RedNote-Vibe: A Lifestyle Social Media Dataset for Detecting AI-Generated Text and Analyzing Its Impact

Anonymous ACL submission

Abstract

We introduce **RedNote-Vibe**, a dataset spanning five years (pre-LLM to July 2025) sourced from lifestyle platform RedNote (Xiaohongshu), capturing the temporal dynamics of content creation and is enriched with comprehensive engagement metrics. To address the detection challenge posed by RedNote-Vibe, we propose the **PsychoLinguistic AIGT Detection Framework (PLAD)**. Grounded in cognitive psychology, PLAD leverages deep psychological signatures for robust and interpretable detection. Our experiments demonstrate PLAD's superior performance and reveal insights into content dynamics: (1) human content continues to outperform AI in emotionally resonant domains; (2) AI content is more homogeneous and rarely produces breaking posts, however, this human-AI gap narrows for arousing higher-investment interactions; and (3) most interestingly, a small group of users who strategically utilize AI tools can achieve higher engagement outcomes.

1 Introduction

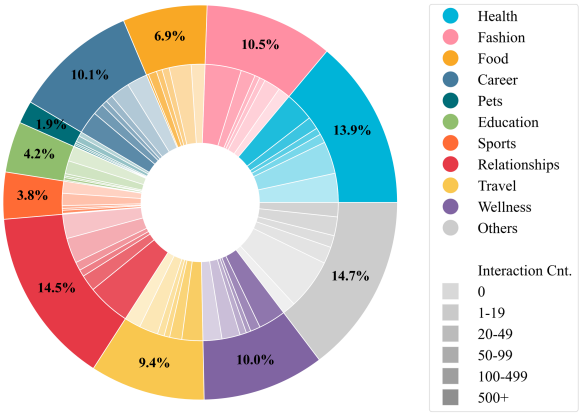
Large Language Models (LLMs) (Achiam et al., 2023; Mallick and Kilpatrick, 2025; Guo et al., 2025) have revolutionized digital content creation, leading to a surge in AI-Generated Text (AIGT). Consequently, developing robust detection methods has become a critical research frontier, with a focus on classification (Gui et al., 2025; Hu et al., 2023) and source attribution (Sun et al., 2025a). However, existing research primarily focus on knowledge-sharing and formal corpora (e.g., news, academic writing). There remains a lack of research on lifestyle-sharing social media platforms, where content is deeply intertwined with personal narratives and commercial influence.

In this work, we identify two critical and unaddressed challenges for AIGT research on social media. First, unlike formal text, where factual accuracy is prioritized, **social media rewards content that maximizes user engagement**, such as likes, comments, and shares (Chung et al., 2023; Cascio Rizzo et al., 2024). On social media platforms that value sharing real-life experiences, LLMs can be prompted to generate sensationalized or controversial content to increase interactions, which undermines connections and trust within the community. Second, **over time, AI and human content are interacting more frequently on social media platforms.** This increasing mixture of content sources presents challenges for detection methods, as linguistic patterns may shift over time. Existing datasets, typically static snapshots from informational sources (e.g., Reddit, Quora), fail to capture these temporal dynamics or the engagement patterns of commercialized social content.

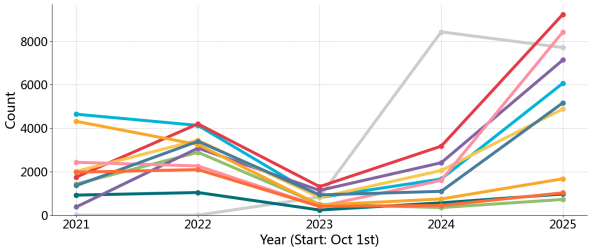
To bridge this gap, we introduce RedNote-Vibe, the first dataset designed for studying AIGT in a dynamic, lifestyle-centric social media context. RedNote-Vibe is collected from RedNote (Xiaohongshu), a leading Chinese platform combining social networking with e-commerce. Our data spans a wide timeline (Oct 2020 - July 2025), covering the transition from the pre-LLM era to the post-LLM era. The dataset features: (1) rich metadata including engagement metrics (likes, comments, collections); (2) parallel AIGT variants generated by 17 diverse LLMs based on human seeds; and (3) an exploration set of real-world post-LLM content. It provides a natural testbed for observing the evolution of AIGT and its impact on user interaction.

Furthermore, facing the challenge that as LLMs evolve, surface-level linguistic artifacts (e.g., syntax patterns) will inevitably be neutralized. We propose the PsychoLinguistic AIGT Detection Framework (PLAD). Instead of relying solely on statistical patterns, PLAD leverages linguistic features grounded in psycholinguistic theories to construct interpretable detection models.

Our experiments demonstrate PLAD not only achieves superior performance compared to black-



(a) Domain and interaction count distribution.



(b) Year distribution across domains.

Figure 1: RedNote-Vibe dataset distribution.

box baselines but also offers insights into how specific psychological signatures correlate with social media engagement. The main contributions of this paper are as follows:

- We introduce RedNote-Vibe, a large-scale Chinese social media AIGT dataset. To our knowledge, it is the first dataset from the RedNote platform.
- We propose PLAD, an interpretable framework leveraging psycholinguistic theories, offering robust detection and explaining the linguistic style of AIGT.
- We provide a comprehensive analysis of the dataset, uncovering temporal trends in AI adoption and revealing how AI-generated content differs from human authorship in driving social interactions.

2 RedNote Engagement Dataset

RedNote (Xiaohongshu)[1] stands as one of the most influential Chinese social media platforms, serving over 300 million monthly active users. This platform emphasizes personal experiences and lifestyle sharing, which makes it particularly vulnerable to AIGC infiltration, as it undermines the authenticity of the content. Despite its influence, RedNote has remained largely unexplored in academic research due to the absence of a publicly available dataset. In this section, we present our RedNote-Vibe dataset, the first publicly available dataset from this platform. It also captures temporal dynamics and engagement patterns, specifically designed for research on AIGT detection and the impact of AIGC on social media.

Domain	#	Length	#Tags	Likes	Comm.	Colls.
Health	19.2k	333.6	7.6	2814.3	97.1	2468.8
Fashion	17.7k	191.2	10.9	1850.5	62.5	681.3
Food	10.5k	398.4	3.0	66.2	11.9	45.0
Career	11.7k	406.9	10.0	2047.2	102.6	1445.9
Pets	3.9k	433.2	2.7	57.5	12.6	34.1
Education	5.9k	496.0	3.7	55.9	6.1	38.3
Sports	17.7k	528.8	3.2	60.1	9.8	36.4
Relation.	22.5k	308.2	7.1	2663.6	275.8	593.9
Travel	14.8k	452.8	10.7	1102.4	86.6	662.0
Wellness	16.0k	396.5	10.5	2348.6	217.7	1025.8
Others	16.9k	220.9	8.9	208.8	1266.2	3225.0

Table 1: Statistics of the RedNote posts across content categories. The post count (#) is presented in thousands (k). Comm. and Colls. refer to the average number of comments and collections, respectively.

2.1 Data Collection and Statistics

Our data collection methodology is grounded in the *2024 RedNote User Behavior Report* [2], which identifies ten dominant content categories: *Career, Wellness, Travel, Health, Food, Pets, Education, Sports, Fashion, and Relationships*. Our dataset focuses on the Chinese language. To ensure our dataset reflects real-world distribution, we employ a two-step retrieval strategy: We first extracted the official popular search keywords provided in the report for each category. For instance, the Relationships category includes specific seed tags such as "Should couples go dutch?" (情侣之间应该AA制吗) and "First meeting with a blind date" (相亲第一次见面). Then, we expand these tag lists based on search recommendations, resulting in about 500 tags for different domains. We use these tags as search keywords to collect the notes.

