C-ReD: A Comprehensive Chinese Benchmark for AI-Generated Text Detection Derived from Real-World Prompts

Chenxi Qing[1*], **Junxi Wu**[2,1*], **Zheng Liu**[1*], **Yixiang Qiu**[1], **Hongyao Yu**[1], **Bin Chen**[3,4†], **Hao Wu**[1,5†], **Shu-Tao Xia**[1,4]

[1] Tsinghua University [2] Nankai University

[3] Harbin Institute of Technology, Shenzhen

[4] Peng Cheng Laboratory [5] Shannon InfoTech

qcx25@mails.tsinghua.edu.cn, chenbin2021@hit.edu.cn

Abstract

Recently, large language models (LLMs) are capable of generating highly fluent textual content. While they offer significant convenience to humans, they also introduce various risks, like phishing and academic dishonesty. Numerous research efforts have been dedicated to developing algorithms for detecting AI-generated text and constructing relevant datasets. However, in the domain of Chinese corpora, challenges remain, including limited model diversity and data homogeneity. To address these issues, we propose **C-ReD**: a comprehensive **C**hinese **Re**al-prompt AI-generated text **D**etection benchmark. Experiments demonstrate that C-ReD not only enables reliable in-domain detection but also supports strong generalization to unseen LLMs and external Chinese datasets—addressing critical gaps in model diversity, domain coverage, and prompt realism that have limited prior Chinese detection benchmarks. We release our resources at https://github.com/HeraldofLight/C-ReD.

1 Introduction

Large language models (LLMs) have quickly become integral to everyday life and professional workflows. Leading models such as ChatGPT (Brown et al., 2020) and DeepSeek (Guo et al., 2025) generate highly fluent and contextually relevant responses, greatly boosting productivity and user experience. However, this same capability can be exploited for malicious or unethical purposes (Fang et al., 2025)—such as phishing, academic dishonesty (Tang et al., 2024), plagiarism (Lee et al., 2023), and the dissemination of disinformation (Mitchell et al., 2023). Compounding the problem, human readers often find it difficult to reliably tell AI-generated text apart from human-written content (Mitchell et al., 2023), which further exacerbates these risks.

To address this challenge, substantial research efforts have focused on developing detectors for AI-generated text (Solaiman et al., 2019; Mitchell et al., 2023; Yang et al., 2024b; Bao et al., 2024; Chen et al., 2025). Concurrently, several benchmark datasets have been released to support detector training and evaluation (Uchendu et al., 2021; Li et al., 2023; He et al., 2024; Wu et al., 2024). However, these datasets predominantly consist of English corpora.

Chinese presents unique challenges for detection due to its complex word segmentation (Tsang et al., 2025), context-sensitive semantics, and abundant use of cultural idioms, metaphors, and informal abbreviations. These linguistic properties render direct adaptation of English-centric methods ineffective and underscore the need for native Chinese detection benchmarks. Unfortunately, prior Chinese datasets suffer from three key limitations: (1) **Limited model diversity**: most rely solely on ChatGPT (Macko et al., 2023; Wang et al., 2023), overlooking widely adopted domestic Chinese LLMs; (2) **Data homogeneity**: texts are often restricted to simple QA formats (Guo et al., 2023), lacking representation from professional domains such as journalism or academic writing; (3) **Unrealistic prompt design**: failing to reflect real-word application scenarios of LLMs.

To bridge these gaps, we introduce **C-ReD**: a comprehensive **C**hinese **Re**al-prompt AI-generated text **D**etection benchmark. C-ReD comprises Chinese human-written texts curated from five distinct domains, paired with AI-generated counterparts produced by nine LLMs—including five leading Chinese domestic models—under five carefully designed, real-world-inspired prompt types.

Our analysis using C-ReD reveals that detection performance varies substantially across domains and generators, with fluent, reasoning-intensive

*Equal contribution.

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42703–42733

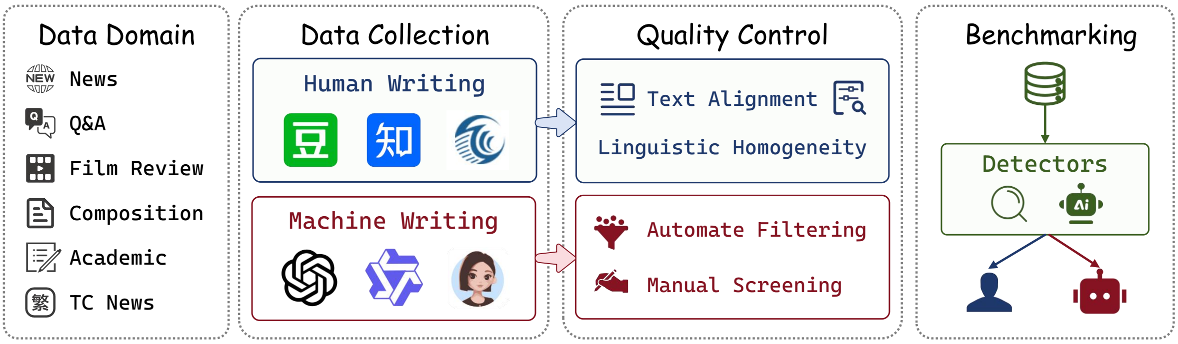


Figure 1: The overview of C-ReD.

models like Deepseek-R1 being particularly challenging to detect. Fine-tuning on C-ReD not only boosts in-domain accuracy but also enables strong generalization to unseen models and external datasets, demonstrating its effectiveness as a representative and scalable foundation for Chinese AI-generated text detection.

2 Related Work

2.1 Detection Methods

Supervised Methods. The goal of supervised methods is training a classifier, which typically leverages neural representations from pre-trained models. RoBERTa (Liu et al., 2019), in particular, has been widely adopted for this purpose. Studies fine-tune it on labeled datasets, use its contextual embeddings to train binary classifiers (Solaiman et al., 2019; Guo et al., 2023), and leverage its semantic capture ability to achieve effective detection in specific domains. ImBD (Chen et al., 2025) identified machine-revised text by imitating the machine-style token distribution by style preference optimization. Recent works incorporate token-level probability distributions with pre-trained model embeddings (Verma et al., 2024; Shi et al., 2024), utilizing both semantic and statistical information effectively. MoSEs (Wu et al., 2025) introduces a flexible framework that models profession-specific writing styles and addresses the limitation of static threshold by employing conditional threshold estimation to enable adaptive detection across diverse domains. The performance of supervised methods is fundamentally based on the quality and representativeness of training data. High-quality datasets, with diverse text types, comprehensive source models, and accurate labels, can alleviate overfitting and improve generalization.

Zero-shot Methods. Existing zero-shot detectors mostly rely on constructing statistical features through pre-trained large language models, without requiring task-specific training. Early approaches rely on token-level likelihood–based metrics such as entropy (Gehrmann et al., 2019), perplexity (Lavergne et al., 2008), log-likelihood (Solaiman et al., 2019), and top-k probability buckets (Gehrmann et al., 2019). DetectGPT (Mitchell et al., 2023) pioneered the paradigm of comparing perturbed texts with original ones for detection, inspiring related methods such as DetectNPR (Su et al., 2023) and DNA-GPT (Yang et al., 2024b). However, these methods suffer from high time costs. Fast-DetectGPT (Bao et al., 2024) improved efficiency via conditional probability curvature, expanding the application scope of training-free detection methods. Lastde (Xu et al., 2025) introduced time series analysis for better accuracy and DNA-DetectLLM(Zhu et al., 2025) introduced the mutation-repair paradigm. Note that zero-shot methods typically utilize labeled data to perform probability calibration.

2.2 Related Datasets

AI-generated Text Benchmarks. A growing body of research has addressed the challenges posed by large language model (LLM)-generated text, leading to the development of several benchmark datasets. TuringBench (Uchendu et al., 2021) pioneered this direction by introducing a dataset consisting of 200k human-written and AI-generated news articles produced by 19 distinct LLMs. However, it is limited by its reliance on models no more advanced than GPT-3 and by the narrow diversity of its data sources. Building on this foundation, MAGE (Li et al., 2023) presents a large-scale testbed comprising outputs from 27 LLMs across seven diverse writing tasks. MGTBench (He et al., 2024) further advances the field by offering a unified framework for both detection and attribution of machine-generated text, enabling systematic evaluation of various methods across multiple datasets. To better reflect real-world deployment scenarios, DetectRL (Wu et al., 2024) introduces four evaluation tasks covering high-risk domains such as social media and news writing—contexts particularly susceptible to synthetic text abuse. Despite these advances, existing benchmarks remain predominantly English-centric and often lack coverage of Chinese-language LLMs, highlighting a critical gap for non-English detection research.

Chinese Corpora. Several studies have included Chinese AI-generated text in their benchmarks, yet significant limitations remain. HC3 (Guo et al., 2023) provides 40k question–answer pairs in both English and Chinese, covering diverse domains such as computer science, law, and medicine. However, it is restricted to a single QA format and relies solely on ChatGPT for generation. MULTITuDE (Macko et al., 2023) is a multilingual detection benchmark encompassing 74k AI-generated texts across 11 languages—including Chinese—produced by eight LLMs. Notably, the Chinese subset was used only for evaluation and excluded from model training, limiting its utility for developing Chinese-specific detectors. M4 (Wang et al., 2023) introduces a multi-generator, multi-domain, and multilingual benchmark that includes Chinese data sourced from Baike and Web QA. Nevertheless, its Chinese AI-generated texts were produced exclusively by ChatGPT and GPT-3.5, omitting widely deployed domestic Chinese LLMs. Collectively, these Chinese datasets suffer from narrow domain coverage, homogeneous text formats, and limited model diversity—particularly the absence of state-of-the-art Chinese foundation models. Coupled with their relatively small scale and constrained scenario representation, these shortcomings render existing resources insufficient for robust, real-world Chinese AI-generated text detection.

3 C-ReD Dataset

3.1 Data Sourcing

We collected human-written texts from five domains where LLMs are commonly deployed. The data include: (1) **News**: 3,000 articles from THUCNews (Li et al., 2006), evenly distributed across five categories (sports, politics, finance, entertainment, education); (2) **Q&A**: 2,956 user answers from Zhihu via Zhihu-KOL (GeeYangML and Ray, 2023); (3) **Film Review**: 2,960 user reviews from Douban, sourced from ChineseNlpCorpus (Aesop et al., 2023); (4) **Composition**: 1,081 model essays from China's National College Entrance Examination, collected via web scraping; (5) **Academic Writing**: 500 Chinese papers (title, keword, abstract, introduction) from ChinaXiv (ChinaXiv, 2025). For news domain, we also consider Traditional Chinese (TC) news for extra dataset and collect 2,000 articles from News-Collection-Zhtw (Oscar), covering four categories (article, tech, science, daily-weekly). Further information on data sources and statistics is provided in Appendix A.1.

3.2 Model Sets

We adopt a diverse set of representative LLMs to generate AI-written texts. To better reflect real-world Chinese AI writing scenarios, we include advanced models from both Chinese and international providers. Our model set includes nine LLMs: **OpenAI** (`gpt-3.5-turbo`, `gpt-4o`) (OpenAI; Hurst et al., 2024), **Google** (`Gemini-2.5-Flash`) (Comanici et al., 2025), **Anthropic** (`Claude-3.5-Haiku`) (Anthropic, 2024), **Deepseek** (`Deepseek-V3`, `Deepseek-R1`) (Liu et al., 2024; Guo et al., 2025)—with Deepseek-R1 employed in chain-of-thought mode—, **Qwen** (`Qwen2.5`, `Qwen3`) (Yang et al., 2024a, 2025), and **Doubao** (`Doubao-1.5-Pro`) (ByteDance, 2025). All models are accessed exclusively via APIs, this black-box setting closely mirrors real-world usage but also increases the difficulty of detection, further details are provided in the Appendix A.2.

3.3 Prompt Design

To enhance the alignment between LLM-generated texts and real-world applications, we designed five domain-specific prompt strategies. **News**: we curated category-specific prompt templates with LLM assistance and combined them with real news headlines to simulate AI-assisted news writing. **Q&A**: we first created general-purpose prompt templates and then specialized them by injecting domain-specific questions, mimicking real-world QA systems. **Film Review**: we developed structured review guidelines using LLMs and paired them with film titles to form final prompts, reflecting practical AI-powered critique generation. **Composition**: we directly utilized descriptions from Gaokao along with specific topics to construct the final prompts. **Academic Writing**: for introduction generation, we fused paper titles, keywords, and abstracts with instructional prompts. Separately, we also designed prompts to generate abstracts solely from titles and keywords. **TC News**: analogous to the News setup, we adapted similar template design methodology to Traditional Chinese journalism. Sample prompts for each domain are provided in Appendix A.3.

Table 1: Core schema shared by all samples in the dataset. Fields below the second midrule are only present for AI-generated samples.

Field	Type	Description
id	int	Unique sample identifier
text	str	Human-written or AI-generated text
label	int	1 = human, 0 = AI
type	int	Encoded text domain
length	int	Character count
attribution	str	Source of the text: model name for AI outputs, or human for human-written samples
original_id	int	ID of the original human-written sample used for AI generation (AI only)
prompt	str	The exact prompt provided to the language model to generate this AI text (AI only)

3.4 Quality Control

To ensure high-quality and consistent data, we performed systematic preprocessing on both human-written and AI-generated texts.

Human-written Text. Human-authored content—sourced from diverse platforms—often contained extraneous elements such as metadata (e.g., headlines, author names, source URLs), non-textual components (e.g., figures, tables, citations), and formatting artifacts introduced during web scraping or PDF conversion. To align these with the clean, paragraph-level format of machine-generated outputs, we removed all such noise, standardized text structure into single coherent paragraphs, and applied length constraints appropriate to each genre. Additionally, we filtered out samples with excessive English content or inconsistent language use to maintain linguistic homogeneity. Full details of the cleaning protocols for each data type are provided in Appendix A.1.

AI-generated Text. To ensure the quality and consistency of AI-generated text, we implemented length constraints in the prompt templates to control output within predefined ranges for each domain. We adopted a dual-measure quality assurance approach: (1) real-time automated filtering that monitors Chinese character ratio, repetitive character sequences, text length, empty outputs, and factual inconsistencies, while also stripping extraneous formatting such as Markdown syntax, section headings, bullet points, and other non-paragraph elements; (2) rigorous manual screening by domain experts to identify and remove low-quality or anomalous samples. Specific parameter values for these quality controls are provided in Appendix A.4.

All texts (human-written and AI-generated) were normalized into plain, single-paragraph format (more details for text length in Appendix A.5). The final dataset is stored as a CSV file with a unified core schema (Table 1); domain-specific fields are detailed in Appendix A.6.

3.5 Statistics

As shown in Figure 2,our dataset comprises a total of 128,610 texts, including 12,997 human-written and 115,613 AI-generated samples across five domains (including extra dataset). The full distribution across domains and large language models is provided in Appendix A.7. Examples of C-ReD are included in Appendix J.

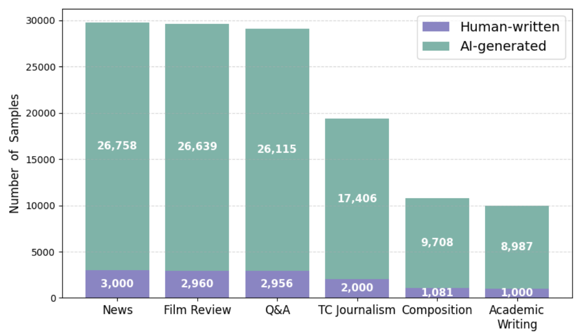


Figure 2: Distribution of samples in C-ReD.

