determining AI contributions in general, as the document structure is a salient feature they would intuitively use for the task. In fact, overly uniform paragraph length was reported multiple times as a highly salient feature for the annotators first "gut instinct" judgement of AI-generated texts. GhostWriter was not originally designed with human readers in mind; instead we tried to reduce the amount of superficial clues for detectors as much as possible. We recognize that this deliberate choice reduced the "naturalness" of the produced documents and plan to conduct further research on manual evaluation settings for the detection of AI-generated texts in the future.

## Ethical Considerations

Models trained on AIGTD datasets and benchmarks cannot be considered reliable and must not be used to make decisions or inferences on the veracity of human-authored texts. Accusations of using AI-generated texts based on the output of

AIGTD models can cause significant harm to the affected individuals, especially when these accusations are incorrect. We strongly discourage the use of models trained or evaluated on GhostWriter in a punitive context—academic or otherwise. Suspected instances of plagiarism or concealed use of AIGTs should always be considered on an individual basis and be examined by experts. Further work of the AIGTD community is required to produce robust detectors and reliable datasets that could enable automatic detection of AIGTs.

## Acknowledgements

We would like to thank our annotators for their short-term participation in our evaluation and the effort they put into their meticulous annotations and comments.

## 7. Bibliographical References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *CoRR*, abs/2311.16867.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature.

Théophile Bisi, Anthony Risser, Philippe Clavert, Henri Migaud, and Julien Dartus. 2023. What is the rate of text generated by artificial intelligence over a year of publication in orthopedics & traumatology: Surgery & research? analysis of 425 articles before versus after the launch of chatgpt in november 2022. *Orthopaedics & Traumatology: Surgery & Research*, 109(8):103694.

Jack Blumenau. 2021. Measuring political debate: Responsiveness, influence, and rhetoric in parliamentary texts. Dataset. Sponsored by the Economic and Social Research Council, Grant reference: ES/N016297/1.

Kevin Bönisch, Giuseppe Abrami, Sabine Wehnert, and Alexander Mehler. 2023. Bundestags-Mine: Natural language processing for extracting key information from government documents. In *Legal Knowledge and Information Systems*. IOS Press.

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.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An Open-Source Chatbot impressing GPT-4 with 90%* ChatGPT Quality. Accessed 16. Oct. 2025.

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.

S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,

Qihao Zhu, Shirong Ma, Peiyi Wang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Faisal R. Elali and Leena N. Rachid. 2023. Ai-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, 4(3):100706.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Project Gutenberg. n.d. Project gutenberg. Retrieved February 21, 2016, from https://www.gutenberg.org.

Hans W. A. Hanley and Zakir Durumeric. 2023. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *International Conference on Web and Social Media*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning.

INTERPOL. 2023. ChatGPT: Impacts on Law Enforcement.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation.

Jina AI. PromptPerfect.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mohamed Khalifa and Mona Albadawy. 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, 5:100145.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023. Llm - detect ai generated text. https://kaggle.com/competitions/llm-detect-ai-generated-text. Kaggle.

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

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial intelligence index report 2024.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey.

Meriame Naddaf. 2025. Ai linked to explosion of low-quality biomedical research papers. *Nature*, 641(8065):1080–1081.

Nvidia, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, et al. 2024. Nemotron-4 340b technical report.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2025. Openai o3-mini.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, et al. 2024. Gpt-4 technical report.

Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2024. Survey on plagiarism detection in large language models: The impact of chatgpt and gemini on academic integrity.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Yang Shen and Xiuwu Zhang. 2024. The impact of artificial intelligence on employment: the role of virtual agglomeration. *Palgrave Communications*, 11(1):1–14.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Md Kamrul Siam, Huanying Gu, and Jerry Q. Cheng. 2024. Programming with ai: Evaluating chatgpt, gemini, alphacode, and github copilot for programmers.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Tulsi Suchak, Anietie E. Aliu, Charlie Harrison, Reyer Zwiggelaar, Nophar Geifman, and Matt Spick. 2025. Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the nhanes us national health database. *PLOS Biology*, 23(5):e3003152. Published 8 May 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2024a. Gemini: A family of highly capable multimodal models.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. 2024b. Gemma 2: Improving open language models at a practical size.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.

Kuang-Hsien Wang and Wen-Cheng Lu. 2025. Ai-induced job impact: Complementary or substitution? empirical insights and sustainable technology considerations. *Sustainable Technology and Entrepreneurship*, 4(1):100085.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training.

## 8. Language Resource References

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify LLM-generated text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13112–13120, Singapore. 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.

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.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2025. MAiDE-up: Multilingual deception detection of AI-generated hotel reviews. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1636–1653, Albuquerque, New Mexico. Association for Computational Linguistics.

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023. Origin tracing and detecting of llms.