We collected 148,000 notes published between October 2020 and July 2025. To maintain the quality of the text content, we filter out notes with fewer than 100 Chinese characters, which were mainly image-based. Finally, our data source contains samples with comprehensive metadata including: 1) Content: note title, text content and tags; 2) Temporal information: publication timestamp; 3) Engagement metrics: likes, comments and collections; 4) Topic domain.

[1] https://www.xiaohongshu.com

[2] http://xhslink.com/o/3ZtMUSbgoCl

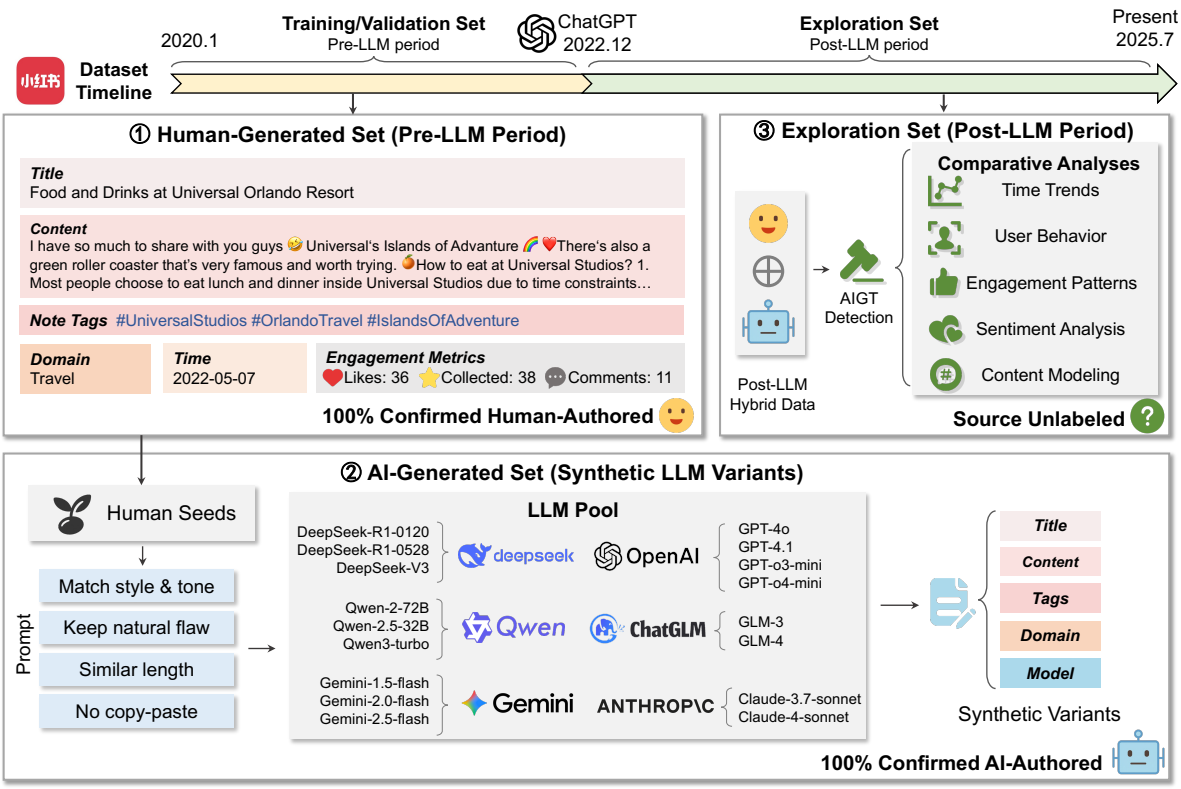


Figure 2: Overview of the RedNote-Vibe dataset construction. **(1) Human-Generated Set** is collected in the pre-LLM period before 2022.12, therefore labeled as human-authored. **(2) AI-Generated Set** is created by prompting a diverse LLM pool with the human seeds to produce synthetic variants. **(3) Exploration Set** contains post-LLM social posts, enabling AIGT detection and extensive analyses comparing AI and human content.

Table 1 presents detailed statistics across domains, which shows that different domains exhibit distinct properties and engagement dynamics. Fig. 1-(a) visualizes the distribution of the total engagement (defined as the sum of all three metrics), Fig. 1-(b) presents the distribution of time span. It can be seen that our dataset covers different levels of interaction metrics, reflecting real-world user interaction behavior on social media.

2.2 AI Samples Generation

To create a comprehensive AIGT detection benchmark, we construct parallel AI-generated versions of RedNote posts. Unlike existing AIGT datasets that mainly direct prompt LLMs to generate AI samples, we employ a pipeline approach to simulate AIGT challenges in real-world scenarios. We design the generation method to be difficult for humans to distinguish. The detailed prompt is shown in Appendix C.1.

We create a model pool comprising 17 representative LLMs from 6 providers. All models receive identical prompts with JSON-formatted output. Notably, we exclusively select seed notes published before November 2022 (pre-LLM period) to ensure their human-authored property. These seeds are randomly distributed across the LLM pool, with each model generating at least 1,000 samples.

Figure 2 illustrates our data construction pipeline and the selected LLM pool. This approach yields a training and validation set with verified human/AI labels. Additionally, we compile an exploration set containing posts from the post-LLM period (2023-2025). While lacking ground-truth labels, this subset enables researchers to investigate real-world content evolution and analyze the emerging linguistic landscape shaped by widespread AI adoption.

In Table 2, we compare our dataset with existing publicly available datasets for AIGT detection in social media. While the existing datasets are valuable, RedNote-Vibe offers unique time spans and timestamps that support AIGT evolution and trend studies in Chinese context. Compared with most related SM-D (Sun et al., 2025b), its sources are rooted in knowledge sharing and community discussions (e.g., Quora, Reddit), RedNote is a lifestyle-sharing platform integrated with commercial activities. This presents distinct challenges:

Dataset	Data Source(s)	#	Time Span	Timestamps	Language(s)	LLMs	Interact.
SM-D (2025b)	Medium, Quora, Reddit	2.4M	Jan 2022 - Oct 2024	✓	English	12	✓
RAID (2024)	News, Books, Reddit, etc.	6.2M	Not Applicable	×	English	11	–
MultiSocial (2024)	Twitter, WhatsApp, etc.	472k	Mostly Pre-2022	×	Multi (22)	7	–
MULTITuDE (2023)	News	74k	Pre-2021	×	Multi (11)	8	–
SAID (2023)	Zhihu	72k	Mar 2023 - Oct 2023	✓	Chinese	–	–
RedNote-Vibe	RedNote	148k	Oct 2020 - Jul 2025	✓	Chinese	18	✓

Table 2: Comparison of RedNote-Vibe with prior social media AIGT datasets. Interact.: SM-D provides the number of followers, comments and likes.

AIGT on RedNote is often crafted into personal narratives (e.g., product reviews, travel experiences, etc.), making it significantly harder to detect compared to informational text. Furthermore, given the platform's commercial attributes, such content poses a more direct influence on consumer behavior. Therefore, RedNote-Vibe not only fills the gap of data source from the RedNote platform but also provides a novel testbed for studying AI-generated content in high-stakes, commercially driven social environments.

2.3 Task Definition

Leveraging our dataset's rich structure, we define three hierarchical classification tasks that reflect real-world AIGT detection scenarios with increasing granularity:

- **AIGT Classification (binary)**: Human vs. AI-generated text detection task, which requires models to distinguish human-written content from any AI-generated text. This task establishes the baseline capability for AIGT detection.
- **AI Provider Identification (6-way)**: A task focusing on the AI-generated subset, where models identify the source among six major AI providers (OpenAI, Google, Anthropic, etc.). This task tests whether detection methods can capture family-level patterns, as models from the same provider often share similar training methodologies and corpus.
- **Model Identification (17-way)**: A fine-grained AI model identification task, where we distinguish between 17 specific AI models within the AI-generated content, representing the most challenging scenario that requires detecting subtle model-specific characteristics.

