Human or Machine? Contrastive Learning for Detecting AI-Generated Chinese E-Commerce Reviews with a Custom Dataset

Anonymous ACL submission

Abstract

AI-generated content proliferation in Chinese e-commerce platforms faces challenges in integrity and consumer trust. While existing detection methods show promising performance within specific domains, their cross-domain robustness remains largely unexplored for Chinese e-commerce reviews. We present the first systematic cross-domain robustness evaluation for Chinese AI-generated text detection, constructing a high-fidelity benchmark dataset using a systematic data generation approach and developing a progressive out-of-distribution evaluation framework. Through extensive experiments across multiple detection approaches, we provide systematic analysis of cross-dataset generalization patterns. Our evaluation reveals that fine-tuned large language models, particularly Qwen-2.5-7B, achieve superior performance across all scenarios (94.8% F1-score in-domain, 63.9% in extreme cross-domain conditions), while contrastive learning approaches show significant performance degradation under distribution shifts (F1-score declining from 86.8% to 35.1%). These findings provide crucial insights into detection paradigm trade-offs and cross-domain robustness challenges in practical deployment.

1 Introduction

In 1950, Alan Turing proposed a test to evaluate machine intelligence: if a human evaluator could not distinguish between responses from a machine and a human, the machine could be considered intelligent (Turing, 1950). Today, as large language models (LLMs) achieve unprecedented fluency, we face the inverse challenge: distinguishing AI-generated content from human writing has become increasingly difficult, raising critical concerns about the integrity of digital information.

The advancement of modern LLMs, from GPT-3's few-shot learning capabilities (Brown et al., 2020) to GPT-4's enhanced reasoning (OpenAI, 2023), has fundamentally transformed content creation capabilities, generating text that closely mirrors human writing in coherence, contextual relevance, and stylistic nuance. As recent comprehensive evaluations demonstrate (Chang et al., 2024), this progress presents significant challenges for AI-generated text detection, where existing methods suffer from poor generalization ability when deployed across different domains, models, or text styles. Recent work has highlighted these fundamental limitations (Doughman et al., 2025), showing that traditional approaches relying heavily on binary classification frameworks often overfit to specific training distributions and fail when confronted with out-of-distribution scenarios, as the detection task extends far beyond simple "AI vs humans" classification (Ji et al., 2024).

The proliferation of AI-generated content poses particular risks in Chinese e-commerce platforms, where authenticity directly impacts consumer trust and market fairness, as consumer reviews significantly influence purchasing decisions through their linguistic and emotional expression patterns (Kronrod and Danziger, 2013). Chinese e-commerce reviews present unique challenges including complex linguistic characteristics (emojis, internet slang, dialectal variations), text brevity (100-500 characters), and domain-specific terminology that differ substantially from general text domains. Current detection methods inadequately address these characteristics, and existing datasets primarily focus on English academic or news domains. While datasets like ASAP (Bu et al., 2021) (46,730 Chinese restaurant reviews with aspect sentiments) provide valuable Chinese review resources for sentiment analysis, they lack the AI-generated counterparts necessary for detection research, leaving a significant gap for Chinese e-commerce applications.

Most importantly, the literature lacks systematic evaluation of cross-dataset generalization capabilities for Chinese AI text detection. While benchmarks like MGTBench (He et al., 2023) and M4GT-Bench (Wang et al., 2024) have established evaluation frameworks for AI-generated text detection, and recent work like MAGE (Li et al., 2024) has highlighted the critical importance of out-of-distribution robustness in detection systems, their focus remains primarily on English domains. The robustness of detection approaches under out-of-distribution conditions, particularly across different generation models, text styles, and domains, remains largely unexplored in the Chinese context.

To address these critical gaps, our work makes the following contributions:

- **First comprehensive Chinese e-commerce AI detection benchmark with systematic out-of-distribution (OOD) evaluation.** We construct the first high-fidelity Chinese e-commerce AI detection dataset using a systematic data generation approach, with systematic LLM evaluation establishing optimal generation models. Our multi-level OOD evaluation framework provides foundational infrastructure for fine-grained cross-dataset generalization analysis.

- **Systematic comparative evaluation revealing performance trade-offs across detection paradigms.** We provide the first systematic comparison of contemporary detection approaches including fine-tuned pre-trained language models (PLMs) and contrastive learning methods under progressive OOD scenarios. Our evaluation reveals that fine-tuned models, particularly Qwen-2.5-7B, achieve superior robustness across all scenarios, while contrastive learning approaches show significant limitations under extreme distribution shifts.

