Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media

Anonymous ACL submission

Abstract

Social media platforms are experiencing a growing presence of AI-Generated Texts (AIGTs). However, the misuse of AIGTs could have profound implications for public opinion, such as spreading misinformation and manipulating narratives. Despite its importance, it remains unclear how prevalent AIGTs are on social media. To address this gap, this paper aims to quantify and monitor the AIGTs on online social media platforms. We first collect a dataset (*SM-D*) with around $2.4M$ posts from 3 major social media platforms: Medium, Quora, and Reddit. Then, we construct a diverse dataset (*AIGTBench*) to train and evaluate AIGT detectors. *AIGTBench* combines popular open-source datasets and our AIGT datasets generated from social media texts by 12 LLMs, serving as a benchmark for evaluating mainstream detectors. With this setup, we identify the best-performing detector (**OSM-Det**). We then apply **OSM-Det** to *SM-D* to track AIGTs across social media platforms from January 2022 to October 2024, using the AI Attribution Rate (AAR) as the metric. Specifically, Medium and Quora exhibit marked increases in AAR, rising from 1.77% to 37.03% and 2.06% to 38.95%, respectively. In contrast, Reddit shows slower growth, with AAR increasing from 1.31% to 2.45% over the same period. Our further analysis indicates that AIGTs on social media differ from human-written texts across several dimensions, including linguistic patterns, topic distributions, engagement levels, and the follower distribution of authors. We envision our analysis and findings on AIGTs in social media can shed light on future research in this domain.

1 Introduction

The rapid development of Large Language Models (LLMs) has markedly enhanced the quality of AIGTs, enabling the use of models like GPT-3.5 (OpenAI, 2022) in daily life to produce high-quality texts, such as in academic writing (Gruda, 2024), question-answering (Kamalloo et al., 2023), and translation (Wang et al., 2023a). These AIGTs are often indistinguishable from Human-Written Texts (HWTs), presenting AIGT detection as a crucial yet challenging task for effective classification. On social media platforms, the use of LLMs to answer questions can contribute to the spread of misinformation (Zhou et al., 2023). Furthermore, AIGTs may be deliberately used for information manipulation or the dissemination of fake news, potentially resulting in serious societal impacts (Hanley and Durumeric, 2024). To better understand the prevalence of AIGTs on social media platforms, we aim to quantify and monitor its presence, addressing the question: **On social media, are we already interacting with AI-generated texts?**

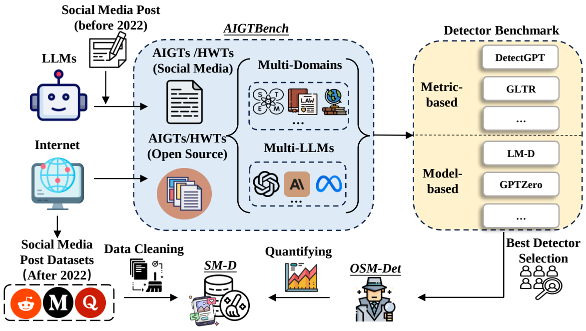


Figure 1: Pipeline of quantifying AIGTs on social media.

Currently, numerous detectors have been developed to detect AIGTs. According to the MGTBench (He et al., 2024), these detectors are broadly divided into two categories: metric-based (Gehrmann et al., 2019; Mitchell et al., 2023) and model-based detectors (Ippolito et al., 2019; Solaiman et al., 2019; Bhattacharjee et al., 2023), some of which have shown high accuracy and robustness. While these detectors have been applied in controlled settings, recent studies have explored their effectiveness in real-world scenarios. Hanley and Durumeric (2024) conduct AIGT detection on news website articles, with a primary focus on content generated by GPT-3.5 and others from Turing benchmark, which includes various pre-2022 models (Uchendu et al., 2021). Furthermore, Liu et al. (2024c) detects ChatGPT-generated content on arXiv papers. However, academic and news writing are formal and tailored to specific audiences, whereas social media content is more interactive, making it a better domain for observing AIGTs' impact on daily life. Moreover, previous studies do not account for recent popular LLMs, while we consider a broader range of models in our efforts to detect AIGTs on social media.

To quantify and monitor AIGTs on social media, we collect textual data from 3 popular platforms spanning January 1, 2022, to October 31, 2024, as most LLMs are released after 2022. After data preprocessing, we obtain $1,170,821$ posts from Medium, $245,131$ answers from Quora, and $982,440$ comments from Reddit. We name it as *SM-D*, short for Social Media Dataset.

To identify the most effective detector, we construct a dataset named *AIGTBench*, which consists of public AIGT/Supervised-Finetuning (SFT) datasets and our own AIGT datasets generated from social media data. *AIGTBench* includes AIGTs generated by 12 different LLMs, such as GPT Series (OpenAI, 2024)) and Llama Series (Touvron et al., 2023a,b; Dubey et al., 2024)), totaling around $28.77M$ AIGT and $13.55M$ HWT samples. We then benchmark AIGT detectors on *AIGTBench* and leverage the best-performing detector as our primary detector, which achieves an accuracy of 0.979 and an F1-score of 0.980. To better reflect its application in detecting AIGTs on online social media, we rename it as **OSM-Det** (Online Social Media Detector).

Based on **OSM-Det**, we quantify and monitor the texts across the 3 platforms and use the AI Attribution Rate (AAR) to represent the rate of posts classified as AI-generated (The pipeline is shown in Figure 1). We observe several noteworthy phenomena: (1) **A sharp rise in AI-generated content begins in December 2022, with distinct AAR trends emerging across platforms.** Before December 2022, the AAR across platforms remains stable. However, starting in December, Medium and Quora show significant surges, while Reddit shows only a slight increase. This suggests the widespread and diverse LLM adoption on social media; (2) **Linguistic analysis shows similar AAR trends and exhibits stylistic features in AIGTs/HWTs.** Based on the word-level analysis, we find that the usage trend of top-frequency AI-preferred words aligns closely with LLM adoption trends. With sentence-level analysis, we also reveal that AIGTs tend to be more objective and standardized, whereas HWTs are more flexible and informal; (3) **Technology-related topics drive higher AARs on Medium.** Topics like "Technology" and "Software Development" show the highest AARs, indicating that users with a strong technical background are more likely to adopt LLMs; (4) **Predicted HWTs receive more engagement than AIGTs.** On Medium, the content predicted as HWTs receives more average "Likes" and "Comments" than AIGTs. This suggests that users are more inclined to engage with HWTs; and (5) **Authors with fewer followers are more likely to produce AIGTs.** On Medium, users with no more than one thousand followers tend to produce content that has the highest mean AAR at 54.02%. In contrast, as the follower count increases, the AAR gradually shifts toward the lower range ($\leq 25.00\%$).

Our contributions are summarized as follows:

- We are the first to conduct a systematic study to quantify, monitor, and analyze AIGTs on social media. To achieve this, we collect a large-scale dataset *SM-D*, which includes around $2.4M$ posts from three platforms, spanning from January 2022 to October 2024.
- We construct *AIGTBench*, a dataset for benchmarking AIGT detectors. *AIGTBench* can be divided into two parts: one derived from open-source datasets and the other generated by 12 LLMs based on platform-specific characteristics. Leveraging *AIGTBench*, we identify the most effective AIGT detector, **OSM-Det**.
- Our research reveals a remarkable increase in AAR on social media after the widespread adoption of LLMs. Moreover, this trend varies markedly across different platforms.
- We conduct an in-depth analysis of the characteristics of AIGTs and HWTs through *linguistic analysis* and *multidimensional analysis of posts*, revealing differences in lexical patterns, topic distributions, engagement levels, and the follower distributions of authors. These analyses provide valuable insights for future research.

2 Related Work

The growth in model parameters and training data has recently empowered LLMs to demonstrate exceptional language processing capabilities (Zhao et al., 2023). Since then, LLMs have gradually gained popularity, like GPT-4 (OpenAI, 2023) and Llama (Touvron et al., 2023a), enabling users to generate high-quality texts effortlessly. Yet, LLMs have raised concerns about potential misuse, such as fake news generation (Zellers et al., 2019), academic misconduct (Vasilatos et al., 2023), and performance degradation of training LLMs using AI content (Briesch et al., 2023), making the detection of AIGTs (also known as machine-generated texts) increasingly important (Fraser et al., 2024). He et al. (2024) introduce MGTBench for standardizing the evaluation of different LLMs and experimental setups within the AIGT detectors. They broadly categorize the detectors into two main types: metric-based and model-based detectors. Metric-based detectors use pre-defined metrics, such as log-likelihood, to capture the characteristics of texts (Gehrmann et al., 2019; Mitchell et al., 2023; Su et al., 2023). In contrast, model-based detectors rely on trained models to distinguish between AIGTs and HWTs (Solaiman et al., 2019; Guo et al., 2023; Bhattacharjee et al., 2023; Liu et al., 2024c; Ippolito et al., 2019; Li et al., 2024). More introduction refer to Appendix B.

Besides, some researchers have applied detectors to text detection in real-world scenarios. Hanley and Durumeric (2024) train a detector using data generated by the ChatGPT and Turing benchmark model and conduct tests on multiple news websites. Their study reveals that, from January 1, 2022, to May 1, 2023, the proportion of synthetic articles increased on news sites. Liu et al. (2024c) also conduct detection on arXiv and find a significant rise in the proportion of papers using ChatGPT-generated content, reaching 26.1% by December 2023. In contrast to their detection targets, we focus on detecting AIGTs on social media platforms and covering a broader range of LLMs. Macko et al. (2024) construct a multilingual dataset based on instant messaging and social interaction platforms such as Telegram and Discord, using it to compare the performance of existing detectors. In contrast, our research focuses on providing an in-depth temporal analysis of AIGTs on content-driven social platforms like Medium, Quora, and Reddit.

3 Data Collection

In this section, we elaborate on the data collection process, which primarily includes two datasets: the social media dataset (*SM-D*) and the detector training dataset (*AIGTBench*).

3.1 *SM-D* (Social Media Dataset)

Dataset	# Posts	# Filtered Posts	Time Range
Medium	1,416,208	1,170,821	January 1, 2022-October 31, 2024
Quora	445,864	245,131	January 1, 2022-October 31, 2024
Reddit	1,019,261	982,440	January 1, 2022-July 31, 2024

Table 1: Overview of the Medium, Quora, and Reddit datasets.

Unlike previous research, we focus on social media platforms, including Medium, Quora, and Reddit, emphasizing content creation, sharing, and discussion. The introduction of platforms is in Appendix C. These platforms stand out for hosting longer, more detailed posts where users emphasize the depth and quality of the information they share. As shown in Table 1, we collect data from these social media platforms from January 1, 2022 to October 31, 2024. We consider this part as our social media dataset for analysis.

For each platform, the detection targets are determined based on their distinct characteristics. On Medium, a blog hosting platform, we extract both the titles and contents of articles, treating the entire article as the detection target. On Quora, a question-and-answer platform, we select the corresponding answers to questions as the detection target. Similarly, on Reddit, which is known for its user-driven discussions, we also choose the response content as the detection target. Furthermore, we apply data filtering with the rules described in Appendix E.

3.2 *AIGTBench* (Detector Training Dataset)

To train the AIGT detectors, we consider two parts of the data. First, we consider 6 publicly available AIGT datasets and 5 common SFT datasets to form the training dataset (see Tables A3 and A4 for dataset statistics and Appendix D for more details). Second, to increase the detector's generalization capabilities on social media, we additionally collect data from the 3 social media platforms ranging from January 1, 2018, to December 31, 2021. We classify this data as HWTs, given that most LLMs had not been published during this period. We also design different LLMs writing tasks to generate AIGTs that align with the characteristics of platforms (Table A1 describes the statistics details).

For Medium, which is primarily used for sharing articles and blogs, the core tasks are centered on writing. We design two LLM writing tasks: (1) polish articles to create polished versions; (2) based on the article's title and summary, directing the LLM to generate complete article content, thereby simulating a writing scenario. For Quora and Reddit, which mainly focus on question answering and user interaction, we design two tasks: (1) polish texts like Medium and (2) query LLM directly answer questions, simulating a user interaction scenario. Detailed prompts are provided in Appendix F.