3 PsychoLinguistic AIGT Detection Framework

Existing approaches to AIGT detection primarily rely on computational linguistic features or supervised training of PLMs, achieving significant performance. However, as AI continues to evolve, the linguistic features such as style and syntax that currently distinguish AIGT will inevitably be neutralized. We hypothesize that detection methods grounded in established psycholinguistic dimensions (e.g., emotional expression patterns, narrative coherence, and lexical diversity) may be more robust to model updates, as these dimensions reflect fundamental differences in how text is generated. Consequently, we propose the PsychoLinguistic AIGT Detection Framework (PLAD), which leverages constructs from psycholinguistics as measurable features for interpretable and potentially more stable AIGT detection.

Our framework quantifies a total of 31 features categorized into four dimensions of human language expression: *Emotional and Social Grounding*, *Cognitive Architecture*, *Lexical Identity and Stylistic Signature*, and *Cohesion and Textual Flow*. These dimensions are rooted in psychological and cognitive theories, capturing aspects from emotional depth to narrative complexity. Table 3 presents the definition of each dimension and the corresponding features with their references.

Our feature set can be categorized into two extraction approaches. (1) Directly computable statistical features that can be obtained using straightforward computational methods, such as emoji density, type-token ratio and other structural measures. (2) Semantically-based features that require evaluation criteria and text analysis tools for assessment. In contrast to traditional approaches to psychological text analysis such as LIWC (Pennebaker et al., 2015), which employs frequency-based word analysis, we adopt a more sophisticated approach using a proxy LLM for psychological text analysis. Specifically, following related work (Rathje et al., 2024; Ghatora et al., 2024), we design evaluation rubrics that convert theoretical constructs into measurable criteria for characteristics. These rubrics are then presented to a proxy LLM, which is instructed to

Dimension	Features & Theoretical Basis
Emotional & Social Grounding *Evaluates emotional depth, social connection, and sensory richness. rooted in personal experience and social awareness.*	• **Emotional Intensity** (Gross, 1998); **Personal Emotional Grounding** (Conway and Pleydell-Pearce, 2000) • **Sensory Detail Richness** (Levine et al., 2002); **Social Connectedness** (Giles and Powesland, 1975) • **Empathetic Engagement** (Baron-Cohen, 1997); **Interactive and Dialogic Stance** (Bakhtin, 2010) • **Unique Emoji Ratio** (Felbo et al., 2017); **Emoji Density** (Felbo et al., 2017)
Cognitive Architecture *Assesses perspectival complexity, reasoning, and narrative structure.*	• **Perspectival Complexity** (Baker-Brown et al., 1992); **Narrative Structure Flexibility** (Labov and Waletzky, 1997) • **Dialectical Argumentation Strength** (Toulmin, 2003); **Self-Correction** (Hyland, 2005) • **Axiological Coherence** (Graham et al., 2009); **Temporal Orientation and Integration** (Reichenbach, 1947) • **Sentence/Word/Character Count** (Manning and Schutze, 1999)
Lexical Identity & Stylistic Signature *Captures stylistic idiolect, rhythmic patterns, and natural imperfections.*	• **Lexical-Stylistic Personalization** (Argamon et al., 2003); **Prosodic Rhythm Consistency** (Halliday and Matthiessen, 2013) • **Imperfection** (Clark and Tree, 2002); **Rhetorical Sophistication** (Grice, 1975) • **Punctuation Ratio**, **Number Ratio** (Manning and Schutze, 1999) • **Type-Token Ratio** (Templin, 1957); **Word Frequency Entropy** (Shannon, 1948); **Word Burstiness** (Gries, 2008)
Cohesion & Textual Flow *Evaluates the organization, semantic progression, and repetition patterns.*	• **Lexical Cohesion** (Halliday and Hasan, 2014); **Inter-Sentential Sentence Similarity** (Foltz et al., 1998) • **Immediate Repetition Density**, **Phrasal Repetition Frequency** (Stamatatos, 2009) • **Sentence Burstiness** (Gries, 2008)

Table 3: Feature list of the PLAD framework, categorized by psychological dimensions. Each feature is grounded in established psycholinguistic theories.

evaluate the input text according to the specified dimensions and provide quantitative scores.

Considering that using a proxy LLM for feature extraction may bring model biases into the feature set. However, we adopt this approach based on several reasons: (1) For complex psycholinguistic constructs, LLMs currently represent the most versatile and scalable assessment method; (2) Recent studies have demonstrated the robustness of LLM-based psychological assessment, and our validation of LLMs against human annotations aligns with previous results. The detailed implementation and evaluation are shown in Appendix C.

Based on the extracted feature vector $\mathbf{f}(x) \in \mathbb{R}^{31}$, we train a supervised classifier to predict the text's label. To ensure the framework's interpretability, we utilize tree-based models such as XGBoost and CatBoost, which provide clear feature importance rankings. The classification task is formally defined as finding the label $\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{f}(x))$, where $\mathcal{Y}$ is the set of possible labels. The model is trained by minimizing the cross-entropy loss, $\mathcal{L}_{\text{CE}} = -\sum_i y_i \log(p_i)$.

4 Experiments

4.1 Benchmarking RedNote-Vibe

Experiment Setup. We compare AIGT detectors on our proposed tasks, with three categories of approaches. (1) **Statistics-based methods**: StyloAI (Opara, 2024), Binoculars (Hans et al., 2024) and the method of Ullah et al. (2024), which serve as representative feature-driven baselines for AIGT detection. (2) **Model-based methods**: This category covers strong baselines, including (i) fine-tuning pre-trained text classification models such as BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and ALBERT-base (Lan et al., 2019), which represent strong baselines (Gritsai et al., 2024; Li et al., 2024), and (ii) state-of-the-art AIGT detection pipelines such as Sniffer (Li et al., 2023), POGER (Shi et al., 2024), and LLM-Idiosyncrasies (Sun et al., 2025a). (3) **PLAD**: For comparison, we evaluate our proposed PLAD framework with different classifiers, i.e., CatBoost (Prokhorenkova et al., 2018) and XGBoost (Chen and Guestrin, 2016). We follow the original training protocols of all methods and implement them on our dataset. All methods share the same dataset and are split into training, validation, and testing sets in an 8:1:1 ratio.

Results. Table 4 summarizes the performance across three tasks. For *AIGT classification*, competitive results are observed, with LLM-Idiosyncrasies achieving the highest recall while PLAD-CatBoost attains the best precision and accuracy. In the more granular *provider* and *model identification* tasks, PLAD consistently outperforms baselines. These

Method	Model Identification (17-way)			Provider Identification (6-way)			AIGT Classification (binary)		
	Precision	Recall	Acc.	Precision	Recall	Acc.	Precision	Recall	Acc.
Statistics-based Methods									
StyloAI (2024)	21.68	22.82	23.20	37.33	37.24	41.21	75.23	72.55	77.90
Ullah et al. (2024))	23.83	25.29	26.01	41.05	41.04	46.41	75.91	73.83	78.59
Binoculars (2024)	21.13	22.67	24.05	40.48	39.14	43.72	72.07	74.17	77.89
Model-based Methods									
FT. BERT (2019)	32.97	32.71	33.76	44.38	42.41	50.47	85.91	88.29	88.24
FT. ALBERT (2019)	32.26	31.74	31.65	41.98	40.61	44.34	83.59	79.63	84.58
FT. RoBERTa (2019)	35.20	31.78	32.21	38.22	40.20	47.64	87.85	88.88	89.52
Sniffer (2023)	30.63	30.77	30.04	44.39	42.88	45.65	82.55	81.30	84.45
POGER (2024)	31.28	30.16	31.88	45.41	44.68	47.16	79.17	82.03	84.89
LLM-Idiosyncrasies (2025a)	32.91	33.16	32.31	49.81	45.71	49.96	88.09	**90.15**	89.07
PLAD Framework with Different Classifiers									
PLAD-XGBoost	32.11	33.51	34.04	**50.73**	**48.77**	**53.30**	86.45	85.31	87.79
PLAD-CatBoost	**35.87**	**36.45**	**36.94**	50.06	47.34	51.89	**88.70**	87.28	**89.62**

Table 4: Comparison of PLAD with existing methods across detection tasks derived from RedNote-Vibe. FT. denote fine-tuning. Precision, recall, and accuracy are computed as macro-averages.