- **Empirical insights into cross-domain robustness patterns and practical deployment considerations.** Through comprehensive experiments, we establish performance benchmarks and identify fundamental challenges in cross-domain AI text detection for Chinese e-commerce applications, providing crucial insights for practical system deployment and future research directions.

2 Related Work

The detection of AI-generated text has evolved from early statistical methods to sophisticated deep learning approaches, with comprehensive surveys documenting this rapid progression (Wu et al., 2025). Current state-of-the-art methods primarily rely on fine-tuning pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for binary classification, with recent approaches integrating linguistic features to enhance classification performance (Yadav and M C, 2024). However, a fundamental challenge in this field is the poor out-of-distribution (OOD) generalization of detection systems. Models trained on specific domains or generation models often fail when tested on different distributions, as they learn spurious correlations rather than fundamental authorship markers (Wang et al., 2024). Recent studies have further highlighted the vulnerability of these detectors to adversarial perturbations (Huang et al., 2024), while promising work on restricted embeddings shows potential for improving robustness (Kuznetsov et al., 2024).

The generalization problem is particularly severe in AI text detection, where traditional detectors tend to overfit to specific training distributions and perform poorly when confronted with different generation models, text styles, or domains. Research on syntactic template detection has revealed that generated texts often exhibit systematic patterns that may not generalize across different contexts (Shaib et al., 2024). While benchmarks like M4GT-Bench (Wang et al., 2024) have highlighted these issues, and multilingual detection efforts have emerged (Agrahari et al., 2025), systematic evaluation frameworks for cross-domain robustness remain limited, especially for non-English languages.

Contrastive learning has shown promising results in various NLP tasks by learning robust representations through contrasting positive and negative sample pairs (Jaiswal et al., 2021). Recent work has explored contrastive learning for AI text detection, with DetectGPT (Mitchell et al., 2023) demonstrating zero-shot detection capabilities. Notably, He et al. (Guo et al., 2024) proposed DeTeCtive, a multi-level contrastive learning framework that argues the key to AI text detection lies in distinguishing writing styles rather than simple binary classification.

Despite significant attention to AI text detection, research specifically targeting Chinese text remains limited. The foundation for Chinese NLP has been strengthened by advances in pre-trained models like ERNIE (Sun et al., 2019) and Chinese BERT with whole word masking (Cui et al., 2021), yet their application to AI text detection remains underexplored. Chinese e-commerce reviews present unique challenges including complex morphological structures, widespread use of emojis and internet slang, and domain-specific terminology. While research on utilizing reviews for recommendation justification (Ni et al., 2019) demonstrates the importance of review authenticity in e-commerce contexts, existing datasets like HC3 (Guo et al., 2023) primarily focus on academic domains, leaving a significant gap for Chinese e-commerce applications that our work addresses.

3 Methodology

In this work, we systematically address the critical challenge of cross-domain robustness in Chinese AI-generated text detection. Our methodology comprises two principal contributions: (1) the construction of a high-fidelity benchmark dataset with systematic out-of-distribution evaluation framework, and (2) comprehensive comparative evaluation of multiple detection paradigms including fine-tuned models and contrastive learning approaches. We provide the first systematic evaluation of detection methods under progressive OOD conditions, enabling fine-grained analysis of generalization patterns in Chinese AI text detection.

3.1 Dataset Construction and Evaluation Framework

A key limitation of existing detection benchmarks is the "fidelity gap" that arises from using generic prompts, which fail to replicate the nuanced, aspect-driven nature of real-world product reviews. To overcome this, we developed a "controlled synthesis" pipeline to generate a parallel corpus where AI-generated reviews are thematically and sentimentally aligned with human-written counterparts, compelling the detector to learn stylistic rather than topic-based cues.

Realistic Prompt Generation.

The dataset construction process begins with the ASAP corpus (Bu et al., 2021), a large-scale collection of Chinese product reviews annotated with fine-grained aspect and sentiment labels. For each human-written review, the corresponding metadata, including overall star rating and aspect-level sentiment polarities (e.g., "Taste#Flavor," "Service#Queueing"), are extracted to form a comprehensive "emotional-semantic profile." These structured profiles are then transformed into diverse, natural-language prompts using a rule-based generator with multiple template variations, ensuring stylistic diversity and naturalness. This approach guarantees that the resulting AI-generated reviews are not only fluent but also maintain a semantically isomorphic relationship with their human counterparts in both topical coverage and sentiment distribution, thereby greatly reducing the adverse impact of structural and emotional imbalance on the AI-generated text detection task.