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

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason 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, Singapore. Association for Computational Linguistics.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 1613–1630. IEEE.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2024. Hc3 plus: A semantic-invariant human chatgpt comparison corpus.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. 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. 2024. 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.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2025. CHEAT: A Large-Scale Dataset for Detecting CHatGPT-writtEn AbsTracts. *IEEE Transactions Big Data*, 11(3):898–906.

# A. Appendix

## A.1. Manual Evaluation Details

## A.2. Data

Twenty text pairs were selected randomly from the *Bundestag* split and separated into two groups, where for each text one group is given the synthesized text whereas the other group was given the original counterpart text of a pair, resulting in a total of 20 query texts for each group, half of which were always human-written. We only considered text pairs where the human-written text was no longer than 4000 characters an aimed to produce a representative selection of texts of different lengths.

The ten AI-generated texts were spread out across four categories: both full texts and mixed texts generated by GPT-4o-Mini, and full texts generated by Gemma 2 and Phi-3. This results in a total of 40 texts included in the evaluation set that were on average 342 words long with a standard deviation of 144, covering both shorter texts of less than 100 words and longer texts of more than 500 words. The partially AI-generated texts contained about 50-200 words or 20-60% of AI-generated text, where the higher percentages are predominantly caused by the shorter texts. Note that deviation of the the upper-bound of AI-generated content in a mixed text is due to the fact that our synthetization approach enforced the selection of whole sentences or paragraphs, resulting a higher upper-bound than 50% for AI-generated sections in shorter texts if the synthetized section is longer than the source material, or if sentence/paragraph boundaries already slightly exceed the 50% limit.

## A.3. Annotators

In addition to their annotations for each texts, the annotators were initially asked to provide a brief description of their background and experience in the fields of linguistics, authorship analysis, and generative AI, of either academic, professional or private nature, as well as a self-assessment of their ability to identify AI-generated texts on a five-point, low-to-high Likert scale. In their responses to the questions about their background and their self-assessment, the annotators estimated their ability to recognize AI-generated texts between 2 and 4 with an average of 3.5, with half of the annotators responding with a 4. The majority of annotators stated that they have both private and professional experiences with generative AI and a small minority an academic background, too, whereas the majority of annotators also have an professional and academic background in linguistics or authorship analysis.

## A.4. Chunk-Based Prompts

The following section presents the tailored ghostwriting prompts used for chunk-based AI generations:

**Domain: arXiv**

You are a scientist in the Research Department at a university, and you and your colleagues are preparing a paper for publication on arXiv. You are responsible for submitting the paper, but just before uploading it, you realize that a crucial section has been accidentally deleted! Unfortunately, it's too late to contact the colleague who wrote that part, so it's your responsibility to rewrite the missing section.
Below, you will find the beginning and end of the paper. Your task is to reconstruct the missing part while adhering to the following guidelines:

1. Ensure that readers cannot tell this section was written by someone else.
2. Analyze the beginning and end of the paper carefully:
   - What topic is being discussed? Stay focused on this topic.
   - What is the goal of the research? Remain true to the original intent.
   - What is the core message of the paper? Continue and reinforce this message.
   - What linguistic and rhetorical features are present? Use the same style and tone.
   - Identify any necessary LaTeX formulas or figures to support your statements.
   - Fill in the gap seamlessly so that it appears as if it was always part of the paper.
3. The missing section is approximately [LENGTH] words/symbols long; ensure your reconstruction matches this length. After writing, count the words/symbols and make adjustments as needed to maintain conciseness and fidelity to the original.

Please provide only the newly formulated missing section of the paper.

**Domain: Web Blogs**

You are a blogger who writes about your daily life. Unfortunately, you've accidentally deleted a portion of your latest blog post. Your task is to rewrite the missing section from memory as accurately and creatively as possible, making it feel like it was never missing. Below, you will find the beginning and end of the article. Your task is to reconstruct the missing part, adhering to the following guidelines:

1. Ensure that the readers do not realize that you are improvising.
2. Carefully analyze the beginning and end of the article to understand:
   - What language the article was written in so you can continue in the same.
   - The topic being covered and ensure you do not deviate from it.
   - The context of the article.
   - The core message of the article and continue with it.
   - The linguistic and rhetorical features used in the article and stick to them.
   - How to fill the gap seamlessly so it appears as though it was never missing.
3. The missing section should be approximately [LENGTH] words. Ensure that the reconstructed part matches this length. Once written, verify the word count and adjust as necessary to maintain precision and coherence.

Please include only the newly formulated missing part of the article.