	Method	Seen Acc.	Unseen Recall
GPT-o3	PLAD	46.46	**56.34**
	BERT-base	42.42	25.35
	ALBERT-base	42.73	28.31
	LLM-Idiosyncrasies	44.34	22.54
Gemini-2.5	PLAD	43.72	**58.46**
	BERT-base	45.23	52.31
	ALBERT-base	42.61	50.00
	LLM-Idiosyncrasies	41.71	53.85

Table 5: Recall on unseen target models (GPT-o3 and Gemini-2.5) in Provider Identification (6-way) task.

results suggest that our psycholinguistic features effectively capture the subtle stylistic signatures of different LLMs, offering both superior detection capability and interpretability.

4.2 Generalization to Unseen Models

In this experiment, we evaluate the generalizability of detection methods on unseen AI models. Given the observation that model providers often share similar training methodologies, data sources, and architectures across different versions. Therefore, a model family inherits a specific style imprint (Spiliopoulou et al., 2025). We design a zero-shot experiment by excluding the latest GPT-o3 and Gemini-2.5 from the training data while retaining other models from their respective providers (OpenAI and Google). This setup mimics the scenario that detection systems encounter newly released models that are not available during training. We set up experiments on the reduced dataset, then evaluate their accuracy on seen models and recall on unseen target models. As results are shown in Table 5, our PLAD framework significantly outperforms other approaches in identifying unseen models, demonstrating that the psycholinguistic features extracted by PLAD can capture more robust and generalizable traces of model families. In the context of rapid iteration of LLMs, this detection capability of new models makes our framework more practical than model-based methods.

4.3 Ablation and Feature Importance Study

To understand how PLAD achieves its detection performance, we conduct an analysis combining a dimension-level ablation study and evaluation of feature importance using CatBoost classifier. Table 6 present the ablation study result. It reveals that the largest performance drop occurred when the Lexical Identity and Stylistic Signature features are removed, leading to a substantial decrease across all three tasks. We further identify the two most important features, *Prosodic Rhythm Consistency* and *Type-Token Ratio*, suggest that AI-generated text tends to exhibit smoother and more uniform rhythmic patterns, whereas human writing often contains irregularities caused by cognitive processing. The third most important feature,

Configuration	17-way	6-way	binary
CatBoost (Full)	**36.16**	**48.66**	**87.98**
w/o Emotional Dim.	27.68	43.22	86.32
w/o Cognitive Dim.	31.40	43.56	86.44
w/o Lexical Dim.	26.81	38.66	85.00
w/o Cohesion Dim.	32.29	44.76	86.57

Table 6: Ablation experiment result of each feature dimension. We report the macro-average F1-score (%) on three classification tasks.

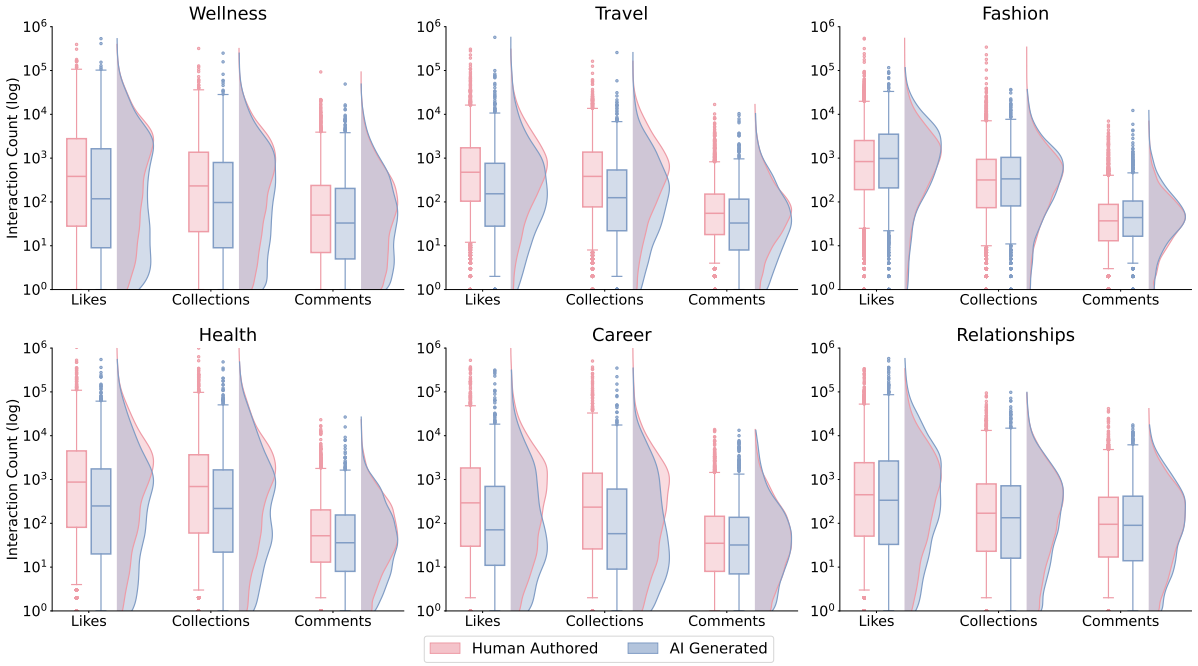


Figure 3: Engagement metrics comparison between Human-authored and AI-generated posts across 6 topical domains.

Imperfection, also from this dimension, which detects the absence of human-like hesitations, self-corrections, and other disfluencies that LLMs are optimized to avoid. Detailed feature importance and the differences between human and AI text analysis are presented in Appendix B.

5 Analysis

In the previous experiments, we demonstrated that our proposed PLAD framework achieves state-of-the-art performance in detecting AIGT on the RedNote dataset, outperforming existing black-box baselines. We now apply PLAD to the real-world posts from the post-LLM era to investigate the temporal evolution of AIGT and its impact on social media engagement.

Considering that the exploration set (90,000 post-LLM posts) is inherently unlabeled, as only the authors themselves know whether AI was used, and human annotation is also infeasible. Therefore, we use pre-LLM posts to evaluate model reliability. We adopt binary AIGT classification and evaluate the model on pre-LLM validation set, which consists of posts written by humans and independent from the exploration set used in downstream analyses. To prioritize precision and minimize false positives, we set a conservative threshold of 0.9 based on validation performance, achieving a false positive rate (FPR) of 4.9%. Note that this threshold is determined solely from the labeled validation set before any analysis is conducted. After that, we apply the model with the pre-determined threshold to the naturally unlabeled exploration set for subsequent analyses. This process follows a validation-then-deployment pipeline and is consistent with previous AIGT analysis studies (Sun et al., 2025b; Liang et al., 2025).

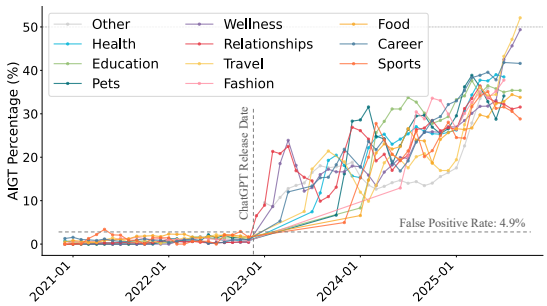


Figure 4: Temporal evolution of AIGT percentage across different domains (2021-2025).