Algorithm 1: Realistic Prompt Generation for Controlled Synthesis

```
procedure GeneratePrompts($D_{human}$)
  Input: $D_{human}$, dataset of human reviews with aspect sentiments
for each review $R_i$ in $D_{human}$ do
  Step 1 - Create Profile:
  $P \leftarrow$ Extract aspect sentiments from $R_i$
  $key \leftarrow$ Sort and join elements of $P$
  $R_i.profile \leftarrow$ Simplify($key$)
  Step 2 - Generate Prompt:
  $details \leftarrow$ Parse($R_i.profile$)
  $T_{aspect}, T_{sent}, T_{other} \leftarrow$ Generate templates using $details$, $R_i.star$
  if $T_{aspect} \neq \emptyset$ and random() $< 0.7$ then
    $prompt \leftarrow$ RandomChoice($T_{aspect}$)
  else
    $prompt \leftarrow$ RandomChoice($T_{aspect} \cup T_{sent} \cup T_{other}$)
  end if
  $R_i.prompt \leftarrow prompt$
end for
return StratifiedSample($D_{human}$, size=1000)
```

Table 1: Realistic Prompt Generation Algorithm for Controlled Synthesis

Generation Model Selection and Chinese LLM Benchmark. To ensure the fidelity and diversity of the AI-generated corpus while minimizing model-specific artifacts, we conducted the first systematic evaluation of mainstream LLMs for Chinese e-commerce review generation. This benchmark evaluation represents a significant contribution, providing authoritative performance rankings of contemporary LLMs on Chinese text generation tasks.

Four state-of-the-art large language models (GPT-4o-mini, deepseek-chat-v3-0324, qwen-turbo, and gemini-2.0-flash) were each tasked with generating 1,000 reviews using a standardized prompt set. This systematic comparison provides the first comprehensive assessment of these models' capabilities in generating authentic Chinese e-commerce content.

The resulting corpora were evaluated with the metrics reported in Table 2: Self-BLEU (Alihos-

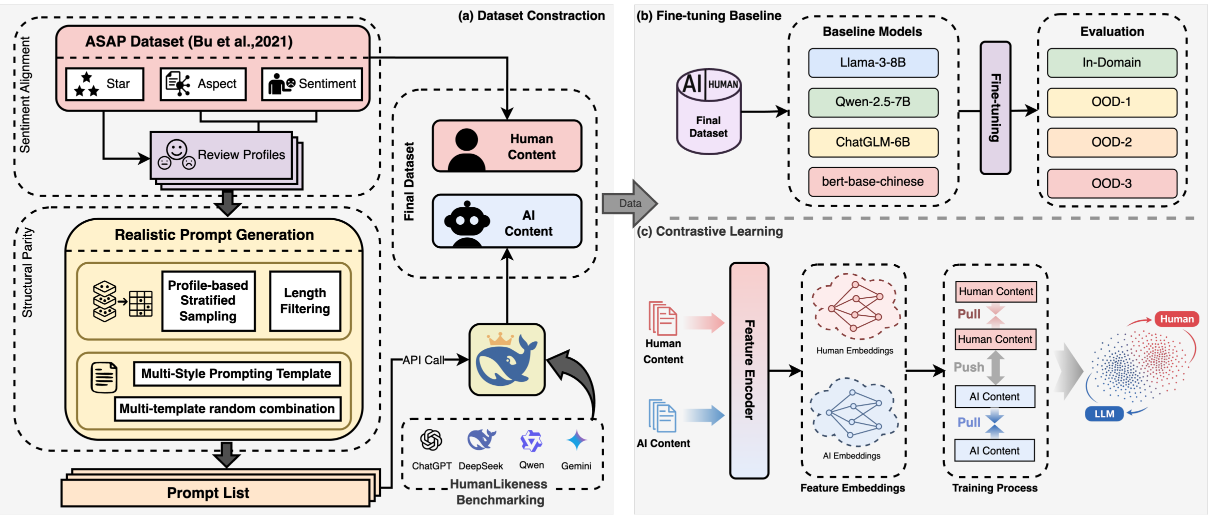


Figure 1: An overview of the Dataset Generation pipeline

Model	TTR	Avg Length	Self-BLEU	Self-BERTScore	JSD vs Human	Rank
Human (Reference)	0.0923	283.23	0.1214	0.0007	0.0000	–
deepseek-chat-v3-0324	**0.0376**	**202.64**	**0.5258**	**0.0074**	**0.4131**	**1**
qwen-turbo	0.0281	183.11	0.5932	0.0097	0.4188	2
gemini-2.0-flash-001	0.0285	102.40	0.5842	0.0100	0.4196	3
gpt-4o-mini	0.0216	157.39	0.6281	0.0146	0.4527	4

Table 2: Comprehensive benchmark evaluation of mainstream LLMs for Chinese e-commerce review generation.