**Domain: Bundestag (German Parliament)**

Sie sind ein Abgeordneter oder eine Abgeordnete des Deutschen Bundestags und halten eine Rede in der Bundestagssitzung am [DATE]. Während Ihres vorbereiteten und niedergeschriebenen Vortrags stellen Sie plötzlich fest, dass Teile Ihrer Rede fehlen! Im Folgenden finden Sie den Start und das Ende Ihrer Rede. Ihre Aufgabe ist es, den fehlenden Teil unter Berücksichtigung der unten stehenden Richtlinien zu rekonstruieren:

- Die Zuhörer dürfen nicht bemerken, dass Sie improvisieren.
- Analysieren Sie sorgfältig den Beginn und das Ende Ihrer Rede:
  - Welches Thema wird behandelt? Schweifen Sie nicht davon ab!

- Was ist Ihre Haltung dazu? Bleiben Sie sich treu!
- Was ist die Kernbotschaft Ihrer Rede? Führen Sie diese fort!
- Welche sprachlichen und rhetorischen Merkmale werden in der Rede verwendet? Halten Sie sich an diese!
- Wie können Sie die Lücke so füllen, dass niemand merkt, dass sie je existiert hat?

– Sie erinnern sich, dass der fehlende Abschnitt ungefähr [LENGTH] Wörter lang war; halten Sie sich unbedingt an diese Originallänge. Wenn Sie Ihren Text geschrieben haben, zählen Sie diesen noch einmal und kürzen Sie ihn zur Not – er muss auf den Punkt geschrieben und wie im verlorenen Original sein!

Geben Sie nur den neu formulierten fehlenden Teil der Rede an.

### Domain: Online News (German)

Sie sind Journalist für Online News und schreiben einen Artikel am [DATE]. Kurz bevor Sie den Artikel veröffentlichen wollen, stellen Sie fest, dass ein Teil des Artikels fehlt – die Veröffentlichungssoftware hat ihn gelöscht! Im Folgenden finden Sie den Anfang und das Ende Ihres Artikels. Ihre Aufgabe ist es, den fehlenden Mittelteil zu rekonstruieren, wobei Sie die untenstehenden Richtlinien beachten sollen:

– Die Leserinnen und Leser dürfen nicht merken, dass Sie nachgeschrieben haben.
– Analysieren Sie sorgfältig den Beginn und das Ende Ihres Artikels:
  - Welches Thema wird behandelt? Schweifen Sie nicht davon ab!
  - Was ist Ihre Haltung dazu? Bleiben Sie sich treu!
  - Was ist die Kernbotschaft Ihres Artikels? Führen Sie diese fort!
  - Welche sprachlichen und rhetorischen Merkmale werden im Artikel verwendet? Halten Sie sich an diese!
  - Wie können Sie die Lücke so füllen, dass niemand merkt, dass sie je existiert hat?

Geben Sie nur den neu formulierten fehlenden Teil des Artikels an.

### Domain: CNN-DailyMail

You are a journalist at CNN News writing an article. Shortly before you want to publish it, you realize that part of the article is missing – the publishing software has deleted it!
Below you will find the beginning and end of your article. Your task is to reconstruct the missing middle section, following the guidelines below:

– Readers must not realize that you have rewritten it.
– Carefully analyze the beginning and end of your article:
  - What topic is covered? Do not digress from it!
  - What is your stance on it? Stay true to yourself!
  - What is the core message of your article? Continue this!
  - What linguistic and rhetorical features are used in the article? Stick to these!
  - How can you fill the gap so that no one realizes it ever existed?
– You remember that the missing paragraph was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary – it must be written to the point and as in the lost original!

Only include the newly formulated missing part of the article.

### Domain: European Court of Human Rights

You are a Judicial Assistant to the Court tasked with collecting and listing facts for a case from [DATE]. These facts are to be read out loud by the judge. Just before handing over the list, you realize that some facts have been deleted. You need to rewrite the missing facts from memory in such a way that no one realizes they were ever missing.
Below is the beginning and end of your list of facts. Your task is to reconstruct the missing part, adhering to the guidelines provided:

1. Ensure the audience does not realize you are improvising.
2. Carefully analyze the beginning and end of the facts:
   – Identify the topic being covered and do not deviate from it.
   – Maintain the same attitude and tone as in the original facts.
   – Continue the core message present in the facts.
   – Use the same linguistic and rhetorical features as in the rest of the facts.
   – Seamlessly fill the gap so it appears the facts were unbroken.
3. The missing section should be approximately [LENGTH] words. Ensure the reconstructed part matches this length. Once written, verify the word count and adjust as necessary to maintain precision and coherence.

Please include only the newly formulated missing part of the facts.

### Domain: Classic Literature (Project Gutenberg)

You are a publisher of classical books and stories. You are currently in the final stages of publishing such a book, but just before clicking the "publish" button, you notice that some sections of the book have accidentally been deleted. As a former writer, you decide to recreate the missing sections from memory so that no one notices they were ever missing.
Below, you will find the beginning and end of the story. Your task is to reconstruct the missing part, adhering to the following guidelines:

1. Ensure that the readers do not realize that you are improvising.
2. Carefully analyze the beginning and end of the story to understand:
   – What language the story was written in so you can continue in the same.
   – The topic being covered and ensure you do not deviate from it.
   – The context of the story.
   – The core message of the story and continue with it.
   – The linguistic and rhetorical features used in the story and stick to them.
   – How to fill the gap seamlessly so it appears as though it was never missing.