5.1 Temporal Evolution of AI Content

Figure 4 illustrates the monthly percentage of detected AI content across different domains. Since the emergence of LLMs, the proportion of AI-generated content has exhibited a steady growth across all domains. This trend indicates that **AI tools are increasingly utilized for lifestyle sharing**. This finding aligns with recent studies on other datasets (Sun et al., 2025b; Liang et al., 2025), re-

inforcing the observation that the adoption of generative AI is a widespread phenomenon.

5.2 Metrics of AI vs. Human Posts

To investigate the impact of content source on interaction, we compare engagement metrics across six representative domains in Figure 3. It reveals that **Human content exhibits superior emotional resonance.** Human-authored posts consistently outperform AI-generated ones in median engagement. This gap is particularly visible in domains requiring personal experience and emotional resonance, such as *Travel*, *Career*, and *Relationships*.

Compared to human content, the interaction distribution for AI posts is more concentrated, with notably fewer high-engagement outliers. This indicates that while AI can generate grammatically correct text, it lacks the novelty and high-arousal qualities necessary to create breakout content. Looking into the variations between interaction types reveals the cognitive investment from users. For instance, a like is a passive and lightweight social signal, whereas a comment requires active and conscious engagement. Interestingly, the human-AI gap in comments is smaller than in likes. This suggests that **while AI content receives lower overall engagement, the performance gap narrows for interactions requiring higher cognitive investment**.

5.3 Author-Level AI Usage and Engagement

In addition to content-level analysis, we conduct an author-centric study to understand how different AI usage strategies correlate with user interaction behavior. We select authors with at least four posts in our dataset to estimate their AI usage, resulting in 829 authors. We compute each author's AI ratio by their posts classified as AIGT by PLAD. Based on this, we categorize authors into three groups: (1) *Traditional Authors* (AI ratio = 0%), who compose entirely without AI assistance (570 authors, 68.7%); (2) *AI-reliant Authors* (AI ratio = 100%), who fully rely on AI for all their posts (230 authors, 27.7%); and (3) *AI-augmented Authors* (0% $<$ AI ratio $<$ 100%), who strategically combine human and AI-generated content (29 authors, 3.5%). For each author group, we present their engagement metrics in Figure 5.

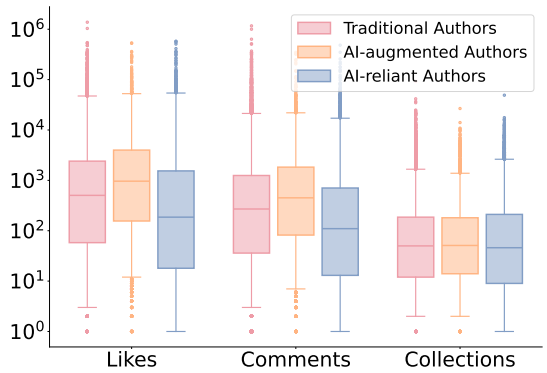


Figure 5: Author-level analysis of AI usage patterns and engagement outcomes.

The results of Traditional Author and AI-reliant Author groups are consistent with the findings in the previous section. However, the AI-augmented Authors group achieves relatively higher engagement. We hypothesize that these authors leverage AI for efficiency while applying human judgment to refine the content quality. This might suggest that **a small group of users who strategically utilize AI tools can achieve higher engagement outcomes**, outperforming both those who avoid AI entirely and those who rely on it exclusively.

6 Discussion and Conclusion

Rethinking AIGT Detection in the Wild. Existing detection methods often report high accuracy (e.g., $> 98\%$). However, we argue that this performance is inflated by the stylistic gap between naive LLM outputs and human writing. In this work, we optimized our generation pipeline to produce content more indistinguishable from human posts. It makes RedNote-Vibe more challenging, reflecting real-world adversarial conditions.

Longitudinal Interactions Between Human and AI Content. As AIGT becomes increasingly prevalent on social media, users are exposed to both human- and AI-generated content with growing frequency. LLMs are trained on human-produced corpora, while their outputs may subsequently influence human writing patterns. Our dataset provides a resource for investigating these temporal dynamics, which we leave for future work.

Conclusion. In this paper, we introduce a novel dataset that fills the gap in Chinese lifestyle social media resources. We also propose an AIGT detection method as the baseline for our dataset and tasks. Our exploratory analysis reveals some AIGT and user behavior patterns of social media. We emphasize that our core contribution lies in the dataset itself, we believe it provides a valuable data source for future AI/human behavior research.

Ethical Statement

All data utilized in this study originates from publicly accessible posts on the RedNote platform. We do not collect any private information. For the author-centric study, we hashed the usernames for grouping, and we only performed statistical analysis, not on individuals.

Regarding the AI-generated variants, they are produced using commercial LLM APIs with the sole objective of constructing a robust detection benchmark. We explicitly prohibit the misuse of this dataset for generating misleading content or spam.

Limitations

While our work provides comprehensive insights into AIGT on social media, we acknowledge several limitations that future work should address:

Language and Platform Specificity. Our study focuses exclusively on RedNote (Xiaohongshu), a platform with dominant of Chinese-language users. Consequently, the linguistic features and engagement patterns observed here may not directly generalize to other platforms. Furthermore, although our PLAD framework is designed based on general psycholinguistic principles and is theoretically language-agnostic, we have not empirically verified its efficacy across other languages in this study.

Limited Coverage of Generative Models. The AIGT variants in our dataset are generated using a subset of LLMs available up to mid-2025. Given the rapid iteration of model capabilities (e.g., the emergence of reasoning-heavy models), the generated vibe may continue to evolve.

Reliance on Proxy LLMs. Our feature extraction pipeline relies on a proxy LLM to quantify psychological dimensions. Biases present in the proxy LLM could propagate into the feature set. However, these biases are unlikely to significantly impact the analysis results. In the Appendix C.3, we provide a robustness evaluation, demonstrating that our method outperforms existing approaches in terms of correlation with human judgments.

Cost of Inference. Compared to lightweight end-to-end classifiers (e.g., RoBERTa-based detectors), PLAD involves a multi-step process of feature extraction via LLMs, which incurs higher computational and financial costs. This may limit its real-time application in high-throughput industrial scenarios without further optimization.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & talk*, 23(3):321–346.

Gloria Baker-Brown, Elizabeth J Ballard, Susan Bluck, Brian De Vries, Peter Suedfeld, and Philip E Tetlock. 1992. The conceptual/integrative complexity scoring manual. *Motivation and personality: Handbook of thematic content analysis*, pages 401–418.

Mikhail Mikhauilovich Bakhtin. 2010. *The dialogic imagination: Four essays*, volume 1. University of texas Press.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Giovanni Luca Cascio Rizzo, Francisco Villarroel Ordenes, Rumen Pozharliev, Matteo De Angelis, and Michele Costabile. 2024. How high-arousal language shapes micro-versus macro-influencers' impact. *Journal of Marketing*, 88(4):107–128.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Jaeyeon Chung, Yu Ding, and Ajay Kalra. 2023. I really know you: How influencers can increase audience engagement by referencing their close social ties. *Journal of Consumer Research*, 50(4):683–703.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Martin A Conway and Christopher W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological review*, 107(2):261.

Wanyun Cui, Linqiu Zhang, Qianle Wang, and Shuyang Cai. 2023. Who said that? benchmarking social media ai detection. *arXiv preprint arXiv:2310.08240*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A

shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Pawanjit Singh Ghatora, Seyed Ebrahim Hosseini, Shahbaz Pervez, Muhammad Javed Iqbal, and Nabil Shaukat. 2024. Sentiment analysis of product reviews using machine learning and pre-trained llm. *Big Data and Cognitive Computing*, 8(12):199.

Howard Giles and Peter F Powesland. 1975. *Speech style and social evaluation.* Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Stefan Th Gries. 2008. Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, 13(4):403–437.