4 Experimental Setup

4.1 Detectors

Using C-ReD, we conduct a comprehensive evaluation of a diverse set of state-of-the-art AI-generated text detection methods, spanning both traditional paradigms—zero-shot and supervised methods. In particular, we further investigate the emerging paradigm of leveraging Large Language Models (LLMs) themselves as detectors, exploring their potential to identify AI-generated text without task-specific training.

4.1.1 Zero-shot Detectors

We evaluate a range of zero-shot detection methods. These include likelihood-based metrics: **Log-Likelihood** (Solaiman et al., 2019), **Entropy** (Gehrmann et al., 2019), and **Log-Rank** (Mitchell et al., 2023). We also assess **LRR** (Su et al., 2023), which combines log-likelihood and log-rank into a normalized score; **Fast-DetectGPT** (Bao et al., 2024), which leverages conditional probability curvature; **Lastde** and **Lastde++** (Xu et al., 2025), which analyze likelihood sequences as time-series signals; and **DNA-DetectLLM** (Zhu et al., 2025), which enhances robustness through a mutation–repair mechanism. Full descriptions and details are provided in Appendix B.1.

4.1.2 Supervised Detectors

We evaluate several representative methods, including the **OpenAI Detector** (Solaiman et al., 2019), a RoBERTa-based classifier trained on GPT outputs; **RADAR** (Hu et al., 2023), which jointly trains an AI-generated text detector by adversarial learning; **ReMoDetect** (Lee et al., 2024), which improves detection performance by enhancing the reward model's ability; and **IMBD** (Chen et al., 2025), which aligns a scoring model with machine-like writing styles via Style Preference Optimization (SPO) and detects AI-generated text using a Style-Conditional Probability Curvature. Full descriptions and details are provided in Appendix B.2.

4.1.3 LLM Detectors

We evaluate eight large language models as detectors, each prompted to perform binary classification—distinguishing human-written from AI-generated text—using a standardized instruction. The full list of evaluated models, along with the exact prompt template and inference settings, is provided in Appendix B.3.

4.2 Metrics

We report two primary evaluation metrics: **Accuracy (Acc)** and **Area Under the ROC Curve (AUROC)**. Accuracy measures the proportion of correctly classified samples (human-written vs. AI-generated) under a fixed decision threshold. AUROC evaluates the detector's ranking performance across all possible thresholds, providing a threshold-invariant assessment that is particularly informative for imbalanced or domain-shifted settings.

5 Evaluation Protocol

We evaluate AI-generated text detectors across seven settings within and beyond C-ReD: in-distribution performance, generalization to unseen generators, cross-domain transfer (with fixed generators), prompt complexity effects, LLM-as-detector reliability under multiple prompting strategies, and out-of-distribution validation on external Chinese datasets. For each domain and each source, we pre-split the data into fixed training and test subsets before any experiments.

5.1 Domain- and Generator-Agnostic Evaluation on C-ReD

A major requirement for practical AI-generated text detectors is robustness across diverse writing domains and unknown generation models. To evaluate this, we assess all zero-shot and supervised methods under a unified protocol in C-ReD: at inference time, detectors perform binary classification—distinguishing human-written from AI-generated text—without access to generator identity or domain labels. The test set covers all five domains, each containing AI-generated texts from all nine LLMs paired with human-written counterparts. Results are reported using the AUROC metric. No method-specific tuning or adaptation is applied: zero-shot detectors use fixed reference models, and supervised detectors rely solely on publicly released pre-trained weights without fine-tuning.

5.2 Training on C-ReD: In-Distribution and Out-of-Distribution Evaluation

A key challenge in AI-generated text detection is building models that generalize beyond the specific generators seen during training. To address this, we fine-tune OpenAI Detector and IMBD on a diverse, domain-balanced training set from seven LLMs in C-ReD, except Claude-3.5-Haiku and Gemini-2.5-Flash. The resulting models are evaluated on test data from all nine LLMs—including these two held-out models—using the AUROC metric to assess both in-distribution performance and robustness to previously unseen generators. This setup directly probes whether training on C-ReD's multi-generator, multi-domain data can improve both reliability and transferability of supervised detectors—a crucial step toward real-world deployment.

5.3 Domain Generalization Under Fixed Generators

An important challenge in AI-generated text detection is whether models trained on one writing domain can generalize to others. To isolate the effect of domain shift from generator variability, we evaluate cross-domain generalization under controlled conditions: using only two representative LLMs (Qwen2.5 and GPT-4o) as fixed generators throughout training and testing. For each generator, we train a supervised detector on human–AI text pairs from one source domain and evaluate it on all five domains without access to domain labels at inference time. This produces a 5×5 evaluation matrix per generator, enabling systematic comparison of in-domain performance against out-of-domain transfer. All results are reported using the AUROC metric under the standard binary classification protocol.

5.4 Evaluating LLMs as AI-Text Detectors

To assess whether large language models can reliably distinguish human-written from AI-generated text, we design a controlled detection task across the five domains in C-ReD. For each domain, we evaluate three representative generators—Qwen2.5, GPT-4o, and Deepseek-R1—against a diverse set of Judge LLMs. We compare three prompting strategies: (1) `normal` (zero-shot), (2) `context` (few-shot with three human-AI example pairs), and (3) `description` (rule-based, using a Qwen3-generated stylistic summary). All prompts enforce a strict binary output format (“机器生成” or “人类生成”) for automatic evaluation. Full prompt templates are provided in Appendix F.1.

5.5 Ablation on Prompt Complexity in Academic Writing

To investigate the effect of prompt complexity on detection difficulty, we generate AI-authored text in the Academic Writing domain using both the original C-ReD prompt and a simplified variant (see Appendix G for design details). For each prompt type, we collect outputs from Qwen2.5 and GPT-4o and pair them with human-written texts from the same topic distribution. Detectors are then evaluated on both sets under the standard binary classification protocol using AUROC, without access to prompt information at inference time.

5.6 Evaluation on External Traditional Chinese News

To assess the robustness of detection methods in a distinct linguistic and stylistic setting, we evaluate all three core protocols on an external dataset of Traditional Chinese news articles. This dataset comprises human-written texts and AI-generated counterparts produced by all nine LLMs in C-ReD, each prompted natively in Traditional Chinese. We replicate the same evaluation setups as in the main benchmark:

- **Domain- and Generator-Agnostic Evaluation**: Detectors are applied without adaptation, with no access to domain or generator labels at inference time.
- **Training and Evaluation on Traditional Chinese Data**: Supervised detectors are fine-tuned on a domain-balanced training set constructed from the Traditional Chinese news data (covering seven of the nine LLMs), and evaluated on the full test set—including the two held-out generators.
- **LLM-as-Detector Evaluation**: Judge LLMs use the three prompting strategies (`normal`, `context`, `description`).

All evaluations follow the standard binary classification protocol using AUROC. This setup provides an independent assessment of detection performance in a real-world, externally collected Traditional Chinese news domain—complementing the main C-ReD and revealing how methods behave under script and stylistic distribution shifts.

5.7 Validation via Transfer Performance on External Chinese Datasets

To assess generalization, we fine-tune OpenAI Detector and IMBD exclusively on C-ReD’s multi-generator Q&A training set (covering seven LLMs). We evaluate these models—before and after fine-tuning—on the Chinese QA subsets of **M4** (Wang et al., 2023), which contain human–AI text pairs generated by **ChatGPT** and **davinci-003**. Although both datasets are QA-oriented, they differ in data source, style, and generation models. Since the M4 Chinese samples were not used in C-ReD’s construction, this provides a strict out-of-distribution test to evaluate whether detectors trained on C-ReD’s Q&A domain generalize to real-world Chinese QA content. Full dataset details are in Appendix I.

6 Results and Discussion

6.1 Domain- and Generator-Agnostic Evaluation on C-ReD

Detection performance varies significantly across domains and LLM generators—two critical factors often overlooked in prior work. Table 2 shows the average AUROC across five domains, with full per-model results in Appendix C. **Domain difficulty is highly non-uniform.** Detection is consistently easier in structured, stylistically constrained domains like Q&A and Film Reviews, where AI-generated text exhibits clear statistical anomalies. In contrast, performance drops in News and Academic writing, where modern LLMs produce fluent, coherent prose that closely mimics human style. **Generator characteristics further modulate detectability.** For instance, Deepseek-R1 frequently employs chain-of-thought (CoT) reasoning, yielding outputs with enhanced logical coherence and structure—properties that obscure typical AI-generation artifacts and make detection more challenging. This effect is evident in the low AUROC scores of multiple methods on Deepseek-R1. Moreover, existing detectors exhibit systemic limitations under generator shift: zero-shot methods relying on a fixed reference model suffer from architecture- or style-mismatch when applied to unseen LLMs, while supervised models like RoBERTa fail on modern generators due to training data bias—having been trained primarily on older, lower-quality AI text.

6.2 Training on C-ReD: In-Distribution and Out-of-Distribution Evaluation

We analyze detection performance under two evaluation settings: pre-trained and fine-tuning on C-ReD. The test set includes nine LLMs, seven of which are part of the C-ReD training distribution (in-distribution, ID), while the remaining two—Claude-3.5-Haiku and Gemini-2.5-Flash—are held out (out-of-distribution, OOD). Full results are reported in Appendix D. Fine-tuning on C-ReD leads to a dramatic improvement across all domains, confirming that supervised detectors heavily rely on domain- and generator-aligned training data. Importantly, this gain extends consistently to OOD generators, demonstrating that **C-ReD's diverse composition enables meaningful generalization beyond the specific models seen during training**. Nevertheless, a small gap between ID and OOD performance remains after fine-tuning, indicating that while C-ReD greatly reduces generator bias, perfect transfer to new commercial models is still challenging. Notably, Deepseek-R1—the hardest ID model to detect—owes its low detectability to its chain-of-thought style, which generates highly coherent, human-like reasoning. This underscores that intrinsic generation traits, not just model identity, fundamentally limit detection reliability.

Table 2: AUROC performance across five domains. Results are averaged over 9 LLMs.

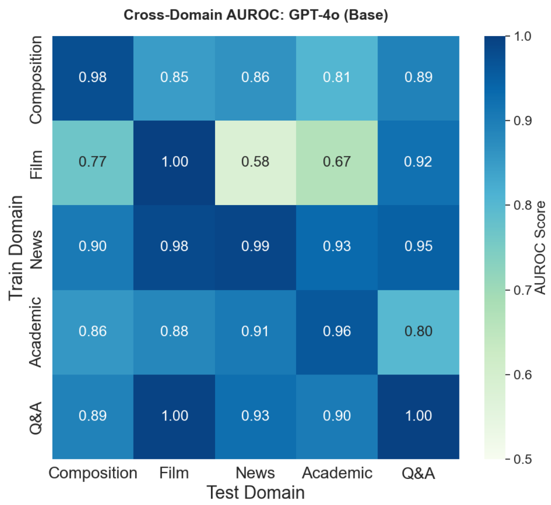
Method	Film	Comp.	Q&A	News	Acad.
Log-Likelihood	0.8344	0.8433	0.9343	0.7373	0.7326
Entropy	0.8427	0.8088	0.9253	0.6873	0.6907
Log-Rank	0.8286	0.8473	0.9301	0.7372	0.7384
LRR	0.7133	0.8473	0.8388	0.7004	0.7231
Fast-DetectGPT	0.6999	0.8952	0.8385	0.7626	0.7132
Lastde	0.5668	0.6500	0.6403	0.6762	0.7468
Lastde++	0.7059	0.8821	0.8148	0.7905	0.7157
DNA-DetectLLM	0.7595	0.9263	0.8439	0.7231	0.6226
RoBERTa-base	0.6461	0.5191	0.5139	0.4937	0.4316
RoBERTa-large	0.6121	0.5393	0.5255	0.3699	0.4059
RADAR	0.8291	0.6338	0.7605	0.4638	0.5167
ReMoDetect	0.9731	0.8731	0.9755	0.8652	0.9126
IMBD	0.8760	0.9140	0.9011	0.7953	0.8056

6.3 Cross-Domain Generalization Analysis Under Fixed Generators

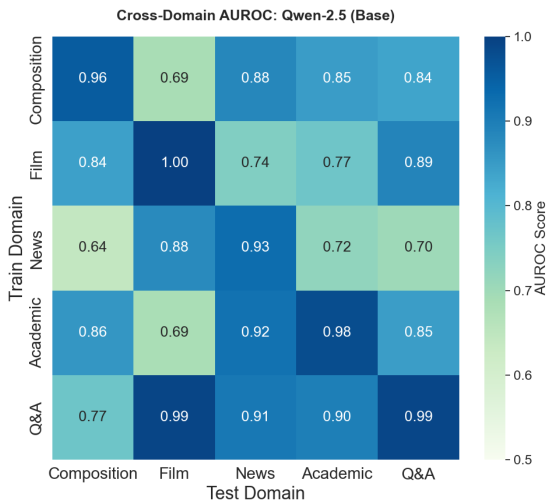
We examine how detection performance varies when models are trained on one domain and evaluated across all five domains. As shown in Figure 3a and 3b, detectors achieve strongest performance on their training domain, but generalization to other domains is highly uneven—depending on both the source training domain and the target test domain. **Training on fluent, information-dense domains like Q&A often leads to stronger cross-domain generalization** than training on highly stylized or opinion-driven domains like Film Review. Full numerical results for all model–domain combinations are provided in Appendix E.

6.4 Analyzing LLMs as AI-Text Detectors

Our evaluation reveals that LLMs generally fail at zero-shot detection (`normal`)—even on text generated by themselves—but achieve substantially higher accuracy when given domain-specific context (`context`) or stylistic descriptions (`description`). This indicates that LLMs' detection performance can be substantially improved through the incorporation of external contextual or stylistic cues—capabilities absent in zero-shot scenarios. Performance gains are consistent across



(a) GPT-4o



(b) Qwen2.5

Figure 3: Cross-Domain AUROC Heatmaps on RoBERTa-base Model.

domains and models, though challenges remain in highly structured domains like academic writing. Full results are in Appendix F.2.

6.5 Ablation on Prompt Complexity in Academic Writing

Our study indicates that AI-generated texts using simplified prompts are nearly as hard to detect as those created with more complex C-ReD prompts. This is consistent for both Qwen2.5 and GPT-4o, showing that these models can produce human-like academic text even with reduced instructions. The minimal difference in detection performance suggests that strong LLMs maintain stylistic and structural coherence regardless of prompt complexity, making their outputs challenging to distinguish.

Table 3: AUROC scores for detecting AI-generated academic writing under original vs. simplified prompts, before and after fine-tuning on C-ReD.

Model	Prompt	Generator	AUROC	
			Baseline	Finetuned
RoBERTa-base	Original	GPT-4o	0.5010	0.9566
	Simplified	GPT-4o	0.4987	0.9201
	Original	Qwen2.5	0.3313	0.9648
	Simplified	Qwen2.5	0.3272	0.9501
RoBERTa-large	Original	GPT-4o	0.4032	0.9868
	Simplified	GPT-4o	0.3937	0.9691
	Original	Qwen2.5	0.3685	0.9906
	Simplified	Qwen2.5	0.3954	0.9829

Full results are reported in Table 3 and results of bootstrap-based statistical analysis are reported in Appendix G.