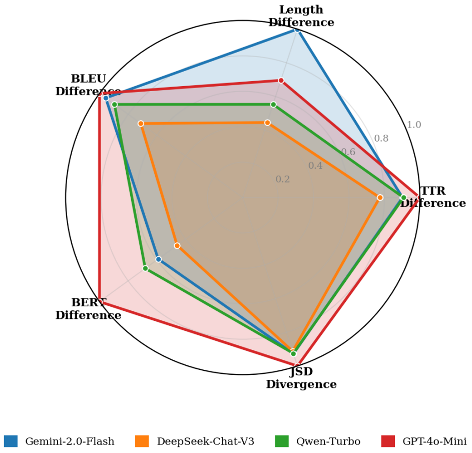


Figure 2: Model Performance Comparison (Lower values indicate closer similarity to human writing)

seini et al., 2019) (diversity), Type-Token Ratio (TTR, lexical richness), average review length, Self-BERTScore (Zhang et al., 2020) (semantic diversity), and Jensen-Shannon Divergence (JSD, distributional similarity to human writing). The detailed descriptions of these metrics are provided in Table 3.

As shown in Figure 2 and Table 2, deepseek-chat-v3-0324 consistently achieves the best overall balance across these criteria, producing text that most closely matches human review characteristics in both lexical and structural aspects.

Dataset Composition and Statistics. The final dataset comprises training, validation, and testing splits, plus three progressive out-of-distribution (OOD) test sets for systematic robustness evaluation. The training, validation, and standard test sets contain AI-generated texts produced by the selected model (DeepSeek-Chat-v3) for in-domain evaluation. The three OOD test sets (OOD-1, OOD-2, OOD-3) are designed to incrementally stress-test different aspects of model generalization: model variation, stylistic robustness, and domain transfer. Detailed statistics are provided in Table 4, with the progressive OOD evaluation design presented in Table 5.

The progressive nature of these test sets allows for fine-grained analysis of performance degradation patterns, enabling identification of specific vulnerabilities in detection models and providing insights into the relative importance of different robustness factors.

Metric	Description
TTR	Type-Token Ratio, measuring lexical richness. Higher values suggest more varied vocabulary.
Avg. Len.	Average review length (in characters). Closer to human indicates better fidelity.
Self-BLEU (n=4)	Measures intra-text similarity within the generated corpus. Lower scores indicate higher diversity.
Self-BERTScore	Semantic similarity among generated reviews. Lower values indicate higher semantic diversity.
JSD vs Human	Jensen-Shannon Divergence between the linguistic feature distributions of generated and human corpora. Lower values indicate higher fidelity.
Rank	Overall ranking based on weighted aggregation of all metrics.

Table 3: Evaluation metrics for the generation model selection.

Dataset	Human	Machine	Total
Training	8,000	8,000	16,000
Validation	1,000	1,000	2,000
Testing	1,000	1,000	2,000
Testing (OOD-1)	1,000	1,000	2,000
Testing (OOD-2)	1,000	1,000	2,000
Testing (OOD-3)	1,000	1,000	2,000
Overall	**13,000**	**13,000**	**26,000**

Table 4: Statistics of the constructed dataset.

3.2 Detection Framework and Baseline Methods

We evaluate multiple detection approaches to provide comprehensive comparative analysis of different paradigms and their relative performance under cross-domain conditions.

Baseline Detection Models. We compare against representative detection paradigms including fine-tuned pre-trained language models (PLMs) and contemporary large language models:

- **Fine-tuned PLMs:** We evaluate bert-base-chinese and chinese-roberta-wwm-ext, specifically designed for Chinese text understanding, as well as multilingual models including Llama-3-8B and Qwen-2.5-7B. These models represent the current state-of-the-art in supervised fine-tuning approaches for text classification.

- **Training Strategy:** All baseline models are fine-tuned using standard binary classification objectives on our training data, where AI-generated content is labeled as class 1 and human-written content as class 0. We ensure fair comparison by using consistent hyperparameters across different model architectures. Detailed training configurations are presented in Table 6.

