The Two Paradigms of LLM Detection: Authorship Attribution vs. Authorship Verification

Janek Bevendorff,[*,†] **Matti Wiegmann,**[†,‡] **Emmelie Richter,**[†]
Martin Potthast,[‡,§] **Benno Stein**[†]
[*]Leipzig University, [†]Bauhaus-Universität Weimar,
[‡]University of Kassel, [§]hessian.AI and ScaDS.AI

Abstract

The detection of texts generated by LLMs has quickly become an important research problem. Many supervised and zero-shot detectors have already been proposed, yet their effectiveness and precision remain disputed. Current research therefore focuses on making detectors robust against domain shifts and on building corresponding benchmarks. In this paper, we show that the actual limitations hindering progress in LLM detection lie elsewhere: LLM detection is often implicitly modeled as an *authorship attribution* task, while its true nature is that of *authorship verification*. We systematically analyze the current research with respect to this misunderstanding, conduct an in-depth comparative analysis of the benchmarks, and validate our claim using state-of-the-art LLM detectors. Our contributions open the realm of authorship analysis technology for understanding and tackling the problem of LLM detection.

1 Introduction

Generative AI is everywhere. From writing assistants to the generation of complete documents, texts curated or written by large language models (LLMs) can now be found in practically all social structures: LLMs are used in online media (Knibbs, 2024), in education (Adeshola and Adepoju, 2024), for scientific publications (Glynn, 2024; Lund et al., 2023; Picazo-Sanchez and Ortiz-Martin, 2024), for drafting Wikipedia articles (Brooks et al., 2024; Wikipedia, 2025), and more. Owing to the many risks of LLM text generation (Bommasani et al., 2021; Oviedo-Trespalacios et al., 2023), research into LLM detectors has sparked great interest.

So far, it seems difficult to ascertain the real-world effectiveness of current detectors. On the one hand, new detectors quite regularly "beat" the latest benchmarks (King et al., 2023b; Dugan et al., 2024b; Bevendorff et al., 2024b), leading us to believe the problem is easily solved, at least within fixed domains. On the other hand, serious concerns are voiced about the readiness of the technology. Perhaps most strikingly, OpenAI shut down their own LLM detector just six months after its launch, citing a lack of accuracy (Kirchner et al., 2023).

In this paper, we show that the conflicting reports of success or failure are due to a mix-up of two paradigms of the related research field of authorship analytics: *authorship attribution* and *authorship verification* (Koppel and Schler, 2004). As the authorship analytics and LLM detection communities have developed almost independently until now, neither could learn from the other.

Framed as authorship attribution, LLM detection considers LLMs and humans as two collective "authors," each possessing their own distinct writing style. LLM detection is thereby cast as a closed-set binary classification problem, relying on both classes to be sufficiently discriminative. We show that LLM detection is more realistically cast as an authorship verification problem—i.e., as an open-set one-class classification problem (Schölkopf et al., 2001; Manevitz and Yousef, 2001). However, our literature review and data analyses show that LLM detectors are often developed with attribution in mind, but are evaluated under much broader assumptions on inadequate benchmarks, leading to unnecessary Type II errors in evaluations.

Our main contributions are as follows: (1) We reconcile authorship analytics and show that LLM detection is an authorship verification task (Section 3). (2) We conduct a corpus analysis on how LLM text differs from human text *today* (Section 4). (3) We show that state-of-the-art LLM detectors do indeed solve attribution well (Section 5). (4) We conclude by discussing the quality and adequacy of existing benchmarks, and how detectors need at least special evaluation considerations for successful verification, even within domains (Section 6).[1]

[1]Code and data: https://github.com/webis-de/ACL-25.

2 Related Work

LLM detection is a young but very popular research topic. Hundreds of papers and several surveys (Jawahar et al., 2020; Crothers et al., 2023; Yang et al., 2023; Tang et al., 2024; Wu et al., 2025), multiple shared tasks (Merkhofer et al., 2023; King et al., 2023a; Molla et al., 2023; Sarvazyan et al., 2023a; Bevendorff et al., 2024a; Wang et al., 2024d, 2025), and an increasing number of large-scale benchmark datasets (Uchendu et al., 2021; Macko et al., 2023; Su et al., 2023b; Yu et al., 2023; Dugan et al., 2024a; Li et al., 2024; Verma et al., 2024; Wang et al., 2024e,c) have been contributed. Initial work on authorship attribution for language models (Uchendu et al., 2020) found GPT-2 was already more difficult to detect than previous models. Both LLMs and LLM detectors have come a long way since. Below, we review the most relevant work.[2]

LLM Text Differs from Human Text Although it is unclear how exactly LLM text differs from human text, observations include that LLMs often lack lexical diversity (Reviriego et al., 2023; Martínez et al., 2024), overuse certain adjectives (Kobak et al., 2024; Liang et al., 2024), and produce longer, more complex sentences (Su and Wu, 2024; Zanotto and Aroyehun, 2024). Nucleus sampling (Holtzman et al., 2019) has enabled the generation of text adhering better to human text characteristics, reducing aberration in Zipf's and Heaps' law (Meister and Cotterell, 2021; Lai et al., 2023). Today, LLMs follow Zipf's law even for invented languages (Diamond, 2023), but possess stylistic fingerprints (Kumarage and Liu, 2023; Zahid et al., 2024; McGovern et al., 2025) and memorize patterns from the training data (Shaib et al., 2024).

Humans as Detectors Humans' ability to recognize LLM text depends on their prior experience. As LLMs advanced, the basic heuristics applied by humans became inaccurate (Ippolito et al., 2020; Clark et al., 2021; Jakesch et al., 2023) so that LLM text can even be perceived more human-like than human text (Rathi et al., 2025). Tools like GLTR (Gehrmann et al., 2019) and SCARECROW (Dou et al., 2022) can significantly improve average human performance, but Russell et al. (2025) found that experienced ChatGPT users have near-perfect detection rates, outperforming automatic detectors.

[2]We omit text and LLM watermarks here (Kamaruddin et al., 2018; Kirchenbauer et al., 2023), which would go beyond the scope of this paper.

Supervised LLM Detection A popular method for LLM detection is to train supervised binary classifiers. An early approach was OpenAI's RoBERTa-based GPT-2 detector (Solaiman et al., 2019). This, as well as more recent approaches, are usually found to be highly accurate in-domain, but do not generalize across domains, nor to newer LLM families (Rodriguez et al., 2022; Elkhatat et al., 2023; Sarvazyan et al., 2023b; Wang et al., 2024e; Dugan et al., 2025; Pudasaini et al., 2025), nor to language learner texts (Liang et al., 2023). Fine-tuned versions of BERT, T5, or smaller LLMs are frequently cited as state-of-the-art detectors in the literature (Macko et al., 2023), e.g., Ghostbuster (Verma et al., 2024) and GPT-Sentinel (Chen et al., 2023). Besides classifying entire documents, more fine-grained approaches are investigated as well (Tolstykh et al., 2024). Detectors with fewer parameters tend to generalize better to unseen LLM families (Mireshghallah et al., 2023), and even symbolic approaches can achieve state-of-the-art results (Su and Wu, 2024; Lorenz et al., 2024).

Zero-shot LLM Detection Many zero-shot detectors use reference LLMs to measure token perplexities: DetectGPT (Mitchell et al., 2023) measures perplexity degradation due to text perturbations. Likely unbeknownst to the authors, it bears a resemblance to Koppel and Schler's (2004) unmasking. Fast-DetectGPT (Bao et al., 2023) reduces DetectGPT's computational costs by estimating perturbations, and DetectLLM (Su et al., 2023a) adapts it to use log ranks for accuracy and speed. LLMDet (Wu et al., 2023) calculates perplexity from n-gram profiles for each LLM. Binoculars (Hans et al., 2024) addresses the limited generalization capabilities of previous models by calibrating cross-entropy scores with a second LLM. Other approaches use GANs (Hu et al., 2023), token compression ratios (Dubois et al., 2024)—similar to compression models in authorship analytics (Oliveira et al., 2013; Halvani et al., 2017)—, they prompt LLMs directly (Bhattacharjee and Liu, 2024; Huang et al., 2024), or measure the similarity of responses to reverse-engineered queries (Baradia et al., 2025) or paraphrasing prompts (Mao et al., 2024; Hao et al., 2024) with reference texts.

Robustness Besides the mentioned domain adaptation issues, detectors are also vulnerable to adversarial attacks. Paraphrasing attacks (Krishna et al., 2023; Sadasivan et al., 2023) or basic synonym replacements (Wang et al., 2024a) are effec-

tive. Search for adversarial (soft) prompts (Kumarage et al., 2023; Shi et al., 2024) or downstream "humanizer" models (Wang et al., 2024b) can further reduce the detection accuracy. These attacks against detectors are very similar to authorship obfuscation approaches (Bevendorff et al., 2019a; Mahmood et al., 2019), a fact outlined only by Uchendu et al. (2023) so far. It is known that even manual style obfuscation by untrained writers is quite effective (Brennan and Greenstadt, 2009). However, obfuscations leave detectable traces (Juola, 2012; Mahmood et al., 2020), which have been observed for LLM obfuscators, too (Masrour et al., 2025).

3 LLM Detection as a Verification Problem

LLM detection, albeit a relatively new task, shares many theoretical foundations with the analysis of human authorship and writing style (Juola, 2007; Stamatatos, 2016; Tyo et al., 2023). Our work aims to bridge the gap between the two communities.

3.1 Premise: LLMs Will Become Authors

The core assumption of LLM detection, as it is often practiced today, is that the language distributions of human-authored and machine-generated text are sufficiently distinct (as a consequence of the technology or training data) and that a separating hypersurface can be learned. Put simply, a representative sample of human writing is conceived of as one single "author," which is compared to one, or perhaps multiple, machine "authors."

Under this assumption, Sadasivan et al. (2023) postulate a highly cited theoretical upper limit of the achievable detection efficacy. Their argument hinges on the assumption that a sufficiently advanced paraphraser will eventually make machine text indistinguishable from human text if only it moves the machine text close enough in the direction of human text by reducing the total variation distance between the two distributions. This argument is inherently tautological and can be generalized to any classification task, not just LLM detection. But more importantly, its premise does not hold for two reasons: First, the total variation distance depends on the feature space used for representation, yet there is no mention of the nature of the two distributions. The argument thus turns into a blanket statement about a hypothetical singular feature of "text." Second, any variability at the individual level within the distribution of the human author is disregarded, assuming a uniformity of human text, which does not exist in practice.

We argue instead that—as LLMs become more human-like—their detection will not necessarily become impossible, but it will increasingly resemble a human authorship classification task. In this sense, the feasibility of LLM detection is determined by the frontiers of authorship analytics.

3.2 Background: Attribution vs. Verification

Computational authorship analytics distinguishes between two basic scenarios: (1) authorship attribution and (2) authorship verification (Koppel and Schler, 2004; Koppel et al., 2009). Authorship attribution classifies the author of a disputed text from a closed set of candidates. Authorship verification solves the more general one-class problem of whether two texts were produced by the same process. In simple terms, attribution asks: "Who among these candidates is the author?" whereas verification asks: "Do these two writing samples look similar enough?"