6.6 Evaluation on External Traditional Chinese News

We evaluate all three core protocols on an external Traditional Chinese news dataset to assess detection robustness under a different linguistic and stylistic setting. The results closely mirror those observed in the main Simplified Chinese benchmark: supervised models achieve strong in-domain performance but still show a slight drop on held-out generators; LLM-as-detector approaches remain highly sensitive to prompting strategy and generator style. This consistency across writing systems suggests that key detection challenges—such as the impact of domain structure and reasoning-intensive generation—are not language-specific but reflect broader, cross-lingual patterns. Full results are provided in Appendix H.

6.7 Validation via Transfer Performance on External Datasets with Chinese Content

As shown in Table 4, detectors trained solely on C-ReD's Q&A domain exhibit strong transferability to external Chinese datasets. Both OpenAI Detector and IMBD show substantial improvements after fine-tuning, demonstrating that C-ReD provides effective supervision even for out-of-distribution generators. Performance on davinci-003 samples is lower but still improves significantly, suggesting greater stylistic divergence from C-ReD's training distribution. These results confirm that fine-tuning on C-ReD enables robust generalization to real-world Chinese AI-text detection scenarios.

Table 4: Transfer performance on **M4** (Wang et al., 2023). Detectors are trained only on C-ReD's Q&A domain (7 LLMs). Results report AUROC score.

Detector	ChatGPT		davinci-003	
	Pre	Post	Pre	Post
Roberta-base	0.6055	**0.8169**	0.6354	**0.6466**
Robeata-lagre	0.5684	**0.8890**	0.3990	**0.7445**
IMBD	0.9751	**0.9918**	0.9756	**0.9818**

7 Conclusion

In this work, we present **C-ReD**: a comprehensive **C**hinese **Re**al-prompt AI-generated text **D**etection benchmark. Spanning five domains and nine LLMs—including leading domestic Chinese models—C-ReD enables rigorous evaluation of detection methods across in-domain performance, cross-domain generalization, out-of-distribution robustness, and cross-lingual transfer. Our extensive experiments reveal that detection difficulty is strongly influenced by both domain structure and generator characteristics, with reasoning-intensive models like Deepseek-R1 posing the greatest challenge. Crucially, fine-tuning on C-ReD yields substantial gains not only on seen generators but also on unseen commercial models and external datasets, demonstrating its representativeness and practical utility. We hope C-ReD will serve as a reliable foundation for future research and deployment of AI-generated text detectors in Chinese contexts.

Limitations

Despite its breadth, C-ReD has several limitations. First, it focuses exclusively on Chinese text; while our preliminary results on Traditional Chinese suggest that key detection challenges generalize across writing systems, broader multilingual coverage remains future work. Second, although C-ReD includes nine LLMs, the rapid pace of model development means that new architectures or reasoning paradigms may emerge that are not represented in the current release. Third, our prompt designs, while inspired by real-world use cases, cannot capture the full spectrum of user behaviors—particularly adversarial or highly customized prompting strategies aimed at evading detection. Finally, our selection of human-written reference texts is limited in scope, and may not fully represent the diversity of writing styles, registers, or domains present in real-world Chinese content.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant 62301189, 62576122, 62571298, Guangdong Basic and Applied Basic Research Foundation under grant 2026A1515011139.

References

Chow Aesop, Clauss Christian, Ding Li, and Zhu Su. 2023. Chinesenlpcorpus. https://github.com/ InsaneLife/ChineseNLPCorpus.

Sonnet Anthropic. 2024. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. *URL https://api. semanticscholar. org/CorpusID*, 273639283.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

ByteDance. 2025. Doubao 1.5pro - doubao team. https://seed.bytedance.com/zh/special/doubao_1_5_pro.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, and 1 others. 2025. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23559–23567.

ChinaXiv. 2025. Chinaxiv.org. https://chinaxiv.org.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. 2025. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8613.

GeeYangML and Wang Ray. 2023. Zhihu-kol. https://github.com/wangrui6/Zhihu-KOL.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

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*.

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*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2251–2265.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

Hyunseok Lee, Jihoon Tack, and Jinwoo Shin. 2024. Remodetect: Reward models recognize aligned llm's generations. *Advances in Neural Information Processing Systems*, 37:2886–2913.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

Jingyang Li, Maosong Sun, and Xian Zhang. 2006. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization. In *proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 545–552.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Mage: Machine-generated text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel 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. *arXiv preprint arXiv:2310.13606*.

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.

OpenAI. Gpt-3.5 turbo model | openai api.

Li Oscar. News-collection-zhtw. https://huggingface.co/datasets/liswei/news-collection-zhtw.

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, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

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*.

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

Yiu-Kei Tsang, Ming Yan, Jinger Pan, and Megan Yin Kan Chan. 2025. A corpus of chinese word segmentation agreement. *Behavior Research Methods*, 57(1):1–15.

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*.

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

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Junchao Wu, Runzhe Zhan, Derek Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia Chao. 2024. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 37:100369–100401.

Junxi Wu, Jinpeng Wang, Zheng Liu, Bin Chen, Dongjian Hu, Hao Wu, and Shu-Tao Xia. 2025. Moses: Uncertainty-aware ai-generated text detection via mixture of stylistics experts with conditional thresholds. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5797–5816.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. 2025. Training-free LLM-generated text detection by mining token probability sequences. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. 2024b. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. In *The Twelfth International Conference on Learning Representations*.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. 2025. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. *arXiv preprint arXiv:2509.15550*.

A Dataset Construction

A.1 Data Source

News We selected 3,000 news articles from THUCNews (Li et al., 2006), a dataset derived from Sina News RSS feeds (2005–2011). The sample is balanced across five categories—sports, politics, finance, entertainment, and education—with 600 articles per category. To ensure suitability for Chinese-language detection, we filtered out articles with high English content, removed metadata (e.g., headlines, publisher information), and merged each article into a single paragraph. All texts were further constrained to a maximum length of 800 words.

Q&A We sourced 2,956 human-written answers from Zhihu-KOL (GeeYangML and Ray, 2023), a dataset collected from Zhihu—a major Chinese Q&A and knowledge-sharing platform. The sample covers 2,892 distinct questions and includes answers with lengths uniformly distributed between 100 and 850 words, ensuring diversity in response length. To enhance data quality and language consistency, we filtered out responses with high English content, repaired formatting issues (e.g., irregular spacing and missing punctuation), and merged each answer into a single coherent paragraph.

Film Review We focus on film reviews as a representative short-text scenario where LLMs are widely applied. We collected 2,960 user reviews from Douban—a leading Chinese film review platform—sourced from the ChineseNlpCorpus (Aesop et al., 2023), which contains reviews for 28 movies. To ensure consistency and suitability for detection tasks, we selected reviews with lengths between 100 and 200 words. We then cleaned the data by removing excessive line breaks, restoring missing punctuation, and merging each review into a single paragraph.

Composition We include Gaokao model essays as a representative genre of Chinese writing frequently generated by LLMs. We collected 1,081 high-scoring essays from China's National College Entrance Examination by crawling publicly available websites. Given the heterogeneous sources, we standardized the data by removing titles, author names, source URLs, editorial comments, and other metadata, retaining only the main essay body. We also cleaned formatting artifacts such as tab characters and excessive whitespace to ensure textual consistency, resulting in a clean and uniform dataset.

Academic Writing To cover the academic writing domain, we extracted 500 Chinese-language research papers from ChinaXiv (ChinaXiv, 2025), an open repository for Chinese scientific literature. From each paper (originally in PDF format), we parsed the title, keyword, abstract, and introduction sections. We removed non-textual elements such as figures, tables, and citations, and further cleaned the extracted text by fixing irregular whitespace, spurious line breaks, and other formatting artifacts, resulting in a clean and standardized academic text dataset.

TC News In addition to Simplified Chinese, we also include Traditional Chinese content by selecting 2,000 news articles from News-Collection-Zhtw (Oscar), covering four categories: article, tech, science, and daily-weekly. To ensure balance, we sampled 500 articles per category. We applied the same cleaning procedure as used for the Simplified Chinese news dataset—removing metadata, filtering out high-English-content articles, and merging text into single paragraphs—and further constrained article lengths to 100–600 words.

A.2 Generative LLMs

All AI-generated texts in this work were produced by querying APIs of nine large language models. This black-box setup closely mirrors real-world usage, where end users interact with LLMs solely through cloud-based interfaces without access to internal weights or decoding details. To ensure a fair and controlled comparison across models, we applied an identical generation protocol for each data type. Same prompt templates were used for all models within a given domain and we fixed decoding parameters where supported by the API. Below is a brief overview of the nine models included in our study:

OpenAI (`gpt-3.5-turbo`, `gpt-4o`): Proprietary models with strong multilingual capabilities, including fluent Chinese generation.

Google (`Gemini-2.5-Flash`): A fast, lightweight model optimized for low-latency applications while maintaining decent Chinese text quality.

Anthropic (`Claude-3.5-Haiku`): The fastest model in the Claude 3.5 series, designed for instant responses and robust instruction following in Chinese.

Deepseek (`Deepseek-V3`, `Deepseek-R1`): Open-weight bilingual models; Deepseek-R1 is enhanced for reasoning via chain-of-thought prompting.

Qwen (`Qwen2.5`, `Qwen3`): State-of-the-art open-source models from Alibaba, excelling in Chinese comprehension and generation.

Doubao (`Doubao-1.5-Pro`): ByteDance's proprietary assistant model, fine-tuned for Chinese writing tasks such as exam essays and creative composition.

A.3 Prompt Design and Examples

To make the AI-generated topics comparable to human-written text, we first analyzed the human corpus for each domain (e.g., examining all movies covered in the review dataset, or breaking down news by category), then used an LLM to generate multiple candidate prompts for each domain, and manually screened them through iterative refinement to select the highest-quality ones.

A.3.1 Composition

Our prompt design is based on official Gaokao composition questions, covering three common formats. We adopt the standard instruction phrasing from recent exams, parameterized by the composition title and target word count. A representative example is shown below:

```
请以{composition title}为题，写一篇作文。要求：自选角度，自行立意；除诗歌外，文体不限；字数控制在{word count}左右。
```

The prompt structure closely follows official Gaokao guidelines, enabling LLMs to generate essays in a manner consistent with how users would interact with them for composition practice in real-world educational contexts.

A.3.2 Film Review

We leveraged LLMs to develop a set of structured review guidelines tailored to common film genres in our dataset. From these, we selected 10 representative templates that capture diverse critical perspectives (e.g., emotional response, thematic analysis). Each prompt is parameterized by the film title and target word count to guide LLM generation. A representative example is shown below:

```
以第一人称视角描述观看{film title}的体验，包括最触动你的1-2个场景及原因以及电影如何引发你的思考。要求：字数为{word count}字左右。
```

This design mirrors real-world practices where AI-generated reviews are typically produced by specifying film-specific instructions—such as perspective, focus, and length—to steer LLM output, thereby enhancing relevance and personalization.

A.3.3 Q&A

We developed structured answer templates using LLMs, informed by an analysis of question-answer patterns in our dataset. From this, we selected 10 representative templates that support diverse response styles. Each prompt combines a specific question and target word count with a standardized instruction to guide coherent and contextually appropriate answers. A representative example is shown below:

```
你是一位资深的网络话题评论员，目前你在网络上遇到了以下话题：{question}，请从一位用户的角度来对此话题进行分析以及探讨，要求内容字数在{word count}左右。
```

This approach reflects a common real-world practice in which LLMs are guided by both the question content and explicit formatting constraints to generate user-like responses—enabling us to simulate authentic human-AI interaction scenarios in online discussion contexts.

A.3.4 News

We developed a total of 25 structured news writing templates—five for each of five major news categories—using LLMs. These templates were derived through iterative analysis of real articles in our dataset. This ensures that LLM-generated content adheres not only to factual context but also to genre-specific stylistic norms. Our instruction-based approach better reflects real-world AI-assisted journalism workflows, where LLMs are typically provided with explicit guidelines rather than asked to continue from a headline alone. A representative example from the sports category is shown below:

```
请以体育新闻报道的风格，撰写一篇关于{title}的文章。内容要包括比赛过程、运动员表现和专业分析，语言要生动有激情，使用体育专业术语，字数控制在{word count}左右。
```

A.3.5 Academic Writing

We implement two distinct generation modes to simulate different stages of academic writing assistance:

Abstract Drafting from Title and Keywords In early-stage ideation, researchers often draft abstracts based on a working title and preliminary keywords. To reflect this scenario, we construct prompts using only the paper title and keywords, instructing the LLM to generate a concise abstract that outlines the research background, core problem, methodology, key findings, and significance.

```
请根据以下信息撰写一篇通用型学术论文摘要。
文章标题：{title}
文章关键词：{keyword}
摘要应简明扼要地说明研究背景、核心问题、研究方法、主要发现（或论点）及研究意义。要求语言严谨、逻辑连贯、术语准确，字数控制在{word count}字左右，适合投稿至各类中文学术期刊。
```

Introduction Generation from Abstract Once an abstract is available, researchers typically expand it into a full introduction. We model this common workflow by providing the LLM with the title, keywords, and abstract to generate an introduction that frames the scientific question and motivates the study.

```
请以提出关键科学问题的方式撰写引言内容，引导读者思考研究价值。
文章标题：{title}
文章关键词：{keyword}
文章摘要：{abstract}
字数控制在{word count}字左右，语言流畅，兼具学术性与可读性。
```

This dual-mode design captures realistic AI-assisted writing practices across different phases of scholarly composition.

A.3.6 TC News

We applied the same template-based approach as in simplified Chinese news to traditional Chinese (TC) journalism, developing 20 structured prompts across four categories: general news, technology, science, and weekly digests (5 templates per category). Each prompt combines a headline and target word count with genre-specific instructions to guide LLMs in producing authentic TC journalistic content. An example from the weekly digest category:

```
根據標題{title}，製作一份每週要點回顧。要求以分項列舉的形式（如‘重點事件’、‘數據解讀’、‘觀點彙總’等）清晰呈現一週的核心信息，語言需精煉、準確，字數控制在{word count}左右。
```

A.4 AI Generation Constraints

To ensure that AI-generated texts faithfully reflect the characteristics of human-written counterparts in each domain, we derived key generation constraints directly from the full empirical distribution of the human data. Specifically:

- **Length (words)**: The min–max range for each domain was set to cover 100% of the human-written samples' lengths, allowing slight extrapolation while preserving genre-typical scale.

Table 5: Quality control thresholds for AI-generated texts by domain.

Domain	Length (words)	Min. Ch Ratio (%)	Max. Rep (chars)
News	80–1100	80	20
Q&A	80–1000	80	25
Film Review	80–300	80	15
Composition	100–2000	80	25
Academic Writing	80–1600	60	40
TC News	80–900	80	20

- **Minimum Chinese Ratio**: Defined as the percentage of Chinese characters among all characters, this threshold filters out generations with excessive English or symbolic noise, aligning with the linguistic purity of cleaned human texts.

- **Maximum Repetition Length**: The maximum allowed length (in characters) of any repeated substring within the generated text. This threshold is set based on the longest duplicated span observed in human-written samples, with a small safety margin; generations containing longer repeated sequences are typically indicative of model degeneration or redundant hallucination.