Test Set	Distribution Shift Factors
In-Domain	**Baseline:** Same generation model (Deepseek-Chat-v3), domain (food reviews), and original text style as training data.
OOD-1	**+ Model Shift:** Multiple generation models (GPT-4.1, Gemini-2.5-Flash, Deepseek-Chat-v3, Qwen-Turbo) while maintaining food domain and original style.
OOD-2	**+ Style Shift:** OOD-1 conditions plus AI-based paraphrasing through the same models to introduce stylistic variations and increase detection difficulty.
OOD-3	**+ Domain Shift:** OOD-2 conditions plus cross-domain transfer from food reviews to general e-commerce categories (electronics, clothing, home products) using Amazon multilingual review corpus (Keung et al., 2020).

Table 5: Progressive out-of-distribution evaluation framework with cumulative distribution shift factors.

Parameter	Value
Learning Rate	2e-5
Batch Size	8
Max Epochs	5
Temperature τ	0.07
Contrastive Weight λ	0.1
Max Sequence Length	512
Optimizer	AdamW

Table 6: Training configuration for all models.

Contrastive Learning Framework. We employ a contrastive learning approach adapted from the DeTeCtive framework (Guo et al., 2024), using the `hfl/chinese-bert-wwm-ext` (Cui et al., 2021) model as our text encoder. This model is specifically chosen for its superior performance on Chinese text understanding tasks and its proven effectiveness in capturing nuanced linguistic patterns in Chinese content.

Following the multi-level contrastive learning paradigm, this approach provides an alternative perspective for distinguishing writing styles through representation learning rather than simple binary

Test Scenario	Metric	Llama-3-8B	Qwen-2.5-7B	BERT-Chinese	Chinese-RoBERTa	Contrastive Learning
In-Domain Test	Accuracy	80.65	**94.85**	88.15	85.40	86.95
	Precision	85.92	**94.32**	89.69	87.72	89.15
	Recall	80.40	**94.57**	88.15	85.40	86.95
	F1-Score	79.90	**94.84**	88.03	85.17	86.76
	AUC-ROC	0.926	**0.996**	0.961	0.954	0.943
OOD-1	Accuracy	79.40	**94.35**	83.90	80.05	80.10
	Precision	83.03	**94.69**	87.08	84.07	85.21
	Recall	80.40	**94.35**	83.90	80.05	80.10
	F1-Score	78.41	**94.33**	83.52	79.41	79.31
	AUC-ROC	0.919	**0.992**	0.934	0.932	0.911
OOD-2	Accuracy	58.85	**82.60**	68.95	63.85	64.80
	Precision	70.85	**85.15**	78.81	74.95	77.53
	Recall	58.85	**78.99**	68.95	63.85	64.80
	F1-Score	50.46	**82.07**	66.04	59.33	60.20
	AUC-ROC	0.819	**0.976**	0.860	0.850	0.805
OOD-3	Accuracy	53.45	**67.30**	58.80	50.10	50.80
	Precision	68.24	**74.98**	75.59	75.03	75.20
	Recall	53.45	**62.40**	58.80	50.10	50.80
	F1-Score	40.57	**63.90**	50.72	33.56	35.09
	AUC-ROC	0.739	**0.843**	0.807	0.713	0.748

Table 7: Comprehensive performance comparison across all evaluation metrics and test scenarios. Shows results for five models across four test datasets with five evaluation metrics. Bold indicates best performance for each metric and test scenario.

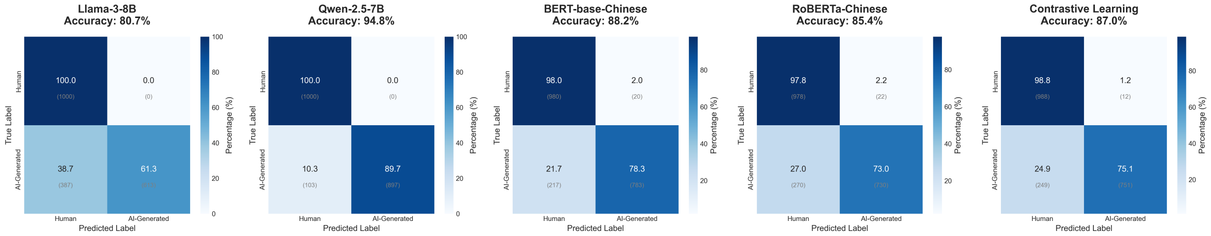


Figure 3: Confusion matrix comparison of all evaluated models (Llama-3-8B, Qwen-2.5-7B, bert-base-chinese, chinese-roberta-wwm-ext, Contrastive Learning) on the in-domain test set. Each matrix shows true/false positive and negative counts and percentages. Qwen-2.5-7B achieves the highest accuracy and lowest misclassification, while other models show varying error patterns.

classification. The contrastive learning objective is formulated as:

$$\mathcal{L}_{cl} = -\log \frac{\exp(\text{sim}(h_i, h_j^+)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(h_i, h_k)/\tau)} \quad (1)$$

where h_i is the representation of the anchor sample, h_j^+ is the positive sample representation, τ is the temperature parameter (set to 0.07), and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. We construct positive pairs from text samples of the same category (both human or both AI) and negative pairs from different categories.