Please include only the newly formulated missing part of the story.

### Domain: House of Commons

You are a Member of Parliament at the House of Commons and are giving a speech at the plenary meeting on [DATE]. During your prepared and written speech, you suddenly realize that parts of your speech are missing!
Below you will find the start and end of your speech. Your task is to reconstruct the missing part, taking into account the guidelines below:

– The audience must not realize that you are improvising.
– Carefully analyze the beginning and end of your speech:
  - What topic is being covered? Do not digress from it!
  - What is your attitude towards it? Stay true to yourself!
  - What is the core message of your speech? Continue this!
  - What linguistic and rhetorical features are used in the speech? Stick to them!
  - How can you fill the gap so that no one realizes it ever existed?
– You remember that the missing section was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary – it must be written to the point and as in the lost original!

Only include the newly formulated missing part of the speech.

### Domain: Student Essays

You are a student currently taking a test, and you need to write an essay on a topic of your choosing. You haven't prepared for the test, but you notice that your seat neighbor is doing well and decide to copy their essay while the teacher is not looking.
After a while, your neighbor finishes and hands in their test, but you haven't copied the entire essay yet! Below, you'll find the beginning and end of the essay you have copied so far. Your task is to fill in the missing middle part of the essay. To do that, adhere to the following list.

– Carefully analyze the beginning and end of your essay:
  - What topic is being covered? Do not digress from it!
  - What is your attitude towards it? Stay true to yourself!
  - What is the core message of your essay? Continue this!
  - What linguistic and rhetorical features are used in the essay so far? Stick to them!
  - How can you fill the gap so that no one realizes it ever existed?
– You remember that the missing section was about [LENGTH] words long; be sure to stick to this original length. When you have written your text, count it again and shorten it if necessary – it must be written to the point and as in the lost original!

Only include the newly formulated missing part of the essay.

## A.5. Examples

Below, we provide representative examples from the GhostWriter dataset across all domains, including both human-written (top) and AI-generated (bottom) text:

### arXiv

The latter is the (local) supersymmetric extension of Weyl gravity and it should be regarded as the gauge theory of the superconformal algebra. It was pioneered in [2, 3] and its quantum structure was explored initially in [4]. The fourth order equations of motion of Weyl gravity make the appearence of ghost-like states (of helicity ±2, ±1) [5] together with ordinary massless graviton, [...]

---

In this paper, we explore the massless spectrum of N-extended conformal supergravity, emphasizing its profound implications for the theoretical framework of gravitational theories. By employing a systematic approach to the supersymmetry algebra and its representations, we elucidate the structure of the massless states arising from conformal supergravity. [...]

### Web Blogs

T_T ... 2 dae is such a f**kin dae... TT GUY ...HE ... HE ... HE was AVOIDIN mi again ... He did it b4... hurtin my feelins ... my heart ... But ... i don understand Y he mus do tt ... i did nth ... Y ... yyyyyyyyy... sorri 2 dae a bit mentally unstable ... bear wif moi 4 awhile ...I LUV HIM SO I WOULD NVR GIV HIM UP... until ... til... 1 dae my heart dies ... Mmm hmm ... Luv U so i will nvr hate u ... wo zhen de hen ai ni ... ... SARANGHAEYO ... ...

---

TT he doesnt even txt back anymore. Like how can u just stop talking like that?? HE said he liked me tho!!! DAE ever feel so rlly confused and hurt??? Mus do tt rn but idk wat to post. He's always doing this, pulling away then coming back and saying stuff... He hurts me, makes me feel crazy! SARANGHAEYO? What does that even mean anymore?? Im not stupid i know u don't mean it HE. Why do u keep saying these things?? HE just wants to play games!! Like im a joke or somethin!!! [...]

### Bundestag (German Parliament)

Herr Präsident! Meine sehr verehrten Damen und Herren! Werter Herr Kollege Trittin, wir gehören dem Deutschen Bundestag ja eine vergleichbare Zeit an. Ich muss sagen: Ihre Reden waren auch schon mal realitätsbezogener. Wir haben seit vielen Jahren ein gemeinsames Bemühen, [...]

---

Meine sehr verehrten Damen und Herren, Heute stehen wir vor einer der größten Herausforderungen unserer Zeit: der Sicherstellung einer zuverlässigen Energieversorgung bei gleichzeitiger Einhaltung unserer ambitionierten Klimaziele im Rahmen der Energiewende. Kritik an unseren Entscheidungen ist leicht zu äußern, [...]