These thresholds were enforced during both real-time generation monitoring and post-generation filtering. As a result, the AI-generated corpus maintains strong distributional alignment with human-written texts across structural and linguistic dimensions. The exact parameter values for each domain are summarized in Table 5.

A.5 Text Length Statistics

We observed that different large language models exhibit varying length biases when generating text—some tend to produce longer outputs than human-written text, while others generate shorter ones, indicating that text length alone cannot reliably distinguish between AI-generated and human-written content. Full details are showed in Table 18.

A.6 CSV Schema

While Table 1 defines the core fields shared across all samples, each domain includes additional metadata relevant to its source or generation context. These domain-specific fields are listed below.
Film Review
`film_title` (str) — Title of the reviewed film.

Composition
`composition_title` (str) — Title of composition.
`year` (int) — Year the composition was written.
News
`news_title` (str) — Headline of the news article.
`news_category` (str) — Editorial category.
Q&A
`question_title` (str) — The original question.
Academic Writing
`paper_title` (str) — Title of the source paper.
`keyword` (str) — Comma-separated keywords.
`category` (str) — `abstract` or `introduction`.
`extra_text` (str) — Contextual supplement: empty string (`""`) for abstracts; contains the full abstract text when the sample is an introduction.

All domain-specific fields are stored as string or integer types and appear only in samples belonging to the respective domain. They are absent in other domains.

A.7 Statistics

Our dataset consists of 128,610 texts across five domains, with contributions from ten text sources (including one human reference and nine large language models). The complete distribution is shown in Table 6.

All models generate roughly the same number of samples as the human reference within each domain, ensuring a fair comparison. Minor discrepancies arise from post-generation filtering, particularly for models like *Gemini-2.5-Flash*, which shows slightly fewer samples in certain domains.

B Detectors

B.1 Zero-shot Detectors

Zero-shot detectors refer to methods that distinguish AI-generated text from human-written text without any task-specific training or labeled examples. Instead, they rely solely on the statistical properties of text under a pre-trained language model (often the same model used for generation, or a comparable reference model). These approaches typically compute scores based on token-level likelihoods, prediction ranks, entropy, or other intrinsic signals derived during forward passes through the model. Below we summarize the zero-shot detection methods evaluated on C-ReD:

- **Log-Likelihood** (Solaiman et al., 2019): Computes the average log-probability of tokens in a text under a reference language model. AI-generated text typically exhibits higher likelihood than human-written text due to exposure during training.

- **Entropy** (Gehrmann et al., 2019): Measures the entropy of the predicted token distribution at each position. Human-written text tends to have higher entropy (more unpredictable) compared to AI output.

- **Log-Rank** (Mitchell et al., 2023): Uses the rank of each ground-truth token in the model's vocabulary prediction sorted by probability. AI-generated tokens often appear at higher ranks.

- **Log-Likelihood Log-Rank Ratio (LRR)** (Su et al., 2023): A zero-shot detection score that combines log-likelihood and log-rank. By taking the ratio of these two quantities, LRR better distinguishes AI-generated text—which tends to have high likelihood but even more strongly concentrated token ranks—from human-written text.

- **Fast-DetectGPT** (Bao et al., 2024): An efficient alternative to DetectGPT (Mitchell et al., 2023) that introduces conditional probability curvature as its core metric and uses a faster sampling approach.

- **Lastde / Lastde++** (Xu et al., 2025): A training-free detection method that treats the sequence of token probabilities generated by a language language model as a time series. By analyzing this sequence, Lastde and Lastde++ identify distinctive patterns characteristic of AI-generated text.

- **DNA-DetectLLM** (Zhu et al., 2025): A zero-shot method that constructs an ideal AI-generated sequence and measures how much an input text must be "repaired" to match it. The repair effort—higher for human text—serves as the detection score.

For all methods except DNA-DetectLLM, we use `EleutherAI/gpt-j-6b` as the scoring model to ensure a consistent basis for comparison. Specifically, Fast-DetectGPT and Lastde++ also adopt GPT-J-6B as their reference model. For DNA-DetectLLM, following its dual-model design, we set the observer model to `tiiuae/falcon-7b` and the performer model to `tiiuae/falcon-7b-instruct`, as recommended in the original implementation.

Domain	Human	GPT-3.5-Turbo	GPT-4o	Deepseek-V3	Deepseek-R1	Qwen2.5	Qwen3	Doubao-1.5-pro	Gemini-2.5-Flash	Claude-3.5-Haiku	Total
Composition	1,081	1,079	1,073	1,080	1,081	1,081	1,077	1,078	1,078	1,081	10,789
Film Review	2,960	2,959	2,960	2,960	2,960	2,960	2,960	2,960	2,960	2,960	29,599
Q&A	2,956	2,933	2,864	2,943	2,869	2,952	2,942	2,939	2,719	2,954	29,071
News	3,000	2,997	2,988	2,999	2,969	2,995	2,999	2,966	2,846	2,999	29,758
Academic Writing	1,000	991	1,000	1,000	999	1,000	1,000	999	998	1,000	9,987
TC News	2,000	1,998	2,000	1,996	1,972	1,970	1,964	1,989	1,517	2,000	19,406
Total	12,997	12,957	12,885	12,978	12,850	12,958	12,942	12,931	12,118	12,994	128,610

Table 6: Distribution of the dataset across domains and text sources (number of samples).

B.2 Supervised Detectors

Supervised detectors are data-driven approaches that learn to discriminate between human-written and AI-generated text by training on labeled datasets containing examples from both sources. The performance of supervised detectors heavily depends on the quality, diversity, and recency of the training data. Below we provide brief technical descriptions of the methods evaluated on C-ReD:

- **OpenAI Detector** (Solaiman et al., 2019): A detection classifier based on the pretrained RoBERTa model (Liu et al., 2019), originally trained to detect GPT-2 outputs. It takes raw text as input and predicts the probability of it being AI-generated. Despite being trained on older models, it remains a widely used baseline.

- **RADAR** (Hu et al., 2023): It trains a robust AI-generated text detector through adversarial learning between two language models: a paraphraser that rewrites AI-generated text to evade detection, and a detector that aims to correctly identify such rewritten outputs. By casting detection as a two-player game, RADAR enhances the detector's generalization to adaptive or obfuscated AI-generated text without requiring access to the original generator.

- **ReMoDetect** (Lee et al., 2024): A detection framework that enhances a reward model's ability by continually fine-tuning the reward model with experience replay to prevent overfitting, while also incorporating a mixed-preference dataset—created by partially rephrasing human-written text with an LLM—as an intermediate signal to refine the decision boundary.

- **Imitate Before Detect (ImBD)** (Chen et al., 2025): It first employs Style Preference Optimization (SPO) to align a scoring model with machine-like writing styles, then uses a Style-Conditional Probability Curvature (Style-CPC) metric for detection.

All supervised detectors are either used in their officially released form or re-implemented following the original papers, and fine-tuned on human–AI text pairs from the C-ReD training split. For RADAR and ReMoDetect, we utilized pre-trained models directly from HuggingFace without fine-tuning. For the OpenAI Detector and IMBD, we performed fine-tuning using specific hyperparameters. All models were tested with a maximum sequence length of 512. Detailed implementation settings and parameter configurations are listed in Table 7.

B.3 LLM Detectors

We evaluate the following eight large language models as detectors: **Qwen2.5** (Yang et al., 2024a), **Qwen3** (Yang et al., 2025), **Deepseek-V3** (Liu et al., 2024), **Doubao-1.5-pro** (ByteDance, 2025), **Gemini-2.5-Flash** (Comanici et al., 2025), **Claude-3.5-Haiku** (Anthropic, 2024), **GPT-4o** (Hurst et al., 2024), and **GPT-3.5-Turbo** (OpenAI).

For each model, we use a unified prompting strategy. The system message is fixed as:

```
请按照要求完成任务，注意，最后的输出内容只有
机器生成 和 人类生成 两种。不需要其他内容。
```

The user message contains the input text to be classified. All models are queried with identical generation settings: `temperature=0.7` and `max_tokens=100`. The final prediction is taken as the stripped output string. This setup ensures a fair and consistent comparison across different LLMs.

Table 7: Implementation Details and Hyperparameter Settings for Supervised Detectors

Method	Model / Base	Configuration / Hyperparameters
RADAR	TrustSafeAI/ RADAR-Vicuna-7B	**Mode:** Pre-trained (No fine-tuning) **Inference:** padding=True, truncation=True, max_length=512
ReMoDetect	hyunseoki/ ReMoDetect-deberta	**Mode:** Pre-trained (No fine-tuning) **Inference:** padding=True, truncation=True, max_length=512
OpenAI Detector	openai-community/ roberta-base/large	**Inference:** padding=True, truncation=True, max_length=512 **Fine-tuning:** lr=2e-5, epochs=5, warmup=100, weight_decay=0.01
IMBD	EleutherAI/ gpt-j-6B	**PEFT Type:** LORA (r=8, alpha=32, dropout=0.1) **Target Modules:** ["q_proj", "v_proj"] **Task:** CAUSAL_LM

C Results of Domain- and Generator-Agnostic Evaluation

Table 11 reports the complete AUROC scores of all evaluated detection methods across the five domains in C-ReD (Film Review, Composition, Q&A, News, and Academic Writing) and nine LLM generators. Results are grouped by domain for clarity, with each row showing performance on texts generated by a specific model under native Chinese prompts. The table includes both zero-shot detectors and supervised models. Notably, zero-shot methods exhibit strong sensitivity to generator fluency—particularly struggling on DeepSeek-R1 and Qwen3 in composition and academic writing. This comprehensive result matrix serves as a reference for fine-grained analysis of model- and domain-specific detection behavior.

D Detection Performance Before and After Fine-Tuning on C-ReD

Table 12 reports the AUROC scores of RoBERTa-base, RoBERTa-large, and IMBD detector across five domains in C-ReD. For each domain, results are split into two blocks: (i) *before fine-tuning* and (ii) *after fine-tuning on C-ReD*. The nine columns correspond to AI-generated text from seven in-distribution models (Deepseek-R1 through Qwen3) and two out-of-distribution models (Claude-3.5-Haiku and Gemini-2.5-Flash). The consistent performance gains—especially on OOD generators—demonstrate that fine-tuning on C-ReD not only improves in-domain detection but also enhances generalization to unseen Chinese-capable LLMs.

E Results of Domain Generalization Under Fixed Generators

Table 13 reports the resulting AUROC scores. Each block corresponds to a different training domain, with rows showing results for base and large variants of GPT-4o and Qwen-2.5 as the underlying generators. Strong in-domain performance (diagonal entries) is consistently observed, but generalization varies significantly across domain pairs. These results highlight both the potential and challenges of building domain-agnostic detectors for Chinese AI-generated text.

Table 8: Example of a description.

Description for News domain (Human vs. Deepseek-R1)
人类生成文本通常语言自然、信息具体、结构紧凑，侧重于客观陈述事实，细节真实且符合事件逻辑，行文风格平实，具有新闻报道或纪实特征；而机器生成文本则往往语言更华丽、修辞更丰富，倾向于使用排比、比喻等修辞手法，情感色彩更浓，有时会加入主观评论或引申意义，结构上更具文学性或论述性，但偶尔存在细节失真或过度渲染的问题。

F LLM Detect Evaluation Setup and Results

F.1 Prompting Strategies

We implement three distinct prompting strategies to probe the detection capabilities of Judge LLMs. All prompts enforce a strict output format (“机器生成” or “人类生成”) to enable automatic evaluation:

- `normal` (zero-shot): The judge receives only a basic instruction without any auxiliary information.

```
请判断下列文段为机器生成还是人类生成,根据判断结果最终输出 机器生成 或者 人类生成。文段如下: {content}
```

- `context` (few-shot with examples): Three paired examples of human and AI text from the same domain are included directly in the prompt.

```
下面给出三个人类生成文本以及对应的机器生成文本示例。
人类文本1：{human example 1}
机器文本1：{ai example 1}
人类文本2：{human example 2}
机器文本2：{ai example 2}
人类文本3：{human example 3}
机器文本3：{ai example 3}
请判断下列文段为机器生成还是人类生成,根据判断结果最终输出 机器生成 或者 人类生成。文段如下：{content}
```

- `description` (rule-based): Instead of raw examples, a natural-language description of stylistic differences—automatically generated by Qwen3 based on example pairs—is inserted into the prompt.

```
下面给出人类生成文本以及机器生成文本的特点描述：{summary description}
请判断下列文段为机器生成还是人类生成,根据判断结果最终输出 机器生成 或者 人类生成。文段如下：{content}
```

The {summary description} is a concise summary produced by Qwen3, capturing perceived distinctions such as fluency, repetition, emotional depth, and structural patterns. For illustration, Table 8 shows a representative description generated for the News domain in the Human vs. Deepseek-R1 setting.

F.2 Full Experiment Results

Table 14 presents the complete results of our experiments across five domains (Film Review, Composition, Q&A, News, and Academic Writing) and three target AI generators (Qwen2.5, GPT-4o, Deepseek-R1). For each Judge LLM and prompting strategy (`normal`, `context`, `description`), we report two metrics: **Human Accuracy (H)** — the percentage of human-written texts correctly identified, and **AI Accuracy (A)** — the percentage of AI-generated texts correctly identified. Higher values in both columns indicate better discrimination ability. Notably, most Judge LLMs struggle under the `normal` setting, while the `context` and `description` strategies consistently improve detection performance—demonstrating the value of domain-specific stylistic cues.

G Setup and Statistical Analysis of Ablation on Prompt Complexity

To investigate the impact of prompt complexity on detection difficulty, we design two broad categories of simplified generation scenarios that reflect common stages in AI-assisted academic writing. Within each category, we implement multiple prompt variants that provide minimal guidance—significantly less than the original C-ReD prompts, which include detailed structural, stylistic, and contextual instructions. Below, we present one representative example from each category.

Abstract Drafting from Title This category simulates early-stage ideation, where users request an abstract based only on high-level metadata. The following is a typical prompt instance:

```
请根据以下信息撰写一篇通用型学术论文摘要。
文章标题：{title}
字数控制在{word count}字左右。
```

Introduction Generation from Title and Keywords This category models later-stage expansion, where title and keywords are used to generate a full introduction. Prompts in this group vary in the amount of provided context, but all omit explicit methodological or rhetorical guidance. An example prompt is:

```
请根据以下信息撰写一篇学术论文的引言部分内容。
文章标题：{title}
文章关键词：{keyword}
要求字数约为{word count}字。
```

These simplified prompts collectively reduce task specificity and stylistic constraints, resulting in more uniform and predictable model outputs.

Statistical Analysis For the prompt complexity ablation study, we conducted bootstrap-based statistical analysis (n=1,000) to compare detection difficulty between original and simplified prompts (results are showed in Table 9 and Table 10). For Qwen2.5, no statistically significant difference was observed across both detectors (p=0.21 for RoBERTa-base; p=0.23 for RoBERTa-large). For GPT-4o, while statistical significance was reached (p=0.01 for base; p=0.03 for large), the absolute AUROC gap is small (1.76%–3.73%) and both conditions yield high detection performance (AUROC > 0.92), suggesting that prompt simplification does not substantially impair detectability.