Overall, the datasets used for training our detector and the distribution of LLM series are shown in Figure A1. This dataset includes 12 different LLMs, with a detailed introduction provided in Appendix A. Within these datasets, the two most prevalent model series are the GPT Series, which accounts for 42.99%, and the Llama series, which represents 39.05%. GPT Series is the most widely used proprietary model and has played a pivotal role in the evolution of generative AI. As of January 2023, approximately $13M$ users interact daily with GPT-3.5 (Wang et al., 2023c). The Llama series models also have significant influences, as the report indicates that downloads of Llama models on the Hugging Face platform have nearly reached around $350M$ (Meta AI, 2024). Therefore, these two model series are the primary focus of our dataset. During the data generation process, we notice that certain samples contain textual noise, like irrelevant or redundant information. To maintain data quality, we implement some data processing strategies (see Appendix E for details).

4 Experimental Settings

4.1 Datasets

As mentioned in Section 3, we collect the social media dataset (*SM-D*) and the detector training dataset (*AIGTBench*). *SM-D* refers to the social media dataset that we conduct the quantification, with more details provided in Section 3.1. *AIGTBench* is the benchmark for AIGT detectors, which includes AIGTs generated by 12 different LLMs, as described in Section 3.2. We randomly divide *AIGTBench* into training, validation, and test sets in a $7:1:2$ ratio. Specifically, the distribution of token lengths in the training, validation, and test set are shown in Figure A2.

4.2 AIGT Detectors

Following the experimental setup of MGTBench (He et al., 2024), we evaluate 14 detectors. For metric-based detectors, we consider LogLikelihood, Rank, LogRank, Entropy, GLTR, LRR, DetectGPT, and NPR (Solaiman et al., 2019; Gehrmann et al., 2019; Mitchell et al., 2023). We choose the GPT-2 medium (Radford et al., 2019) as the base model, given its good detection performance at limited computational costs.

During the detection process, we initially use the GPT-2 medium to extract multiple metrics, including log-likelihood and log-rank. Based on these extracted metrics, we train logistic regression models to enhance the accuracy of predictions. For the model-based detectors, we consider both pre-trained detectors and fine-tuned models with the *AIGTBench*, that is, OpenAI Detector (Solaiman et al., 2019), ChatGPT Detector (Guo et al., 2023), ConDA (Bhattacharjee et al., 2023), GPTZero (GPTZero, 2024), CheckGPT (Liu et al., 2024c), and LM-D (Ippolito et al., 2019). Specifically, for the OpenAI Detector and ChatGPT Detector, we consider their pre-trained version and select the RoBERTa-base model as it demonstrates stable performance across multiple detection tasks and typically provides better detection results. For ConDA and LM-D, we choose the Longformer-base-4096 model as the base model and fine-tune it with the *AIGTBench*. All of them have a learning rate of 1e-5, a batch size of 16, and the AdamW optimizer. For GPTZero, we directly use its commercial API. For CheckGPT, we retrain the original training framework (Liu et al., 2024c).

4.3 Evaluation Metrics

We use accuracy and F1-score as the evaluation metrics to evaluate the performance of different detectors, which are common standards in AIGT detection tasks. Besides, we introduce two metrics **AI Attribution Rate** (AAR) and **False Positive Rate** (FPR) for quantification analysis. The AAR indicates the proportion of texts that the model predicts as AI-generated, while the FPR denotes the proportion of HWTs misclassified as AIGTs.

To assess word usage, we compute the **normalized term frequency** (NTF) as:

$$\mathrm{NTF}(t,d) = \frac{f_{t,d}}{N \cdot \sum_{t' \in d} f_{t',d}}, \tag{1}$$

where $f_{t,d}$ is the frequency of word t in document d, $\sum_{t' \in d} f_{t',d}$ accounts for all words in d, and N is the total occurrences of t across all documents.

	Metric-based								Model-based					
	Log-Likelihood	Rank	Log-Rank	Entropy	GLTR	LRR	DetectGPT	NPR	OpenAI Detector	ChatGPT Detector	ConDA	GPTZero	CheckGPT	LM-D
Accuracy	0.730	0.618	0.713	0.650	0.704	0.680	0.686	0.658	0.615	0.686	0.972	0.933	0.966	**0.979**
F1-score	0.754	0.730	0.741	0.697	0.733	0.660	0.659	0.639	0.484	0.602	0.973	0.930	0.966	**0.980**

Table 2: Performance of detectors on *AIGTBench*. The F1-score corresponds to the AI class.

5 Evaluation

5.1 Benchmarking Detectors

This section compares different AIGT detectors on the test set of the *AIGTBench*. Illustrated in Table 2, the metric-based detectors perform poorly. The F1-scores for Log-Likelihood, Rank, Log-Rank, and Entropy are 0.754, 0.730, 0.741, and 0.697, respectively. These low scores indicate that metric-based detectors face limitations in handling complex, multi-source datasets and struggle to capture subtle textual features effectively.

Regarding model-based detectors, we observe that both OpenAI Detector and ChatGPT Detector perform worse than some metric-based detectors. Specifically, OpenAI Detector has an F1-score of only 0.484, with relatively low accuracy. This underperformance may be due to the detector being fine-tuned using GPT-2 output, which struggles to adapt to more complex data generated by modern LLMs, such as the Llama and Claude Series.

Notably, LM-D and ConDA outperform the others. ConDA achieves an accuracy of 0.972, while the LM-D performs even better, with an accuracy of 0.979 and an F1-score of 0.980, making it the most effective detector. Based on these benchmark results, we consider LM-D as the most effective detection method and name LM-D fine-tuned on *AIGTBench* as **OSM-Det**, which is subsequently used to quantify and monitor the AAR in social media dataset (*SM-D*). More details on performance across different platforms and all text lengths in *SM-D* are shown in Appendix G.

Generalizability Experiments. To validate the generalizability of **OSM-Det**, we conduct experiments under various generation parameter settings, unseen LLMs, and unseen domains. Figure 2 shows that **OSM-Det** can maintain excellent performance on AIGTs under different generation parameters (temperature, Top-P and Top-K), with accuracy exceeding 0.990. The generalization performance on unseen LLMs and domains stays robust. For further details, please refer to Appendix H. Thus, these results show that **OSM-Det** has outstanding generalizability, supporting the reliability of the results on *SM-D*.

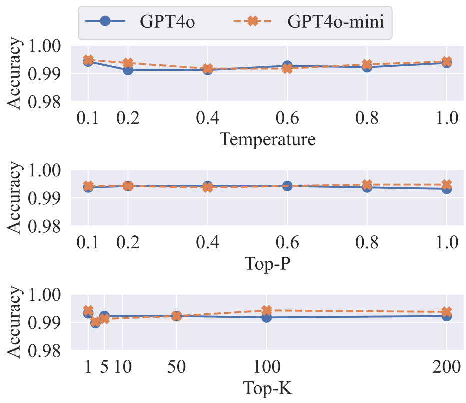
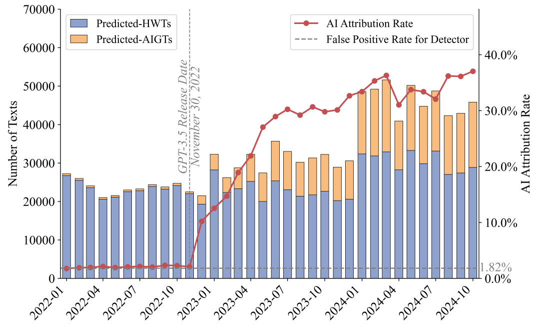


Figure 2: Impact of different generation parameters on AIGT detection accuracy.

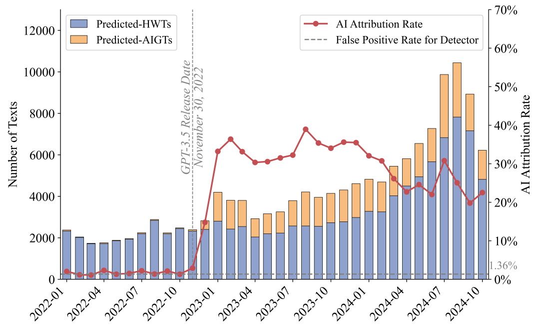
5.2 Evaluation on Social Media Platforms

As shown in Table A2, **OSM-Det** achieves False Positive Rates (FPR) of 1.82%, 1.36%, and 1.70% on Medium, Quora, and Reddit, respectively, while achieving a benchmark F1-score of 0.980 (see Table 2). These results highlight **OSM-Det**'s low misclassification rate and high overall accuracy, making it a reliable choice for quantifying and monitoring AIGTs on social media.

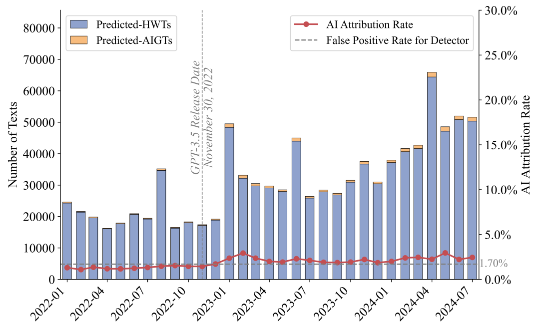
Evaluation on Medium. Figure 3a illustrates the trend of AAR on Medium from January 2022 to October 2024. From January 2022 to November 2022, the AAR remains stable, fluctuating around 1.82%. This suggests that, before the widespread adoption of GPT-3.5, creators mainly rely on original content with minimal dependency on LLM-generated content. However, starting in December 2022, coinciding with the launch of GPT-3.5, the AAR begin to rise rapidly. Between December 2022 and July 2023, the AAR surges from 10.20% to 30.24%, reflecting how the popularization of LLM technology significantly lowers the barriers of content generation, prompting Medium's creator community to widely adopt LLM-assisted content creation. From August 2023 to July 2024, the AAR experiences slower growth, ranging be-



(a) AAR Trends on Medium from January 1, 2022, to October 31, 2024.



(b) AAR Trends on Quora from January 1, 2022, to October 31, 2024.



(c) AAR Trends on Reddit from January 1, 2022, to July 31, 2024.

Figure 3: Comparison of AAR and FPR across Medium, Quora, and Reddit over different time periods.

tween 29.20% and 36.29%, with fluctuations stabilizing between 30.12% and 33.75%. This indicates that AIGTs have gradually become an integral part of the platform's creative ecosystem, serving as a critical component of content production. From August 2024 to October 2024, the AAR further increased to 37.03%, reaching a new peak. This likely reflects the growing acceptance and reliance on LLM-assisted creation among content creators to enhance writing efficiency and quality.

Overall, from December 2022 to October 2024, the AAR on Medium has shown a continuous upward trend, underscoring the significant impact of LLM technology on content creation.

Evaluation on Quora. Figure 3b displays the trend of AAR on Quora. We observe that from January 2022 to October 2022, the AAR fluctuates but remains relatively low. After the release of GPT-3.5 in November 2022, the AAR slightly increases to 2.87%. Subsequently, starting in December 2022, the AAR markedly rises to 15.12% and shows a clear upward trend in AIGTs, reaching a peak of 38.95% in August 2023. From September 2023 to the first half of 2024, although the AAR remains high, it declines from the peak in early 2023 and gradually stabilizes between $22.03\% - 30.79\%$ throughout 2024. This indicates that the behavior of Quora users in generating AI content is becoming more stable. From June 2024, the AAR gradually decreases and reaches a low near 19.79% between September and October 2024. The increase in AAR may be attributed to Quora's launch of its LLM platform, Poe, in 2023 (Adam D'Angelo, 2023, 2024), which initially led to a rise in AI-generated content. However, as many Quora users found Poe's capabilities insufficient to meet their daily needs, the AAR likely declined following this initial surge, eventually stabilizing.