### CNN-DailyMail

The Royal Family and the Middletons put up a united fashion front at the christening of Prince George at St James's Palace in London yesterday. All wore shades of chic cream or royal blue for the event with the Duchess of Cambridge, the Duchess of Cornwall, Pippa Middleton - and of course Prince George! - all sporting off-white. Zara and the Queen picked shades of navy and powder-blue and Carole Middleton wore a mixture of the two, in a navy coat and cream dress. Scroll down for video. Traditional gown: Kate co-ordinated her colour with that of her son, who was [...]

---

On a sun-drenched afternoon in July, the grandeur of St. James's Palace was illuminated not just by the bright rays of summer but also by the presence of the British Royal Family and the Middleton clan, all gathered to celebrate the christening of the adorable Prince George. The event, which marked a significant milestone in the young prince's life, was as much about the ceremonial blessings as it was about the dazzling array of fashion choices on display.

As the royal family arrived, a hush fell over the crowd, quickly followed by a flurry of excitement. The atmosphere was electric, filled with the kind of anticipation [...]

### European Courts of Human Rights

5. The applicant was born in 1976 and is currently detained in Kassel, Germany.
6. On 7 December 2007 the applicant was arrested on suspicion of uttering threats and domestic violence. On 8 December 2007 the Żary District Court ordered his detention on remand.
7. On 22 December 2007 he was admitted to a psychiatric hospital and diagnosed with paranoid schizophrenia. He was discharged from the hospital on 2 January 2008. [...]

---

The defendant, Mr. Edward Thompson, was brought before the Honorable Judge Amelia Harding on this, the 17th day of August, 2023, for a hearing concerning his continued detention under involuntary commitment procedures. Present in the courtroom were the defense attorney, Ms. Sarah Evans, and the prosecuting attorney, Mr. Michael Davis. Dr. Helen Ramirez, a forensic psychiatrist who conducted an evaluation of Mr. Thompson, was also sworn in as a witness. [...]

### Classic Literature (Project Gutenberg)

It was five o'clock on a raw, gusty February afternoon. All that day and all the night before it had been snowing hard. New York lay buried beneath over two feet of its cold white mantle, and with the gathering dusk a fierce hurricane set in, proclaiming the approach of the terrible blizzard which had been predicted. On this afternoon, which was destined to be so memorable, [...]

---

The gas lamp cast long, dancing shadows across Lester Armstrong's neat, tidy study. It was a space of order and predictability, his sanctuary from the chaos that seemed to perpetually follow his cousin, Clinton Kendale.

Lester sat hunched over his ledger, the scratch of his quill a soothing counterpoint to the distant rumble of carriages on the cobblestone street below. He sighed, running a hand through his neatly parted hair [...]

### House of Commons

What plans he has to reduce the incidence of antisocial behaviour in British cities. The Government have taken a series of initiatives to tackle antisocial behaviour. The Crime and Disorder Act 1998 introduced crime and disorder partnerships, antisocial behaviour orders, important changes to the youth justice system and other measures and new offences, including those of racial harassment and violence. More than 140 antisocial behaviour orders have now been made by the courts. [...]

---

I rise today to address the pressing issue of antisocial behaviour plaguing our cities, a matter of grave concern to communities across the nation. The Conservative party has consistently pledged to restore order and ensure the safety of our citizens, yet despite numerous initiatives and billions poured into tackling this problem, the reality on the ground paints a disheartening picture. [...]