The total training objective combines contrastive learning with cross-entropy classification:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cl} \quad (2)$$

where λ is a weighting parameter (set to 0.1) and $\mathcal{L}_{ce}$ is the standard cross-entropy loss. This approach provides an additional evaluation dimension by learning distinctive features through representation learning, offering insights into how different methodological approaches handle cross-domain generalization and stylistic pattern recognition.

4 Experiments

4.1 Experimental Design

We evaluate five detection approaches across our progressive OOD framework: Llama-3-8B, Qwen-2.5-7B, bert-base-chinese, chinese-roberta-wwm-ext, and our contrastive learning method. All experiments are conducted on Google Colab with hyperparameters kept as consistent as possible across different model architectures (Table 6). We report accuracy, precision, recall, F1-score, and AUC-ROC, with results averaged over 5 independent runs to ensure reliability.

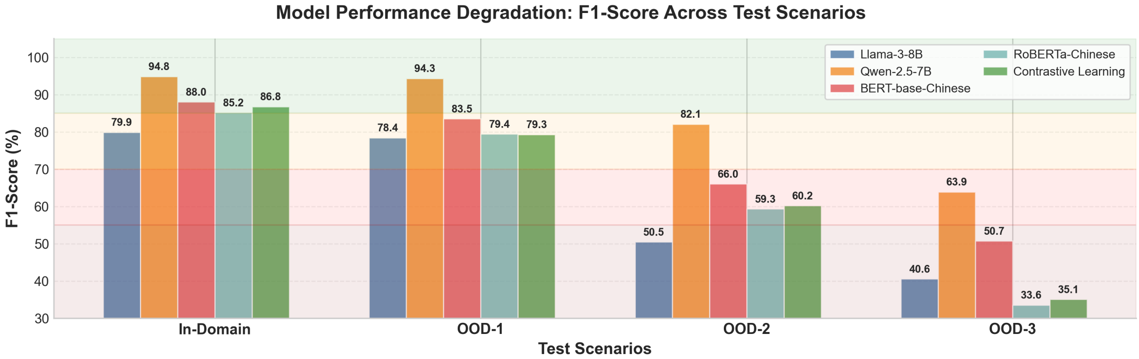


Figure 4: F1-score degradation across progressive out-of-distribution scenarios (In-Domain, OOD-1, OOD-2, OOD-3). Error bars represent standard deviation across 5 independent runs.

4.2 Results

Table 7 presents comprehensive results across all evaluation scenarios. The results demonstrate clear performance hierarchies across different approaches. Qwen-2.5-7B achieves the strongest overall performance, maintaining 94.8% F1-score in-domain and demonstrating superior cross-domain robustness with 63.9% F1-score in the most challenging OOD-3 scenario.

Our comparative analysis reveals distinct degradation patterns across approaches: fine-tuned models like Qwen-2.5-7B show moderate degradation (94.8% → 63.9%), while contrastive learning exhibits steeper decline (86.8% → 35.1%). bert-base-chinese and chinese-roberta-wwm-ext show intermediate patterns, dropping to 50.7% and 33.6% respectively in OOD-3 conditions. Statistical significance testing using paired t-tests across 5 independent runs confirms these performance differences ($p < 0.001$ for all pairwise comparisons between Qwen-2.5-7B and other methods in OOD scenarios).

As shown in Figure 4, this systematic evaluation reveals distinct degradation patterns, with the progression from OOD-1 (model variation) to OOD-2 (stylistic shifts) to OOD-3 (domain transfer) demonstrating the relative impact of each distribution shift factor on detection robustness. The confusion matrix analysis (Figure 3) confirms Qwen-2.5-7B's superior classification accuracy and lowest misclassification rates across all test scenarios.

4.3 Visualization Analysis

To better understand the representational properties of our contrastive learning approach, we conduct comprehensive interpretability analysis combining lexical feature importance and embedding visualization. We begin with word cloud analysis to identify the most discriminative linguistic features, followed by t-SNE dimensionality reduction to visualize learned representations.