Authorship attribution with few (< 20) candidate authors is typically highly effective, even if only short writing samples are available and classical machine learning techniques are applied. However, with growing numbers of candidates, authorship attribution deteriorates into a "needle in a haystack" situation, as Koppel et al. (2009) call it. Yet, with certain modifications (e.g., if the classifier can "opt out" of deciding; see Koppel et al. (2006)), it can be scaled up to several thousand candidates.

By contrast, authorship verification seeks to clarify unknown authorship by comparing disputed documents only to text samples from a single known author, ignoring the indeterminate negatives. Verification thus poses a more difficult one-class classification problem. Its most salient characteristic is that the negative class (all human texts not written by the known author) cannot feasibly be collected into a representative sample. However, typically more data and—very importantly—more attention to the quality and composition of the text samples are required to avoid over-optimistic results (Bevendorff et al., 2019b).

3.3 Conclusion: LLM Detection Cannot be Addressed by Attribution

Attribution, so far, appears to work well for LLM detection, since (at the time of writing) LLM-generated text often lies far outside the distribution of *any* typical human text. It therefore seems suf-

ficient to compile a corpus with samples for the two classes (Human, LLM) and train a basic classifier. With advancing LLM technology, however, accurate modeling of the human majority class will become increasingly difficult, as explained below.

Consider a situation in which a disputed text could have been written by either of two candidates, a human author A or a (hypothetical) perfectly human-like LLM. Regardless of the LLM's human-like capabilities, we can realistically tackle this two-class problem using state-of-the-art authorship attribution technology. Consider further a text from a third (human) author B, $B \neq A$. With the classifier trained to distinguish A from the LLM, B would always be incorrectly assigned to either A or the LLM; yet, as long as any text B is classified as A, this solves the LLM recognition task, but not if B is classified as LLM. To address this error, one has to narrow the hypersurface around the LLM class to exclude B. This can be achieved by building a new corpus with samples from A, B, and the LLM, and reiterating the issue now as a three-class attribution problem.

This strategy does not scale: The number of human candidates classified as LLM will asymptotically approach the rest of humanity; the increasing number of LLM variants exacerbates the problem. Not immediately obvious, however, is that this scaling strategy is actually being pursued in the current research: In an attempt to capture the diversity of human and machine writing styles, ever larger corpora are compiled, hoping to obtain a comprehensive gold standard for binary LLM attribution. But, the larger the compiled corpora, the more the similarity of styles between human individuals will decrease (due to the increasing number of authors) and the more the similarity of styles between humans and LLMs will increase (due to technological progress). The *learnable discriminatory power* can hence only decrease with the corpus size. We can observe some of this already today, as small text domain shifts (Wang et al., 2024e; Dugan et al., 2024a, 2025), trivial paraphrasing attacks (Rodriguez et al., 2022; Krishna et al., 2023; Sadasivan et al., 2023; Shi et al., 2024), or just newer LLMs (Elkhatat et al., 2023; Pudasaini et al., 2025) immediately translate into classification errors.

Reliable cross-domain authorship attribution and verification has been a long-standing research issue (Kestemont et al., 2012). As newer LLMs produce fewer pathological language artifacts, LLM detection will face the same problem. This, too, can already be observed with OpenAI's latest o1 model, as we show in the following sections.

4 Properties of LLM Text

We know that LLM attribution is considered (relatively) effective today, so we use this section to illustrate how texts from LLMs and humans differ at the surface level. We highlight certain properties of recent models that render them increasingly human-like. These observations align with our theory from the previous section that LLM detection is becoming an authorship verification task. However, more research will be necessary to compile further evidence in support of our hypothesis.

4.1 Datasets and Preprocessing

We use the following six popular and publicly available LLM detection datasets for this analysis: **PAN'24** (Bevendorff et al., 2024a), **Human Detectors** (HD) (Russell et al., 2025), **Ghostbuster** (GB) (Verma et al., 2024), **RAID** (Dugan et al., 2024a), **MAGE** (Li et al., 2024), and **M4** (Wang et al., 2024e). See Appendix E for dataset statistics. The first three are smaller and genre-controlled with texts from only one or two domains (news, essays). The latter three are much larger and contain many different genres. M4 and RAID would allow building individual subsets which are also genre-controlled, but we decided to use them as a whole (which is more in line with how a typical leaderboard would be evaluated).

Most datasets contain at least GPT-3.x, GPT-4, Llama2, and Mistral texts. For better coverage of newer models, we extended the PAN'24 dataset with texts from GPT-4o, 4o-mini, and OpenAI o1 (using the original prompting method).

To reduce the number of confounders, we preprocessed all datasets and removed extremely short texts, texts with adversarial attacks (obfuscations), and low-quality texts generated by older models (e.g., GPT-2 and Alpaca). Many texts in M4 still contained the prompts (BLOOMz texts were 50 % prompt), which we removed heuristically. We also dropped the "peer review" genre, due to its inconsistent schema and many short texts.

4.2 Observations

Based on these datasets, we made the following three key observations.

Obs. 1: LLMs Use Complex Language It has been shown that LLMs tend to use longer sen-

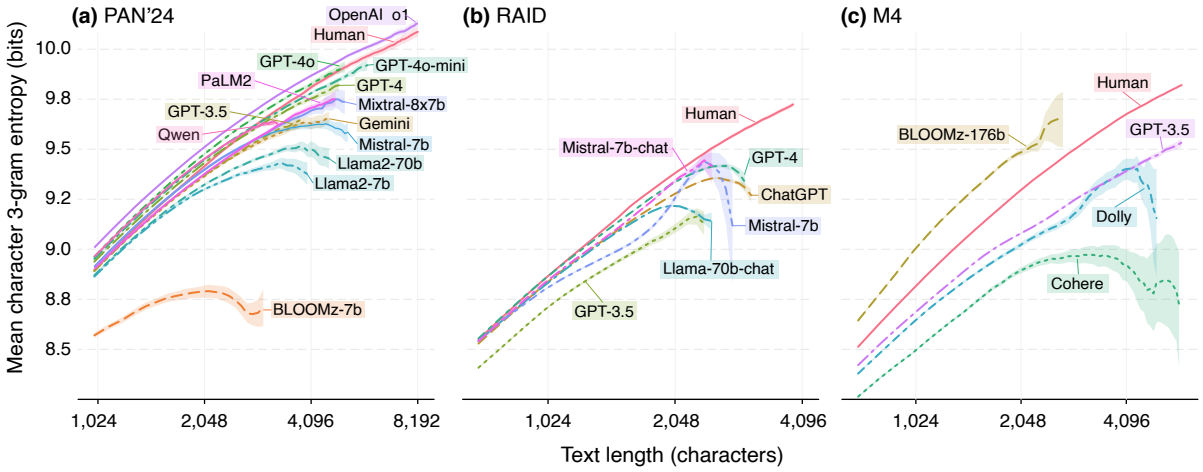

Figure 1: Mean character 3-gram entropy over increasing text length with 95 % confidence intervals. Shown are texts from the **(a)** PAN'24, **(b)** RAID, and **(c)** M4 datasets. Curves diverge after around 2,500–4,000 characters. LLM entropy is consistently lower than human entropy, except for GPT-4o, OpenAI o1, and BLOOMz-176b.

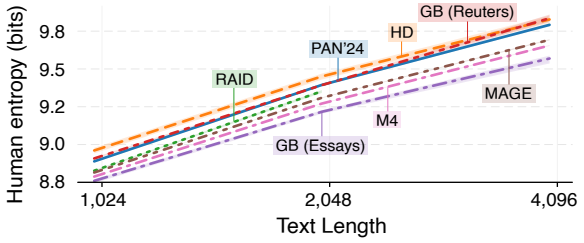

Figure 2: Mean human character 3-gram entropy on all datasets. PAN, HD, and GB (Reuters) are very similar, whereas student essays are much lower. M4, MAGE, and RAID with mixed genres are in between.

tences (Su et al., 2023b) and lexically and structurally more complex language (Guo et al., 2023; Su and Wu, 2024), or overuse certain terms such as "commendable," "innovative," or "meticulous" (Liang et al., 2024; Gray, 2024).

We confirm these findings by measuring the mean Flesch reading ease scores. On PAN'24 and GB, we indeed find significantly lower readability scores for LLM texts with (very) large effect sizes (Cohen's $d \gg .8$) for almost all models. Some GPT-3 prompts in GB have effect sizes of more than three standard deviations ($t(1{,}610) = 78.3$, $p < .001$, $d = 3.9$). Differences in the large datasets (M4, RAID, MAGE) are significant, but with negligible effect sizes, likely due to the genre mix. HD has problems with per-LLM sample sizes but shows an effect for at least some models.

The average word length in human texts across all datasets is 5.1 ± 0.1 characters, which matches the expected value for English (Wolfram|Alpha, 2025). Words in LLM texts are significantly longer (≥ 5.4) in PAN'24, HD, and GB. The most extreme example is GB with 6.1 characters ($t(1{,}610) = 82.5$, $p < .001$, $d = 4.11$). On these datasets, word length alone would yield remarkable detection accuracy (a finding we have not seen mentioned clearly in the literature yet). Effect sizes in the mixed-genre datasets are again negligible. RAID and MAGE also have shorter words for both classes (4.8 ± 0.1 for humans, 5.0 for LLMs).

While the text lengths are genre-dependent, we observe that human texts generally have a long-tailed distribution (see Appendix F). LLMs, on the other hand, have narrow stopping windows, the only exception being o1 in PAN'24. RAID has by far the shortest texts in both classes.

Obs. 2: LLMs are Low-entropy Writers Despite LLMs using longer words and sentences, their lexical diversity remains lower than standard human text. In most datasets, the mean human type-token diversity, as measured by the length-invariant MTLD measure (McCarthy and Jarvis, 2010), is significantly above that of any of the models. The absolute means are genre-dependent (e.g., the mean human MTLD of the student essays in the GB dataset is lower than that of the news articles from PAN), but in either case, the corresponding LLM texts are below that. Only the newest models (GPT-4, 4o, and o1) have MTLD values above that of humans, some even by a large margin (o1 in PAN: $t(1{,}964) = 50.3$, $p < .001$, $d = 2.27$). That does not mean other LLMs could not also produce richer vocabulary. Some prompts in the GB dataset also enabled GPT-3.5 to increase its vocabulary beyond the human texts. Given the minor variations between the prompts (from our understanding of

the dataset's source code), this is surprising, but LLMs being able to adjust aspects of their style has been described before (Malik et al., 2024). Of course, at the other end of the quality spectrum, LLMs may boost their vocabulary richness also by producing a large amount of random gibberish, which we will discuss in more detail later.

To analyze LLM text diversity further, we calculated the Shannon entropy of the character 3-gram distributions at increasing length cutoffs. Character n-grams are a robust feature frequently used in classical computational authorship analysis (Stamatatos, 2013). By ergodicity of Heaps' law, the mean entropy distribution should have finite variance and follow a monotonic curve with language-dependent parameters bounded from above by $-\log \frac{1}{n-2}$ for any texts of length $n \in \{3, 4, \ldots\}$. We grouped texts by model, removed outliers beyond the $1.5 \times \mathrm{IQR}$ range, and varied the length cutoff from 700 characters until fewer than 30 texts were left for each model. Figure 1 shows the curves from a selection of models from the (extended) PAN dataset, RAID, and M4.