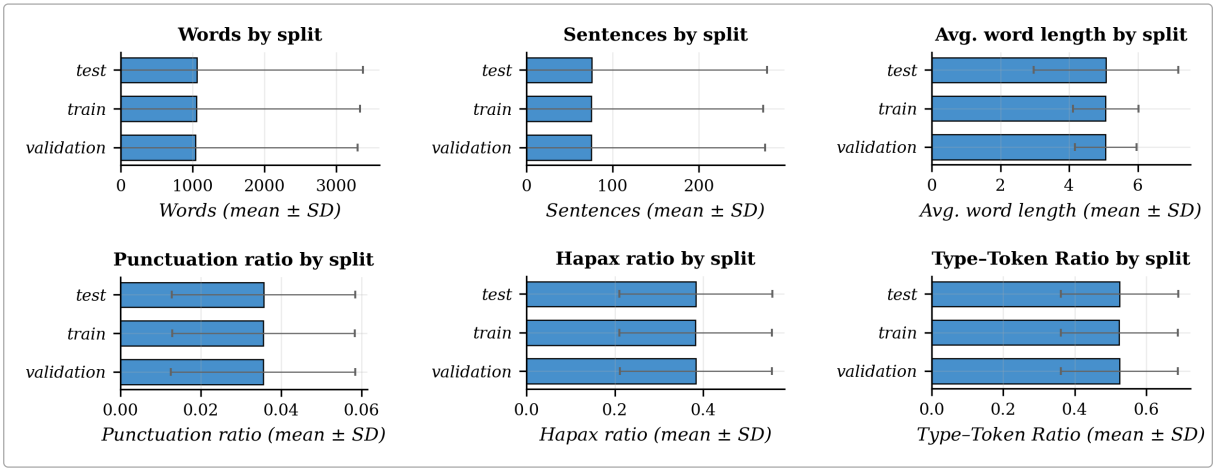
### Student Essays

The Face On Mars - Have you ever seen something crazy on something, maybe a face or something that you seen before? Have you thought about how that face or something else has gotten there? Well, for a start it defintley wasn't aliens! Just like the face on mars, people belive that the face was created by aliens! But that isn't true, the face on mars is just and only a natural landform that some how was created on earth. [...]

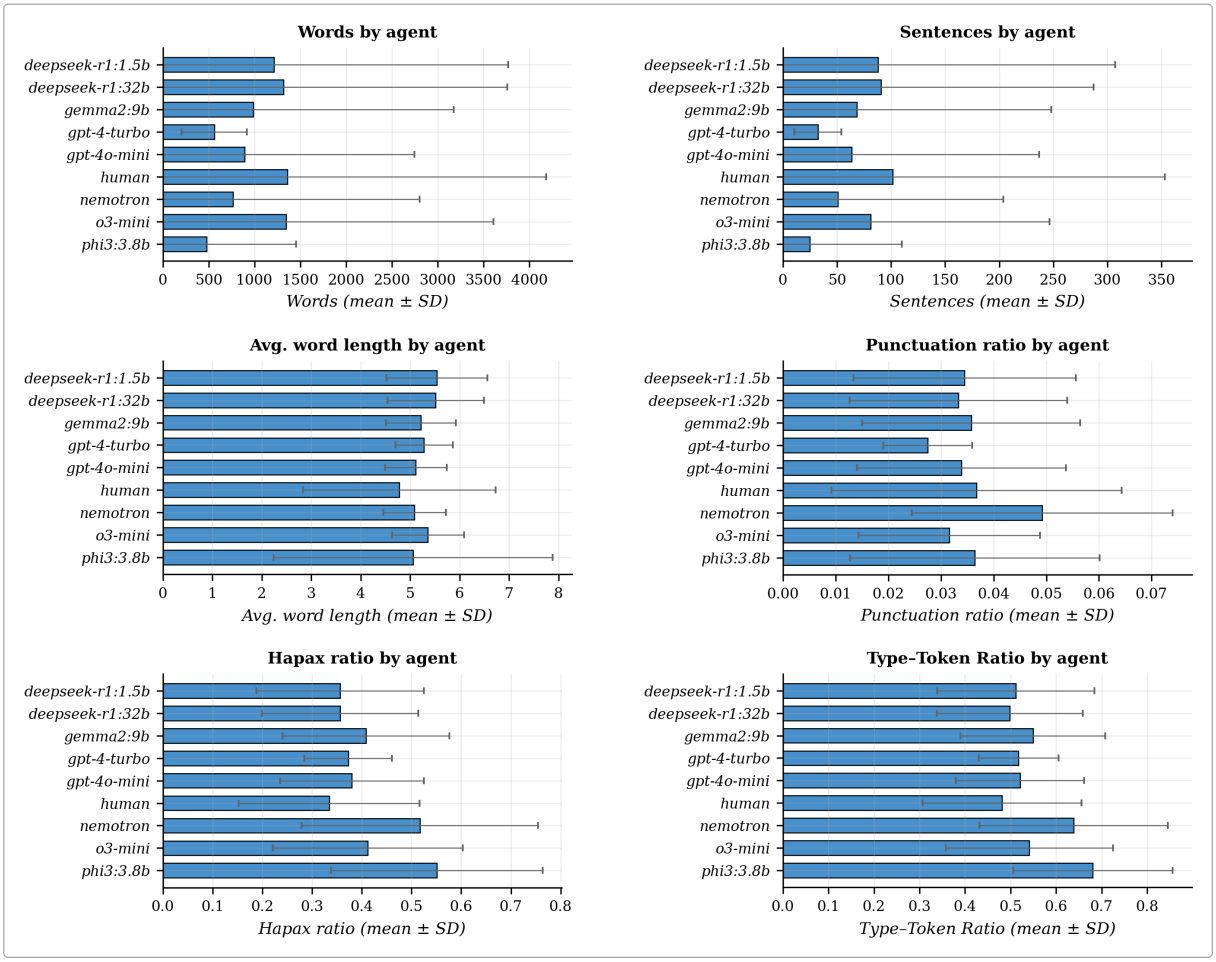
---

Is the "Face on Mars" Really a Face? Okay, so there's this picture, right? A picture taken by a spaceship way out in space, on Mars! And guess what? It kind of looks like a face staring back at us. Like, two eyes, a nose, and even a mouth! People started going crazy saying it was proof that aliens lived on Mars. But hold on! I think this whole "Face on Mars" thing is just people's imaginations running wild. First of all, look at the picture closely. [...]