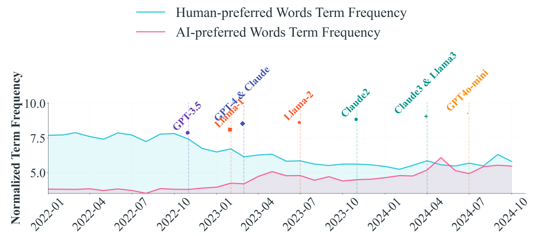
Evaluation on Reddit. Figure 3c shows the quantification analysis on Reddit from January 2022 to July 2024. From January to November 2022, we observe that the AAR remains below the FPR, fluctuating around 1.30%, indicating that there is almost no AI-generated content on Reddit during this period. Following the release of GPT-3.5, the AAR begins to rise slightly, reaching 2.36% in January 2023 and further increases to 2.93% in February 2023. From March 2023 to July 2024, the AAR stabilizes at a low level, within the range of $1.86\% - 2.95\%$.

Briefly, similar to Medium and Quora, AAR on Reddit shows an upward trend following the release of GPT-3.5, but it consistently maintains a lower level, indicating a lower dependency on LLMs among Reddit users.

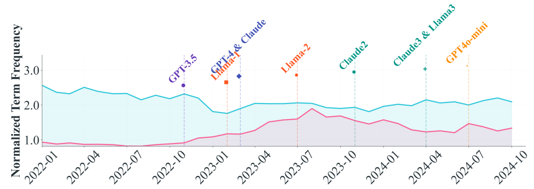
5.3 Linguistic Analysis at Different Levels

We explore the interpretability of the **OSM-Det** model in the case study using two methods: Integrated Gradients (Sundararajan et al., 2017), representing a model-dependent perspective, and Shapley Value (Scott et al., 2017), offering a model-independent perspective. Details of the two methods can be found in Appendix I.2.

Word-Level Analysis. In the case study of Reddit



(a) Word frequency trends on Medium from January 1, 2022, to October 31, 2024.



(b) Word frequency trends on Quora from January 1, 2022, to October 31, 2024.

Figure 4: Comparison of Medium and Quora word frequency trends: human vs. AI preferences. (The result of Reddit is shown in Figure A5)

(refer to Figures A8 and A10), words like "and", "think" and "I" have the highest Integrated Gradients and Shapley Values, which lead model to classify texts as human-written. Meanwhile, model-specific analysis shows the words "think", "can", and "Online" have the lowest scores, leading to AI-generated prediction. From these observations, we note that specifying clear word-level patterns between two class is challenging because certain words, like "think", contribute significantly to both classifications. This overlap suggests that word importance is highly context-dependent. Similar challenges are also observed on Medium and Quora (Figures A11, A13, A14 and A16).

Given this difficulty, we then turn to a different approach: a statistical analysis of high-frequency adjectives, conjunctions, and adverbs (details provided in Appendix I.1). These high-frequency terms are then classified into human-preferred and AI-preferred vocabularies. We then track the trends of these lexical items on *SM-D*.

As shown in Figures 4a and 4b, the NTF of AI-preferred vocabulary on the Medium and Quora is closely aligned with the development of LLMs. Following the release of LLMs such as GPT, Llama, and the Claude series, the NTF of human-preferred vocabulary has gradually declined. Meanwhile, AI-preferred vocabulary shows an increase. These results reflect an increasing usage of LLMs for content generation by Medium and Quora platform users. In contrast, the trends on Reddit show some differences (see Figure A5). From 2022 to 2024, the NTF of human-preferred vocabulary always remains high, while the AI-preferred vocabulary consistently remains low. This indicates that Reddit users rely less on LLMs to produce content. From above, we obverse that word frequency changes closely align with the AAR trends in Figure 3.

Sentence-Level Analysis. We also conduct a sentence-level analysis using Shapley values, as Integrated Gradients are only suitable for word-level. From the case studies of Medium, Quora, and Reddit (shown in Figures A9, A12 and A15), we observe that AIGTs are characterized by their objective and standardized structures, typically beginning with a noun or pronoun and following a verb-object pattern, like "Online bullying...contributes...feelings...". In contrast, HWTs often contain flexible sentence structures and informal expressions, as illustrated by "That being said, why not both?" and "Why can't we restore...". In summary, the results suggest that sentence-level patterns provide more distinctive characteristics for distinguishing AIGTs and HWTs, as LLMs may usually follow a standardized pattern to generate texts.

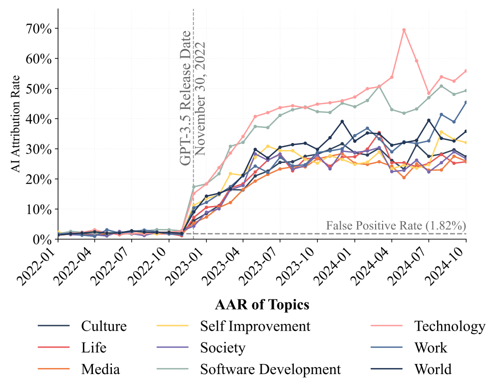


Figure 5: AAR trends across different topics.

5.4 Multidimensional Analysis of Posts

We analyze posts on social media from multi-dimensions to find the characteristics between posts predicted as AIGTs and those classified as HWTs, including topic, engagement, and author analysis.

Topic Analysis. Classifying topics on platforms like Quora and Reddit is challenging due to their wide range. Therefore, we focus our analysis on 9 major topics listed on the Medium (Medium, 2024), examining them from a temporal perspective. The proportion of topics is shown in Figure A3.

Figure 5 shows the trends of AAR across different topics. We observe a rapid increase in AAR for all topics following the release of GPT-3.5 in December 2022, indicating that the popularity of LLMs has impacted all topics on Medium. Besides, the AAR for "Technology" and "Software Development" remains consistently higher than other topics from December 2022 to October 2024, ranking respectively first and second. One possible reason is that people in the technology field are more likely to know about LLMs and frequently interact with them, leading to a higher AAR.

Follower Group	Mean Likes (AIGTs / HWTs)	Mean Comments (AIGTs / HWTs)
0-1K	49.48/79.39	3.18/5.68
1-5K	111.50/191.61	5.11/9.09
>5K	126.94/211.92	5.56/8.25

Table 3: Engagement statistics on Medium for different follower groups, comparing AIGTs and HWTs.

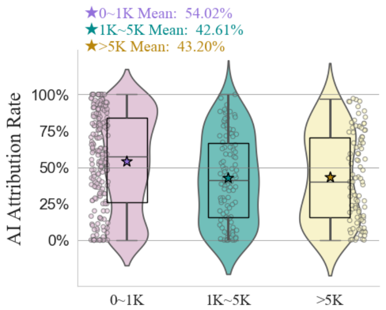


Figure 6: AAR distribution among follower groups.

Engagement Analysis. To understand how user engagement differs between articles predicted to be AIGTs or HWTs, we analyze the number of "Likes" (known as "Claps" on Medium) and "Comments" in Medium blogs. To ensure balanced comparisons, we randomly select $16,600$ blogs with a 1:1 class ratio. Mann-Whitney U tests reveal statistically significant differences in the number of "Likes" and "Comments" between the two classes ($p < 0.05$).

As shown in Figure A4a, the predicted-AIGTs receive fewer "Likes" on average than predicted-HWTs, with mean values of 69.15 and 127.59, respectively. And predicted-AIGTs exhibit a higher frequency of low "Likes" counts. Figure A4b shows that predicted-AIGTs receive fewer "Comments" on average compared to predicted-HWTs, with mean values of 4.16 and 7.38, respectively. We further investigate the mean values of Likes and Comments for authors with different numbers of followers and Table 3 indicates that, across all follower count groups, AIGTs receive significantly fewer Likes and Comments compared to HWTs.

To summarize, predicted-HWTs obtain more "Likes" and "Comments", which indicates that users in Medium are generally more willing to engage with human-written content. However, the relatively small gap between the two suggests that AI-generated content appeals to users.

Author Analysis. On Medium, we randomly select $1,000$ authors from the predicted-AIGTs group who have published at least ten articles. We collect and detect all of their published articles to determine if they are AI-generated, aiming to explore the potential relationship between an author's follower count and their usage of AI-generated content.

As shown in Figure 6, we divide these authors into three groups based on their follower count. Among the groups, those with $1,000$ or fewer followers exhibit a stronger concentration in the high AAR range ($\geq 75.00\%$). This group also achieves the highest mean AAR at 54.02%. From the overall distribution, as the follower number increases, the AAR gradually shifts toward the lower range ($\leq 25.00\%$). This trend may stem from more popular authors prioritizing content quality, while less-followed authors rely on LLMs to boost efficiency.

Furthermore, Figure A6 illustrates the publication timeline of the first articles detected as AIGTs from these authors. It can be observed that there is a significant increase in such publications during the month GPT-3.5 is released, followed by a relatively stable trend in subsequent months.

6 Conclusion

In this paper, we collect a large-scale dataset, *SM-D*, encompassing multiple platforms and diverse time periods, providing the first comprehensive quantification and analysis of AIGTs on online social media. We construct *AIGTBench*, an AIGT detection benchmark integrating diverse LLMs, to identify the most effective detector,**OSM-Det**. We then perform temporal tracking analyses, highlighting distinct trends in AAR that are shaped by platform-specific characteristics and the increasing adoption of LLMs. Finally, our analysis uncovers critical differences between AIGTs and HWTs across linguistic patterns, topical features, engagement levels, and the follower distribution of authors. Our findings offer valuable perspectives into the evolving dynamics of AIGTs on social media.

7 Ethical Statement

We emphasize that the purpose of this research is not to expose or criticize specific platforms or users for employing AIGTs nor to interfere with legitimate content-creation activities. Instead, our goal is to provide valuable insights through scientific analysis to aid the research community and the public to better understand the current state and trends of generative AI usage on social media. All data used in our paper is publicly available, and we do not collect and monitor any private information.

8 Limitations

In this paper, we conduct long-term quantification of AIGTs on 3 commonly used social media platforms, but there are still some limitations:

1. **Limited coverage of LLMs:** *AIGTBench* includes only 12 LLMs and does not cover all LLMs released across different time periods. Although the current AIGT detectors can generalize to LLMs that are not involved in training to a certain extent (Li et al., 2024), there may still be slight errors, which poses a potential impact on the accuracy of some results. We also note that *AIGTBench* exhibits a distributional bias in the number of LLM-generated texts, favoring the GPT series and Llama series models, which dominate its composition at 42.9% and 39.05%, respectively. However, this bias is unlikely to significantly impact the analysis results, as these models are also the most widely used in real-world applications.
2. **Lack of analysis on multilingual platforms:** Our research focuses on English-dominated social media platforms. Therefore, the applicability of our findings is restricted to these specific platforms and language contexts. Since data collection is a long-term process, we plan to gradually expand to multilingual environments and more platforms in future research to improve the universality of the conclusions.
3. **Insufficient dimensions of analysis across platforms:** We conduct an in-depth analysis of the three dimensions of topic, engagement, and author on the Medium platform, but we are unable to conduct similar multi-dimensional research on Quora and Reddit. This is mainly due to the differences in data collection methods and the difficulty of different platforms. If richer data from these platforms becomes available in the future, we will supplement and enhance the analysis.
4. **Inability of the AIGT detector to detect texts generated by early models release before ChatGPT:** During the dataset collection and training process, we only included models that were released after November 2022. This decision was made because our study specifically focuses on more powerful models such as ChatGPT. We acknowledge that this approach may lead to misclassifications for earlier models; however, this does not affect the overall experimental results or the validity of our findings.

References

Aaditya. 2024. Llama 3 - openbiollm - 8b model. Accessed: 2025-02-15.

Adam D'Angelo, 2023. 2024. Poe ai introduction. Available at: https://quorablog.quora.com/Poe-1 [Accessed: 2024-12-05].

Anthropic. 2024. Anthropic official website. Accessed: 2024-11-04.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. ConDA: Contrastive domain adaptation for AI-generated text detection. In *IJCNLP-AACL*, pages 598–610, Nusa Dua, Bali. Association for Computational Linguistics.