Table 9: Performance comparison on RoBERTa-base

Generator	Prompt	AUROC	95% CI	Difference	Diff 95% CI	P-value
GPT-4o	Original	0.9566	[0.9366, 0.9730]	—	—	—
GPT-4o	Simplified	0.9201	[0.8912, 0.9415]	0.0373	[0.0068, 0.0727]	0.0100
Qwen2.5	Original	0.9648	[0.9490, 0.9789]	—	—	—
Qwen2.5	Simplified	0.9501	[0.9277, 0.9664]	0.0160	[-0.0062, 0.0410]	0.2100

Table 10: Performance comparison on RoBERTa-large

Generator	Prompt	AUROC	95% CI	Difference	Diff 95% CI	P-value
GPT-4o	Original	0.9868	[0.9764, 0.9945]	—	—	—
GPT-4o	Simplified	0.9691	[0.9542, 0.9818]	0.0176	[0.0024, 0.0349]	0.0320
Qwen2.5	Original	0.9906	[0.9834, 0.9959]	—	—	—
Qwen2.5	Simplified	0.9829	[0.9703, 0.9932]	0.0077	[-0.0051, 0.0213]	0.2320

H Results on Traditional Chinese News

We provide full evaluation results on an external Traditional Chinese news dataset. Table 16 shows AUROC scores of various detection methods across nine LLM generators. Table 17 presents detection performance before and after fine-tuning. Table 15 reports the accuracy of LLM-as-a-judge under different prompting strategies.

I Setup and of Transfer Performance on External Chinese Datasets

We construct our external Chinese test set by randomly sampling 600 human–AI text pairs from each of two generation models (**ChatGPT** and **davinci-003**) in the M4 dataset (Wang et al., 2023), specifically from its *Baike/Web QA* domains. This yields a balanced test set of 1,200 samples, ensuring coverage of both factual and open-domain question-answering styles commonly found in real-world Chinese applications. All AI-generated texts in this set are produced by models not included in C-ReD's training distribution, enabling a strict out-of-distribution evaluation of transfer performance.

J Examples of C-ReD

We provide representative examples from the C-ReD dataset to illustrate the characteristics of human-written and AI-generated texts across different domains. Each example includes: (1) the domain and model used for generation, (2) the original prompt, (3) the human-written text, and (4) the corresponding AI-generated content. Table 19 provides examples for each domain in the dataset.

Table 11: Full AUROC detection performance across five domains and nine LLM generators in C-ReD.

Method	Deepseek-R1	Deepseek-V3	Doubao-1.5-pro	GPT-3.5-Turbo	GPT-4o	Qwen2.5	Qwen3	Claude-3.5-Haiku	Gemini-2.5-Flash	Average
Film Review										
Log-Likelihood	0.5848	0.6801	0.7074	0.9737	0.9705	0.9692	0.8257	0.9247	0.8737	0.8344
Entropy	0.6239	0.7103	0.7235	0.9634	0.9697	0.9716	0.8567	0.9116	0.8533	0.8427
Log-Rank	0.5864	0.6811	0.6887	0.9714	0.9658	0.9652	0.8136	0.9200	0.8654	0.8286
LRR	0.5598	0.6310	0.5626	0.8538	0.8153	0.8420	0.6657	0.7706	0.7192	0.7133
Fast-DetectGPT	0.4144	0.5126	0.5741	0.9092	0.8639	0.8329	0.6149	0.8112	0.7659	0.6999
Lastde	0.6218	0.5585	0.6725	0.5191	0.5094	0.4496	0.6463	0.6589	0.4651	0.5668
Lastde++	0.4688	0.5295	0.5874	0.8882	0.8572	0.8244	0.6418	0.7843	0.7715	0.7059
DNA-DetectLLM	0.4411	0.5978	0.6816	0.9526	0.9375	0.8729	0.7284	0.8131	0.8104	0.7595
RoBERTa-base	0.6951	0.5306	0.7522	0.6708	0.6579	0.5345	0.6479	0.5839	0.7419	0.6461
RoBERTa-large	0.6014	0.6451	0.7016	0.6263	0.5346	0.5388	0.6470	0.6564	0.5575	0.6121
RADAR	0.8096	0.8100	0.8478	0.7911	0.8244	0.8427	0.8300	0.8387	0.8672	0.8291
ReMoDetect-deberta	0.9333	0.9603	0.9516	0.9972	0.9983	0.9979	0.9902	0.9912	0.9378	0.9731
IMBD	0.6610	0.7775	0.8156	0.9703	0.9663	0.9472	0.8889	0.9557	0.9016	0.8760
Composition										
Log-Likelihood	0.2331	0.8778	0.9747	0.9939	0.9905	0.9739	0.6178	0.9667	0.9618	0.8433
Entropy	0.2341	0.8080	0.9369	0.9866	0.9753	0.9649	0.5367	0.9268	0.9098	0.8088
Log-Rank	0.2501	0.8808	0.9748	0.9944	0.9911	0.9736	0.6314	0.9668	0.9626	0.8473
LRR	0.3939	0.8393	0.9526	0.9943	0.9870	0.9416	0.6559	0.9251	0.9359	0.8473
Fast-DetectGPT	0.3786	0.9580	0.9925	0.9992	0.9997	0.9173	0.8370	0.9775	0.9973	0.8952
Lastde	0.8035	0.5908	0.5451	0.7979	0.6449	0.5814	0.7020	0.6863	0.4979	0.6500
Lastde++	0.4020	0.9305	0.9902	0.9964	0.9987	0.8550	0.8204	0.9515	0.9943	0.8821
DNA-DetectLLM	0.5115	0.9857	0.9952	1.0000	1.0000	0.9679	0.8806	0.9985	0.9976	0.9263
RoBERTa-base	0.5519	0.4270	0.6066	0.5793	0.4623	0.3253	0.6041	0.3994	0.7157	0.5191
RoBERTa-large	0.6754	0.5625	0.5122	0.5445	0.5373	0.4993	0.5404	0.5374	0.4449	0.5393
RADAR	0.6731	0.5302	0.6216	0.6677	0.6561	0.6260	0.6480	0.6596	0.6216	0.6338
ReMoDetect-deberta	0.5727	0.8287	0.9675	0.9931	0.9944	0.9840	0.6625	0.9640	0.8912	0.8731
IMBD	0.5668	0.9234	0.9910	0.9999	0.9999	0.9761	0.7791	0.9966	0.9929	0.9140
Q&A										
Log-Likelihood	0.8210	0.8598	0.9212	0.9928	0.9865	0.9699	0.9331	0.9681	0.9563	0.9343
Entropy	0.8438	0.8650	0.9042	0.9825	0.9728	0.9574	0.9228	0.9478	0.9316	0.9253
Log-Rank	0.8092	0.8514	0.9162	0.9919	0.9862	0.9690	0.9262	0.9661	0.9541	0.9301
LRR	0.6541	0.7146	0.8264	0.9495	0.9370	0.8978	0.8019	0.8792	0.8886	0.8388
Fast-DetectGPT	0.5308	0.6506	0.8730	0.9806	0.9713	0.8531	0.8188	0.9131	0.9548	0.8385
Lastde	0.5281	0.6165	0.6237	0.8345	0.7046	0.6200	0.6061	0.7665	0.4626	0.6403
Lastde++	0.5470	0.6127	0.8460	0.9601	0.9547	0.8153	0.7915	0.8731	0.9326	0.8148
DNA-DetectLLM	0.5611	0.6774	0.8664	0.9907	0.9778	0.8169	0.8366	0.9299	0.9387	0.8439
RoBERTa-base	0.5403	0.4533	0.5539	0.5676	0.4993	0.4586	0.5381	0.4147	0.5997	0.5139
RoBERTa-large	0.5364	0.5688	0.5290	0.5302	0.5302	0.4725	0.5209	0.5204	0.5207	0.5255
RADAR	0.7259	0.7585	0.7935	0.7755	0.7619	0.7586	0.7881	0.7598	0.7230	0.7605
ReMoDetect-deberta	0.9582	0.9629	0.9642	0.9992	0.9976	0.9943	0.9764	0.9934	0.9338	0.9755
IMBD	0.7218	0.7696	0.8707	0.9935	0.9935	0.9365	0.9088	0.9799	0.9353	0.9011

Method	Deepseek-R1	Deepseek-V3	Doubao-1.5-pro	GPT-3.5-Turbo	GPT-4o	Qwen2.5	Qwen3	Claude-3.5-Haiku	Gemini-2.5-Flash	Average
News										
Log-Likelihood	0.4525	0.5441	0.7869	0.9214	0.8805	0.8133	0.6680	0.8128	0.7562	0.7373
Entropy	0.4665	0.5349	0.7068	0.8527	0.8056	0.7676	0.6312	0.7448	0.6752	0.6873
Log-Rank	0.4451	0.5387	0.7915	0.9270	0.8841	0.8158	0.6614	0.8120	0.7590	0.7372
LRR	0.4365	0.5107	0.7604	0.9080	0.8479	0.7634	0.6022	0.7454	0.7294	0.7004
Fast-DetectGPT	0.4108	0.5544	0.9019	0.9775	0.9674	0.8066	0.7242	0.8891	0.9020	0.7926
Lastde	0.6623	0.6272	0.7331	0.7485	0.6831	0.6607	0.6475	0.7191	0.6045	0.6762
Lastde++	0.4652	0.5492	0.8876	0.9678	0.9589	0.8002	0.7329	0.8582	0.8944	0.7905
DNA-DetectLLM	0.3566	0.5378	0.8106	0.9619	0.9234	0.6466	0.6254	0.8360	0.8096	0.7231
RoBERTa-base	0.5132	0.4070	0.5558	0.5596	0.4974	0.3992	0.4972	0.4919	0.5218	0.4937
RoBERTa-large	0.4647	0.4688	0.3370	0.3233	0.3299	0.2686	0.4005	0.3652	0.3710	0.3699
RADAR	0.4754	0.4487	0.4552	0.5229	0.4620	0.4299	0.4794	0.4494	0.4515	0.4638
ReMoDetect-deberta	0.6840	0.7762	0.8831	0.9686	0.9620	0.9359	0.8488	0.9389	0.7898	0.8652
IMBD	0.5624	0.5657	0.8809	0.9565	0.9655	0.8198	0.7059	0.8640	0.8373	0.7953
Academic Writing										
Log-Likelihood	0.5166	0.7011	0.8225	0.8416	0.8510	0.6917	0.5715	0.7778	0.8198	0.7326
Entropy	0.5594	0.6820	0.7627	0.7437	0.7611	0.6692	0.5890	0.7069	0.7429	0.6907
Log-Rank	0.5258	0.7033	0.8221	0.8509	0.8588	0.6983	0.5780	0.7810	0.8279	0.7384
LRR	0.5685	0.6736	0.7594	0.8350	0.8310	0.6844	0.5975	0.7421	0.8161	0.7231
Fast-DetectGPT	0.3648	0.6488	0.8238	0.8996	0.9099	0.6303	0.4551	0.8070	0.8793	0.7132
Lastde	0.5545	0.7243	0.8159	0.8672	0.8345	0.7328	0.5825	0.8360	0.7735	0.7468
Lastde++	0.4191	0.6412	0.8093	0.8776	0.9023	0.6206	0.4943	0.7886	0.8881	0.7157
DNA-DetectLLM	0.2721	0.5437	0.7377	0.8959	0.8613	0.5116	0.2794	0.7970	0.7048	0.6226
RoBERTa-base	0.4315	0.4115	0.4281	0.5194	0.5010	0.3313	0.4060	0.3946	0.4608	0.4316
RoBERTa-large	0.4551	0.4738	0.3653	0.3934	0.4032	0.3685	0.4149	0.3674	0.4113	0.4059
RADAR	0.4706	0.5217	0.5439	0.5758	0.5384	0.5285	0.4649	0.5450	0.4611	0.5167
ReMoDetect-deberta	0.7414	0.9028	0.9322	0.9816	0.9846	0.9557	0.8386	0.9589	0.9173	0.9126
IMBD	0.5424	0.7402	0.8326	0.9509	0.9442	0.8059	0.6217	0.9328	0.8798	0.8056

Table 12: AUROC detection performance before (upper block per domain) and after (lower block per domain) fine-tuning on C-ReD. Columns 1–7 show results on in-distribution generators; Columns 8–9 show out-of-distribution (OOD) generalization to unseen models. Fine-tuning substantially boosts both in-domain accuracy and OOD robustness across all domains.

Method	Deepseek-R1	Deepseek-V3	Doubao-1.5-pro	GPT-3.5-Turbo	GPT-4o	Qwen2.5	Qwen3	Claude-3.5-Haiku	Gemini-2.5-Flash	Average
Film Review										
RoBERTa-base	0.6951	0.5306	0.7522	0.6708	0.6579	0.5345	0.6479	0.5839	0.7419	0.6461
RoBERTa-large	0.6014	0.6451	0.7016	0.6263	0.5346	0.5388	0.6470	0.6564	0.5575	0.6121
IMBD	0.6610	0.7775	0.8156	0.9703	0.9663	0.9472	0.8889	0.9557	0.9016	0.8760
RoBERTa-base	0.9997	0.9998	0.9974	0.9968	1.0000	0.9986	0.9975	0.9878	0.9906	0.9965
RoBERTa-large	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9995	0.9999
IMBD	0.9981	0.9989	0.9986	0.9997	0.9997	0.9995	0.9994	0.9990	0.9959	0.9988
Composition										
RoBERTa-base	0.5519	0.4270	0.6066	0.5793	0.4623	0.3253	0.6041	0.3994	0.7157	0.5191
RoBERTa-large	0.6754	0.5625	0.5122	0.5445	0.5373	0.4993	0.5404	0.5374	0.4449	0.5393
IMBD	0.5668	0.9234	0.9910	0.9999	0.9999	0.9761	0.7791	0.9966	0.9929	0.9140
RoBERTa-base	0.8900	0.9409	0.8834	0.9218	0.9544	0.9234	0.8597	0.9768	0.8572	0.9120
RoBERTa-large	0.9530	0.9648	0.9467	0.9490	0.9785	0.9347	0.9350	0.9823	0.9389	0.9537
IMBD	0.9103	0.9898	0.9935	0.9975	0.9976	0.9920	0.9727	0.9974	0.9968	0.9831
Q&A										
RoBERTa-base	0.5403	0.4533	0.5539	0.5676	0.4993	0.4586	0.5381	0.4147	0.5997	0.5139
RoBERTa-large	0.5364	0.5688	0.5290	0.5302	0.5302	0.4725	0.5209	0.5204	0.5207	0.5255
IMBD	0.7218	0.7696	0.8707	0.9935	0.9935	0.9365	0.9088	0.9799	0.9353	0.9011
RoBERTa-base	0.9896	0.9929	0.9853	0.9942	0.9945	0.9849	0.9916	0.9921	0.9493	0.9860
RoBERTa-large	0.9976	0.9988	0.9965	0.9993	0.9990	0.9971	0.9987	0.9975	0.9843	0.9965
IMBD	0.9875	0.9893	0.9918	0.9998	0.9997	0.9975	0.9968	0.9992	0.9944	0.9951
News										
RoBERTa-base	0.5132	0.4070	0.5558	0.5596	0.4974	0.3992	0.4972	0.4919	0.5218	0.4937
RoBERTa-large	0.4647	0.4688	0.3370	0.3233	0.3299	0.2686	0.4005	0.3652	0.3710	0.3699
IMBD	0.5624	0.5657	0.8809	0.9565	0.9655	0.8198	0.7059	0.8640	0.8373	0.7953
RoBERTa-base	0.9796	0.9804	0.9789	0.9693	0.9809	0.9765	0.9717	0.9898	0.9528	0.9755
RoBERTa-large	0.9893	0.9922	0.9790	0.9656	0.9899	0.9898	0.9841	0.9907	0.9733	0.9838
IMBD	0.9292	0.9512	0.9936	0.9961	0.9959	0.9891	0.9785	0.9946	0.9858	0.9793
Academic Writing										
RoBERTa-base	0.4315	0.4115	0.4281	0.5194	0.5010	0.3313	0.4060	0.3946	0.4608	0.4316
RoBERTa-large	0.4551	0.4738	0.3653	0.3934	0.4032	0.3685	0.4149	0.3674	0.4113	0.4059
IMBD	0.5424	0.7402	0.8326	0.9509	0.9442	0.8059	0.6217	0.9328	0.8798	0.8056
RoBERTa-base	0.9419	0.9396	0.9196	0.8821	0.9565	0.9648	0.9523	0.9720	0.9083	0.9375
RoBERTa-large	0.9810	0.9753	0.9357	0.9666	0.9829	0.9888	0.9595	0.9792	0.9515	0.9689
IMBD	0.9570	0.9890	0.9817	0.9948	0.9978	0.9908	0.9583	0.9953	0.9888	0.9837

Table 13: Cross-domain AUROC results when training on a single domain and evaluating across all five domains, using texts generated by GPT-4o or Qwen-2.5. For each block, the training domain is fixed (left column), and columns show performance on Composition, Film Review, News, Academic Writing, and Q&A. Diagonal entries indicate in-domain performance; off-diagonal entries reflect generalization capability.