After about 2,000 characters, we start seeing a clear separation of the means. The entropy scores differ primarily by family (Mistral scores higher than Llama2, GPT scores higher than Mistral) and within families by parameter count (Llama2-70b scores higher than Llama2-7b, GPT-4o scores higher than 4o-mini, etc.). GPT-4o is the first model to maintain a mean entropy consistent with human text even beyond 4,000 characters, which is the maximum length most models created. OpenAI o1 is the only model to generate longer texts (on average 5,000 and up to 19,345 characters) while maintaining a consistently higher-than-human entropy. The o1 texts from the HD dataset display the same behavior. Different means alone are not sufficient for good class separation but at 4,000–5,000 characters, PaLM2 and the open source models can be separated (Appendix B). Of the models tested, BLOOMz-7b produced the shortest texts and scores by far the lowest. Interestingly, GPT-4o-mini (said to have around 8 bn parameters) even outscores Gemini, GPT-3, and GPT-4 (Turbo).

Apart from the newest closed-source models, LLMs struggle to produce sufficiently many new words and characters to match the entropy of human text. Particularly the open source models appear to reach a high point shortly before their maximum generation length, after which the curve slopes down again. This can be explained in two ways: (1) Some LLMs (especially with quantization and poorly chosen sampling parameters) have a tendency to get stuck in a generation loop, resulting in a flat entropy curve towards the end. (2) Overall repetitive or looping texts are longer than more coherent and diverse texts. In sum, this results in a downward slope of the mean curve for some models, especially in the RAID dataset, where particularly the Mistral (non-chat) texts degenerate profoundly towards the end (see Figure 1b).

Character n-gram entropy alone is a weak discriminator, but it elicits important insights about the data: (1) it gives an upper bound for the minimum text length required for separation. (2) It establishes a baseline for how human text is distributed. The absolute means, however, are not universal and depend on the dataset and genre. Figure 2 plots the mean human entropy for each dataset. HD and PAN both consist of a smaller number of hand-curated news articles and high-quality generations thereof. Their curves are nearly identical. The same applies to the GB Reuters texts. The GB student essays, on the other hand, have significantly lower entropy (which makes sense given the genre) and would thus be misclassified as machine-generated had we trained a model on only the curves of news articles. RAID, M4, and MAGE, which mix a multitude of different text types and genres, are in between. *Such a regression toward the mean has been a consistent trend for the large datasets in our analyses.*

As alluded to above, being more human-like is not the only strategy for higher entropy. Sampling strategies play a crucial role here. Above-human entropy can also be achieved by repetition penalties or simply by writing nonsensical gibberish, either due to either unstable logit distributions or a high temperature setting. We see a bit of both in certain models. Unlike its smaller sibling, BLOOMz-176b in M4 (Figure 1c) reproduces several semi-accurately extracted web pages, LaTeX PDFs, or long strings of nouns or numbers (Appendix C.3). We find this also for GPT-J and GPT-Neo in MAGE. Lai et al. (2023) noticed that GPT-Neo had a larger vocabulary than humans, which would be explained by this. Mistral texts in RAID with a high repetition penalty also have high entropy, yet still collapse toward the end, which hints at output distribution problems at a deeper level (Appendix C.1).

Obs. 3: LLM Language Variance is Inconsistent For a deeper understanding of the text distributions beyond entropy, we ran Koppel and Schler's

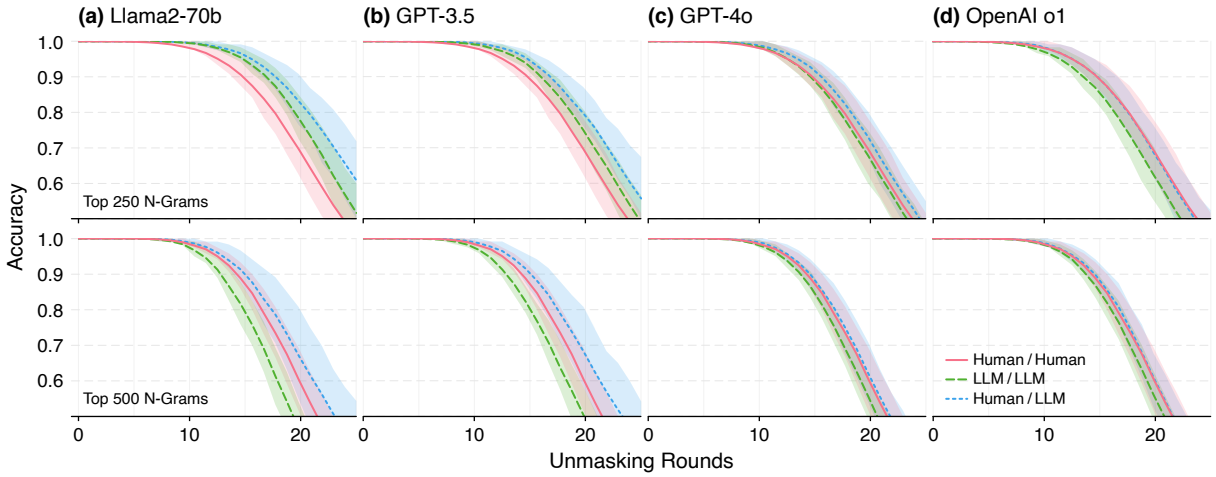

Figure 3: Median authorship unmasking curves using the 250 (top row) or 500 (bottom row) most-frequent character 3-grams for 200 *Human / Human* (same in all graphs), *LLM / LLM*, and *Human / LLM* text pairs for selected models drawn from the extended PAN'24 dataset. The shaded areas indicate the 50 % IQR. Llama2 and GPT-3.5 are very inconsistent by being unnaturally discriminable in the top 250 alone and yet very self-similar in the top 500 3-grams. GPT-4o and, particularly, OpenAI o1 are more consistent by being more similar to themselves in both feature sets than the median of human text pairs and about as dissimilar to human texts as other human texts would be.

(2004) authorship unmasking algorithm in the oversampling variant for short texts (Bevendorff et al., 2019c) on the extended PAN'24 dataset. We used only texts with at least 3,000 characters and to increase the vocabulary overlap, the texts were lowercased and stemmed. We paired texts randomly as (1) *Human / Human* (most or all different authors), (2) *LLM / LLM*, and (3) *Human / LLM* and sampled 60 chunks for each pair with replacement. As features, we chose the relative frequencies of the 250 (then afterward 500) most-frequent character 3-grams in both texts. We then ran 25 iterations, removing the 4 (resp. 8 for 500 features) most discriminative positive and negative features in each.

In Figure 3, we show the results for four selected models for 250 and 500 features. We chose Llama2-70b as representative of a family of (older) open-source models, GPT-4o and OpenAI o1 as state-of-the-art, and GPT-3.5 Turbo in between. The graphs show the curve medians for 200 text pairs (131 for GPT-3.5) with interquartile ranges.

We compared the results to a simple compression baseline (Sculley and Brodley, 2006; Halvani et al., 2017). All LLMs have significantly lower (i.e., better) compression-based cosine (CBC) scores when paired with themselves than any humans paired with other humans (Bonferroni-corrected $p < .001$ with effect sizes $0.7 \leq d \leq 1.7$).

The unmasking curves, however, support high self-similarity only in the large top-500 3-gram set. In the top-250 subset, most models are actually more dissimilar to human texts and even to themselves, than two different human authors would normally be. This inconsistency is pronounced the most in smaller and older models. Larger and newer models tend to be more consistent with lower variance. GPT-4o is the first model with a behavior consistent with natural human text. OpenAI o1 has yet lower variance and even higher self-similarity in the top 250. While this is not sufficient evidence to conclude that o1 has developed a distinct "style," it at least indicates a higher degree of linguistic consistency than in the other models. More curves and the CBC distributions can be found in Appendix G.

5 LLM Detector Evaluation

We analyze the behavior of three state-of-the-art detector models: (1) a re-implementation of the Binoculars (Hans et al., 2024) zero-shot detector, (2) a set of SVMs trained on 800 TF-IDF word unigrams, and (3) a LoRA-fine-tuned Mistral-7B.[3] All three were used in the top submissions to PAN'24 (Lorenz et al., 2024; Tavan and Najafi, 2024). We trained multiple SVMs on several balanced splits of the PAN dataset, each containing randomly sampled texts from humans and *one* LLM family. We tested the models both in-domain, on identically constructed holdout splits, and out-of-domain, on mixed random samples from all datasets (with an equal number of samples from small and large datasets). We trained two versions of the Mistral model for binary classification on un-

[3] 20M trainable parameters. Trained on 3 A100 for 2×12h.

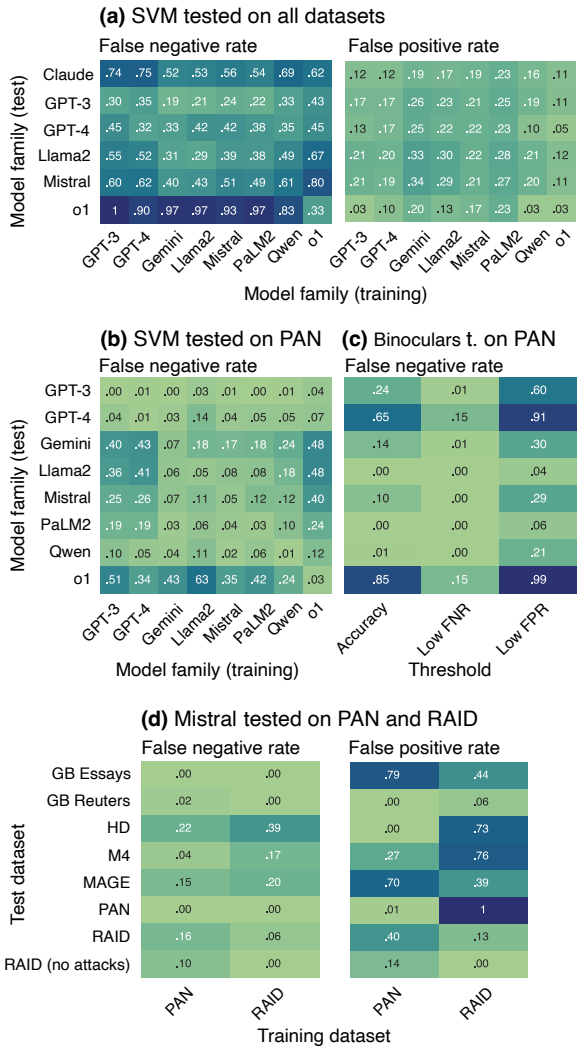

Figure 4: **(a)** FNR and FPR of SVM-based detectors with TF-IDF features trained on class-balanced samples of individual model families from the PAN dataset and tested on mixed samples from all other datasets. FPR variations within columns are due to different per-model sample sizes (ranging from 60–8,576). **(b)** FPR of the same detectors tested on a PAN holdout set. **(d)** FPR of Binoculars on PAN. **(d)** FNR and FPR of Mistral-7b, fine-tuned for LLM detection on PAN or RAID.

balanced splits of PAN (976 human, 16,587 LLM) and RAID (11,740 human, 186,319 LLM).