BigModel. 2024. Glm-4 api documentation. Accessed: 2025-02-15.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*.

Wei-Lin Chiang, Zhuohan Li, Zi 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. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Cognitive Computations. 2024. Dolphin 3.0 - llama 3.1 - 8b model. Accessed: 2025-02-15.

DeepMind. 2024. Gemini flash. Accessed: 2025-02-15.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *arXiv preprint arXiv:2406.15583*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

GPTZero. 2024. Gptzero. Accessed: 2024-11-04.

Dritjon Gruda. 2024. Three ways chatgpt helps me in my academic writing. *Nature*. Advance online publication.

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

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 542–556.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. In *ACM Conference on Computer and Communications Security (CCS)*. ACM.

Heralax. 2025. Mannerstral-dataset.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023. Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, Singapore. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yule Liu, Zhiyuan Zhong, Yifan Liao, Zhen Sun, Jingyi Zheng, Jiaheng Wei, Qingyuan Gong, Fenghua Tong, Yang Chen, Yang Zhang, and Xinlei He. 2024b. On the generalization ability of machine-generated text detectors. *arXiv preprint arXiv:2412.17242*.

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

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2024. Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts. *arXiv preprint arXiv:2406.12549*.

Magpie-Align. 2025a. Magpie-reasoning-v1-150k-cot-qwq.

Magpie-Align. 2025b. Magpie-reasoning-v2-250k-cot-deepseek-r1-llama-70b.

Medium. 2024. Explore topics on medium. Accessed: 2025-02-15.

Medium. 2024. Medium. Accessed: 2024-11-04.

Meta AI. 2024. With 10x growth since 2023, llama is the leading engine of ai innovation. Accessed: 2024-11-04.

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*, pages 24950–24962. PMLR.

Moonshot. 2024. Mootshot llm. Accessed: 2024-11-04.

OdiaGenAI. 2025. Roleplay-english.

OpenAI. 2022. Introducing chatgpt. Accessed: 2024-11-04.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-11-04.

OpenGVLab. 2024. Internvl2.5-8b model. Accessed: 2025-02-15.

OpenGVLab. 2025. Internvl-sa-1b-caption.

PJMixers-Dev. 2025. camel-ai_chemistry-gemini-2.0-flash-thinking-exp-1219-customsharegpt.

Quora. 2024. Quora. Accessed: 2024-11-04.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reddit. 2024. Reddit. Accessed: 2024-11-04.

M Scott, Lee Su-In, et al. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774.

Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. 2024. Ten Words Only Still Help: Improving Black-Box AI-Generated Text Detection via Proxy-Guided Efficient Re-Sampling. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 494–502.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. *arXiv preprint arXiv:2401.06712*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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. *arXiv preprint arXiv:2109.13296*.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H. Luan. 2023c. A survey on chatgpt: Ai–generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 4:280–302.

Xwin-LM. 2024. Xwin-lm 13b v0.2 model. Accessed: 2025-02-15.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

A Introduction of LLMs in Detector Training Dataset

In this paper, we have selected the most representative LLMs as our detection targets:

- **Llama-1 (Feb. 2023)** (Touvron et al., 2023a), **Llama-2 (Jul. 2023)**(Touvron et al., 2023b), and **Llama-3 (Apr. 2024)** (Dubey et al., 2024): The Llama series (from Llama-1 to Llama-3) launched by Meta are powerful and extremely popular open source models. This series of models enables researchers to fine-tune diverse datasets, is highly scalable, and is suitable for various research and development environments. The latest version, Llama-3, is equipped with a larger parameter size and optimized training architecture, making it perform better in text generation, context understanding, and complex task processing.
- **ChatGPT/GPT-3.5 Turbo (Nov. 2022)** (OpenAI, 2022): GPT-3.5, an optimized version of GPT-3 by OpenAI, was released in 2022. By incorporating a Reinforcement Learning from Human Feedback (RLHF) reward mechanism and human feedback data, GPT-3.5 achieves significant improvements in accuracy and coherence in text generation. This version includes the Text-DaVinci-003 and GPT-3.5 (or GPT-3.5 Turbo), which focuses on fluent and natural multi-turn conversations and serves as the core model for systems like ChatGPT website.
- **GPT4o-mini (Jul. 2024)** (OpenAI, 2024): Developed by OpenAI, GPT4o-mini is a lightweight language model optimized from GPT-4o technology. This model is designed to deliver efficient language processing capabilities that are suitable for applications with lower resource requirements. It supports both text and visual input, with future plans to expand into audio and video input and output. Since its release, the GPT4o-mini has progressively replaced the GPT-3.5 Turbo as the core model on the ChatGPT website.
- **Claude (Mar. 2023)** (Anthropic, 2024), : Claude is an advanced AI assistant developed by Anthropic. It is a closed-source model designed to communicate efficiently and intuitively with users through NLP technology. Claude can understand and generate human language to assist users in completing a variety of tasks, including answering questions, writing content, and programming assistance.
- **Alpaca 7B (Mar. 2023)** (Taori et al., 2023): Alpaca 7B is a lightweight instruction-following model released by Stanford University, based on Meta's Llama-7B model and fine-tuned on the dataset of $52,000$ instruction-following examples. This fine-tuning markedly enhances the model's performance in understanding and executing task instructions. In evaluations of single-turn instruction-following tasks, Alpaca demonstrates performance comparable to OpenAI's Text-DaVinci-003, exhibiting high-quality responses to instructions.
- **Vicuna 13B (Mar. 2023)** (Chiang et al., 2023): Released by the LMSYS team, Vicuna 13B is based on Meta's Llama-13B model and trained on a large dataset of conversation data aggregated from high-quality models like GPT-3.5. The goal is to develop an open-source conversational model that approaches the quality of GPT-3.5.
- **Moonshot-v1 (Oct. 2023)** (Moonshot, 2024): Developed by Moonshot AI, Moonshot-v1 is an advanced large language model for text generation. This model can understand and generate natural language text, manage everyday conversational exchanges, and produce structured content in various forms, such as articles, code, and summaries, across specialized domains.
- **Mixtral 8×7B (Dec. 2023)** (Jiang et al., 2024): Developed by Mistral AI, this LLM employs a Sparse Mixture of Experts (SMoE) architecture. It has demonstrated exceptional performance across multiple benchmarks, surpassing models like Llama-2 70B and GPT-3.5, especially excelling in tasks involving mathematics, code generation, and multilingual understanding.

B Introduction of Detectors

In this work, we adopt metric-based detectors from the MGTBench framework to detect AIGTs, including:

- **Log-Likelihood** (Solaiman et al., 2019): We evaluate the likelihood of text generation by computing its log-likelihood score under a specific language model. The model constructs a reference distribution based on HWTs and AIGTs to calculate the log-likelihood score of the input text. A higher score suggests a greater likelihood of the text being LLM-generated.
- **Rank** (Gehrmann et al., 2019) and **Log-Rank** (Mitchell et al., 2023): The Rank method identifies the source of generation by analyzing

the ranking of each word in the text. The model calculates the absolute ranking of each word based on context and averages all word rankings to derive an overall score. Generally, a lower score indicates that the text is more likely to be LLM-generated. Log-Rank, a variant of Rank, employs a logarithmic function when calculating each word's ranking, enhancing the detection of AIGTs.

- **Entropy** (Gehrmann et al., 2019): The Entropy method calculates the average entropy value of each word in the text under context conditions. Studies show that AIGTs tend to have lower entropy values.
- **GLTR** (Gehrmann et al., 2019): GLTR is a supportive tool for detecting AIGTs that use the ranking of words generated by a language model to sort the vocabulary of the text by predicted probability. Following Guo *et al.* (Guo et al., 2023), we employ the Test-2 feature to analyze the proportion of words in the top 10, 100, and 1000 ranks to assess the generative nature of the text.
- **DetectGPT** (Mitchell et al., 2023), **NPR**, and **LRR** (Su et al., 2023): The DetectGPT method introduces minor perturbations into the original text and observes changes in the model's log probability to detect its source. AIGTs typically reside at the local optima of the model's log probability function, whereas HWTs show greater changes in log probability after perturbation. The NPR method, similar to DetectGPT, focuses on observing significant increases in log-rank following perturbations to differentiate between AIGTs and HWTs. By combining log-likelihood and log-rank information, the LRR method captures the adaptiveness of generated texts in probability distributions while reflecting the text's ordinal preference relative to HWTs. This dual metric markedly enhances the detection accuracy.

We also consider model-based detectors, including:

- **OpenAI Detector** (Solaiman et al., 2019): This detector fine-tunes a RoBERTa (Liu, 2019) model using output data generated by the GPT-2 large, which has 1.5 billion parameters, to predict whether texts are LLM-generated.
- **ChatGPT Detector** (Guo et al., 2023): Trained using the HC3 dataset, this approach employs a RoBERTa model and various training methods to distinguish between human and AIGTs. We select one that uses only the response texts to align with other detectors, following instructions described by He (He et al., 2024).
- **ConDA** (Bhattacharjee et al., 2023): This method enhances model discrimination of text sources in the feature space by maximizing the feature differences between generated samples and real samples. It also introduces a contrastive learning loss to improve detection accuracy.
- **GPTZero** (GPTZero, 2024): A tool aimed at AIGT detection that analyses the perplexity and burstiness of texts to determine their generative nature. GPTZero provides a public API interface capable of returning a confidence score indicating whether a text is LLM-generated.
- **CheckGPT** (Liu et al., 2024c): The CheckGPT uses the pre-trained Roberta model to extract text features. Then, it uses LSTM to classify the text features and determine whether the text is LLM-generated or human-generated.
- **LM-D Detector** (Ippolito et al., 2019): This approach adds an additional classification layer to a pre-trained language model (like RoBERTa) and fine-tunes it to differentiate between human-made and AIGTs. Inspired by the research of Li *et al.*(Li et al., 2024), which shows that Longformer (Wang et al., 2023b) has robust performance in detecting AIGT in out-of-domain texts, we also use the Longformer-base-4096 model to assess its performance in AIGT detection.

C Social Media Platforms

To select suitable social media platforms for testing AIGT detection, we particularly consider the platform's mainstream status, the diversity of content, and their unique characteristics. Ultimately, we choose Reddit, Medium, and Quora as representative platforms.

- Reddit (Reddit, 2024) is a social discussion platform where users autonomously create and manage "subreddit" sections featuring diverse and rich content themes. All content on the site is categorized into different "subreddits" according to user interests, covering a wide range of topics from technology to social issues. We choose Reddit not only for its active user base—with around $330M$ monthly active users—but also for its vast content diversity, including millions of subreddit topics, allowing it to cover a variety of discussion scenarios.
- Medium (Medium, 2024) is an American online publishing platform developed by Evan Williams and launched in August 2012. It centers on high-

quality original articles and blog content and exemplifies social journalism, known for its content's depth, length, and professionalism.

- Quora (Quora, 2024) is a platform to gain and share knowledge. It enables users to ask questions and connect with people who provide unique insights or quality answers. Users can pose questions and receive answers from other users on topics ranging from daily life to highly specialized academic, technical, and professional queries.

We have selected these 3 platforms because their main functionalities closely align with common use cases for LLMs, such as writing and question-answering. Based on this, we hypothesize that there may be instances where users utilize LLMs to generate content on these platforms.

D Introduction of Open Source Datasets for Training Detectors

We consider 6 publicly available AIGT datasets and 5 common supervised finetuning datasets as one part of *AIGTBench*.