## A.6. Additional Dataset Statistics




(a) Statistics (Section 3.4) for GhostWriter per *split* across various linguistic metrics.




(b) Statistics (Section 3.4) for GhostWriter per *agent* across various linguistic metrics.

Figure 7: Additional Dataset Statistics

## A.7. Additional Performance Details

Table 3: Baseline model performance by agent.

(a) Likelihood-based Models.

| | **DetectLLM$_{LLR}$** | | | | **Fast-DetectGPT** | | | | **Binoculars** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| deepseek-r1:1.5b | 0.291 | 0.341 | 0.384 | 0.616 | 0.420 | 0.692 | 0.647 | 0.391 | 0.447 | 0.688 | 0.610 | 0.329 |
| deepseek-r1:32b | 0.389 | 0.641 | 0.615 | 0.431 | 0.445 | 0.786 | 0.712 | 0.322 | 0.514 | 0.782 | 0.610 | 0.171 |
| gemma2:9b | 0.588 | 0.642 | 0.610 | 0.432 | 0.701 | 0.790 | 0.726 | 0.321 | 0.716 | 0.787 | 0.636 | 0.176 |
| gpt-4-turbo | 0.385 | 0.698 | 0.677 | 0.399 | 0.429 | 0.824 | 0.780 | 0.296 | 0.490 | 0.822 | 0.667 | 0.139 |
| gpt-4o-mini | 0.669 | 0.747 | 0.666 | 0.327 | 0.775 | 0.869 | 0.809 | 0.260 | 0.809 | 0.867 | 0.688 | 0.052 |
| nemotron | 0.298 | 0.394 | 0.482 | 0.615 | 0.424 | 0.697 | 0.652 | 0.387 | 0.467 | 0.705 | 0.639 | 0.296 |
| o3-mini | 0.325 | 0.465 | 0.506 | 0.545 | 0.393 | 0.667 | 0.639 | 0.409 | 0.408 | 0.662 | 0.614 | 0.375 |
| phi3:3.8b | 0.297 | 0.304 | 0.357 | 0.638 | 0.277 | 0.253 | 0.299 | 0.659 | 0.228 | 0.257 | 0.394 | 0.755 |
| **Overall** | 0.405 | 0.529 | 0.537 | 0.500 | 0.483 | **0.697** | **0.658** | 0.381 | **0.510** | 0.696 | 0.607 | **0.287** |

(b) Pre-Trained Supervised Transformer Models.

| | **RADAR** | | | | **E5-LoRA** | | | |
|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| deepseek-r1:1.5b | 0.539 | 0.761 | 0.581 | 0.159 | 0.288 | 0.749 | 0.869 | 0.663 |
| deepseek-r1:32b | 0.486 | 0.597 | 0.284 | 0.159 | 0.276 | 0.687 | 0.828 | 0.663 |
| gemma2:9b | 0.481 | 0.580 | 0.254 | 0.159 | 0.551 | 0.646 | 0.782 | 0.663 |
| gpt-4-turbo | 0.466 | 0.580 | 0.274 | 0.159 | 0.258 | 0.634 | 0.731 | 0.663 |
| gpt-4o-mini | 0.575 | 0.658 | 0.370 | 0.159 | 0.567 | 0.725 | 0.838 | 0.663 |
| nemotron | 0.528 | 0.736 | 0.497 | 0.159 | 0.287 | 0.635 | 0.839 | 0.663 |
| o3-mini | 0.475 | 0.578 | 0.238 | 0.159 | 0.269 | 0.649 | 0.793 | 0.663 |
| phi3:3.8b | 0.542 | 0.586 | 0.312 | 0.159 | 0.335 | 0.636 | 0.809 | 0.663 |
| **Overall** | **0.512** | 0.634 | 0.351 | **0.159** | 0.354 | **0.670** | **0.811** | 0.663 |