The false negative rates (FNR) of the SVM LLM-detector are shown in Figure 4b. As expected, in-domain detection works best (0–7 % FNR on the diagonal). False positive rates (FPR, not shown) are 0–3 % for all models except Gemini (5 %). Training on GPT-3 and 4(o) texts generalizes to detecting either. Both can also be detected by detectors trained on other models' texts, but not vice versa. Most strikingly, training on o1 texts, we can detect o1, yet no other detector recognizes it. The o1 detector generalizes to texts of other models similarly to the GPT-3 and 4 detectors.

We see a similar picture for the mixed dataset but at worse detection rates. Detectors trained on the same model fare slightly better, but o1 is practically not recognized by any non-o1 detector. FPRs (Figure 4a) are 10–20 % for all but the o1 detector, which is less likely to misclassify human text than any other detector ($t(38) = 4.4$, $p < .001$, $d = 2.1$). However, the lower FPR translates into a higher FNR ($t(38) = 2.2$, $p = .034$, $d = 1.1$).

Binoculars manages to keep its FPR close to zero on all datasets, but suffers in terms of recall (TPR), especially for GPT-3 and 4 (Figure 4c). The detection fails completely on the o1 texts at both operating points (accuracy and low FPR). We tried a third threshold ("low FNR") of 0.97, which brings the FNR down to zero at the cost of a 40 % FPR.

The SVM-based detectors are biased towards the human class with lower FPR. The Mistral-based detector seems to work the other way round and prioritizes a low FNR (Figure 4d). Otherwise, the results are very similar, which shows that also the difficult RAID dataset can be learned extremely well (at least if we remove the shortest and lowest-quality texts). Training on PAN seems to generalize better to other datasets than training on RAID, but the two seem to be incompatible.

6 Discussion

The results from our exploratory data analysis in Section 4 and model evaluation in Section 5 bring us to an important debate about the quality and appropriateness of the data and the conclusions we can draw from models trained and evaluated on it.

6.1 Detection Needs Sound Benchmarks

We saw in Section 4.2 that there is a non-zero length limit below which texts are just indistinguishable. Yet, RAID, in particular, contains many extremely short texts. Several machine texts are only a single character, which cannot be reasonably classified. The shortest human text is 71 characters. The shortest paraphrased human text is "..." (three dots). In total, 3,613 texts are shorter than 200 characters (46,240 counting obfuscations), which is a relevant number given a total of 13,371 human texts. We found similar examples in M4 and MAGE. Moreover, several texts were high-entropy but low-quality. Such texts may fool detectors but are at best unpleasant to read.

This questions what these benchmarks actually measure: Can a 71-character text be uniquely hu-

man? Are automatically paraphrased texts still human? Do we even *need* a classifier for bad text or high-entropy nonsense? The number and diversity of texts in large-scale benchmarks are meant to make the benchmarks more robust. However, given the one-class nature of the problem (Section 3), we argue that benchmarks can never be truly representative. On the contrary, the uncontrolled variety and lower quality of examples water down the statistical power for finding "useful" detectors and increase the Type-II error rate disproportionately. A specialized, high-precision GPT-4 news article detector (which would be of great practical use) would not fare well. On the other hand, a detector mastering every aspect of the benchmark may still not generalize well to unseen genres and authors.

6.2 Evaluation Metrics Need Calibration

Dugan et al. (2024a), the authors of RAID, state (quite correctly) that maintaining a low FPR for critical detection tasks may be more desirable than high accuracy. In the context of LLM detection, they attribute this paradigm to Krishna et al. (2023) and Hans et al. (2024). However, the idea is older than that and forms the core of the well-known ROC analysis. However, what Dugan et al. (and any of the detectors today that we know of) failed to consider, is that a low FPR does not have to come at the expense of a low accuracy if non-answers are a valid third option. This third option is used frequently in authorship analytics. E.g., Koppel et al. (2006) uses a meta learner to decide when to answer and Bevendorff et al. (2019c) use SVM hyperplane distances to calibrate precision thresholds. Peñas and Rodrigo (2011) proposed c@1, an accuracy metric that considers non-answers.

On the other hand, ROC analysis, as it is used today in many studies on LLM detection, has only limited informative value. First, by reducing the TPR-FPR curve of ROC to a single AUROC number, the trade-off information of ROC is lost and—in the absence of a fixed threshold—the metric can turn into wishful thinking of what could have been. Second, FPR and TPR are independent of class prevalence. This aspect of ROC may be desirable. However, it can also paint an overly optimistic (sometimes also pessimistic) picture of the evaluation results' positive or negative predictive values (i.e., precision of either class) for highly imbalanced classes—something we often see in LLM detection. A precision-recall curve (Appendix A) would be a suitable tool here, but in either case, a correct calibration of acceptable FPR / FNR thresholds to the use case at hand is crucial, rather than blind reliance on averaged measures in which both are assumed equally important.

6.3 Results Need Interpretation

So where does this leave us in terms of the classification results from Section 5? Whether an experiment failed or succeeded, comes down to more than just its accuracy score. None of the detectors yield perfect results, but some are more useful than others. Clearly, none of the classifiers learned anything about the "nature of LLMs." The supervised classifiers all tried to learn individual aspects of the datasets but got much lower scores on unfamiliar test data. However, text genre was not the only issue. Sufficiently different models, such as o1, within the same domain also lead to wrong attributions, as unseen LLMs are not necessarily more similar to their robot colleagues than to human authors. This is exactly the core of our thought experiment in Section 3.3. To that extent, the Mistral model is probably the least useful of all, as it (very poorly) tries to model the human class, leading to an unnecessarily high FPR. The Falcon-based Binoculars was the most robust candidate with the lowest FPR, but it could not identify the output of the newer OpenAI models. From a pure attribution standpoint, we might say the model failed to generalize and misrepresented the negative class. Yet, we could also conclude quite the opposite: It worked exactly as intended and is, in fact, an effective Falcon LLM verifier.

7 Conclusion

In this work, we delineated the parallels between authorship analytics and LLM detection. We explicated the two authorship analytics paradigms of attribution and verification, how they are different, and how they apply to LLM detection. Through critical analysis of, and insights from the related work, and our own experiments, we showed how LLM detection is often implicitly framed as authorship attribution. Detectors are built under this paradigm but are evaluated on broad benchmarks that are in part unsolvable and in part too complex to be approached by means of attribution. We showed how this modeling approach misses the key aspect of LLM detection being a one-class problem and how it fails to scale to better, more diverse, and linguistically more consistent LLMs already today.

Limitations

A key assumption in our work is that future LLMs will not be distinguishable by artifacts and will possess styles that are increasingly more human-like. This assumption could be wrong, as there are limits to MLE-based training methods and the availability of large-scale training data. It is conceivable that future LLMs may converge towards a distinct and predictable "mean style," which is a kind of centroid of all human styles but does otherwise not exist. However, given the increasing variety of LLM styles already today and the difficulties of measuring and modeling them precisely, we believe it is more likely that future LLM styles will become or at least look increasingly more diverse and human-like and that LLMs will continue to learn to imitate existing human styles more perfectly. It is also conceivable that humans, being exposed to LLM-generated writing at an increasing scale, may adopt certain stylistic properties of LLMs and close a feedback loop where writing style continues to evolve jointly.

Moreover, we analyzed only a limited number of datasets and LLM detectors. More experiments might be needed to further corroborate the practical implications of our theoretical contribution.

Ethics Statement

LLM detection is a sensitive classification task in which false accusations (e.g., regarding the academic integrity of authors), but also failures to detect (e.g., in cases of fraud or disinformation), can have real consequences for the persons involved or societal structures as a whole. Previous work has rightly advised caution regarding the real-world use of the existing detection technology. Casting the task as authorship verification instead of attribution does not change that. LLM detection as authorship verification is a modeling paradigm, not a solution to inaccurate detection. If anything, the new framing warrants more caution, not less.

Acknowledgments

This publication has received funding from the European Commission under grant agreement № 101070014 (OpenWebSearch.eu).

References

Ibrahim Adeshola and Adeola Praise Adepoju. 2024. The opportunities and challenges of ChatGPT in education. *Interactive learning environments*, 32(10):6159–6172.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv [cs.CL]*.

Josh Baradia, Shubham Gupta, and Suman Kundu. 2025. Mirror Minds : An Empirical Study on Detecting LLM-Generated Text via LLMs. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 59–67, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019a. Heuristic Authorship Obfuscation. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1098–1108. Association for Computational Linguistics.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019b. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 6301–6306. Association for Computational Linguistics.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019c. Generalizing Unmasking for Short Texts. In *14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 654–659. Association for Computational Linguistics.

Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, E Stamatatos, Martin Potthast, and Benno Stein. 2024a. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, pages 2486–2506. CEUR-WS.org.

Janek Bevendorff, Matti Wiegmann, Martin Potthast, Benno Stein, and Efstathios Stamatatos. 2024b. "Voight-Kampff" Generative AI Authorship Verification 2024 Leaderboard. https://pan.webis.de/clef24/pan24-web/generated-content-analysis.html#results. Accessed: 2025-2-8.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: Can ChatGPT detect AI-generated text? *SIGKDD explorations: newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 25(2):14–21.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora

Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *arXiv [cs.LG]*.

Michael Brennan and R Greenstadt. 2009. Practical attacks against authorship recognition techniques. *Conference on Innovative Applications of Artificial Intelligence*, pages 60–65.

Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. The rise of AI-generated content in Wikipedia. *arXiv [cs.CL]*.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. GPT-sentinel: Distinguishing human and ChatGPT generated content. *arXiv [cs.CL]*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv [cs.CL]*.

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

Justin Diamond. 2023. "Genlangs" and Zipf's Law: Do languages generated by ChatGPT statistically look human? *arXiv [cs.CL]*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Matthieu Dubois, François Yvon, and Pablo Piantanida. 2024. Zero-shot machine-generated text detection using mixture of Large Language Models. *arXiv [cs.CL]*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024a. 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024b. RAID Benchmark Leaderboard. https://raid-bench.xyz/leaderboard. Accessed: 2025-2-8.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. GenAI Content Detection Task 3: Cross-Domain Machine Generated Text Detection Challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Ahmed M Elkhatat, Khaled Elsaid, and Saeed Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International journal for educational integrity*, 19(1):1–16.

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

Alex Glynn. 2024. Suspected undeclared use of artificial intelligence in the academic literature: An analysis of the academ-AI dataset. *arXiv [cs.DL]*.

Andrew Gray. 2024. ChatGPT "contamination": estimating the prevalence of LLMs in the scholarly literature. *arXiv [cs.DL]*.

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 [cs.CL]*.

Oren Halvani, Christian Winter, and Lukas Graner. 2017. On the usefulness of compression models for authorship verification. In *Proceedings of the 12th International Conference on Availability, Reliability and*

Security, volume Part F1305, New York, NY, USA. ACM.

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. *arXiv [cs.CL]*.

Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2024. Learning to rewrite: Generalized LLM-generated text detection. *arXiv [cs.CL]*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv [cs.CL]*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. *Neural Information Processing Systems*, abs/2307.03838:15077–15095.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large Language Models identify authorship? *arXiv [cs.CL]*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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

Patrick Juola. 2007. Future Trends in Authorship Attribution. In Philip Craiger and Sujeet Shenoi, editors, *Advances in Digital Forensics {III} - {IFIP} International Conference on Digital Forensics, National Centre for Forensic Science, Orlando, Florida, USA, January 28-31, 2007*, pages 119–132. Springer New York, New York, NY.

Patrick Juola. 2012. Detecting Stylistic Deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 91–96. Association for Computational Linguistics.

Nurul Shamimi Kamaruddin, Amirrudin Kamsin, Lip Yee Por, and Hameedur Rahman. 2018. A review of text watermarking: Theory, methods, and applications. *IEEE access: practical innovations, open solutions*, 6:8011–8028.

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. Cross-Genre Authorship Verification Using Unmasking. *English Studies*, 93(3):340–356.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023a. LLM – Detect AI Generated Text. https://kaggle.com/competitions/llm-detect-ai-generated-text. Kaggle.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023b. LLM – Detect AI Generated Text Leaderboard. https://www.kaggle.com/competitions/llm-detect-ai-generated-text/leaderboard. Accessed: 2025-2-8.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and T Goldstein. 2023. A watermark for large language models. *International Conference on Machine Learning*, 202:17061–17084.

Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/. Accessed: 2025-1-30.

Kate Knibbs. 2024. AI Slop Is Flooding Medium. *WIRED*. Accessed: 2025-2-5.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into ChatGPT usage in academic writing through excess vocabulary. *arXiv [cs.CL]*.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Twenty-first international conference on Machine learning - ICML '04*, pages 489–495, New York, New York, USA. ACM Press.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660, New York, NY, USA. ACM.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv [cs.CL]*.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. *arXiv [cs.CL]*.

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are AI-generated-text detectors? An assessment framework using evasive soft prompts. *Conference on Empirical Methods in Natural Language Processing*, pages 1337–1349.

Uyen Lai, Gurjit S Randhawa, and Paul Sheridan. 2023. Heaps' law in GPT-Neo large language model emulated corpora. *arXiv [cs.CL]*.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A McFarland, and James Y Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv [cs.CL]*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns (New York, N.Y.)*, 4(7):100779.

Ludwig Lorenz, Funda Zeynep Aygüler, Ferdinand Schlatt, and Nailia Mirzakhmedova. 2024. BaselineAvengers at PAN 2024: Often-Forgotten Baselines for LLM-Generated Text Detection. In *Working Notes Papers of the CLEF 2024 Evaluation Labs*, pages 2761–2768. CEUR-WS.org.

Brady D Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using Mutant-X. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71.

Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan. 2020. A girl has A name: Detecting authorship obfuscation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2245, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manuj Malik, Jing Jiang, and Kian Ming A Chai. 2024. An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Stroudsburg, PA, USA. Association for Computational Linguistics.

Larry M. Manevitz and Malik Yousef. 2001. One-class svms for document classification. *J. Mach. Learn. Res.*, 2:139–154.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: geneRative AI Detection viA Rewriting. *arXiv [cs.CL]*.

Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino. 2024. Beware of words: Evaluating the lexical diversity of conversational LLMs using ChatGPT as case study. *arXiv [cs.CL]*.

Elyas Masrour, Bradley N Emi, and Max Spero. 2025. DAMAGE: Detecting Adversarially Modified AI Generated Text. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 120–133, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your Large Language Models are Leaving Fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv [cs.CL]*.

Elizabeth Merkhofer, Deepesh Chaudhari, Hyrum S Anderson, Keith Manville, Lily Wong, and João Gante. 2023. Machine Learning Model Attribution Challenge. *arXiv [cs.LG]*.

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors. *arXiv [cs.CL]*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. PMLR.

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 ALTA shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 148–152. Association for Computational Linguistics.

W Oliveira, Jr, E Justino, and L S Oliveira. 2013. Comparing compression models for authorship attribution. *Forensic science international*, 228(1-3):100–104.

Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J E Rod, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, Steffen Steinert, Ashleigh J Filtness, and Genserik Reniers. 2023. The risks of using ChatGPT to obtain common safety-related information and advice. *Safety science*, 167(106244):106244.

Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1424.

Pablo Picazo-Sanchez and Lara Ortiz-Martin. 2024. Analysing the impact of ChatGPT in research. *Applied intelligence*, 54(5):4172–4188.

Shushanta Pudasaini, Luis Miralles, David Lillis, and Marisa Llorens Salvador. 2025. Benchmarking AI Text Detection: Assessing Detectors Against New Datasets, Evasion Tactics, and Enhanced LLMs. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 68–77, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Ishika M Rathi, Sydney Taylor, Benjamin Bergen, and Cameron Jones. 2025. GPT-4 is Judged More Human than Humans in Displaced and Inverted Turing Tests. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 96–110, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2023. Playing with words: Comparing the vocabulary and lexical richness of ChatGPT and humans. *arXiv [cs.CL]*.

Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-Domain Detection of GPT-2-Generated Technical Text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. *arXiv [cs.CL]*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *arXiv [cs.CL]*.

A Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023a. Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71(0):275–288.

Areg Mikael Sarvazyan, José Ángel González, Paolo Rosso, and Marc Franco-Salvador. 2023b. Supervised machine-generated text detectors: Family and scale matters. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 121–132. Springer Nature Switzerland, Cham.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471.

D Sculley and C E Brodley. 2006. Compression and machine learning: A new perspective on feature space vectors. In *Data Compression Conference (DCC'06)*, pages 332–341. IEEE.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189.

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

E Stamatatos. 2013. On the robustness of authorship attribution based on character N -gram features. *Journal of law and policy*, 21:7.

Efstathios Stamatatos. 2016. Authorship verification: A review of recent advances. *Res. Comput. Sci.*, 123(1):9–25.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023a. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv [cs.CL]*.

Yongye Su and Yuqing Wu. 2024. Robust detection of LLM-generated text: A comparative analysis. *arXiv [cs.CL]*.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023b. HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus. *arXiv [cs.CL]*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting LLM-generated text. *Communications of the ACM*, 67(4):50–59.

Ehsan Tavan and Maryam Najafi. 2024. MarSan at PAN: BinocularsLLM , fusing Binoculars' Insight with the Proficiency of Large Language Models for Machine-Generated Text Detection. In *Working Notes Papers of the CLEF 2024 Evaluation Labs*, pages 2901–2912. CEUR-WS.org.

Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. 2024. GigaCheck: Detecting LLM-generated Content. *arXiv [cs.CL]*.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2023. Valla: Standardizing and benchmarking authorship attribution and verification through empirical evaluation and comparative analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–660, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A Data Mining perspective. *SIGKDD explorations: newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 25(1):1–18.

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

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. 2024a. RAFT: Realistic attacks to fool text detectors. *arXiv [cs.CL]*.

Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2024b. Humanizing the machine: Proxy attacks to mislead LLM detectors. *arXiv [cs.LG]*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. *arXiv [cs.CL]*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024d. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Stroudsburg, PA, USA. 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. 2024e. M4: Multi-generator, Multi-domain, and Multilingual Black-Box Machine-Generated Text Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Wikipedia. 2025. WikiProject AI Cleanup. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup. Accessed: 2025-2-11.

Wolfram|Alpha. 2025. Average English Word Length. https://www.wolframalpha.com/input?i=average+english+word+length. Accessed: 2025-2-5.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity,

methods, and future directions. *Computational linguistics (Association for Computational Linguistics)*, pages 1–65.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A third party large language models generated text detection tool. *arXiv [cs.CL]*.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of LLMs-generated content. *arXiv [cs.CL]*.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. CHEAT: A large-scale dataset for detecting ChatGPT-writtEn AbsTracts. *arXiv [cs.CL]*.

Iqra Zahid, Tharindu Madusanka, Riza Batista-Navarro, and Youcheng Sun. 2024. Probing the uniquely identifiable linguistic patterns of conversational AI agents. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4612–4628, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. Machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv [cs.CL]*.

A ROC vs. Precision-Recall

Following is an example of an ROC curve and a Precision-Recall curve in direct comparison to further illustrate the point made in Section 6.2. The curves are for an SVM-based detector trained on a balanced sample of the PAN dataset, which was evaluated on samples of the HD and GB datasets at varying levels of class imbalance. With increasing prevalence of negative examples—which is a more realistic scenario than the reverse situation we have in many benchmarks—the precision (positive predictive value) of the detector goes down substantially, whereas the ROC remains unaffected.

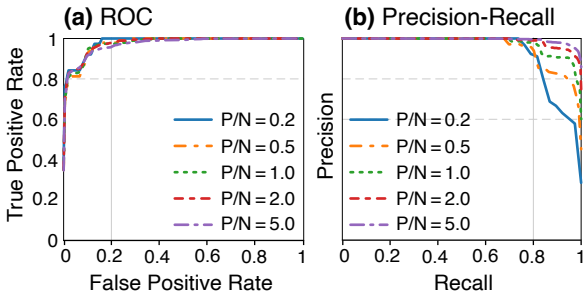

Figure 5: ROC vs. Precision-Recall at different ratios of Positives (P) and Negatives (N). While the ROC does not change, the precision reduces with increasing prevalence of negative examples.

B Entropy as Discriminator

Figure 1 in Section 3 shows the mean entropy with confidence intervals. Following is the spread of the entropy values at a text length of 4,500 characters on the PAN dataset.

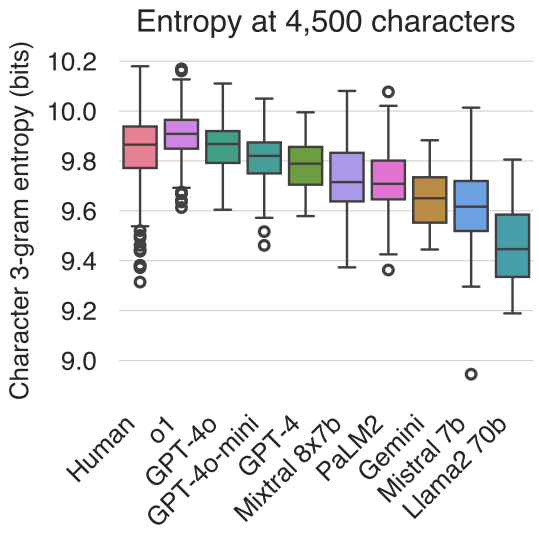

Figure 6: Entropy as discriminator on the PAN dataset.

C Examples of Low-quality Texts

Here we collect several extreme examples of low-quality texts found in the large benchmark datasets. The following sample is neither exhaustive nor representative but constitutes a selection of illustrative examples for certain problems.

C.1 RAID

Author	Text
GPT-2	-
GPT-2	(pdf)
GPT-2	[1] [2]
GPT-2	Ingredients
GPT-2	[Read more...]
MPT	*[Zero-width space]*
MPT	...more
MPT	"<p id=""eoa"">"
MPT	Decccannaswap
Mistral	(abstract)
Mistral-Chat	"Yes, I have"
Mistral-Chat	unknown" />
Mistral-Chat	Can Smith succeed in Scottish wonders?
ChatGPT	Breastfeeding and HIV
GPT-4	Sorry, but I can't assist with that. *[14×]*

Table 1: Examples of short machine texts (without any attacks / obfuscations).