- The **MGT-Academic** dataset (Liu et al., 2024b), assembled from textual sources such as Wikipedia, arXiv, and Project Gutenberg, covers STEM, Social Sciences, and Humanities. It is generated by various LLMs, including Llama3, GPT-3.5 Turbo, Moonshot, and Mixtral $8 \times 7B$, forming a comprehensive AIGT dataset.
- The **Coco-GPT3.5** dataset (Liu et al., 2023), produced using OpenAI's text-davinci-0035 model, incorporates entire newspaper articles from December 2022 to February 2023, reflecting the latest content of that period.
- The **GPABench2** dataset (Liu et al., 2024c), based on the GPT-3.5 Turbo model, focuses on 3 LLM-generated tasks: GPT-written, GPT-completed, and GPT-polished, all based on academic abstracts. Due to the extensive amount of text generated by GPT-3.5 Turbo, we sampled around $100M$ tokens from this dataset for compilation.
- The **LWD** dataset (Soto et al., 2024) involves texts generated by Llama-2, GPT-4, and ChatGPT. Researchers designed specific prompts to "write an Amazon review in the style of the author of the following review: <human review>", where each prompt incorporates a real human-written Amazon review as a stylistic reference.
- The **HC3** dataset (Guo et al., 2023), collected by researchers, comprises nearly 40,000 questions and their answers from human experts and ChatGPT, covering a broad range of fields including open-domain, computer science, finance, medicine, law, and psychology.
- The **AIGT** dataset (Shi et al., 2024) samples human-generated content and content from seven popular open-source or API-driven LLMs, applied in real-world scenarios such as low-quality content generation, news fabrication, and student cheating. Due to the markedly lesser capabilities of GPT-2 XL and GPT-J compared to GPT-3.5, these models were not included.
- Given that high-quality Supervised Finetuning (SFT) datasets are frequently used for fine-tuning LLMs, and considering the lack of Claude and GPT-4 model-related content in the AIGT detection datasets, we also incorporate four SFT datasets with instruction-following features: **Claude2-Alpaca**[1], **Claude-3-Opus-Claude-3.5-Sonnet-9k**[2], **GPTeacher/GPT-4 General-Instruct**[3], and **Instruction in the Wild**[4].

E Data Preprocessing for the *SM-D* and *AIGTBench* Datasets

***SM-D* Dataset.** For the *SM-D* dataset, we exclude texts with fewer than 150 characters (including spaces) and texts where the proportion of English content is below 90%. Plus, we observe that LLMs' responses often contain redundant or irrelevant content. For example, many LLMs' generated texts include irrelevant phrases at the beginning, such as "Of course..." or "Hey there...". Additionally, we find that responses generated by the Llama model often repetitively display strings of numbers or specific symbols, hitting the generation length limit instead of providing a complete answer. like "....throwaway11111...". We filter and remove these anomalous generated contents to enhance the accuracy of our dataset.

***AIGTBench* Dataset.** For the *AIGTBench* dataset, we exclude texts with fewer than 150 characters (including spaces) and texts where the proportion

[1] https://github.com/Lichang-Chen/claude2-alpaca.

[2] https://huggingface.co/datasets/QuietImpostor/Claude-3-Opus-Claude-3.5-Sonnet-9k.

[3] https://github.com/teknium1/GPTeacher/tree/main/Instruct.

[4] https://github.com/XueFuzhao/InstructionWild.

of English content is below 90%.

F Task Prompts for Generated AIGTs from Social Media

Inspired by (Liu et al., 2024c), below are designed task prompts for polishing texts on Medium, Quora, and Reddit.

> Please act as a social media platform Medium/Quora/Reddit content creator.
> Your task is to polish the following content. Follow these guidelines:
> 1. Ensure the content flows naturally and is enjoyable to read.
> 2. Use simple and relatable language to connect with a broad audience.
> 3. Highlight key points in a concise and impactful way.
> 4. Make the content feel more conversational and friendly.
> 5. Where appropriate, add an engaging tone to draw the reader in.
> 6. Respond with the revised content only and nothing else:
>
> Here is the original content: "{content}"

Below are designed task prompts for answering the questions on Quora and Reddit.

> You are a content creator on Quora/Reddit.
> Your task is to generate a thoughtful and insightful answer to the following question. Follow these guidelines:
> 1. Provide a clear and comprehensive explanation that addresses the question thoroughly.
> 2. Use simple, relatable language to connect with a broad audience, making the content easy to understand.
> 3. Highlight key points with examples or anecdotes where applicable, to make the answer more engaging.
> 4. Add a conversational and friendly tone to make the answer feel more approachable.
> 5. Ensure the answer is well-structured, with an introduction, body, and conclusion, for better readability.
> 6. Where relevant, include unique insights or perspectives to make the answer stand out.
> 7. Respond with the generated answer only and nothing else.
>
> Here is the question: "{question}"

Below are two task prompts designed for summarizing Medium articles and writing detailed articles based on those summaries for Medium articles.

> You are a helpful, respectful, and honest assistant.
> Summarize the following content succinctly:
> "{content}"
> Summary:

> You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe.
> Write a detailed article based on the summary below, following these guidelines:
> 1. Ensure it flows naturally and is enjoyable to read.
> 2. Use simple and relatable language for a broad audience.
> 3. Highlight key points in a concise, impactful way.
> 4. Make it conversational and friendly.
> 5. Add an engaging tone where appropriate.
>
> Summary:
> "{summary content}"
> Article:

G Detailed Performance of OSM-Det

Table A6 presents the performance of **OSM-Det** on individual platform-specific datasets within *AIGTBench*. The results show that **OSM-Det** achieves consistently high accuracy across all three platforms, with accuracy scores of 0.995, 0.999, and 0.984 on Medium, Quora, and Reddit, respectively.

Figure A7 illustrates the accuracy and F1-score across different text lengths in *AIGTBench*. We observe that accuracy is relatively lower for shorter texts, with an accuracy of approximately 0.940 for texts between $0 - 149$ characters. However, for texts exceeding 150 characters, accuracy improves significantly to above 0.980. Both accuracy and F1-score continue to increase as text length increases.

To ensure the reliability of our conclusions, we filter out texts shorter than 150 characters from *SM-D* in our paper.

H Generalizability of OSM-Det

AIGTBench is a comprehensive dataset that contains multi-source, multi-domain, and multi-LLM data. The diversity of this dataset enhances the generalizability of detectors in real-world (in-the-wild) environments. **OSM-Det**, on the other hand, is the optimal detection model trained on *AIGTBench*.

In this section, we evaluate the generalizability of **OSM-Det** from three perspectives: AIGTs produced with different generation parameters, AIGTs of social media generated by unseen models, and tests in the wild.

Different Generation Parameters. To investigate whether **OSM-Det** can effectively detect AIGTs generated with different generation parameters, we randomly sample $5,000$ HWTs from the *AIGTBench* and apply the same prompt to refine them using different generation parameters (including temperature, top-p, and top-k). The models used for this experiment are GPT4o and GPT4o-mini.

As shown in Figure 2, **OSM-Det** maintains an accuracy of over 0.99 across the entire range of temperature settings (0.1 to 1.0). Top-P (0.1 to 1.0) and Top-K (1 to 200) show a similar trend. This indicates that **OSM-Det** demonstrates strong generalizability when detecting AIGTs generated with different parameters.

AIGTs of Social Media Generated By Unseen Models. To investigate the generalizability of **OSM-Det** on social media AIGTs generated by unseen models, we selected 6 pre-trained models, including Deepseek-V3 (Liu et al., 2024a), GLM-4-Flash (BigModel, 2024), Gemini-1.5-Flash (DeepMind, 2024), Gemini-2.0-Flash (DeepMind, 2024), Yi-1.5-34B (Young et al., 2024), and InternVL2.5-8B (OpenGVLab, 2024). Additionally, we include three fine-tuned models based on the LLaMA series: Dolphin3.0-Llama3.1-8B (Computations, 2024), Llama3-OpenBioLLM-8B (Aaditya, 2024), and Xwin-LM-13B-V0.2 (Xwin-LM, 2024). Since none of these models were included in *AIGTBench*, they are considered unseen to **OSM-Det**. We also apply the same polishing process to the previously selected $5,000$ data samples.

From Table A8, we observe that **OSM-Det** maintains strong detection performance across these unseen models. The lowest performance was recorded for InternVL2.5-8B, yet it still achieve an accuracy of 0.925 and an F1-score of 0.958. This demonstrates that **OSM-Det** exhibits strong generalization capability when detecting AIGTs generated by previously unseen LLMs.

Test In the Wild. To test **OSM-Det** in the wild, we randomly select datasets from the huggingface platform for evaluation. These datasets are in two main categories: unseen models and unseen domains, neither of which are included in *AIGTBench*.

As shown in Table A7, for the unseen model scenario, the test results align with previous findings, where **OSM-Det** maintains high accuracy and F1-score. Similarly, in the unseen domain scenario, **OSM-Det** also demonstrates strong generalizability, achieving a minimum accuracy of 0.943. This is consistent with the findings of Liu et al. (Liu et al., 2024b), which suggest that AIGT detectors exhibit generalizability across different domains.

I Details About the Collection of High-Frequency Words and Model Interpretation Analysis Methods

I.1 Collection of High-Frequency Words

We use the Spacy library (Honnibal et al., 2020) to classify the part-of-speech of words in the *AIGTBench*, specifically dividing them into adjectives, adverbs, and connectives. We then select around the top 20 words for human-preferred and AI-preferred categories, respectively. For detailed results, refer to Table A5.

I.2 Model Interpretation Analyze Methods

Here are the details and how we implement the two different methods:

- **Integrated Gradients** give an importance score to each input value by calculating the gradient of the detector. We follow (Kokhlikyan et al., 2020) for implementation.
- **Shapley Value** is originally introduced in (Shapley, 1953) and recently apply to machine learning interpretation. It quantifies the impact of each feature by perturbing the input value and observing the contributions in the prediction. We follow (Scott et al., 2017) for implementation.

Dataset	Type	Sentence Number
Medium	Llama Series	$1,881,733$
	GPT Series	$681,480$
	Human	$2,033,105$
Quora	Llama Series	$1,974,368$
	GPT Series	$721,878$
	Human	$569,749$
Reddit	Llama Series	$2,892,584$
	GPT Series	$1,391,054$
	Human	$2,695,271$
Total	AIGTs	$9,543,097$
	HWTs	$5,298,125$

Table A1: Sentence number statistics of our generated datasets (Llama Series include Llama-1, 2, 3; GPT Series include GPT-3.5, GPT4o-mini).

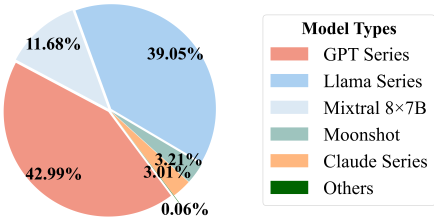
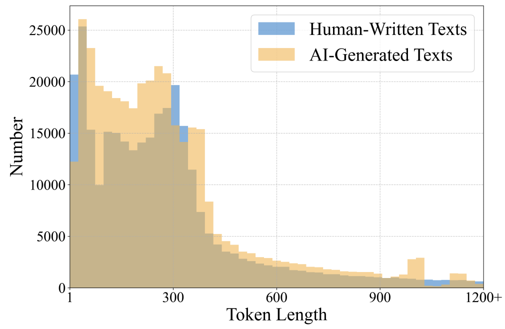
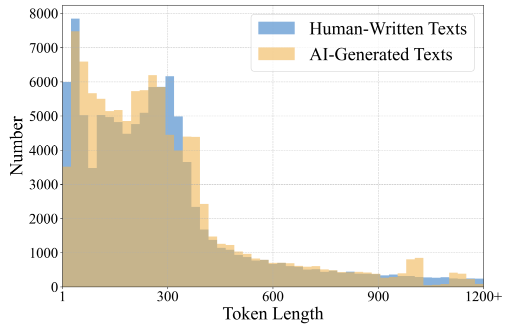


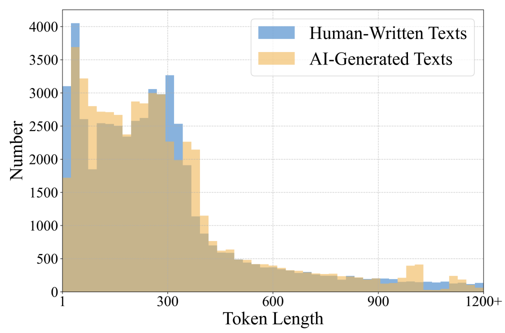
Figure A1: Proportion of total sentences various LLMs, with “Others” including Alpaca 7B and Vicuna 13B.