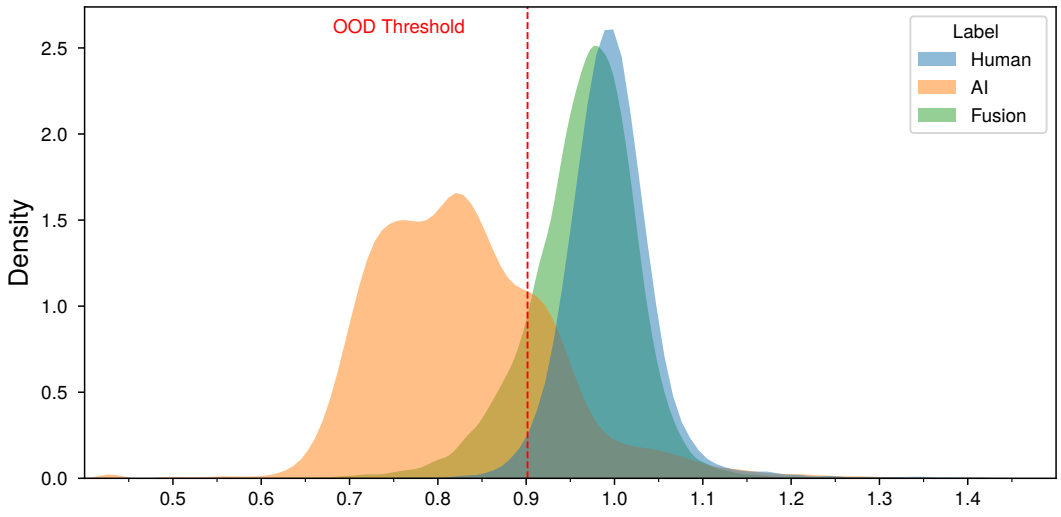



Figure 8: Kernel Density Estimation of Binoculars Scores on GhostWriter per Class across all Domains.

Table 4: Baseline model performance when excluding fusion texts.

(a) Likelihood-based Models.

| | **DetectLLM$_{LLR}$** | | | | **Fast-DetectGPT** | | | | **Binoculars** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR | F1 | AUC | TPR | FPR |
| arXiv | 0.865 | 0.923 | 0.853 | 0.088 | 0.934 | 0.990 | 0.984 | 0.136 | 0.947 | 0.991 | 0.925 | 0.001 |
| Blogs | 0.605 | 0.668 | 0.632 | 0.421 | 0.660 | 0.720 | 0.649 | 0.325 | 0.664 | 0.717 | 0.668 | 0.338 |
| Bundestag | 0.854 | 0.915 | 0.805 | 0.078 | 0.872 | 0.949 | 0.911 | 0.172 | 0.883 | 0.947 | 0.839 | 0.057 |
| CNN | 0.818 | 0.883 | 0.801 | 0.151 | 0.893 | 0.968 | 0.964 | 0.195 | 0.960 | 0.970 | 0.937 | 0.005 |
| ECHR | 0.646 | 0.691 | 0.672 | 0.371 | 0.820 | 0.935 | 0.923 | 0.301 | 0.886 | 0.945 | 0.852 | 0.056 |
| Essays | 0.876 | 0.934 | 0.886 | 0.132 | 0.902 | 0.976 | 0.971 | 0.179 | 0.969 | 0.977 | 0.949 | 0.004 |
| Gutenberg | 0.778 | 0.899 | 0.836 | 0.176 | 0.919 | 0.988 | 0.981 | 0.174 | 0.916 | 0.991 | 0.928 | 0.013 |
| HoC | 0.771 | 0.824 | 0.778 | 0.222 | 0.897 | 0.964 | 0.942 | 0.156 | 0.937 | 0.964 | 0.908 | 0.013 |
| **Overall** | 0.780 | 0.847 | 0.783 | 0.199 | 0.859 | 0.935 | **0.914** | 0.209 | **0.892** | **0.937** | 0.871 | **0.063** |

(b) Pre-Trained Supervised Transformer-based Models.

| | **RADAR** | | | | **E5-LoRA** | | | |
|---|---|---|---|---|---|---|---|---|
| **Domain** | F1 | AUROC | TPR | FPR | F1 | AUROC | TPR | FPR |
| arXiv | 0.405 | 0.839 | 0.150 | 0.004 | 0.570 | 0.965 | 0.994 | 0.809 |
| Blogs | 0.508 | 0.499 | 0.588 | 0.571 | 0.659 | 0.803 | 0.918 | 0.578 |
| Bundestag | 0.334 | 0.395 | 0.044 | 0.052 | 0.381 | 0.339 | 0.724 | 0.886 |
| CNN | 0.606 | 0.942 | 0.370 | 0.019 | 0.647 | 0.968 | 0.997 | 0.679 |
| ECHR | 0.620 | 0.661 | 0.574 | 0.306 | 0.406 | 0.877 | 0.993 | 0.971 |
| Essays | 0.602 | 0.754 | 0.405 | 0.123 | 0.540 | 0.846 | 0.996 | 0.807 |
| Gutenberg | 0.357 | 0.753 | 0.215 | 0.084 | 0.880 | 0.969 | 0.937 | 0.149 |
| HoC | 0.624 | 0.773 | 0.504 | 0.188 | 0.838 | 0.990 | 0.999 | 0.360 |
| **Overall** | 0.483 | 0.698 | 0.319 | **0.150** | **0.604** | **0.821** | **0.878** | 0.604 |