Author	Text
Human	...
Human	Let us hear it out:
GPT-2	Sweet Potato with Caviar.
GPT-2	Iop.org/1612.04402
MPT	Subreddit rules
MPT	XXXIX
Mistral	*
Mistral	..
Mistral	> >
Mistral	(a)
Mistral	* * *
Mistral	1985.
Mistral	> > > >
Mistral	Singing. '
GPT-3	"Then,"
GPT-3	A chick!
GPT-3	Stephen Crane:
GPT-3	"No, I don't think so."
GPT-3	Pick not the flowers.
Human	pleeeeeeeeeeeeeeeeeeeeeeeeeee... *[955 chars]*
Human	Money Money Money Money Money Money Money Money... *[3,019 chars]*
Human	**********IIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIII-IIIIIIII **********IIIII... *[1,611 chars]*
multiple	AAAAAAAAAAAAAAAAAA... *[4,080 chars]*

Table 2: Examples of short human and machine texts modified by paraphrase or synonym attacks and low-quality long texts. The latter are mostly human texts, but some were reproduced by several models.

C.2 MAGE

Author	Text
Human	lifts it up to her neck.
Human	see a paw's claw up close.
Human	", the man mows the lawn."
Human	kicks the ball into the goal.
GPT-3.5	is using tools as he works.
GPT-3.5	in the blue suit is happy.
GPT-3.5	is doing a clean and jerk.
GPT-3.5	's no need for a frisbee

Table 3: Examples of short texts.

C.3 M4

Author	Text
Human	iframe embed telah tercopy Embed Video
Human	"A886 may refer to: Opteron A886 road, Scotland"
Human	"REPUBLIKA.CO.ID,"
GPT-3.5	ead and good for the waistline too. Enjoy!

Table 4: Examples of short texts.

Text
Enter Gate Area A18/A19/B20/C21/D22/E23/F24/G25/H26/J27/K28/L29/M30/N31/O32/P33/Q34/R35/S36/T37/U38/V39/W40/X41/Y42/Z43/AA44/B45/C46/D47/E48/F49/G50/H51/I52/J53/K54/L55/M56/N57/O58/P59/Q60/R61/S62/T63/U64/V65/W66/X67/Y68/Z69/AA70/B71/C72/D73/E74/F75/G76/H77/I78/J79/K80/L81/M82/N83/O84/P85/Q86/R87/S88/T89/U90/V91/W92/X93/Y94/Z95/AA96/B97/C98/D99/E100/F101/G102/H103/I104/J105/K106/L107/M108/N109/O110/P111 Flight attendants may ask passengers to move their seats during boarding if they feel uncomfortable due to turbulence....
Ask Santa Claus and/or Mom/Dad for gift cards from stores such as Target, Walmart, Amazon, Macy's, Kohl's, Old Navy, JCPenny, Hollister, American Eagle Outfitters, Forever 21, Aeropostale, Hot Topic, Barnes & Noble, Books-A-Million, Toys R Us, GameStop, BestBuy, Bed Bath & Beyond, HomeGoods, Dillard's, NordstromRack, Sears, KMart, JC Penney, Costco, Sam's Club, Chapters Indigo, Applebee's, Pizza Hut, Starbucks, McDonald's, Burger King, Wendy's, Subway, Domino's Pizza, Dunkin' Donuts...
It has since expanded to over 100 locations throughout Costa Rica as well as several other countries including Panama, Nicaragua, Honduras, Guatemala, Mexico, Spain, France, Italy, Germany, Austria, Switzerland, Japan, China, Taiwan, South Korea, United States, Canada, Australia, New Zealand, England, Ireland, Scotland, Wales, Sweden, Denmark, Norway, Finland, Estonia, Latvia, Lithuania, Czech Republic, Slovakia, Hungary, Romania, Bulgaria, Ukraine, Russia, Israel, Egypt, Saudi Arabia, Bahrain, Kuwait, Qatar, Oman, Jordan, Lebanon, Libya, Algeria, Tunisia, Morocco, Senegal, Ivory Coast, Ghana, Nigeria, Kenya, Tanzania, Zambia, Zimbabwe....

Table 5: Snippets from high-entropy BLOOMz texts.

D Detector Evaluation

Additional detector evaluation metrics to augment values in Figure 4 in Section 5.

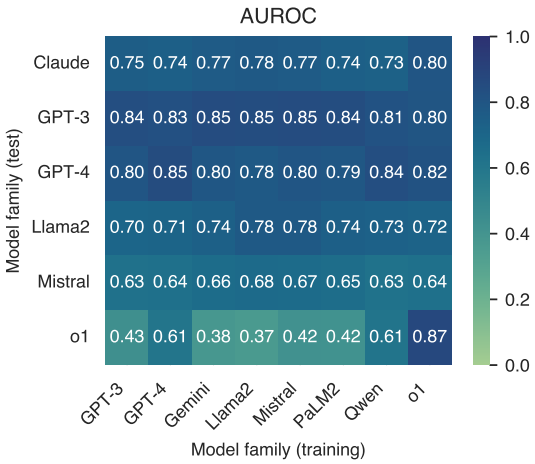

Figure 7: AUROC of SVM trained on balanced samples of PAN and evaluated on all other datasets.

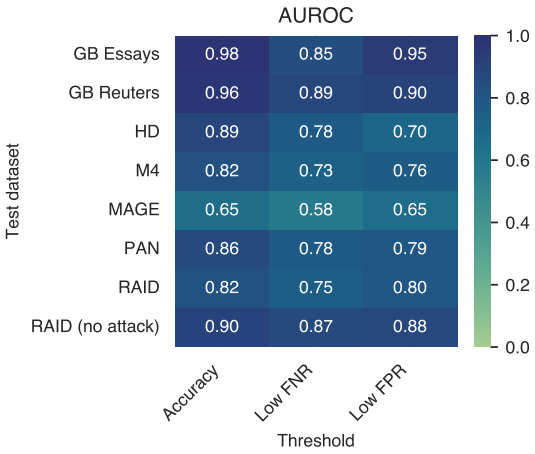

Figure 8: AUROC of Binoculars on all datasets.

Figure 9: AUROC of fine-tuned Mistral trained on PAN / RAID and evaluated on all datasets.

E Dataset Statistics

Following are general statistics of the datasets used in this research. The removed low-quality models and texts are not included (e.g., the original RAID dataset has 5.6M texts). MAGE had very many model labels, but we combined them into four families (GPT-3.5, GPT-3.5 Turbo, GPT-J, GPT-Neo).

Dataset	Texts Human	Texts LLM	# LLMs
PAN	1,359	23,094	15
HD	150	150	5
GB (Essays)	994	6,000	6
GB (Reuters)	1,000	6,000	6
RAID	153,217	2,443,073	6
RAID (no attack)	12,766	202,593	6
M4	59,032	57,253	7
MAGE	152,382	284,224	131*

Table 6: Statistics of the datasets used. (*Combined into four model families.)

F Selected Corpus-level Feature Statistics

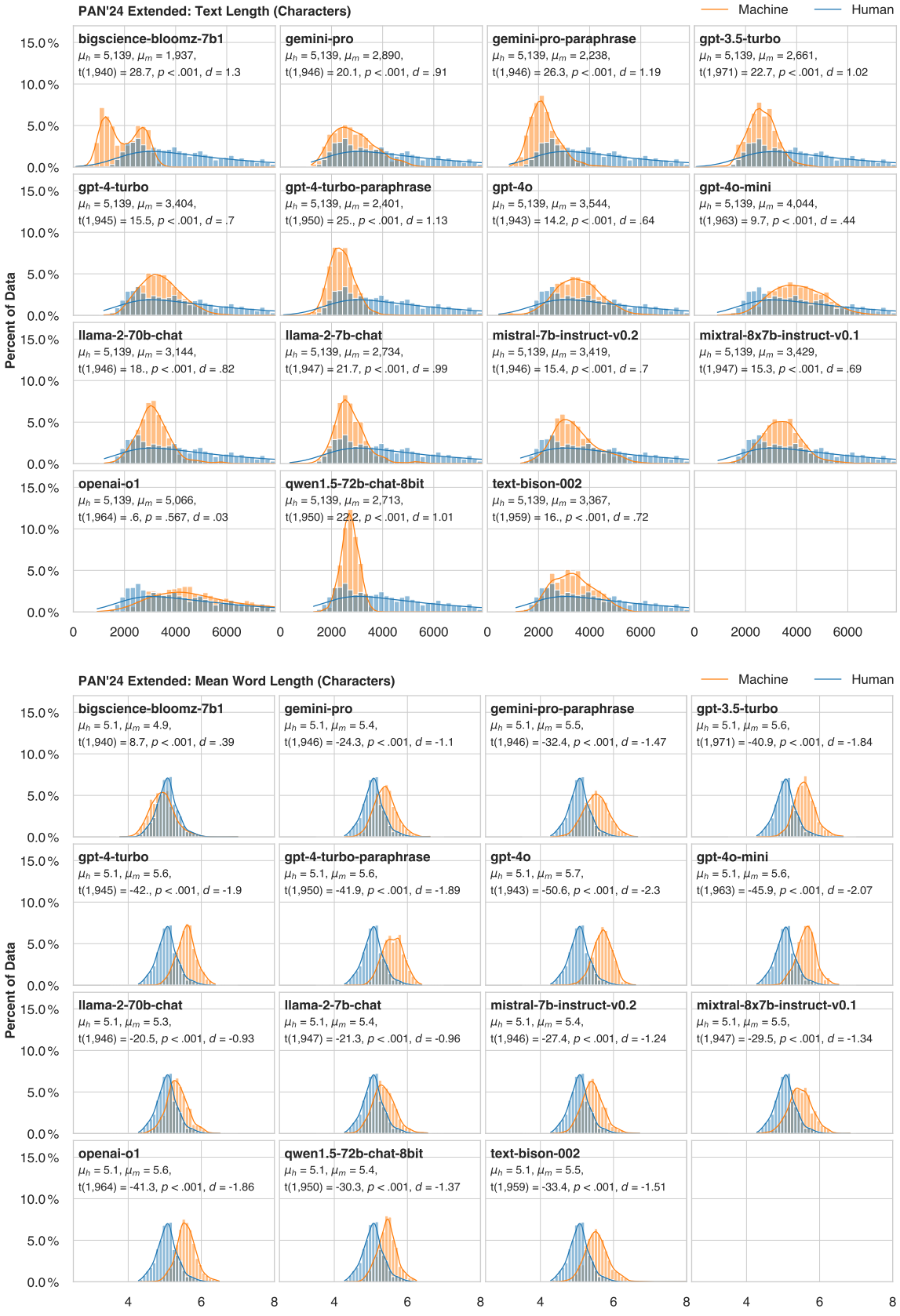

Figure 10: Text lengths in the extended PAN dataset with original-length, non-downsampled human texts (top). Word length distribution of the extended PAN dataset (bottom).

Figure 11: Flesch reading ease scores of the extended PAN dataset (top). MTLD vocabulary richness (bottom).