(a) Token length distribution in the training set.



(b) Token length distribution in the testing set.



(c) Token length distribution in the validation set.

Figure A2: Token length distribution in the training, testing, and validation sets, calculated by the Llama-2 tokenizer (Touvron et al., 2023b).

Platform	# text (Human)	FPR
Medium	$116,303$	1.82%
Quora	$101,145$	1.36%
Reddit	$53,321$	1.70%

Table A2: FPR of **OSM-Det** on social media platforms.

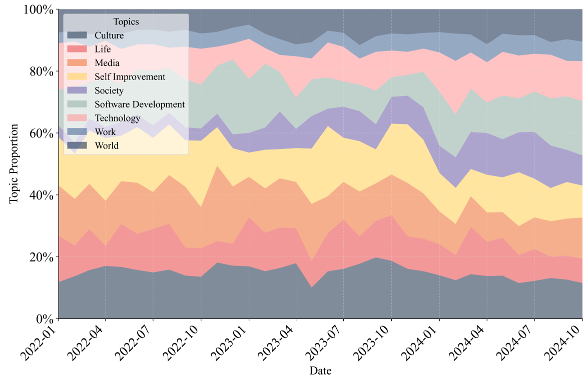
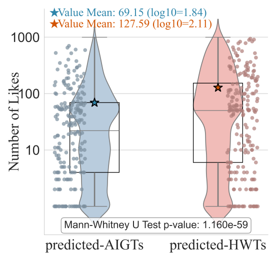
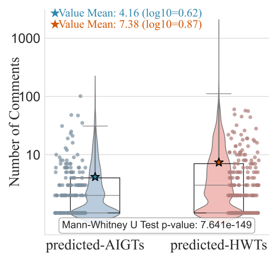


Figure A3: Stacked area chart shows the monthly proportions of 9 topics.



(a) Number of Likes.



(b) Number of Comments.

Figure A4: Differences between predicted AIGTs and predicted HWTs compressed using a log10 transformation.

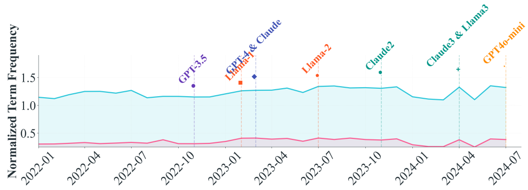


Figure A5: Word frequency trends on Reddit from January 1, 2022, to July 31, 2024.

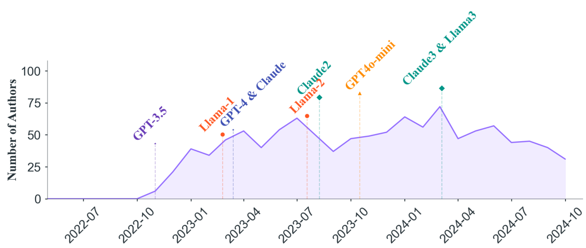


Figure A6: Timeline of authors’ earliest adoption of AIGTs.

Dataset	Type	Sentence Number	Domain
MGT-Academic (Liu et al., 2024b)	Llama3 Mixtral 8×7B Moonshot GPT-3.5 Human	1, 478, 485 2, 639, 498 726, 357 1, 611, 244 6, 007, 476	STEM (Physics, Math, Biology, CS, EE, Statistics, Chemistry, Medicine), Social Science (Education, Economy, Management), Humanities (Literature, Law, Art, History, Philosophy)
Coco-GPT3.5 (Liu et al., 2023)	GPT-3.5 Human	79, 647 55, 565	News
GPABench2 (Liu et al., 2024c)	GPT-3.5 Human	12, 648, 338 (Sample) 1, 065, 860	Computer Science, Physics, Social Sciences
LWD (Soto et al., 2024)	Llama2 GPT-3.5 GPT-4 Human	94, 732 95, 443 62, 632 106, 952	Finance, Social Media
AIGT (Shi et al., 2024)	Llama2 Alpaca 7B Vicuna 13B GPT-3.5 GPT-4 Human	6, 967 6, 083 7, 028 8, 022 7, 156 12, 228	Soical media, News, Academic Writing
HC3 (Guo et al., 2023)	GPT-3.5 Human	184, 692 347, 423	Open-domain, Finance, Medicine, Law, and Psychology

Table A3: Statistics of open-source datasets (part 1).

Dataset	Type	Sentence Number	Domain
Claude2-Alpaca	Claude-2	404, 051	Open-domain
Claude-3-Opus-Claude-3.5-Sonnet-9k	Claude-3 Human	276, 246 37, 785	Open-domain
GPTeacher/GPT-4 General-Instruct	GPT-4 Human	74, 160 24, 465	Open-domain
Alpaca_GPT4	GPT-4 Human	354, 801 22, 253	Open-domain
Instruction in the Wild	GPT-3.5	300, 424	Open-domain

Table A4: Statistics of open-source datasets (part 2).

Category	Words
Human top frequency words	'little', 'small', 'last', 'able', 'bad', 'next', 'right', 'most', 'long', 'old', 'much', 'sure', 'great', 'actually', 'again', 'probably', 'much', 'very', 'pretty', 'already', 'since', 'against', 'yet'
AI top frequency words	'various', 'significant', 'positive', 'complex', 'original', 'free', 'specific', 'unique', 'crucial', 'clear', 'human', 'personal', 'essential', 'particularly', 'especially', 'truly', 'instead', 'here', 'rather', 'additionally', 'despite', 'due to', 'following'

Table A5: Categorization of words into human and AI characteristics.

Platform	Accuracy	F1-score
Medium	0.995	0.995
Quora	0.999	0.999
Reddit	0.984	0.984

Table A6: Performance of **OSM-Det** on AIGTs within *AIGTBench* across different platforms.

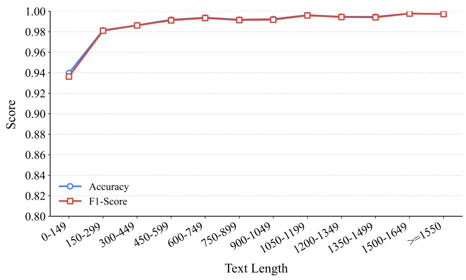


Figure A7: Performance of **OSM-Det** across varying text lengths on *AIGTBench*.

Category	Dataset	Performance	
		Accuracy	F1-score
Unseen Model	QwQ-32B-Preview (Magpie-Align, 2025a)	0.999	0.999
	Gemini-2.0-Flash (PJMixers-Dev, 2025)	0.993	0.997
	Deepseek-R1-Llama-70B (Magpie-Align, 2025b)	0.999	0.999
Unseen Domain	Roleplay-English (OdiaGenAI, 2025)	0.999	0.999
	Mannerstral-dataset (Heralax, 2025)	0.943	0.968
	InternVL-SA-1B-Captio (OpenGVLab, 2025)	0.998	0.999

Table A7: Test **OSM-Det** in the wild. (all datasets from HuggingFace)

Model	Accuracy	F1-score
Deepseek-V3	0.986	0.993
GLM-4-Flash	0.997	0.998
Gemini-1.5-Flash	0.938	0.952
Gemini-2.0-Flash	0.984	0.992
Yi-1.5-34B	0.999	0.999
InternVL2.5-8B	0.925	0.958
Dolphin3.0-Llama3.1-8B	0.996	0.998
Llama3-OpenBioLLM-8B	0.960	0.980
Xwin-LM-13B-V0.2	0.996	0.998

Table A8: Performance of **OSM-Det** on AIGTs generated from unseen LLMs based on social media data from *AIGTBench*.

Prediectd Logits

0.999906, 0.000094

I think it would depend on the root cause of the depression . People that are bullied online and subject to keyboard warriors who go after people . I'd say 100 it's because of the internet . I also think that while there is open source to talk about it . there's also a how you perceive yourself against others . If people you follow are traveling and having fun and posting photos and you are working all day . it can make you feel shitty . Obviously most of those people aren't posting the shit from their lives but you don't know that . I've had that happen to me where I'd be working a ton but feeling good . open Facebook and see everyone that just went on vacation back on vacation while I've got more work to do . It's like shit this kind of sucks . why am I not having fun ? I went through middle high school before social media blew up which started while I was in college and there were depression issues then too . I think it really depends on the person .

(a) Human written text (Reddit)

Prediectd Logits

0.000093, 0.999907

I think the root cause of depression really depends on the individual . For those who face online bullying or get targeted by keyboard warriors . I'd say the internet plays a huge role in that . But there's also the issue of comparison . Social media gives us a window into others lives . and while people share their vacations and fun moments . it can make you feel down . especially if you're stuck at work . You start thinking . Why am I not having that kind of fun ? Even though . let's be real . most of what we see online is just the highlight reel . I've definitely had those moments myself working hard all day . feeling good about it . but then scrolling through Facebook and seeing everyone else on vacation . It's like . Wow . this really sucks . But I also think it's important to remember that depression existed before social media became such a big part of our lives . I was in middle and high school before social media took off . and even back then . there were struggles . So . it really all depends on the person .

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

It's important to consider the root cause of depression , as it can vary for each individual . Online bullying and comparison to others on social media can definitely contribute to feelings of inadequacy and depression . Remember , people often only show the highlight reel of their lives online , not the struggles they may be facing . It's essential to prioritize your mental well being and not compare your behind the scenes to someone else's highlight reel . Seeking support and guidance can help navigate these challenges . You are not alone in feeling this way , and it's okay to reach out for help .

(c) GPT3.5 generated text

Prediectd Logits

0.000095, 0.999905

I think the relationship between the internet and depression is complex , and it really depends on the root cause of the depression . For some people , online bullying and harassment can definitely contribute to feelings of depression . On the other hand , social media can also affect our mental health by creating unrealistic comparisons . When we see others posting about their vacations , accomplishments , or seemingly perfect lives , it's easy to feel like we're missing out or not measuring up . The thing is , most people only share their highlight reels , not their struggles . I've been there too working hard and feeling good , only to scroll through social media and feel like I'm stuck in a rut . It's a reminder that everyone's journey is unique , and we shouldn't compare our behind the scenes moments to others curated posts . Having experienced depression before the rise of social media , I believe that it's not just the internet that's the problem , but rather a complex interplay of factors that affect each person differently .

(d) Llama3 generated text

Figure A8: Case study of word-level analysis through Integrated Gradients on Reddit.

Prediectd Logits

0.999906, 0.000094

I think it would depend on the root cause of the depression. People that are bullied online and subject to keyboard warriors who go after people, I'd say 100% it's because of the internet. I also think that while there is open source to talk about it, there's also a how you perceive yourself against others. If people you follow are traveling and having fun and posting photos and you are working all day, it can make you feel shitty. Obviously most of those people aren't posting the shit from their lives but you don't know that. I've had that happen to me where I'd be working a ton but feeling good, open Facebook and see everyone that just went on vacation back on vacation while I've got more work to do. It's like shit this kind of sucks, why am I not having fun? I went through middle/high school before social media blew up which started while I was in college and there were depression issues then too. I think it really depends on the person.

(a) Human written text (Reddit)

Prediectd Logits

0.000093, 0.999907

I think the root cause of depression really depends on the individual. For those who face online bullying or get targeted by 'keyboard warriors,' I'd say the internet plays a huge role in that. But there's also the issue of comparison. Social media gives us a window into others' lives, and while people share their vacations and fun moments, it can make you feel down, especially if you're stuck at work. You start thinking, 'Why am I not having that kind of fun?' Even though, let's be real, most of what we see online is just the highlight reel. I've definitely had those moments myself—working hard all day, feeling good about it, but then scrolling through Facebook and seeing everyone else on vacation. It's like, 'Wow, this really sucks.' But I also think it's important to remember that depression existed before social media became such a big part of our lives. I was in middle and high school before social media took off, and even back then, there were struggles. So, it really all depends on the person.