Train Domain	Model	Composition	Film Review	News	Academic	Q&A
Composition	GPT-4o (Base)	0.9765	0.8493	0.8641	0.8064	0.8924
	GPT-4o (Large)	0.9876	0.9702	0.9431	0.8740	0.9619
	Qwen-2.5 (Base)	0.9553	0.6861	0.8831	0.8487	0.8363
	Qwen-2.5 (Large)	0.9775	0.9728	0.9754	0.8768	0.9565
Film Review	GPT-4o (Base)	0.7685	1.0000	0.5832	0.6680	0.9207
	GPT-4o (Large)	0.7376	1.0000	0.8080	0.7809	0.9502
	Qwen-2.5 (Base)	0.8435	0.9998	0.7410	0.7665	0.8868
	Qwen-2.5 (Large)	0.8134	1.0000	0.8689	0.7969	0.9098
News	GPT-4o (Base)	0.9049	0.9838	0.9894	0.9328	0.9478
	GPT-4o (Large)	0.9093	0.9604	0.9944	0.9471	0.9499
	Qwen-2.5 (Base)	0.6419	0.8776	0.9295	0.7200	0.7025
	Qwen-2.5 (Large)	0.8619	0.9494	0.9925	0.9005	0.9227
Academic	GPT-4o (Base)	0.8568	0.8830	0.9060	0.9627	0.7995
	GPT-4o (Large)	0.8691	0.9355	0.8851	0.9668	0.8379
	Qwen-2.5 (Base)	0.8583	0.6852	0.9181	0.9799	0.8462
	Qwen-2.5 (Large)	0.8269	0.9050	0.9399	0.7435	0.6394
Question Answer	GPT-4o (Base)	0.8919	0.9956	0.9255	0.9032	0.9974
	GPT-4o (Large)	0.9612	0.9981	0.9488	0.9431	0.9997
	Qwen-2.5 (Base)	0.7692	0.9908	0.9093	0.8981	0.9887
	Qwen-2.5 (Large)	0.8857	0.9972	0.9689	0.9469	0.9973

Table 14: Results of LLM-base Detector Evaluation.

Judge LLM	Human vs. Qwen2.5						Human vs. GPT-4o						Human vs. Deepseek-R1					
	normal		context		description		normal		context		description		normal		context		description	
	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A
Film Review																		
Qwen2.5	99.16	7.26	99.66	88.51	98.82	83.11	99.16	8.61	99.66	89.36	99.49	81.42	99.16	23.99	99.66	87.50	98.65	80.91
Qwen3	89.70	9.46	99.66	83.45	98.48	88.51	89.70	8.28	99.66	92.40	99.16	83.28	89.70	13.01	99.16	90.71	95.44	94.59
Deepseek-V3	96.96	14.19	99.66	91.22	97.30	91.22	96.96	19.43	99.16	96.45	99.83	81.42	96.96	44.76	96.28	92.57	93.92	96.45
Doubao-1.5-pro	99.83	8.61	98.31	91.22	99.16	84.97	99.83	12.84	99.32	87.33	99.66	80.24	99.83	14.86	98.81	85.81	99.66	72.80
Gemini-2.5-Flash	99.83	4.39	100.0	88.01	99.49	87.33	99.83	4.05	100.0	86.66	99.83	81.42	99.83	11.49	100.0	85.14	99.66	79.56
Claude-3.5-Haiku	83.45	34.12	98.99	88.85	99.83	75.84	83.45	45.44	98.82	96.45	99.49	80.91	83.45	47.13	95.44	87.33	99.83	67.57
GPT-4o	100.0	0.17	99.66	89.53	99.66	82.77	100.0	0.00	99.83	90.03	99.66	81.08	100.0	1.69	99.16	91.39	98.48	75.17
GPT-3.5-Turbo	89.36	3.89	91.22	60.47	97.97	74.66	89.36	3.38	90.54	46.28	99.32	49.16	89.36	11.15	68.75	87.50	97.97	63.51
Composition																		
Qwen2.5	97.65	3.76	95.77	66.20	86.38	48.83	97.65	27.70	98.12	70.89	90.61	27.23	97.65	1.41	97.65	75.59	71.36	30.99
Qwen3	94.37	0.94	85.92	72.30	77.46	84.04	94.37	0.47	89.20	65.73	75.12	73.24	94.37	0.00	69.95	68.08	44.13	63.38
Deepseek-V3	95.77	3.76	76.53	95.31	87.79	54.46	95.77	7.98	87.79	87.79	78.87	55.40	95.77	4.23	58.22	81.69	43.19	64.32
Doubao-1.5-pro	95.31	5.63	83.57	90.61	90.14	46.48	95.31	11.27	96.24	67.61	88.26	40.85	95.31	0.47	98.59	62.44	60.09	53.05
Gemini-2.5-Flash	97.65	22.07	100.0	96.71	95.31	63.85	97.65	37.09	98.59	91.55	93.43	65.26	97.65	17.37	98.59	85.45	79.34	59.62
Claude-3.5-Haiku	58.22	41.31	64.32	93.90	82.63	53.52	58.22	49.77	76.06	84.04	84.51	38.97	58.22	22.96	43.66	67.14	59.62	28.64
GPT-4o	100.0	0.00	82.16	97.18	77.93	65.73	100.0	0.00	79.34	97.65	87.32	41.31	100.0	0.00	80.75	98.59	61.03	53.05
GPT-3.5-Turbo	98.12	0.47	93.43	16.43	99.06	3.76	98.12	0.94	96.71	9.39	99.53	0.94	98.12	0.47	91.08	5.16	98.12	0.47
Q&A																		
Qwen2.5	90.06	45.32	92.16	99.43	73.61	99.81	90.06	75.53	92.93	99.62	88.15	97.51	90.06	72.85	94.26	93.31	87.19	89.87
Qwen3	82.98	6.69	98.47	88.72	84.89	99.81	82.98	7.07	98.66	98.09	89.67	99.62	82.98	14.53	97.51	84.13	88.91	95.03
Deepseek-V3	94.46	9.94	95.98	97.9	86.23	99.24	94.46	21.41	96.37	98.85	89.1	99.62	94.46	29.45	96.94	88.72	86.04	91.20
Doubao-1.5-pro	97.13	17.97	93.69	98.47	83.37	98.85	97.13	35.56	94.07	99.62	91.97	98.28	97.13	33.27	97.71	87.95	91.40	86.81
Gemini-2.5-Flash	99.24	21.61	98.85	96.37	96.37	97.90	99.24	32.50	99.43	95.98	98.09	98.47	99.24	26.58	99.24	84.70	99.04	81.84
Claude-3.5-Haiku	44.93	43.02	46.27	99.43	78.59	98.66	44.93	60.8	69.79	99.43	92.73	86.23	44.93	72.66	74.0	94.84	90.63	68.26
GPT-4o	98.85	0.38	93.12	99.43	90.63	97.71	98.85	0.00	91.59	99.62	93.69	97.51	98.85	3.25	95.41	94.26	94.07	82.98
GPT-3.5-Turbo	74.95	4.78	59.08	58.70	79.73	82.22	74.95	9.37	51.63	74.95	73.04	85.66	74.95	14.15	44.55	80.50	79.16	79.54
News																		
Qwen2.5	65.45	52.18	98.36	83.45	97.45	26.18	65.45	65.45	99.64	86.0	97.82	25.64	65.45	68.18	98.91	81.45	98.91	12.73
Qwen3	84.73	17.45	98.73	89.09	94.91	73.45	84.73	15.45	98.55	88.91	96.73	63.45	84.73	21.45	98.73	67.09	96.18	31.64
Deepseek-V3	89.09	10.91	97.82	91.82	89.09	70.91	89.09	15.27	98.18	88.36	90.18	57.64	89.09	21.82	98.55	82.0	89.64	32.0
Doubao-1.5-pro	98.0	12.18	99.64	86.18	98.55	21.27	98.0	15.82	99.82	77.64	98.91	12.18	98.0	12.73	100.0	57.09	99.27	13.27
Gemini-2.5-Flash	96.73	19.09	99.82	88.73	97.64	65.09	96.73	26.18	99.45	79.82	98.18	54.36	96.73	25.09	99.82	85.45	97.09	30.91
Claude-3.5-Haiku	67.09	67.82	53.09	97.64	88.55	48.36	67.09	77.09	74.36	94.55	91.45	46.0	67.09	62.0	59.45	90.55	92.0	28.55
GPT-4o	99.64	0.00	95.09	95.45	97.27	22.73	99.64	0.00	93.09	98.91	97.27	15.82	99.64	0.36	93.64	94.73	98.36	16.73
GPT-3.5-Turbo	72.36	5.09	54.91	28.73	93.82	4.73	72.36	9.27	52.18	32.18	93.45	0.55	72.36	17.45	67.45	26.36	93.27	9.09
Academic Writing																		
Qwen2.5	2.50	97.0	84.50	97.0	48.0	96.50	2.50	99.50	86.50	100.0	62.50	94.0	2.50	100.0	88.0	99.50	38.0	100.0
Qwen3	86.0	4.50	95.0	84.50	58.0	97.0	86.0	11.0	91.0	100.0	72.0	90.50	86.0	13.50	89.0	97.50	37.50	100.0
Deepseek-V3	49.50	54.50	95.50	77.50	66.50	87.0	49.50	63.50	84.50	98.0	58.0	87.50	49.50	70.50	73.0	100.0	33.0	100.0
Doubao-1.5-pro	89.50	14.50	96.50	80.50	51.0	97.0	89.50	22.50	90.0	98.0	80.50	60.50	89.50	14.50	93.50	99.50	38.0	100.0
Gemini-2.5-Flash	78.0	60.50	100.0	93.50	59.0	96.50	78.0	71.0	99.50	98.50	72.0	93.50	78.0	62.0	98.50	99.0	51.50	100.0
Claude-3.5-Haiku	30.50	66.50	33.50	98.0	39.0	96.50	30.50	73.50	65.50	97.0	54.50	76.50	30.50	74.50	83.0	81.0	9.0	100.0
GPT-4o	99.0	1.50	90.0	94.50	69.0	81.0	99.0	1.0	88.0	96.50	82.50	55.50	99.0	2.0	86.0	100.0	66.0	98.0
GPT-3.5-Turbo	69.50	12.50	28.0	79.50	77.50	46.0	69.50	21.50	25.0	86.50	39.0	65.50	69.50	33.50	42.50	81.0	22.50	91.0

Table 15: LLM-as-detector results on Traditional Chinese news. Each cell shows accuracy (%) for distinguishing human (H) vs. AI (A) text under three prompt types. Performance is highly dependent on both judge model and prompt design.

Judge LLM	Human vs. Qwen2.5						Human vs. GPT-4o						Human vs. Deepseek-R1					
	normal		context		description		normal		context		description		normal		context		description	
	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A
Qwen2.5	81.85	71.17	89.68	91.10	61.57	83.27	81.85	81.85	89.32	91.81	54.09	84.34	81.85	77.22	90.04	91.10	43.42	83.27
Qwen3	87.19	23.84	81.85	96.09	67.97	88.26	87.19	23.84	81.49	93.95	56.94	92.53	87.19	16.73	74.38	86.83	46.26	85.41
Deepseek-V3	95.02	31.32	88.97	97.15	77.58	81.85	95.02	43.06	85.41	97.51	68.33	87.54	95.02	34.16	77.58	99.64	47.33	86.83
Doubao-1.5-pro	94.31	30.96	98.22	89.68	61.21	77.94	94.31	44.84	97.51	87.54	49.82	92.17	94.31	27.76	95.37	84.70	46.98	88.61
Gemini-2.5-Flash	98.58	43.77	99.64	91.81	78.65	83.27	98.58	46.26	99.64	96.09	77.22	88.26	98.58	35.23	99.29	90.39	70.46	80.07
Claude-3.5-Haiku	71.17	73.31	31.67	99.29	58.72	75.44	71.17	85.05	66.55	97.86	50.18	82.92	71.17	64.06	68.33	94.31	37.72	88.61
GPT-4o	100.0	1.78	88.61	93.95	76.87	67.26	100.0	1.07	87.19	96.44	64.77	80.07	100.0	7.12	77.22	97.51	56.58	75.44
GPT-3.5-Turbo	83.27	6.05	50.89	73.31	67.62	47.69	83.27	8.54	53.74	79.72	61.21	58.72	83.27	23.84	56.94	71.17	53.02	72.24

Table 16: AUROC detection performance of diverse methods on Traditional Chinese news across nine LLM generators.

Method	Deepseek-R1	Deepseek-V3	Doubao-1.5-pro	GPT-3.5-Turbo	GPT-4o	Qwen2.5	Qwen3	Claude-3.5-Haiku	Gemini-2.5-Flash	Average
Log-Likelihood	0.3826	0.5783	0.8223	0.9659	0.9646	0.9032	0.3991	0.8603	0.8639	0.7489
Entropy	0.4467	0.6077	0.7782	0.9223	0.9254	0.8776	0.4581	0.8152	0.7933	0.7361
Log-Rank	0.3772	0.5643	0.8097	0.9661	0.9613	0.8977	0.3854	0.8563	0.8606	0.7421
LRR	0.4002	0.4940	0.6915	0.9019	0.8791	0.7638	0.3754	0.7531	0.7701	0.6699
Fast-DetectGPT	0.2718	0.4479	0.8455	0.9722	0.9557	0.7352	0.2974	0.8300	0.9080	0.6960
Lastde	0.5375	0.5640	0.5950	0.6894	0.5883	0.5332	0.5192	0.6368	0.4232	0.5652
Lastde++	0.2969	0.4194	0.7948	0.9513	0.9387	0.6871	0.3143	0.7861	0.8718	0.6734
DNA-DetectLLM	0.3330	0.4603	0.8667	0.9038	0.9328	0.5974	0.3450	0.7914	0.7588	0.6655
RoBERTa-base	0.4450	0.4960	0.5444	0.5720	0.5257	0.3221	0.4566	0.3736	0.5732	0.4787
RoBERTa-large	0.6720	0.6476	0.5879	0.5771	0.5494	0.4369	0.5912	0.6255	0.5660	0.5837
RADAR	0.4922	0.5298	0.5483	0.5505	0.5150	0.5606	0.5043	0.5523	0.5564	0.5344
ReMoDetect-deberta	0.7618	0.8729	0.9508	0.9938	0.9951	0.9914	0.7887	0.9702	0.9190	0.9160
IMBD	0.4286	0.6057	0.8736	0.9360	0.9347	0.8543	0.5103	0.8719	0.8437	0.7621

Table 17: Detection performance on Traditional Chinese news before (upper block) and after (lower block) fine-tuning.