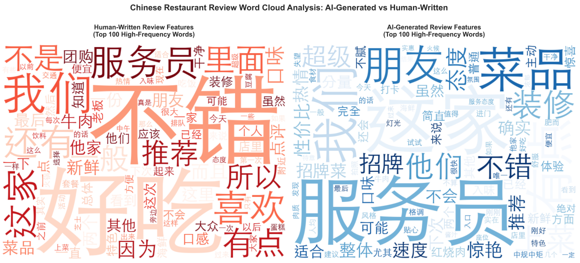
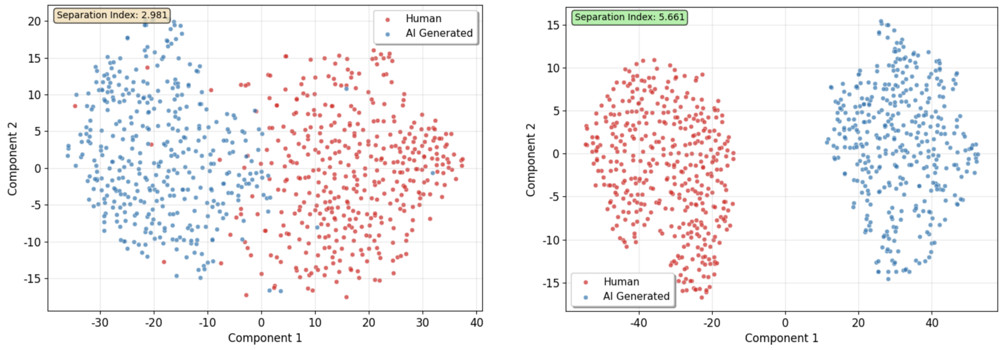


Figure 5: Chinese word cloud visualization showing the most important lexical features for distinguishing human-written and AI-generated e-commerce reviews.

Building on these lexical insights, we conduct embedding visualization using t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction on text embeddings, following established practices in representation learning analysis for AI text detection.

Figure 6 shows the clustering behavior of text embeddings in a 2D space. Before contrastive learning training (left), human and AI-generated texts show overlapping distributions with poor separability (Separation Index: 2.981). After applying our contrastive learning framework (right), the embedding space is reorganized achieving improved class separation (Separation Index: 5.661), demonstrating clearer spatial boundaries between human and AI-generated content. This embedding visualization illustrates how contrastive learning affects representation space organization, complementing the lexical feature analysis shown in Figure 5 and providing insights into the representational changes induced by different training methodologies.



(a) Before contrastive learning (t-sne)

(b) After contrastive learning (t-sne)

Figure 6: t-SNE visualization of text embeddings before and after contrastive learning training. Separation Index improves from 2.981 to 5.661, indicating better class discrimination.

5 Discussion

Our evaluation reveals distinct robustness patterns across detection paradigms. Fine-tuned large language models demonstrate superior cross-domain stability, with Qwen-2.5-7B showing gradual F1-score degradation (OOD-1: 0.5%↓ → OOD-2: 13.5%↓ → OOD-3: 32.6%↓) compared to steeper declines in Chinese-specific models (bert-base-chinese: 5.1%↓ → 25.0%↓ → 42.4%↓) and contrastive learning approaches (8.6%↓ → 30.6%↓ → 59.6%↓). This performance hierarchy suggests that model scale and architectural sophistication contribute significantly to cross-domain robustness.

The substantial decline in OOD-3 scenarios reflects the inherent complexity of cross-domain transfer in Chinese e-commerce contexts. Models trained on food-specific vocabulary and review patterns face challenges when encountering diverse product categories (electronics, clothing, home products) that exhibit different linguistic conventions and evaluation structures. This domain gap represents a fundamental technical challenge in AI text detection, where semantic feature spaces vary significantly across e-commerce domains, requiring domain-adaptive strategies for practical deployment.

6 Conclusion

This work establishes the first systematic benchmark for cross-domain robustness in Chinese AI-generated text detection, providing foundational infrastructure for evaluating detection methods under realistic deployment conditions. Our findings expose significant language-specific adaptation gaps in contemporary multilingual models and highlight the critical need for domain-adaptive training in Chinese commercial contexts. Future research priorities include developing Chinese-specific adversarial training strategies, investigating hybrid architectures that combine fine-tuning robustness with contrastive interpretability, and extending evaluation frameworks to emerging domains. This systematic approach provides methodological foundations and empirical baselines for advancing Chinese AI text detection research toward more robust and deployable systems.