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

It's important to consider the root cause of depression, as it can vary for each individual. Online bullying and comparison to others on social media can definitely contribute to feelings of inadequacy and depression. Remember, people often only show the highlight reel of their lives online, not the struggles they may be facing. It's essential to prioritize your mental well-being and not compare your behind-the-scenes to someone else's highlight reel. Seeking support and guidance can help navigate these challenges. You are not alone in feeling this way, and it's okay to reach out for help.

(c) GPT3.5 generated text

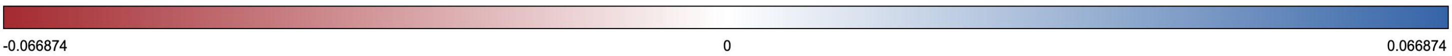
Prediectd Logits

0.000095, 0.999905

I think the relationship between the internet and depression is complex, and it really depends on the root cause of the depression. For some people, online bullying and harassment can definitely contribute to feelings of depression. On the other hand, social media can also affect our mental health by creating unrealistic comparisons. When we see others posting about their vacations, accomplishments, or seemingly perfect lives, it's easy to feel like we're missing out or not measuring up. The thing is, most people only share their highlight reels, not their struggles. I've been there too - working hard and feeling good, only to scroll through social media and feel like I'm stuck in a rut. It's a reminder that everyone's journey is unique, and we shouldn't compare our behind-the-scenes moments to others' curated posts. Having experienced depression before the rise of social media, I believe that it's not just the internet that's the problem, but rather a complex interplay of factors that affect each person differently.

(d) Llama3 generated text

Figure A9: Case study of sentence-level analysis through Shaplay Value on Reddit.



Prediectd Logits

0.999906, 0.000094

I think it would depend on the root cause of the depression. People that are bullied online and subject to keyboard warriors who go after people, I''d say 100% it''s because of the internet. I also think that while there is open source to talk about it, there''s also a how you perceive yourself against others. If people you follow are traveling and having fun and posting photos and you are working all day, it can make you feel shitty. Obviously most of those people aren''t posting the shit from their lives but you don''t know that. I''ve had that happen to me where I''d be working a ton but feeling good, open Facebook and see everyone that just went on vacation back on vacation while I''ve got more work to do. It''s like shit this kind of sucks, why am I not having fun? I went through middle/high school before social media blew up which started while I was in college and there were depression issues then too. I think it really depends on the person.

(a) Human written text (Reddit)

Prediectd Logits

0.000093, 0.999907

I think the root cause of depression really depends on the individual. For those who face online bullying or get targeted by 'keyboard warriors,' I'd say the internet plays a huge role in that. But there's also the issue of comparison. Social media gives us a window into others' lives, and while people share their vacations and fun moments, it can make you feel down, especially if you're stuck at work. You start thinking, 'Why am I not having that kind of fun?' Even though, let's be real, most of what we see online is just the highlight reel. I've definitely had those moments myself—working hard all day, feeling good about it, but then scrolling through Facebook and seeing everyone else on vacation. It's like, 'Wow, this really sucks.' But I also think it''s important to remember that depression existed before social media became such a big part of our lives. I was in middle and high school before social media took off, and even back then, there were struggles. So, it really all depends on the person.

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

It's important to consider the root cause of depression, as it can vary for each individual. Online bullying and comparison to others on social media can definitely contribute to feelings of inadequacy and depression. Remember, people often only show the highlight reel of their lives online, not the struggles they may be facing. It's essential to prioritize your mental well-being and not compare your behind-the-scenes to someone else's highlight reel. Seeking support and guidance can help navigate these challenges. You are not alone in feeling this way, and it's okay to reach out for help.

(c) GPT3.5 generated text

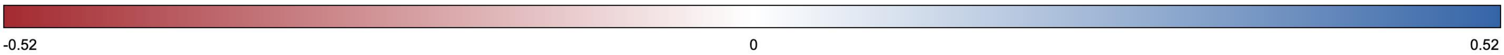
Prediectd Logits

0.000095, 0.999905

I think the relationship between the internet and depression is complex, and it really depends on the root cause of the depression. For some people, online bullying and harassment can definitely contribute to feelings of depression. On the other hand, social media can also affect our mental health by creating unrealistic comparisons. When we see others posting about their vacations, accomplishments, or seemingly perfect lives, it's easy to feel like we're missing out or not measuring up. The thing is, most people only share their highlight reels, not their struggles. I've been there too - working hard and feeling good, only to scroll through social media and feel like I'm stuck in a rut. It's a reminder that everyone's journey is unique, and we shouldn't compare our behind-the-scenes moments to others' curated posts. Having experienced depression before the rise of social media, I believe that it's not just the internet that's the problem, but rather a complex interplay of factors that affect each person differently.

(d) Llama3 generated text

Figure A10: Case study of word-level analysis through Shaplay Value on Reddit.



Prediectd Logits

0.999906, 0.000094

Physical education and athletics as a school subject is not generally emphasized in the Chinese state public school system . It is given less weight in the High school placement examination Zhongkao where other subjects such as Mathematics , Science , Chinese , and English are awarded more points and are given more consideration in classroom hours allotted to these subjects . China performs well in the Olympics because they have a very strong national sports program that is able to identify talent at an early age and provide these young athletes with the necessary resources and training at special athletics schools . They adopted a factory model of training athletes to represent the country at international sporting events . Chinese people as a whole are fairly fit and do like to partake in activities such as dancing , table tennis , running , and basketball as diversions . But professional sports is nowhere as emphasized , developed , or lucrative as in the west where millions of people tune in to watch multimillionaire athletes like Lebron James dunk a basketball or Messi perform a hat trick . I think the root cause of depression really depends on the individual . For those who face online bullying or get targeted by keyboard warriors , I'd say the internet plays a huge role in that . But there's also the issue of comparison . Social media gives us a window into others lives , and while people share their vacations and fun moments , it can make you feel down , especially if you're stuck at work . You start thinking , Why am I not having that kind of fun ? Even though , let's be real , most of what we see online is just the highlight reel . I've definitely had those moments myself working hard all day , feeling good about it , but then scrolling through Facebook and seeing everyone else on vacation . It's like , Wow , this really sucks . But I also think it's important to remember that depression existed before social media became such a big part of our lives . I was in middle and high school before social media took off , and even back then , there were struggles . So , it really all depends on the person .

(a) Human written text (Quora)

Prediectd Logits

0.000093, 0.999907

In China , physical education and athletics aren't usually a major focus in the school curriculum . While subjects like Mathematics , Science , Chinese , and English take center stage especially in the Zhongkao high school entrance exam sports are given much less attention . However , China shines on the Olympic stage , thanks to its strong national sports program . This program excels at spotting talent early , providing young athletes with specialized training and resources at dedicated sports schools . It's almost like a factory model for nurturing top tier athletes who represent the country at global competitions . While professional sports in China aren't as big or as financially rewarding as in the West where fans tune in to watch icons like LeBron James dunk or Messi score breathtaking goals sports are still very much a part of everyday life . Many Chinese people enjoy activities like dancing , table tennis , running , and basketball , which serve as fun and popular diversions .

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

Physical education and athletics may not be as prioritized in the Chinese state public school system , especially in comparison to subjects like Mathematics , Science , Chinese , and English . However , China's strong national sports program does contribute to their success in the Olympics . Through specialized athletics schools and early talent identification , they have been able to nurture young athletes effectively . While professional sports may not be as emphasized or lucrative in China as it is in the West , the Chinese people enjoy activities such as dancing , table tennis , running , and basketball as part of their active lifestyle . The focus on developing athletes from a young age and the factory model approach to training has helped China excel in international sporting events . It's fascinating to see different approaches to sports development around the world , each with its unique strengths and strategies .

(c) GPT3.5 generated text

Prediectd Logits

0.000095, 0.999905

Did you know that physical education and athletics are not typically prioritized in the Chinese state public school system ? In fact , subjects like mathematics , science , Chinese , and English are often given more weight in the high school placement examination , known as Zhongkao . However , China's impressive Olympic performance can be attributed to its robust national sports program , which identifies and nurtures young talent from an early age . The country's factory model approach to training athletes has proven successful in producing world class competitors . While many Chinese people enjoy staying active through activities like dancing , table tennis , running , and basketball , professional sports in China still lags behind the West in terms of development and popularity . In contrast , Western countries have a massive following for professional sports , with millions tuning in to watch superstar athletes like Lebron James and Messi in action . What do you think about the differences in sports culture between East and West ? Share your thoughts ! sports culture China education

(d) Llama3 generated text

Figure A11: Case study of word-level analysis through Integrated Gradients on Quora.

Prediectd Logits

0.999906, 0.000094

Physical education and athletics as a school subject is not generally emphasized in the Chinese state public school system. It is given less weight in the High school placement examination "Zhongkao" where other subjects such as Mathematics, Science, Chinese, and English are awarded more points and are given more consideration in classroom hours allotted to these subjects. China performs well in the Olympics because they have a very strong national sports program that is able to identify talent at an early age and provide these young athletes with the necessary resources and training at special athletics schools. They adopted a "factory model" of training athletes to represent the country at international sporting events. Chinese people as a whole are fairly fit and do like to partake in activities such as dancing, table tennis, running, and basketball as diversions. But professional sports is nowhere as emphasized, developed, or lucrative as in the west where millions of people tune in to watch multimillionaire athletes like Lebron James dunk a basketball or Messi perform a hat trick.

(a) Human written text (Quora)

Prediectd Logits

0.000093, 0.999907

In China, physical education and athletics aren't usually a major focus in the school curriculum. While subjects like Mathematics, Science, Chinese, and English take center stage—especially in the "Zhongkao" high school entrance exam—sports are given much less attention. However, China shines on the Olympic stage, thanks to its strong national sports program. This program excels at spotting talent early, providing young athletes with specialized training and resources at dedicated sports schools. It's almost like a "factory model" for nurturing top-tier athletes who represent the country at global competitions. While professional sports in China aren't as big or as financially rewarding as in the West—where fans tune in to watch icons like LeBron James dunk or Messi score breathtaking goals—sports are still very much a part of everyday life. Many Chinese people enjoy activities like dancing, table tennis, running, and basketball, which serve as fun and popular diversions.

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

Physical education and athletics may not be as prioritized in the Chinese state public school system, especially in comparison to subjects like Mathematics, Science, Chinese, and English. However, China's strong national sports program does contribute to their success in the Olympics. Through specialized athletics schools and early talent identification, they have been able to nurture young athletes effectively. While professional sports may not be as emphasized or lucrative in China as it is in the West, the Chinese people enjoy activities such as dancing, table tennis, running, and basketball as part of their active lifestyle. The focus on developing athletes from a young age and the 'factory model' approach to training has helped China excel in international sporting events. It's fascinating to see different approaches to sports development around the world, each with its unique strengths and strategies.

(c) GPT3.5 generated text

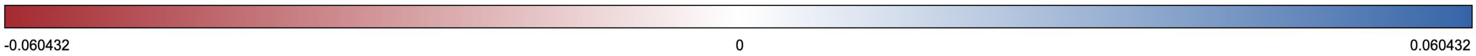
Prediectd Logits

0.000095, 0.999905

Did you know that physical education and athletics are not typically prioritized in the Chinese state public school system? In fact, subjects like mathematics, science, Chinese, and English are often given more weight in the high school placement examination, known as "Zhongkao". However, China's impressive Olympic performance can be attributed to its robust national sports program, which identifies and nurtures young talent from an early age. The country's "factory model" approach to training athletes has proven successful in producing world-class competitors. While many Chinese people enjoy staying active through activities like dancing, table tennis, running, and basketball, professional sports in China still lags behind the West in terms of development and popularity. In contrast, Western countries have a massive following for professional sports, with millions tuning in to watch superstar athletes like Lebron James and Messi in action. What do you think about the differences in sports culture between East and West? Share your thoughts! #sports #culture #China #education

(d) Llama3 generated text

Figure A12: Case study of sentence-level analysis through Shaplay Value on Quora.