Method	Deepseek-R1	Deepseek-V3	Doubao-1.5-pro	GPT-3.5-Turbo	GPT-4o	Qwen2.5	Qwen3	Claude-3.5-Haiku	Gemini-2.5-Flash	Average
RoBERTa-base	0.4450	0.4960	0.5444	0.5720	0.5257	0.3221	0.4566	0.3736	0.5732	0.4787
RoBERTa-large	0.6720	0.6476	0.5879	0.5771	0.5494	0.4369	0.5912	0.6255	0.5660	0.5837
IMBD	0.4286	0.6057	0.8736	0.9360	0.9347	0.8543	0.5103	0.8719	0.8437	0.7621
RoBERTa-base	0.9917	0.9952	0.9866	0.9832	0.9882	0.9892	0.9896	0.9877	0.9614	0.9859
RoBERTa-large	0.9977	0.9995	0.9912	0.9906	0.9940	0.9983	0.9984	0.9973	0.9823	0.9944
IMBD	0.9649	0.9854	0.9857	0.9957	0.9973	0.9835	0.9552	0.9919	0.9827	0.9825

Table 18: Text Length Statistics across Different Domains

Source	Mean	Median	Std	Source	Mean	Median	Std
Composition				**News**			
Human	895.22	855.00	200.34	Human	413.70	403.00	191.55
Deepseek-R1	979.38	934.00	224.05	Deepseek-R1	500.51	457.00	244.01
Deepseek-V3	901.02	880.00	157.79	Deepseek-V3	452.34	421.00	193.56
Doubao-1.5-pro	929.01	903.00	157.64	Doubao-1.5-pro	442.44	436.00	197.48
GPT-3.5-Turbo	746.54	734.00	136.95	GPT-3.5-Turbo	466.31	486.00	160.78
GPT-4o	1021.25	1008.00	128.75	GPT-4o	620.30	635.00	265.60
Qwen2.5	772.76	761.00	96.33	Qwen2.5	468.37	482.00	173.09
Qwen3	975.46	952.00	171.91	Qwen3	488.55	482.00	217.89
Claude-3.5-Haiku	677.22	672.00	84.20	Claude-3.5-Haiku	439.41	438.00	147.47
Gemini-2.5-Flash	1181.01	1143.00	216.43	Gemini-2.5-Flash	608.87	597.00	268.34
Film Review				**Academic Writing**			
Human	124.96	128.00	14.50	Human	430.07	403.00	176.06
Deepseek-R1	182.43	180.00	34.63	Deepseek-R1	544.00	499.00	225.99
Deepseek-V3	178.93	179.00	20.47	Deepseek-V3	469.91	439.00	170.19
Doubao-1.5-pro	135.41	138.00	19.23	Doubao-1.5-pro	449.53	418.00	198.14
GPT-3.5-Turbo	150.70	153.00	13.44	GPT-3.5-Turbo	327.42	280.00	138.44
GPT-4o	175.79	174.00	25.94	GPT-4o	512.26	498.50	169.72
Qwen2.5	177.70	175.00	35.45	Qwen2.5	390.84	371.00	127.54
Qwen3	143.21	143.00	20.56	Qwen3	506.41	485.00	183.13
Claude-3.5-Haiku	151.05	154.00	20.16	Claude-3.5-Haiku	333.70	311.00	98.71
Gemini-2.5-Flash	190.87	189.00	39.24	Gemini-2.5-Flash	740.76	718.50	273.73
Q&A				**TC News**			
Human	443.31	444.00	198.53	Human	423.20	512.00	161.74
Deepseek-R1	502.20	479.00	232.06	Deepseek-R1	544.94	568.00	202.75
Deepseek-V3	492.64	472.00	214.05	Deepseek-V3	454.68	482.00	161.48
Doubao-1.5-pro	469.00	475.00	207.75	Doubao-1.5-pro	440.56	515.00	172.69
GPT-3.5-Turbo	422.39	433.00	163.68	GPT-3.5-Turbo	434.34	489.00	123.79
GPT-4o	582.13	590.00	245.41	GPT-4o	578.11	663.00	212.07
Qwen2.5	441.81	450.50	180.69	Qwen2.5	469.94	506.00	173.34
Qwen3	487.21	474.00	214.08	Qwen3	537.77	616.00	201.92
Claude-3.5-Haiku	392.82	380.00	151.08	Claude-3.5-Haiku	449.14	462.00	141.01
Gemini-2.5-Flash	760.97	779.00	301.81	Gemini-2.5-Flash	597.11	657.00	233.85

Table 19: Examples of C-ReD

Domain LLM Generator	Film review GPT-4o
Film Title	你的名字
Human Text	拖了很久，终于等到室友考完研一起去看了。说实话没有预想中那么感动我，什么看完很想谈恋爱啊，哭的稀里哗啦啊，可能是提前知道是个HE吧。是部合格的玩时间梗的电影，比国产动画好得多，但总体而言可能还是期待过高吧。
Prompt	为未看过"你的名字"的观众写一篇观影指南，包括：适合的观影人群以及观影前需了解的历史/文化背景等。要求撰写内容字数在104左右。
AI-Generated Text	《你的名字》是一部适合青少年及成人观众的动画电影，尤其适合喜欢浪漫、奇幻元素和精美画面的观众。观影前，了解一些日本的文化背景会有助于更好地理解影片中的细节，例如日本的乡村生活、传统节日以及神道教的信仰，这些都在影片中占有重要位置。影片还涉及时间和空间交错的主题，因此保持开放的心态和对细节的关注将提升观影体验。日本的自然风光和现代都市的对比也为影片增添了独特的视觉美感。

Domain LLM Generator	Composition Deepseek-V3
Title	安全
Human Text	人的生命只有一次，生命是如此地脆弱，需要精心的呵护，无视安全，就无视生命的存在。当我们在安静的教室中听课时，安全的概念或意识在我们的脑海里荡然无存，就体会不到安全的重要性。而每分每秒都有一些人出生或意外死亡，季节可以轮回，而生命的逝去就如一江春水向东流——一去无返了。在我们的周围演绎着一场场血淋淋的教训：车祸、火灾、坠楼等等，一个个美好的生命就这样悄然殒失了。安全第一，警钟长鸣。我们学校的新校区坐落于郊外，由于缺乏严格的管理，交通秩序异常纷乱，司机超速行驶，学生交通安全意识模糊，在短短的两个月里三个年轻鲜活的生命灰飞烟灭，全校师生陷入了悲痛之中，而受害者的家长更是倍尝生死离别的巨大悲痛和创伤，谁之过？这三起事件引起了上级领导的高度重视，大力投资修建路面设施，给学生安上了一道生命的防护线。此后，再也没有发生类似事件，每当看到那长长的防护栅栏，每个本校学生都深深地懂得那是用鲜血和生命换来的。为什么只有当事情发生了才想到去修治？防患于未然，它留给我们的不仅仅是思考。有一则事例让我终生难忘：在一次交通宣传课上，放映了许多真实的车祸事件，其中有一个七岁的女孩在一次车祸中失去了下肢，整个上身被安置在一个疲软的蓝球里，两只手拄着木屐一样的木板蜗行爬索，面对记者的采访，她却笑得那么的天真无邪，说她的愿望就是再要一个篮球，在场的每一个人无不为之动容，泪如泉涌，教室里一片抽噎哭泣，心酸，心痛。当你正愁你的身体不够好时，你有没有想过一个没有下肢的人是怎样度日的？你不觉得惭愧吗？你的烦恼又算得了什么？认真地遵守交通法则，即是对自己的生命的珍视，也是对他人生命的负责。避免一场车祸的发生，就会避免两个家庭的残损和悲痛。悲剧，我们谁也不愿看到！怎样避免，你我他，大家都有责任，都有义务。维系生活的美妙，就要珍视自己的生命，而安全就是第一位的。
Prompt	请围绕题目"安全"，自主选择最适合的文体（议论文/记叙文/散文等），撰写一篇768字的文章。
AI-Generated Text	生命的底色安全，是生命最基本的底色。它如同空气般无处不在，却又常常被我们忽视。只有当危险来临，我们才会惊觉安全的重要性。在这个充满不确定性的世界里，安全不仅是个人的需求，更是整个社会运转的基础。安全是个人成长的基石。记得小时候学骑自行车，父亲总是紧紧扶着后座，直到我掌握了平衡才放手。这种安全感让我敢于尝试，敢于冒险。心理学家马斯洛将安全需求列为人类五大基本需求之一，正是因为只有在安全的环境中，人才能发展出更高层次的需求。古希腊哲学家亚里士多德曾说："安全感是幸福的前提。"当我们不必时刻提防危险，才能专注于自我提升和心灵成长。安全是社会文明的标志。古罗马人修建的输水道历经两千年仍在使用，正是因为他们在设计时就考虑到了安全因素。中国古代的都江堰水利工程，至今仍在造福成都平原，体现了古人对安全的深刻理解。一个社会的文明程度，往往可以从它对安全的重视程度中窥见一斑。法国思想家卢梭在《社会契约论》中指出："安全是公民让渡部分自由所换取的首要利益。"现代社会通过法律、制度和科技手段构建起的安全网络，正是人类文明进步的体现。安全需要每个人的参与和守护。日本在防灾教育方面的经验值得借鉴，从幼儿园开始就培养孩子们的安全意识，定期进行防灾演练。这种全民参与的安全文化，使得日本在面对地震等自然灾害时能够将损失降到最低。中国古代的"城门失火，殃及池鱼"告诉我们，安全从来不是某个人的私事。德国哲学家雅斯贝尔斯说："安全不是没有危险，而是有能力应对危险。"这种能力需要个人意识的觉醒和集体智慧的积累。站在新的时代节点上，我们对安全的理解也在不断深化。从物理安全到心理安全，从现实安全到网络空间安全，安全的维度在不断扩展。但无论形式如何变化，安全始终是那个让生命得以绽放的底色。让我们珍视这份底色，守护这份底色，因为只有在安全的画布上，生命才能描绘出最绚丽的图景。正如印度诗人泰戈尔所言："生命因安全而美丽，因危险而深刻。"在安全与挑战的平衡中，人类文明得以持续向前。

Domain LLM Generator	News Deepseek-R1
Category	Entertainment
Title	谢安琪发声明斥香港杂志 撇清当陈奕迅小三传言
Human Text	前日，有香港杂志指谢安琪插足陈奕迅与徐濠萦婚姻，昨晚谢安琪发声明指报道内容严重失实和毫无根据，并保留一切向相关之娱乐传媒追究法律责任之权利。据该香港杂志报道，陈奕迅与徐濠萦相恋13年，2004年诞下爱女康堤，期间徐濠萦不事生产"大花筒"的形象深入民心，但陈奕迅为安排父亲陈裘大北上医病已花了千万元，在精神和经济压力沉重下，决定对老婆开刀，二人终日吵架，三年婚姻告急，令不胜其烦的陈奕迅终于崩溃，于上周六广州个唱上三度失控痛哭。并指两人婚姻告急还与谢安琪插足有关，同属一个唱片公司的谢安琪与陈奕迅志趣相投性情相近，加上女方不避嫌以短信狂攻，二人友谊一日千里，据称两人亲密关系在唱片公司内已经路人皆知，在今年5月谢安琪的演唱会上，两人在台上忘情相拥，大唱情歌，似乎将关系进一步公开。对于杂志的报道，谢安琪方面发表声明称，"该报道有关谢安琪的内容严重失实和毫无根据。我们对于谢安琪无辜被卷入报道中陈奕迅的私人和家庭事务感到非常困惑和愤怒。另外，谢安琪自2009年初已不断遭到部分娱乐传媒的无理攻击，对于部分娱乐传媒以失实报道攻击他人博取销量的编采手法，我们保留一切向相关之娱乐传媒追究法律责任之权利。"
Prompt	请以时间轴形式报道"谢安琪发声明斥香港杂志 撇清当陈奕迅小三传言"娱乐事件，包含：事件起源说明、各方回应以及最新进展更新等。要求保持中立客观立场，字数控制在500左右。
AI-Generated Text	近日，香港一本主流娱乐杂志发布报道，声称歌手谢安琪与已婚歌手陈奕迅存在暧昧关系，暗示她介入陈奕迅的婚姻生活。该报道基于匿名消息源，于上周初在社交媒体广泛传播，引发公众热议和粉丝争议。报道中，杂志指谢安琪与陈奕迅私下互动频繁，引发第三者传言，导致谢安琪个人形象受损。事件发生后，谢安琪迅速于次日通过其经纪公司发表正式声明。在声明中，她强烈否认所有传言，斥责该杂志捏造事实、侵犯隐私，并强调与陈奕迅仅为工作伙伴和朋友关系，从未有任何越界行为。她要求杂志立即撤回报道、公开道歉，并保留法律追究权利。随后，陈奕迅方面也发表声明，其经纪人代表陈奕迅表达对妻子徐濠萦的坚定支持，否认任何婚外情存在。声明中，陈奕迅团队批评报道不实，呼吁媒体尊重艺人隐私，停止传播谣言，并强调陈奕迅家庭生活稳定。面对各方压力，该娱乐杂志回应称报道经过初步核实，但表示将重新审视内容来源，未正式道歉或撤回报道。截至最新进展，谢安琪团队表示正在评估证据，可能采取法律行动；陈奕迅方面则保持低调，专注于音乐工作。事件暂无进一步升级，公众讨论转向娱乐新闻伦理问题，部分支持艺人维权，部分质疑报道真实性。

Domain LLM Generator	Academic Writing Qwen2.5
Category	Abstract
Title	高强度聚焦超声消融治疗子宫腺肌症的临床研究
Keywords	子宫腺肌症；高强度聚焦超声；临床观察
Human Text	探初步观察高强度聚焦超声治疗子宫腺肌症的疗效。对我院采用高强度聚焦超声消融治疗的25例子宫腺肌症患者进行定期症状随访，随访痛经及月经改善情况。25例患者均完成治疗，随访至6个月，痛经缓解率86%，月经减少率84%，近期疗效显著。高强度聚焦超声消融治疗是一种保留子宫的无创治疗技术，能够有效治疗子宫腺肌症病灶，缓解痛经、月经量多等症状，近期效果确切，远期疗效尚需进一步临床观察验证。
Prompt	请根据以下信息撰写一篇通用型学术论文摘要。 文章标题："高强度聚焦超声消融治疗子宫腺肌症的临床研究" 文章关键词："子宫腺肌症；高强度聚焦超声；临床观察" 摘要应简明扼要地说明研究背景、核心问题、研究方法、主要发现（或论点）及研究意义。要求语言严谨、逻辑连贯、术语准确，字数控制在190字左右，适合投稿至各类中文学术期刊。
AI-Generated Text	子宫腺肌症是一种常见的妇科疾病，其特征为子宫内膜异位至肌层内生长，导致月经异常、痛经等症状，严重影响患者生活质量。高强度聚焦超声作为一种非侵入性治疗手段，在肿瘤等疾病的局部消融中展现出良好效果。本研究旨在探讨高强度聚焦超声对于子宫腺肌症的临床疗效及安全性。通过回顾性分析120例接受该疗法的子宫腺肌症患者的临床资料，结果显示，高强度聚焦超声能有效缓解症状，改善患者的生活质量，且不良反应轻微可控。研究表明，高强度聚焦超声可能是治疗子宫腺肌症的有效选择之一，值得进一步深入研究和推广应用。