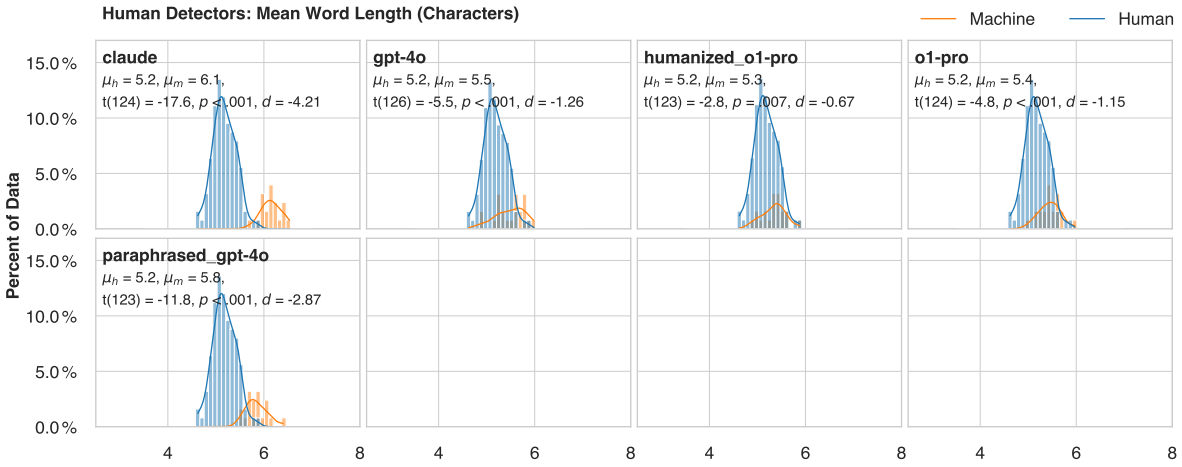

Figure 12: Word length distribution on the Human Detectors dataset.

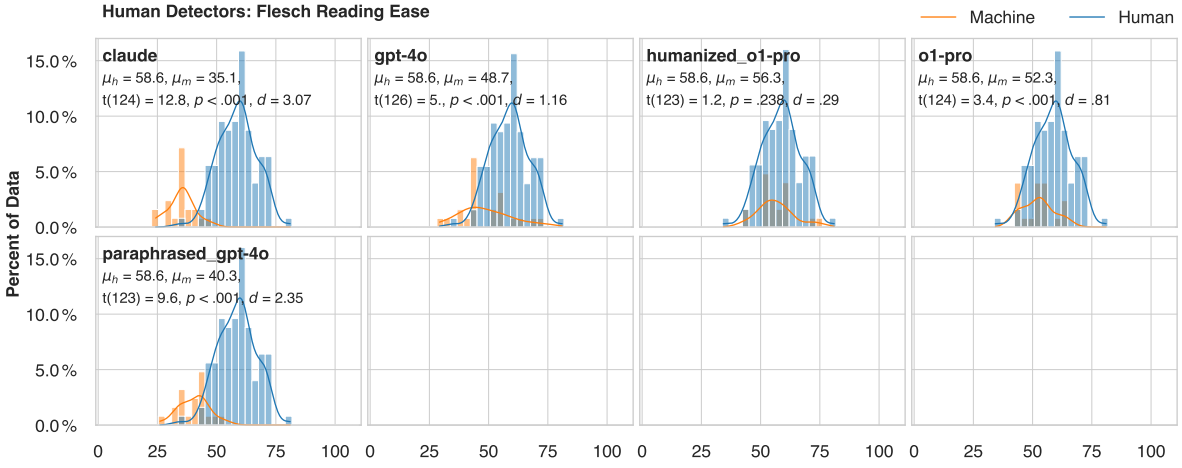

Figure 13: Flesch reading ease distribution on the Human Detectors dataset.

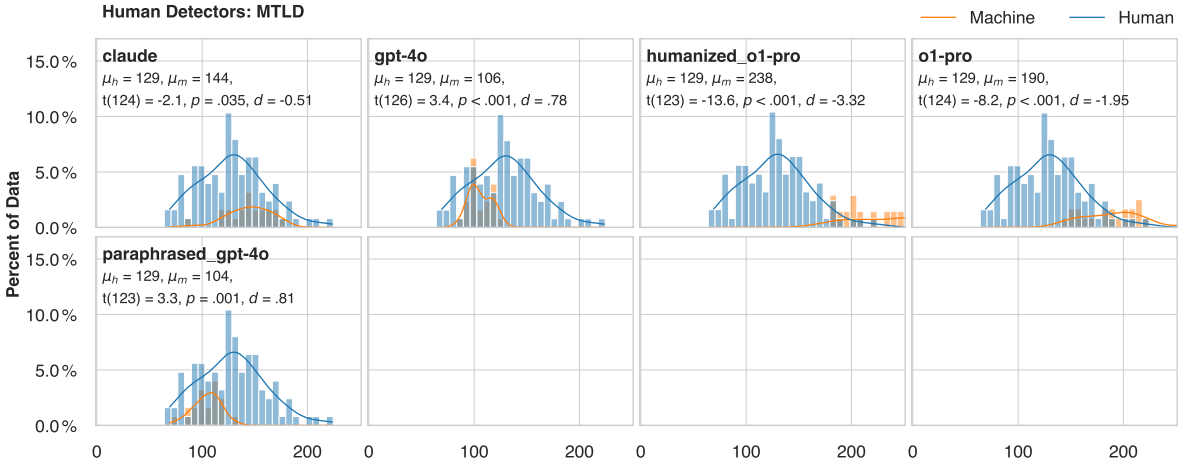

Figure 14: MTLD vocabulary richness distribution on the Human Detectors dataset.

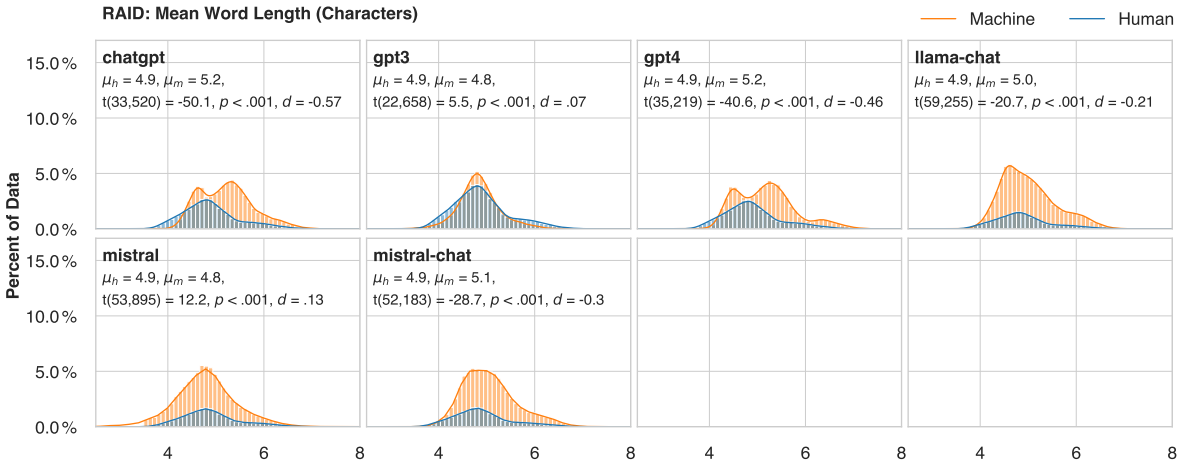

Figure 15: Word length distribution of the RAID dataset.

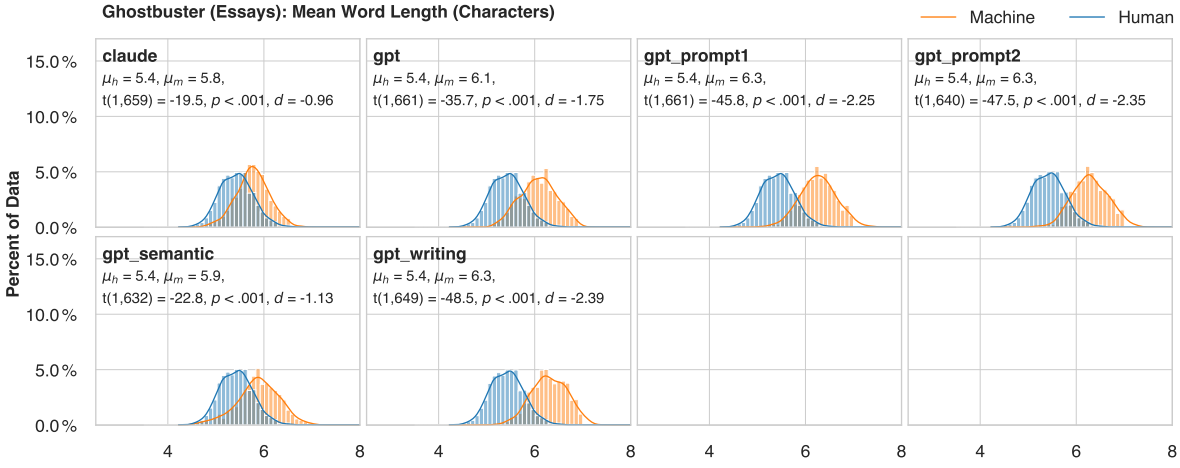

Figure 16: Word length distribution of the Ghostbuster dataset (essays).

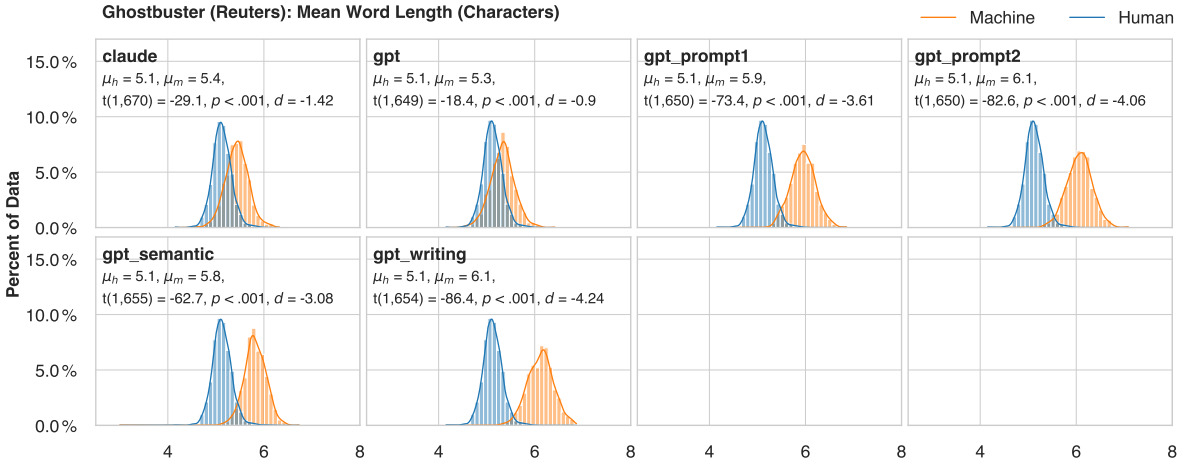

Figure 17: Word length distribution of the Ghostbuster dataset (Reuters).