German Gritsai, Anastasia Voznyuk, Andrey Grabovoy, and Yury Chekhovich. 2024. Are ai detectors good enough? a survey on quality of datasets with machine-generated texts. *arXiv preprint arXiv:2410.14677*.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

Jiayi Gui, Baitong Cui, Xiaolian Guo, Ke Yu, and Xiaofei Wu. 2025. Aider: A robust and topic-independent framework for detecting ai-generated text. In *Proceedings of the 31st international conference on computational linguistics*, pages 9299–9310.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

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

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are ai-generated text detectors robust to adversarial perturbations? pages 6005–6024.

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum, London.

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

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Brian Levine, Eva Svoboda, Janine F Hay, Gordon Winocur, and Morris Moscovitch. 2002. Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychology and aging*, 17(4):677.

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023. Origin tracing and detecting of llms. *arXiv preprint arXiv:2304.14072*.

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.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2025. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, pages 1–11.

Shengchao Liu, Xiaoming Liu, Yichen Wang, Zehua Cheng, Chengzhengxu Li, Zhaohan Zhang, Yu Lan, and Chao Shen. 2024. Does detectgpt fully utilize perturbation? bridging selective perturbation to fine-tuned contrastive learning detector would be better. *arXiv preprint arXiv:2402.00263*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and 1 others. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987.

Shrestha Basu Mallick and Logan Kilpatrick. 2025. Gemini 2.0:flash, flash-lite and pro. Accessed: 2025-05-01.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

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.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. In *25th International Conference on Artificial on Artificial Intelligence in Education*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Xiaoqi Qiu, Yongjie Wang, Xu Guo, Zhiwei Zeng, Yu Yue, Yuhong Feng, and Chunyan Miao. 2024. Paircfr: Enhancing model training on paired counterfactually augmented data through contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11955–11971.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Hans Reichenbach. 1947. Elements of symbolic logic.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Joseph S Schafer, Kayla Duskin, Stephen Prochaska, Morgan Wack, Anna Beers, Lia Bozarth, Taylor Agajanian, Mike Caulfield, Emma S Spiro, and Kate Starbird. 2024. Electionrumors2022: A dataset of election rumors on twitter during the 2022 us midterms. *arXiv preprint arXiv:2407.16051*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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. *arXiv preprint arXiv:2402.09199*.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play favorites: A statistical method to measure self-bias in llm-as-a-judge. *arXiv preprint arXiv:2508.06709*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. 2025a. Idiosyncrasies in large language models. *arXiv preprint arXiv:2502.12150*.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025b. Are we in the ai-generated text world already? quantifying and monitoring aigt on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Ubaid Ullah, Sonia Laudanna, P Vinod, Andrea Di Sorbo, Corrado Aaron Visaggio, and Gerardo Canfora. 2024. Beyond words: Stylometric analysis for detecting ai manipulation on social media. In *European Symposium on Research in Computer Security*, pages 208–228. Springer.

Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. Checked: Chinese covid-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):58.

Zhongping Zhang, Wenda Qin, and Bryan Plummer. 2024. Machine-generated text localization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8357–8371.

Appendix

In the Appendix, we provide related work (Section A) and further analysis, including the differences between human-written text and AI-generated text from the perspective of the PLAD framework (Section B.1) and how these features affect the engagement metrics of posts (Section B.2). In addition, we analyze the correlation between features, reflecting the orthogonality of our framework (Section B.3). Finally, we introduce details of our AI samples generation (Section C.1), feature extraction method (Section C.2) and the validation of our feature quality (Section C.3).

A Related Work

Social media datasets have been developed for AIGC detection, including TweepFake (Fagni et al., 2021) for early Twitter content, SAID (Cui et al., 2023) for modern LLM detection, and Election-Rumors2022 (Schafer et al., 2024) for election-related content analysis. Chinese social media datasets focus primarily on Sina Weibo for tasks like NER (Peng and Dredze, 2015) and fake news detection (Yang et al., 2021), but lack coverage of RedNote platform and temporal dynamics.

AIGC detection methods span three categories: watermarking techniques that embed imperceptible marks during generation (Kirchenbauer et al., 2023; Liu et al., 2024), classifier-based approaches using fine-tuned transformers like BERT and RoBERTa (Hu et al., 2023; Huang et al., 2024), and statistical methods that establish discrimination thresholds (Mitchell et al., 2023; Zhang et al., 2024). Multi-class detection frameworks like Sniffer (Li et al., 2023) and LLM-Idiosyncrasies (Sun et al., 2025a) leverage LLM embeddings for classification. However, existing statistical methods struggle with real-world accuracy (Qiu et al., 2024; Sadasivan et al., 2023), highlighting the need for more robust and interpretable approaches.

B Further Analysis

B.1 Feature Statistics for Human and AI Post

To understand the differences in text style between human and AI, we conduct a comparative analysis of linguistic features extracted from human- and AI-authored texts. We apply statistical tests across the dataset, and we select six representative features for illustration, covering the four dimensions of our framework. As shown in Figure 7, these features

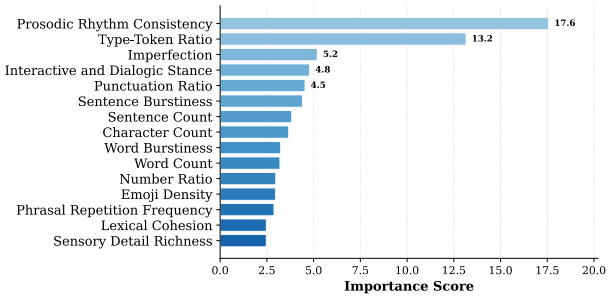


Figure 6: The top-15 important feature scores.

highlight systematic divergences between human and AI writing styles.

The results indicate that AI texts consistently achieve higher values in imperfection, reflecting stable fluency and a lack of surface-level flaws. By contrast, human writing displays a much broader distribution. A similar contrast appears in prosodic rhythm consistency and sentence burstiness. AI-generated texts demonstrate regularity and uniformity in rhythm and fluctuations in sentence patterns, whereas human writing is more dynamic and irregular, often breaking rhythmic patterns. In the lexical level, the type-token ratio results show that AI text tends to maintain higher lexical diversity, reflecting stochastic generation processes that avoid frequent repetition. In contrast, human writers employ more shallow lexical distributions, which is shaped by personal stylistic habits.

In terms of emotion and rhythm, AI can achieve higher scores than human writing. This suggests that AI can mimic or even outperform humans in many human-like writing characteristics, while real human writing is more restrained.

B.2 Feature Analysis for Engagement

To discover the relationship between features and engagement metrics (i.e. `likes`, `collections`, and `comments`), we conducted an analysis using SHAP value. For each metric, we train a `CatBoost` regressor model, denoted by $f(x)$, using its default hyperparameters. Subsequently, we employ SHapley Additive exPlanations (SHAP) to interpret the model's predictions (Lundberg and Lee, 2017). SHAP attributes an importance value to each feature based on principles from cooperative game theory, ensuring properties like local accuracy and consistency.

As shown in Figure 8, we observe distinct patterns of feature influence across different engagement types. For "likes", the most salient predictors are *Punctuation Ratio* and *Word Frequency*

Entropy. This indicates that lightweight forms of engagement are primarily driven by surface-level features such as punctuation density and lexical diversity. In contrast, higher-order psycholinguistic features (e.g., *Perspectival Complexity*, *Axiological Coherence*) play only a secondary role. This suggests that likes are largely sensitive to readability and rhythm rather than deeper cognitive or semantic structures.

In the case of "collections", *Word Frequency Entropy* emerges as the dominant feature, followed by *Phrasal Repetition Frequency* and *Axiological Coherence*. Compared with likes, collections are more strongly associated with content richness and value consistency. *Axiological Coherence* further suggests that users are more inclined to preserve texts that demonstrate coherent values and internal logical alignment. Thus, collections appear to reflect more deliberate and evaluative forms of engagement.