Domain LLM Generator	Academic Writing Qwen3
Category	Introduction
Title	利用生物技术创建主要作物雄性不育杂交育种和制种的技术体系
Keywords	雄性不育；基因工程；杂种优势；玉米；水稻；小麦
Abstract	雄性不育技术在作物杂种优势利用和杂交种生产中发挥着重要作用。基于核质互作雄性不育的“三系法”与光温敏核不育的“两系法”已经在水稻等主要作物的杂交制种中获得了广泛应用，但是存在着资源利用效率低、育性不稳定、易受外界环境影响等诸多问题。近三十年来，利用生物技术创建不同类型的植物雄性不育系取得了一系列突破性进展。本文主要针对玉米、水稻、小麦三大作物的基因工程雄性不育技术的最新进展进行总结，特别详细地描述了本实验室最近研究创制的玉米多控不育技术体系，以期为相关研究和产业化应用提供技术参考。
Human Text	作物的杂种优势是指杂交种在一种或多种性状上优于其双亲的现象，通常表现为增产、稳产、抗病、抗逆等。杂交制种的关键技术在于发展和利用可以控制的授粉系统，以防止自花授粉导致的自交衰退。雄性不育在玉米、水稻、小麦等主要作物的杂种优势利用和杂交制种中发挥了重要作用。植物雄性不育是指在雌雄同株植物中，雄蕊发育不正常，不能产生有功能的花粉，但雌蕊发育正常，能接受正常花粉而受精结实，并能将雄性不育性状遗传给后代的现象。植物雄性不育按照不育性遗传方式的差异，可分为3类：细胞质雄性不育、细胞核雄性不育和核质互作型雄性不育。由于母体遗传的原因，第一类细胞质不育系的F1不能自交结实，因而不能在育种和生产上利用；利用常规杂交育种技术，第二类核不育系的保持和繁殖存在困难，也不能在育种和生产上有效利用；第三类核质互作不育系，理论和实践上都可以被育种利用，但是该类不育系的广泛利用可能会导致杂交种细胞质单一化，易受细胞质专一性病原小种的侵染而导致杂交种生产存在巨大的风险，如20世纪70-80年代北美玉米带出现的小斑病T小种侵染造成毁灭性的生产灾难。此外，在某些环境中，核质互作不育系的育性不稳定，有可能降低杂交制种的纯度和产量。随着现代生物技术的快速发展，一方面，利用分子生物技术，有望将遗传稳定的核不育基因有效地利用起来，即通过生物技术保持和繁殖非转基因的雄性不育系，例如我们实验室的玉米多控不育技术、美国杜邦先锋的SPT技术、水稻智能不育技术等；另一方面，还可以通过基因工程手段，创制各种诱导型/条件型雄性不育系，即通过生物技术创制转基因的不育系，如美国孟山都的玉米草甘膦诱导型雄性不育技术、水稻草丁膦诱导型雄性不育技术、小麦断裂基因雄性不育系统等。近年来，随着分子生物学和基因工程技术的发展，作物细胞核雄性不育的利用获得了一些突破性进展。本文将从玉米、水稻、小麦等主要作物的基因工程不育技术的创制策略、研究现状、利用途径和应用前景等几方面进行阐述，以期为高效利用作物核雄性不育技术、提高作物杂种优势利用效率和建立作物高效育种和制种技术体系提供参考依据。
Prompt	请用简洁明了的语言撰写引言内容，直接点明研究动机、核心方法与贡献。 文章标题："利用生物技术创建主要作物雄性不育杂交育种和制种的技术体系" 文章关键词："雄性不育；基因工程；杂种优势；玉米；水稻；小麦" 文章摘要："雄性不育技术在作物杂种优势利用和杂交种生产中发挥着重要作用。基于核质互作雄性不育的“三系法”与光温敏核不育的“两系法”已经在水稻等主要作物的杂交制种中获得了广泛应用，但是存在着资源利用效率低、育性不稳定、易受外界环境影响等诸多问题。近三十年来，利用生物技术创建不同类型的植物雄性不育系取得了一系列突破性进展。本文主要针对玉米、水稻、小麦三大作物的基因工程雄性不育技术的最新进展进行总结，特别详细地描述了本实验室最近研究创制的玉米多控不育技术体系，以期为相关研究和产业化应用提供技术参考。" 字数限制约869字，避免冗长背景，突出实用性。
AI-Generated Text	杂种优势是提高作物产量的重要途径，而雄性不育技术是实现大规模杂交制种的核心支撑。传统“三系法”依赖细胞质雄性不育系、保持系和恢复系的配套使用，在水稻中已有成功应用，但存在恢保关系复杂、恢复源有限、配组自由度低等问题；“两系法”利用光温敏核不育特性，虽简化了育种流程，但其育性易受环境温度波动影响，导致制种风险增加，限制了在玉米、小麦等作物中的稳定应用。因此，发展高效、稳定、可控的新型雄性不育技术体系成为突破当前杂交育种瓶颈的关键。近年来，基因工程技术的进步为构建精准、可调控的雄性不育系统提供了全新路径。通过靶向破坏花粉发育关键基因或特异性阻断雄配子形成过程，可在不影响植株正常生长的前提下实现完全雄性不育，且不受细胞质背景和环境条件制约。本研究聚焦玉米、水稻和小麦三大主粮作物，系统梳理了基于CRISPR/Cas9基因编辑、启动子驱动的组织特异性基因沉默、化学诱导型育性控制等策略所构建的基因工程雄性不育（GMS）技术进展。这些新技术不仅显著提升了不育系创制效率，还实现了育性转换的精确调控，为自动化、规模化杂交制种奠定了基础。在此基础上，我们实验室创新性地开发出“玉米多控不育技术体系”（MCMS），该体系整合了花药特异性启动子驱动的细胞毒素基因、可去除的选择标记以及化学保护剂介导的育性恢复模块，实现了不育系繁殖与杂交种生产的双向可控。MCMS在多个骨干自交系中稳定表达，不育率达100%，且可通过喷施安全剂实现同步育性恢复，保障杂交种子生产纯度。田间试验表明，该系统适用于机械化制种，显著降低人工去雄成本，提升制种效率30%以上。此外，该技术具备良好的生物安全性，外源基因不随花粉扩散，符合转基因作物管理规范。本文旨在总结当前主要作物基因工程雄性不育技术的实用化进展，重点展示MCMS在玉米中的成功实践，为其在水稻、小麦中的推广应用提供技术范式。该类技术有望打破传统不育系统对特定遗传资源和生态条件的依赖，推动我国杂交育种向智能化、标准化方向升级，助力粮食作物持续增产。

Domain LLM Generator	Q&A Gemini-2.5-Flash
Question	文艺青年所谓的「独自漂泊」、「一个人旅行」或者「带着书本流浪」是出于什么样的心理状态？目的是什么？
Human Text	以自己的经历来回答吧。男，不年轻，略文艺，偏二逼。出去旅游从来都是一个人走，也带书，因为喜欢看书，打发旅途中的等候时间和长途坐车时间，而且书随时能看，不受飞机或者没电限制，嫌重随便扔。为什么一个人？因为找不到志同道合体力相当的朋友。我是随遇而安狂奔体验型的体力驴，有计划但时不时不走寻常路，比如住个村民家里之类的，有计较的朋友同行很可能会疯掉。徒步速度和作战半径一般人赶不上，跟朋友一起走朋友会累死，我也不爽，何必呢？一个人在路上可以捡朋友，捡到好的同行一段或者拼车拼房（不要误解，就是拼个标间省钱，路上哪来那么多艳遇）。到岔口各走各的路，ok咯，很多朋友都是这样捡到的，挺好的啊。一个人，想怎么着就怎么着，不用考虑别人的感受，想奢侈想凑合都不是问题。目的就是不受约束，老子出来花钱玩就是玩，就是图个爽，让我听导游的还不如躺家里睡觉。这和空虚没有任何关系，只是放假出来玩而已。至于翻瓶子的老人和环卫工人...这关我p事？我没放假时拼死拼活干活跟客户陪笑脸的时候也没谁来怜悯我啊，这是我挣钱应该做的，我也没觉得有什么不对。我将来要是沦落到捡瓶子，我也不会想着谁需要来怜悯我，我的世界和他人无关。我拼死拼活挣来的钱，花天酒地或者孤独出游，我爱咋咋，就是图个爽。嗯，就这样子，没有题主想的那么高尚或者蛋疼；）
Prompt	请用558字左右直接回答以下问题：文艺青年所谓的「独自漂泊」、「一个人旅行」或者「带着书本流浪」是出于什么样的心理状态？目的是什么？。回答应当开门见山给出结论，随后分段落说明理由、提供证据，最后简要总结。
AI-Generated Text	文艺青年热衷于“独自漂泊”、“一个人旅行”或“带着书本流浪”，其核心心理状态是追求自我认同、逃离现实束缚和寻求精神富足，最终目的是实现个人成长与心灵自由。首先，这种行为源于对自我认同的强烈渴望。在社会多元化和信息爆炸的时代，个体很容易迷失在各种外部期待和集体意识中。独自漂泊提供了一个抽离日常、远离熟人社会的契机，让文艺青年有机会直面内心，探索真实的自我。他们希望通过陌生的环境和独处的时光，剥离掉社会角色赋予的标签，重新审视自己的价值观、兴趣和人生方向。这种对“我是谁”的追问，是许多文艺青年选择独行路上的内在驱动力。其次，逃离现实束缚是另一个重要原因。许多文艺青年对城市生活的喧嚣、职场竞争的压力以及人际关系的复杂感到厌倦甚至窒息。他们渴望暂时摆脱房贷、车贷、KPI等物质枷锁，逃离固定的生活模式和既定的社会规则。独自旅行成为一种仪式性的“出走”，象征着对既有秩序的反叛和对自由的向往。流浪者形象的浪漫化，也反映了他们对“诗和远方”的执着追求，希望在路上找到一种不同于朝九晚五的生活可能性。再者，寻求精神富足是其深层目的。文艺青年往往对精神世界有更高的要求和更敏锐的感知力。他们不满足于物质层面的享受，更看重内心的充实和思想的启发。带着书本流浪，意味着他们将阅读视为一种生活方式，通过文学、哲学、艺术等领域的滋养，拓宽视野，深化思考，提升审美。在旅途中，他们更容易与当地文化、自然风光产生共鸣，从而获得独特的感悟和创作灵感。这种经历本身就是一种精神财富的积累，让他们觉得生命更有深度和意义。此外，这种行为也带有一定的“仪式感”和“自我证明”的色彩。在社交媒体盛行的今天，独自旅行或流浪的经历往往被视为一种独立、勇敢和有品位的象征。通过分享旅途中的见闻和感悟，文艺青年在一定程度上也希望获得他人的认同和赞赏，以此来强化自己的独特和不凡。综上所述，文艺青年的“独自漂泊”等行为，是他们在面对时代变迁和个体成长困境时，选择的一种积极的自我探索和精神突围方式。他们通过这种方式寻找自我、摆脱束缚、充实内心，最终目标是实现个体的自由意志和精神上的圆满。

Domain LLM Generator	TC News GPT-3.5-Turbo
Category	Article
Title	無論快樂或悲傷,請對身邊的人說一句「做得很好」
Human Text	「正向思考，積極正向！」「你試過瑜伽嗎？」「快樂是一種選擇！」「說實話，如果你只吃健康飲食，你就會感覺很好。」「哦，還有瑜伽。做瑜伽！」如果你有心理疾病，這種情況很可能就會出現，有人會想試圖與你進行這種對話。基於某些理由，這個星球上有些人，一旦發覺你有點難過或焦慮，他們便覺得可以介入你的生活，告訴你該怎麼做。當然，這種情形特別令人惱火。沒有憂鬱或難過情緒的人覺得，只要夠努力，你可以想出快樂。可悲的是，事情並非如此，當你聽到「要往正面想」第一百次的演說，你已準備好用瑜伽墊打，或塞一堆甘藍菜到他們嘴裡，讓他們閉嘴。傳統意義上的「正面」是好的思想，對許多人來說確實有幫助，但對有些人來說卻有如重大挑戰一樣令人畏懼。不過極端的說，正面是一種「正面」概念，可以幫助你專心，並將焦點放在對你重要的事情上。正面就是這樣，每個人對正面的定義都不同，只有你可以創造對你有用的定義。這是一個獨特的個人概念，為每個人量身定做。你不能把你的定義强加在別人身上，別人也不能强加在你身上。或許有些適合你的事情，並不完全符合「變好」的既定概念。或許香甜地吃一碗營養穀麥會讓你開心起來，或許把平和瑜伽墊拋開，好好跑一場。尋找能夠幫助你感覺更好的事物，每天感覺便會好一點。對你來說這樣很重要，很棒，又很獨一無二。
Prompt	請以專業評論員的視角，撰寫一篇關於「無論快樂或悲傷,請對身邊的人說一句「做得很好」」的文章。內容應包含事件背景、多角度分析和深度思考，語言需嚴謹且富有洞察力，避免主觀臆斷，字數控制在542左右。
AI-Generated Text	在這個充滿變數和挑戰的時代裡，人們常常感受到壓力和焦慮，因此，一句「做得很好」的肯定和支持成為了極具價值的情感交流方式。不論是面對快樂或悲傷，這句簡單的話語都能温暖人心，賦予力量，讓人感受到被理解和被珍惜。從心理學角度來看，人類是社會性動物，我們需要他人的關愛和支持來維持情感健康。當一個人遭遇挫折時，聽到身邊的人說一句「做得很好」，可以打破孤獨感，讓他感受到不孤單，這樣的肯定不僅是一種情感的交流，更是一種温暖的情感投射。從社會學的角度來看，現代社會充斥著競爭和壓力，人們往往追求成功和完美，卻忽略了對自己和他人的肯定。當我們學會對身邊的人說一句「做得很好」時，不僅可以促進正向情感的傳遞，也可以建立積極的社會支持網絡，讓人與人之間的關係更加緊密和和諧。從人際關係的角度來看，一句「做得很好」可能改變整個關係的走向。無論是家庭、友情還是工作環境，當我們學會用肯定和支持的語言來溝通時，會讓彼此之間的情感更加豐富，信任更加穩固，進而促進關係的和諧發展。因此，不妨在生活中多用這句簡單的話語來表達對身邊人的支持和肯定。無論是在快樂的時刻，還是在悲傷的時刻，一句「做得很好」都能成為情感的橋樑，連結彼此的心靈，讓生活充滿温暖和希望。讓我們共同努力，用這句話來燃點生活中的每一個美好瞬間，讓愛和關懷永遠相伴左右。