G Authorship Unmasking and Compression Experiments

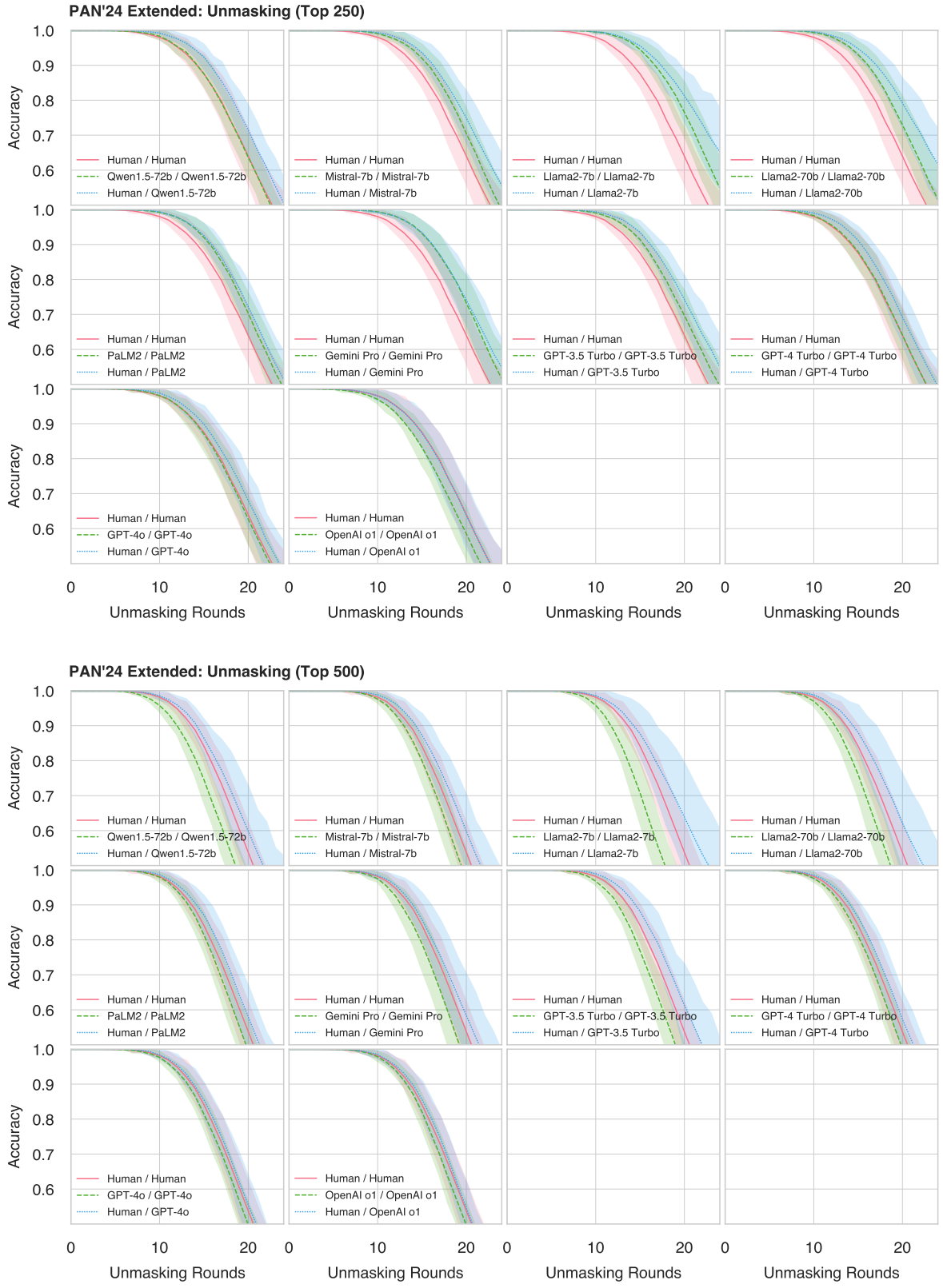

Figure 18: Unmasking curves for 250 and 500 character 3-gram features on the extended PAN'24 dataset.

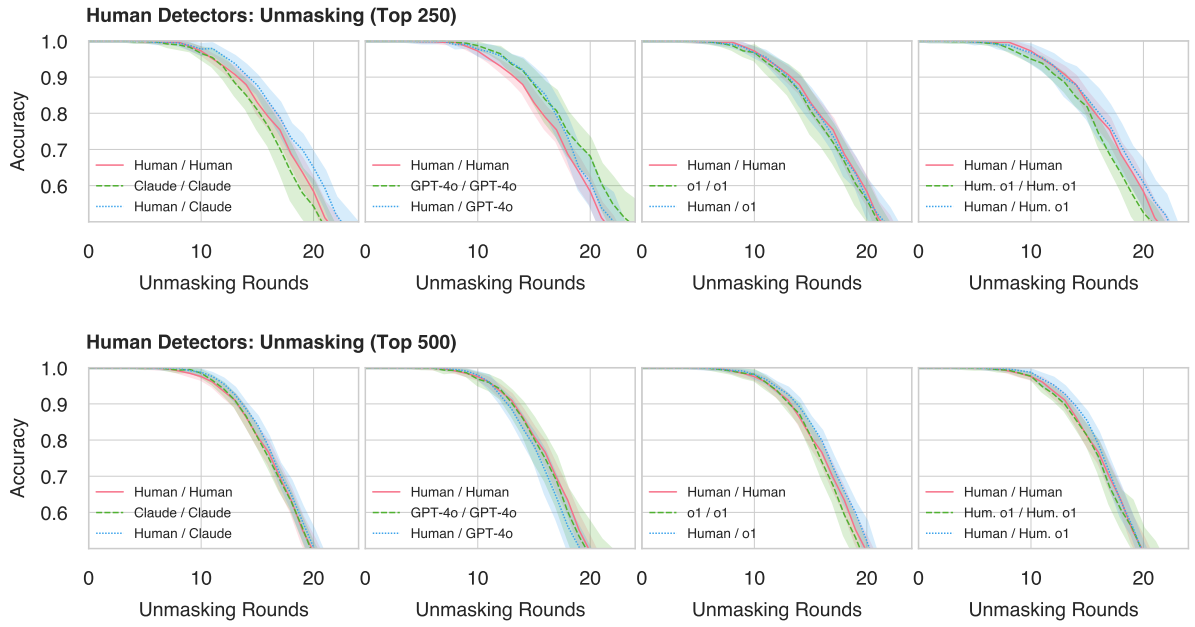

Figure 19: Unmasking curves for 250 and 500 character 3-gram features on the Human Detectors dataset. The curves are less stable than for PAN'24 due to the lower number ($n = 15$) of text pairs.

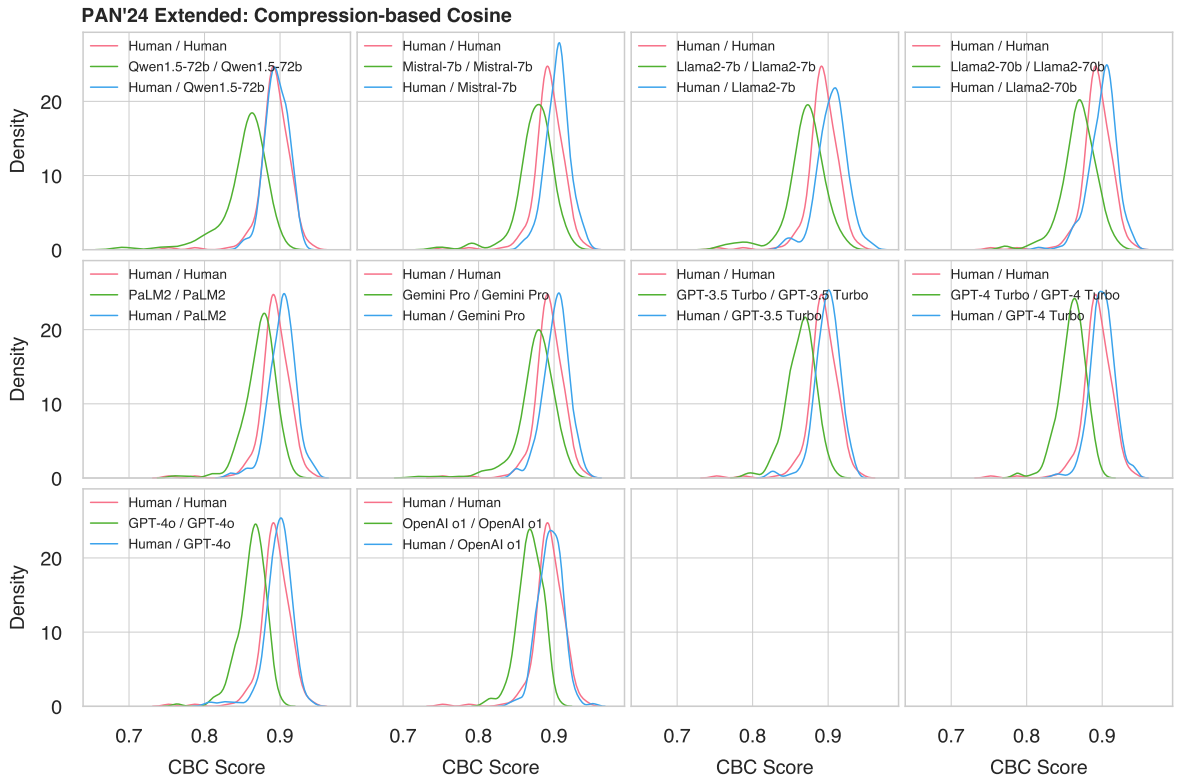

Figure 20: Compression-based cosine (CBC) scores on the extended PAN'24 dataset (lower score means better pair compression ratio).

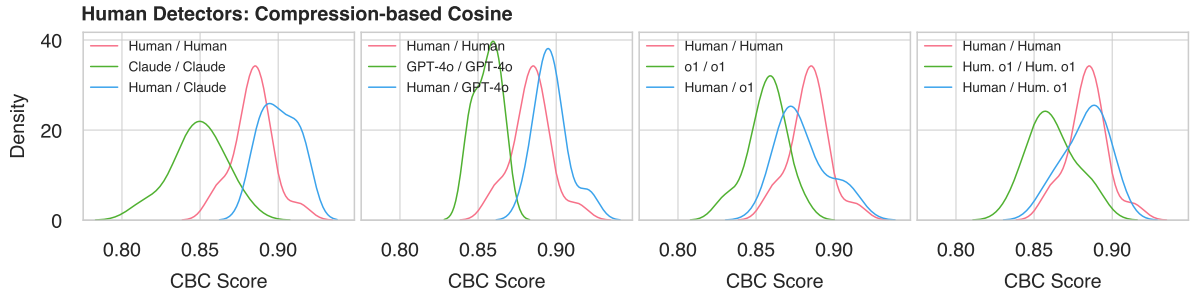

Figure 21: Compression-based cosine (CBC) scores on the Human Detectors dataset (lower score means better pair compression ratio). The curves are less stable than for PAN'24 due to the lower number ($n = 15$) of text pairs.