For "comments", in addition to *Punctuation Ratio*, socially oriented features such as *Lexical-Stylistic Personalization*, *Empathetic Engagement*, and *Interactive and Dialogic Stance* exhibit the strongest influence. Unlike likes or collections, commenting behavior is primarily shaped by interpersonal dynamics, empathy, and argumentative stance. This highlights the role of dialogic and relational features in fostering deeper interactions.

B.3 Feature Correlation Analysis

To ensure that the proposed features capture complementary aspects of text, we conducted a pairwise correlation analysis. Figure 9 illustrates the distribution of Pearson correlation coefficients across all feature pairs ($N = 465$).

Overall, the results confirm that the features are weakly correlated. The average absolute correlation is 0.1884 with a standard deviation of 0.2290, suggesting that the majority of features contribute orthogonal information. Specifically, only 6 feature pairs exhibit strong correlations ($|r| > 0.7$), while 89 pairs fall into the medium range ($0.3 < |r| < 0.7$). The vast majority, 370 feature pairs, remain weakly correlated ($|r| < 0.3$).

This distribution indicates that the feature set avoids redundancy and is well-suited for capturing distinct dimensions of linguistic behavior. Therefore, the framework benefits from a diversified set of signals spanning emotional grounding, cognitive architecture, stylistic identity, and textual cohesion. It provides a robust foundation for interpretable AIGT detection.

C Implementation Details

C.1 AI Sample Generation Details

To ensure reproducibility and transparency, we provide implementation details of our AI sample generation pipeline. The generation process is controlled by a seed-based prompting strategy, where each human-authored note serves as the reference input. Given a seed note and its corresponding domain label, we construct a structured prompt (see pseudocode below) that instructs the LLM to produce a new RedNote-style post. The prompt enforces the following constraints: (1) emulate the stylistic and colloquial properties of the seed (including informal punctuation and minor grammatical imperfections common in social media), (2) maintain thematic and length consistency with the seed, while ensuring semantic novelty, and (3) output results strictly in a predefined JSON schema containing the `title` and `content` fields.

Prompt Template for AI Sample Generation:

```
Reference post: {seed_note}
Please create a new RedNote-style post based on the reference above. Requirements:
1. The new post should have a similar theme and topic domain ({domain}) but must not be identical to the reference.
2. Mimic the writing style of the reference, including colloquial tone, informal punctuation, and possible minor errors typical of social media.
3. Keep the length roughly consistent with the seed note.
4. Output strictly in JSON format as follows:
{"title": "Post title", "content": "Post content"}
Output only the JSON object, without any extra text.
```

Table 7: Pseudocode of the prompt used for AI sample generation.

We employ commercial paid APIs from multiple providers to generate the AI samples. The generation script incorporates error handling, including automated detection of extraneous markdown wrappers (i.e., `json`) and recovery via JSON string extraction. Invalid or unparsable generations are discarded to ensure dataset integrity.

Each LLM in our pool receives identical prompts and seed distributions, guaranteeing fairness across providers. Importantly, we only select seed notes

published prior to November 2022 to exclude potential AI-generated contamination.

C.2 Feature Extraction

We develop an automated pipeline to extract features from each text entry using a standardized LLM-based scoring system. Each feature is defined through structured JSON criteria that specify evaluation dimensions, scoring rubrics, and key indicators. The features are listed in Table 3. We use the latest qwen-turbo (2025-07-15) as the proxy model.

C.2.1 Scoring Framework

Our approach employs a two-stage process: (1) text preprocessing involving removal of extraneous characters and normalization, and (2) LLM-based evaluation using dynamically assembled prompts. Each of the 31 features is defined by a JSON schema containing:

- **Dimension description**: Definition of the psychological construct.
- **Scoring criteria**: Anchored 0-1 scale with explicit behavioral markers.
- **Key indicators**: Textual evidence to focus evaluation.
- **Few-shot Examples**: A set of text samples paired with their expert-assigned scores. These examples guide the model's in-context learning, calibrating its judgment to align with human evaluation standards.

For each text sample, we dynamically construct evaluation prompts by embedding the target text and relevant feature criteria into a standardized template that instructs the LLM to follow a Chain-of-Thought reasoning process.

C.2.2 Prompt Template Structure

The evaluation prompt is dynamically constructed by assembling five core components into a standardized template. It begins by establishing the Task Context to define the psycholinguistic analysis objective, which identifies the evaluation task. The template then assigns a Role Definition, positioning the LLM as an expert evaluator. Subsequently, the specific Dimension Specification is injected from the JSON file, followed by the preprocessed Target Text for evaluation. Finally, the prompt provides detailed CoT Instruction, guiding the LLM through reasoning steps with metacognitive checks to ensure a rigorous scoring process.

The LLM is instructed to output only a numerical score between 0.0 and 1.0, ensuring standardized quantitative assessment across all features.

C.2.3 Example: Emotional Intensity

To illustrate our approach, we present the JSON example for emotional intensity evaluation:

```
{
  "dimension_id": "emotional_intensity",
  "description": "Evaluates the depth,
      regulation, and contextual
      appropriateness of emotional
      expression. This dimension
      assesses the presence of emotion,
      and its nuance, variability, and
      alignment with the narrative
      events. High scores reflect a rich
      , well-regulated, and contextually
       congruent emotional landscape,
      while low scores indicate
      expressions that are flat, extreme
      , or mismatched with the situation
      .",
  "scoring_criteria": {
    "0_score": "Emotionally flat,
        suppressed, or chaotically
        unregulated. The expression is
        either monotonous (e.g.,
        alexithymic, detached) or
        extreme and overwhelming (e.g.,
        hysterical, disproportionate
        rage). There is a significant
        incongruence between the emotion
         described and the context.",
    "1_score": "Rich, nuanced, and
        contextually appropriate
        emotional expression. The author
         conveys a spectrum of feelings
        using diverse vocabulary.
        Emotional intensity is well-
        regulated, rising and falling in
         a way that is congruent with
        the narrative. Acknowledges
        complex or mixed emotions."
  },
  "key_indicators": [
    "Analyze the ratio and distribution
        of positive (e.g., 'joy', '
        relief') vs. negative (e.g., '
        grief', 'fear') emotion words.
        Assess the mix of high-arousal (
        e.g., 'ecstatic', 'furious') vs.
         low-arousal (e.g., 'serene', '
        content') terms.",
    "Evaluate the richness of the
        emotional lexicon. Does the
        author use a variety of synonyms
         and descriptors for feelings,
        or repeatedly use the same basic
         emotion words?",
    ...
  ],
  "few-shot examples": [
    {
```

```
      "text": "I can't believe she left.
           I'M SO ANGRY! EVERYTHING IS
          AWFUL! I will NEVER be happy
          again, this is the worst thing
           that could ever happen to
          anyone! I hate everything and
          everyone!",
      "score": 0.3,
      "rationale": "While strong emotion
           is present, it is extreme,
          one-dimensional, and
          unregulated. The use of
          absolutes ('NEVER', '
          EVERYTHING') and
          disproportionate intensity
          without nuance suggests a lack
           of emotional modulation,
          mapping to the lower end of
          the scale."
    },
    ...
  ]
}
```

This structured approach ensures consistent, objective evaluation across all 31 psycholinguistic dimensions while maintaining the flexibility to adapt criteria for different psychological constructs.

C.3 Reliability of Proxy LLMs

Psychometrics using LLMs is a rapidly evolving field, and existing research has demonstrated that modern LLMs can be more reliable and stable for psychological text analysis than traditional tools. In this work, we adopt the established practice for feature extraction with targeted prompt engineering. Since prompt comparison has a large search space, a more meaningful comparison is the performance of different proxy models. Therefore, in this section, we utilize a human-annotated dataset of psychological constructs to evaluate and verify the performance of our proxy model.