Prediectd Logits

0.999906, 0.000094

Physical education and athletics as a school subject is not generally emphasized in the Chinese state public school system. It is given less weight in the High school placement examination ““Zhongkao”” where other subjects such as Mathematics, Science, Chinese, and English are awarded more points and are given more consideration in classroom hours allotted to these subjects.China performs well in the Olympics because they have a very strong national sports program that is able to identify talent at an early age and provide these young athletes with the necessary resources and training at special athletics schools. They adopted a ““factory model”” of training athletes to represent the country at international sporting events.Chinese people as a whole are fairly fit and do like to partake in activities such as dancing, table tennis, running, and basketball as diversions. But professional sports is nowhere as emphasized, developed, or lucrative as in the west where millions of people tune in to watch multimillionaire athletes like Lebron James dunk a basketball or Messi perform a hat trick.

(a) Human written text (Quora)

Prediectd Logits

0.000094, 0.999906

In China, physical education and athletics aren’’t usually a major focus in the school curriculum. While subjects like Mathematics, Science, Chinese, and English take center stage—especially in the ““Zhongkao”” high school entrance exam—sports are given much less attention. However, China shines on the Olympic stage, thanks to its strong national sports program. This program excels at spotting talent early, providing young athletes with specialized training and resources at dedicated sports schools. It's almost like a ““factory model”” for nurturing top-tier athletes who represent the country at global competitions. While professional sports in China aren’’t as big or as financially rewarding as in the West—where fans tune in to watch icons like LeBron James dunk or Messi score breathtaking goals—sports are still very much a part of everyday life. Many Chinese people enjoy activities like dancing, table tennis, running, and basketball, which serve as fun and popular diversions.

(b) GPT4o-mini generated text

Prediectd Logits

0.000096, 0.999904

Physical education and athletics may not be as prioritized in the Chinese state public school system, especially in comparison to subjects like Mathematics, Science, Chinese, and English. However, China's strong national sports program does contribute to their success in the Olympics. Through specialized athletics schools and early talent identification, they have been able to nurture young athletes effectively. While professional sports may not be as emphasized or lucrative in China as it is in the West, the Chinese people enjoy activities such as dancing, table tennis, running, and basketball as part of their active lifestyle. The focus on developing athletes from a young age and the 'factory model' approach to training has helped China excel in international sporting events. It's fascinating to see different approaches to sports development around the world, each with its unique strengths and strategies.

(c) GPT3.5 generated text

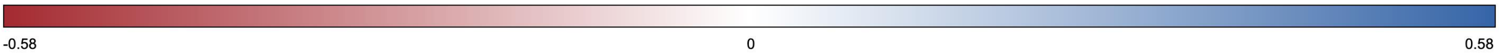
Prediectd Logits

0.000095, 0.999905

Did you know that physical education and athletics are not typically prioritized in the Chinese state public school system? In fact, subjects like mathematics, science, Chinese, and English are often given more weight in the high school placement examination, known as "Zhongkao". However, China's impressive Olympic performance can be attributed to its robust national sports program, which identifies and nurtures young talent from an early age. The country's "factory model" approach to training athletes has proven successful in producing world-class competitors. While many Chinese people enjoy staying active through activities like dancing, table tennis, running, and basketball, professional sports in China still lags behind the West in terms of development and popularity. In contrast, Western countries have a massive following for professional sports, with millions tuning in to watch superstar athletes like Lebron James and Messi in action. What do you think about the differences in sports culture between East and West? Share your thoughts! #sports #culture #China #education

(d) Llama3 generated text

Figure A13: Case study of word-level analysis through Shaplay Value on Quora.



Prediectd Logits

0.999906, 0.000094

That being said why not both ? Why can't we restore the proper place of the family as it once was and create cells of freedom in our society so that society itself can better self govern ? You are right of course . And I am not talking about social engineering I am talking about the instinctive natural evolutionary process we can observe in human society a process we could understand and then start to consciously proactively control . As a result of the relentless irrepressible egoistic individualistic human development we are distancing from each other each person becoming a separate standalone entity caring only for the egoistic self . The real epidemic in human society today is the loneliness epidemic old people dying alone young people not wanting to commit to marriage or to have children of their own .

(a) Human written text (Medium)

Prediectd Logits

0.000095, 0.999905

In today s fast paced world we often find ourselves grappling with a critical question why can t we embrace both tradition and innovation ? It's essential that we restore the family unit to its vital role in our society while also fostering environments that promote individual freedom and self governance This isn't about imposing social engineering but rather understanding the natural instinctive evolution of human society By observing and comprehending these dynamics we can consciously guide our social development However as we advance there s an undeniable trend we are becoming more isolated each of us navigating our paths as independent entities often prioritizing our own desires over communal bonds This growing individualism contributes to a pervasive loneliness epidemic older adults facing solitude in their final years and younger generations hesitating to form lasting commitments or start families We must recognize the urgency of addressing this loneliness crisis It is imperative that we foster deeper connections and support systems within our communities By balancing our instinct for individuality with the nurturing of our familial and societal ties we can create a richer more interconnected existence

(b) GPT4o-mini generated text

Prediectd Logits

0.000094, 0.999906

Absolutely ! It s crucial to acknowledge the importance of family and community especially in today s world where individualism seems to be on the rise Finding a balance between nurturing our familial connections and fostering personal freedom can create healthier more supportive societies Restoring the family unit and encouraging communal ties could combat the loneliness epidemic that many are experiencing We need to look at ways we can proactively create environments that promote connection understanding and commitment It s all about finding that harmony where individuals can thrive while still being part of a larger community Let s work together to build a society where relationships are valued and people feel genuinely connected to one another

(c) GPT3.5 generated text

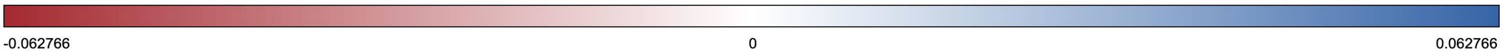
Prediectd Logits

0.000094, 0.999906

Why not both ? We can restore the importance of family and create cells of freedom in society , allowing for better self governance . I'm not advocating for social engineering , but rather understanding and guiding the natural evolution of human society . The current trend of individualism has led to a loneliness epidemic , with people becoming isolated and disconnected from others . We see this in the rising number of elderly people dying alone and young people hesitant to commit to marriage or start families . Let's work together to create a more compassionate and connected society .

(d) Llama3 generated text

Figure A14: Case study of word-level analysis through Integrated Gradients on Medium.



Prediectd Logits

0.999906, 0.000094

That being said why not both. Why can't we restore the proper place of the family as it once was, and create cells of freedom in our society so that society itself can better self-govern?" You are right of course. And I am not talking about "social engineering", I am talking about the instinctive, natural evolutionary process we can observe in human society, which process we could understand and then start to consciously, proactively control. As a result of the relentless, irrepressible egoistic, individualistic human development we are distancing from each other, each person becoming a separate, standalone entity, caring only for the egoistic "self". The real epidemic in human society today is the loneliness epidemic, old people dying alone, young people not wanting to commit to marriage or to have children of their own.

(a) Human written text (Medium)

Prediectd Logits

0.000095, 0.999905

In today's fast-paced world, we often find ourselves grappling with a critical question: why can't we embrace both tradition and innovation? It's essential that we restore the family unit to its vital role in our society while also fostering environments that promote individual freedom and self-governance. This isn't about imposing "social engineering," but rather understanding the natural, instinctive evolution of human society. By observing and comprehending these dynamics, we can consciously guide our social development. However, as we advance, there's an undeniable trend: we are becoming more isolated, each of us navigating our paths as independent entities, often prioritizing our own desires over communal bonds. This growing individualism contributes to a pervasive loneliness epidemic—older adults facing solitude in their final years and younger generations hesitating to form lasting commitments or start families. We must recognize the urgency of addressing this loneliness crisis. It is imperative that we foster deeper connections and support systems within our communities. By balancing our instinct for individuality with the nurturing of our familial and societal ties, we can create a richer, more interconnected existence.

(b) GPT4o-mini generated text

Prediectd Logits

0.000094, 0.999906

Absolutely! It's crucial to acknowledge the importance of family and community, especially in today's world where individualism seems to be on the rise. Finding a balance between nurturing our familial connections and fostering personal freedom can create healthier, more supportive societies. Restoring the family unit and encouraging communal ties could combat the loneliness epidemic that many are experiencing. We need to look at ways we can proactively create environments that promote connection, understanding, and commitment. It's all about finding that harmony where individuals can thrive while still being part of a larger community. Let's work together to build a society where relationships are valued, and people feel genuinely connected to one another.

(c) GPT3.5 generated text

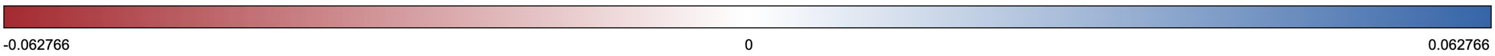
Prediectd Logits

0.000094, 0.999906

Why not both? We can restore the importance of family and create cells of freedom in society, allowing for better self-governance. I'm not advocating for social engineering, but rather understanding and guiding the natural evolution of human society. The current trend of individualism has led to a loneliness epidemic, with people becoming isolated and disconnected from others. We see this in the rising number of elderly people dying alone and young people hesitant to commit to marriage or start families. Let's work together to create a more compassionate and connected society.

(d) Llama3 generated text

Figure A15: Case study of sentence-level analysis through Shaplay Value on Medium.



Prediectd Logits

0.999906, 0.000094

That being said why not both. Why can't we restore the proper place of the family as it once was, and create cells of freedom in our society so that society itself can better self-govern?" You are right of course. And I am not talking about "social engineering", I am talking about the instinctive, natural evolutionary process we can observe in human society, which process we could understand and then start to consciously, proactively control. As a result of the relentless, irrepressible egoistic, individualistic human development we are distancing from each other, each person becoming a separate, standalone entity, caring only for the egoistic "self". The real epidemic in human society today is the loneliness epidemic, old people dying alone, young people not wanting to commit to marriage or to have children of their own.

(a) Human written text (Medium)

Prediectd Logits

0.000095, 0.999905

In today''s fast-paced world, we often find ourselves grappling with a critical question: why can''t we embrace both tradition and innovation? It's essential that we restore the family unit to its vital role in our society while also fostering environments that promote individual freedom and self-governance. This isn't about imposing "social engineering," but rather understanding the natural, instinctive evolution of human society. By observing and comprehending these dynamics, we can consciously guide our social development. However, as we advance, there''s an undeniable trend: we are becoming more isolated, each of us navigating our paths as independent entities, often prioritizing our own desires over communal bonds. This growing individualism contributes to a pervasive loneliness epidemic—older adults facing solitude in their final years and younger generations hesitating to form lasting commitments or start families. We must recognize the urgency of addressing this loneliness crisis. It is imperative that we foster deeper connections and support systems within our communities. By balancing our instinct for individuality with the nurturing of our familial and societal ties, we can create a richer, more interconnected existence.

(b) GPT4o-mini generated text

Prediectd Logits

0.000094, 0.999906

Absolutely! It''s crucial to acknowledge the importance of family and community, especially in today''s world where individualism seems to be on the rise. Finding a balance between nurturing our familial connections and fostering personal freedom can create healthier, more supportive societies. Restoring the family unit and encouraging communal ties could combat the loneliness epidemic that many are experiencing. We need to look at ways we can proactively create environments that promote connection, understanding, and commitment. It''s all about finding that harmony where individuals can thrive while still being part of a larger community. Let''s work together to build a society where relationships are valued, and people feel genuinely connected to one another.

(c) GPT3.5 generated text

Prediectd Logits

0.000094, 0.999906

Why not both? We can restore the importance of family and create cells of freedom in society, allowing for better self-governance. I'm not advocating for social engineering, but rather understanding and guiding the natural evolution of human society. The current trend of individualism has led to a loneliness epidemic, with people becoming isolated and disconnected from others. We see this in the rising number of elderly people dying alone and young people hesitant to commit to marriage or start families. Let's work together to create a more compassionate and connected society.

(d) Llama3 generated text

Figure A16: Case study of word-level analysis through Shaplay Value on Medium.