Limitations

Our evaluation framework requires substantial computational resources and high-quality training data, with potential risks of dataset leakage and annotation errors affecting reliability. The focus on e-commerce reviews limits generalization to other critical domains (social media, academic, news content), while the temporal evolution of Chinese internet language poses ongoing adaptation challenges.

Deployment of such detection systems raises ethical concerns including potential misclassification of authentic human voices, particularly minority dialects or creative writing styles, and possible misuse for content censorship. The adversarial nature of this field presents recursive challenges: as detection methods improve, generation strategies will correspondingly evolve, requiring continuous system updates and validation.

Acknowledgments

References

Shifali Agrahari, Subhashi Jayant, Saurabh Kumar, and Sanasam Ranbir Singh. 2025. EssayDetect at GenAI detection task 2: Guardians of academic integrity: Multilingual detection of AI-generated essays. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 299–306, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901. Curran Associates, Inc.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 1–14.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. MGTBench: Benchmarking machine-generated text detection. *Preprint*, arXiv:2303.14822.

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are AI-generated text detectors robust to adversarial perturbations? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6005–6024, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting machine-generated texts: Not just "ai vs humans" and explainability is complicated. *Preprint*, arXiv:2406.18259.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Ann Kronrod and Shai Danziger. 2013. “it’s a metaphor!”: The effect of figurative language on online consumer reviews. *Journal of Consumer Research*, 40(3):526–538.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Robust AI-Generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17036–17055, Miami, Florida, USA. Association for Computational Linguistics.

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

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.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 25192–25208. PMLR.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Conference on Empirical Methods in Natural Language Processing*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced representation through knowledge integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3570–3596, Bangkok, Thailand. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek Fai Wong, and Lidia Sam Chao. 2025. A survey on LLM-Generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–331.

Abhishek Yadav and Shunmuga Priya M C. 2024. Classifying ai vs. human content: integrating bert and linguistic features for enhanced classification. *Operations Research Forum*, 5(2):77.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Prompt Engineering Strategy

Our controlled synthesis approach employs two complementary prompt generation strategies to maximize stylistic diversity while maintaining semantic coherence. This dual-strategy framework ensures comprehensive coverage of authentic review patterns. The comprehensive statistics for our prompt generation process are presented in Table 8, while Tables 9 and 10 provide detailed examples of the template categories used in our dual-strategy approach.

Prompt Characteristic	Value	Coverage
Total Generated Prompts	10,000	100%
Unique Prompt Variants	8,333	83.3%
Profile-based Templates	5,000	50.0%
Rewrite-based Templates	5,000	50.0%
Distinct Template Patterns	20+	High diversity

Table 8: Prompt generation statistics demonstrating comprehensive template diversity and balanced strategy distribution.

Category	Templates
Direct Requests	写一个餐厅好评/差评 *Write a positive/negative restaurant review* 帮我写个餐厅评价 *Help me write a restaurant evaluation* 给餐厅写个评论/点评 *Write a restaurant comment/review*
Sentiment-Guided	写个正面的餐厅评价 *Write a positive restaurant evaluation* 写一个满意的餐厅评论 *Write a satisfied restaurant review* 给这家餐厅写个好评 *Write a favorable review for this restaurant*
Aspect-Specific	写个餐厅评价，说味道不错 *Write a review mentioning good taste* 评论服务态度很好的餐厅 *Review restaurant with excellent service* 点评价格合理的餐厅 *Review reasonably priced restaurant* 评价环境满意的餐厅 *Review restaurant with satisfactory environment*

Table 9: Profile-based prompt templates for controlled review generation across multiple semantic dimensions.

Strategy	Templates
Style Mimicry	请基于以下真实评论重写一个类似的餐厅评价：{review}。要求：保持相同情感倾向，使用不同表达方式。 *Rewrite a similar restaurant review based on: {review}. Maintain emotional tone with different expressions.*
Content Variation	参考这个餐厅评论的风格，写一个内容不同的评价：{review}。注意：保持评价角度和语气一致。 *Reference this review's style for different content: {review}. Maintain consistent perspective and tone.*
Stylistic Imitation	模仿以下评论的写作风格，创作新的餐厅评价：{review}。要求：语调和表达习惯相似。 *Imitate the writing style to create new review: {review}. Maintain similar tone and expression patterns.*

Table 10: Style-transfer prompt templates for sophisticated linguistic pattern replication.