In Table 8, we report the Spearman correlation of different proxy LLMs and traditional LIWC-based methods against human labels on different psychological constructs.

Method	Concreteness	Sentiment	Formality
Qwen-turbo	0.720	0.785	0.770
DeepSeek-V3	0.716	0.793	0.768
GPT-4o	0.746	0.795	0.788
LIWC-based Method	0.554	0.570	0.429

Table 8: Spearman correlation of different methods against human labels on psychological constructs.

We acknowledge that despite using the best possible methods to extract features, existing methods cannot achieve 100% accuracy. However, the PLAD framework does not rely on absolute values for interpretability; instead, it distinguishes human and AI text based on differences across various feature dimensions.

C.4 Supplementary Temporal Ablation

To quantify the value of RedNote-Vibe's longitudinal coverage, we conduct a year-wise ablation on the binary detection task. Models are trained on the pre-LLM portion of the dataset (2020–2022); we remove one year at a time while keeping the removed amount equal across conditions. As a size-matched control, we also randomly drop the same number of samples across all three years.

Training Data	PLAD-CatBoost	BERT	RoBERTa
whole	**87.98**	**87.08**	**88.36**
random drop	83.66	77.07	77.54
w/o 2020	81.78	73.63	74.17
w/o 2021	81.56	73.44	74.65
w/o 2022	79.92	74.01	72.47

Table 9: Temporal coverage ablation on binary AIGT detection. We report macro-average F1-score (%).

As shown in Table 9, removing any single year causes a larger drop than randomly removing the same amount of data, suggesting that the temporal span contributes beyond dataset size. The largest degradation appears when removing 2022, the year closest to the LLM transition, indicating that recent pre-LLM human writing provides especially informative contrastive signals for detecting early AI-generated content.

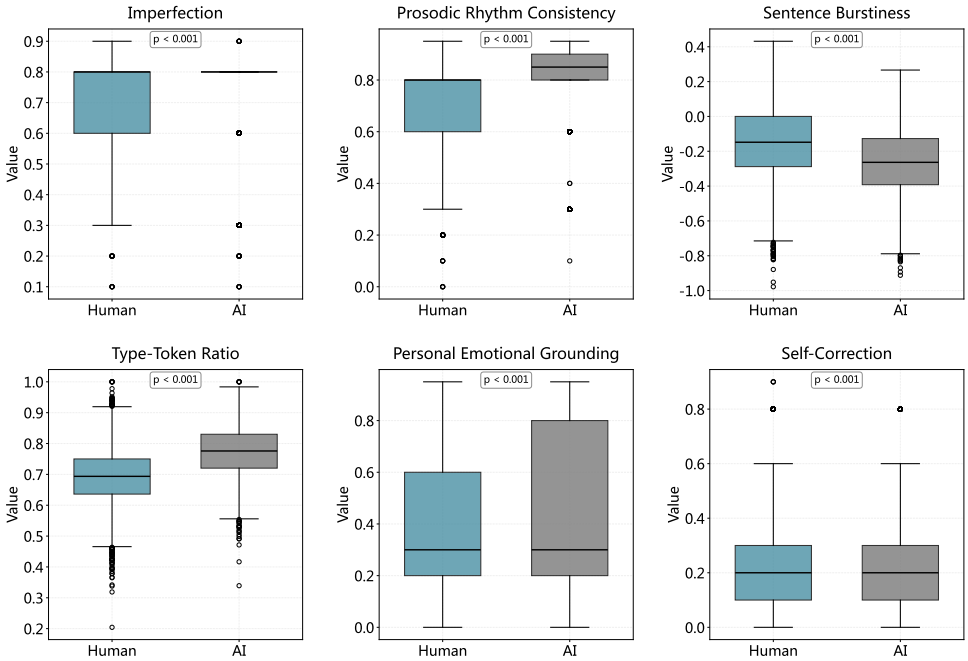


Figure 7: Comparison of feature statistics between human and AI-generated texts.

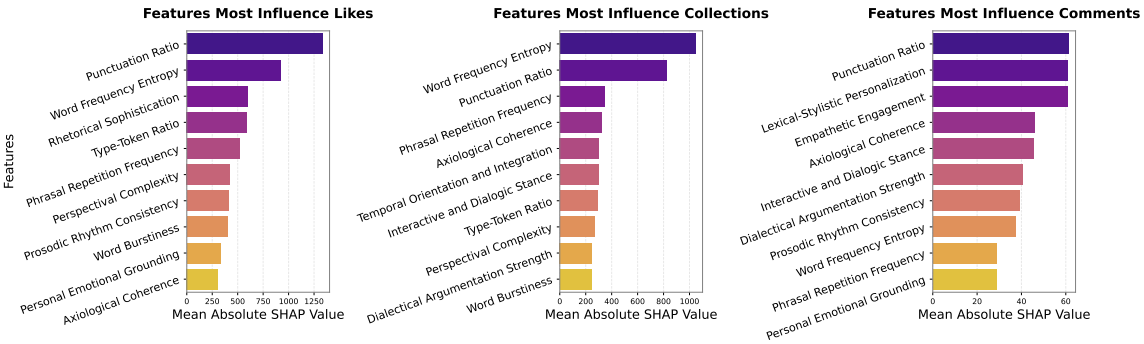


Figure 8: Top-10 most influential features of likes, collections and comments metrics.

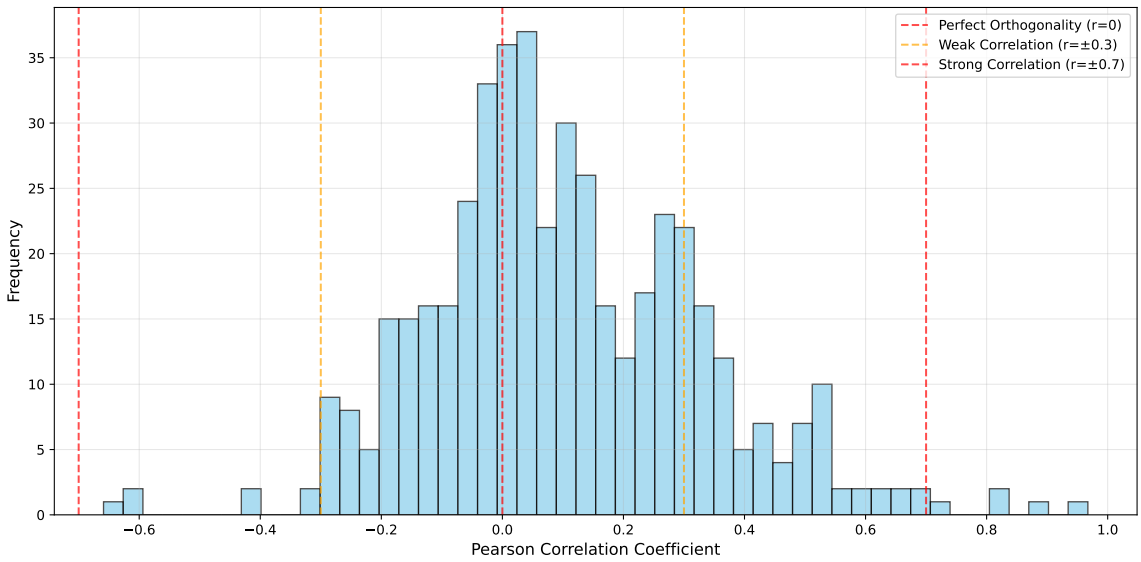


Figure 9: Distribution of pairwise feature correlations. Dashed lines indicate thresholds for weak ($|r| = 0.3$) and strong ($|r| = 0.7$) correlation. Most feature pairs are weakly correlated, demonstrating orthogonality of the feature set.