Benchmarking the Detection of LLMs-Generated Modern Chinese Poetry

Shanshan Wang[1] **Junchao Wu**[1] **Fengying Ye**[1] **Jingming Yao**[2]
Lidia S. Chao[1] **Derek F. Wong**[1*]
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2]Department of Portuguese, Faculty of Arts and Humanities, University of Macau
nlp2ct.{shanshan,junchao,fengying}@gmail.com
{jmyao,lidiasc,derekfw}@um.edu.mo

Abstract

The rapid development of advanced large language models (LLMs) has made AI-generated text indistinguishable from human-written text. Previous work on detecting AI-generated text has made effective progress, but has not involved modern Chinese poetry. Due to the distinctive characteristics of modern Chinese poetry, it is difficult to identify whether a poem originated from humans or AI. The proliferation of AI-generated modern Chinese poetry has significantly disrupted the poetry ecosystem. Based on the urgency of identifying AI-generated poetry in the real Chinese world, this paper proposes a novel benchmark for detecting LLMs-generated modern Chinese poetry. We first construct a high-quality dataset, which includes both 800 poems written by six professional poets and 41,600 poems generated by four mainstream LLMs. Subsequently, we conduct systematic performance assessments of six detectors on this dataset. Experimental results demonstrate that current detectors cannot be used as reliable tools to detect modern Chinese poems generated by LLMs. The most difficult poetic features to detect are intrinsic qualities, especially style. The detection results verify the effectiveness and necessity of our proposed benchmark. Our work lays a foundation for future detection of AI-generated poetry.[1]

1 Introduction

Large language models (LLMs) have made non-negligible progress in various tasks (Lan et al., 2024; Yan et al., 2024; Fang et al., 2023, 2025; Lyu et al., 2025). However, the rapid development of advanced LLMs has made AI-generated text indistinguishable from human-written text (Hayawi et al., 2024; Najjar et al., 2025). This AI-generated content could deceive readers through novel forms of manipulation (Weidinger et al., 2022; Buchanan et al., 2021; Cooke, 2018), which makes reliable detection of AI-generated text a challenge (Wu et al., 2025). This phenomenon extends even to the domain of poetry, traditionally considered the exclusive realm of human creative expression (Bena and Kalita, 2020; Chaudhuri et al., 2024).

AI-generated poetry has become increasingly prevalent in the real world. It is foreseeable that the total amount of poetry generated by AI will surpass the cumulative amount of all poetry created in human history (Huo, 2020). Professional models were shown to have the potential to generate high-quality poems that are engaging to readers (Linardaki, 2022; Köbis and Mossink, 2021). However, humans have been demonstrated to exhibit limited accuracy in distinguishing between AI-generated poetry and human-written poetry (Darda et al., 2023; Jakesch et al., 2023). Empirical studies indicate that humans cannot distinguish poems generated by ChatGPT-3.5 and may misclassify AI-generated poems as human-written literary works (Porter and Machery, 2024).

With the development of literature, modern Chinese poetry has become one of the most popular literary genres in the current Chinese world. Modern Chinese poetry is a unique genre of poetry, which is obviously different from classical Chinese poetry and rhymed English poetry (Wang et al., 2024). Specifically, modern Chinese poetry is free in form and innovative in language, and is not restricted by format, sentence length, rhythm, or meter (Guo, 1957; Wang, 2006; Skerratt, 2013; Xu, 2015; Awan and Khalida, 2015).

Can detection models accurately identify AI-generated modern Chinese poetry? The meaningful evaluation of LLMs must be grounded in cultural context (Lan et al., 2025). Previous work on detecting LLM-generated text has made effective progress (Chen et al., 2025), but these methods are not applicable to modern Chinese poetry. For

*Corresponding Author
[1]Data and code are available at: https://github.com/NLP2CT/AIGenPoetry-Detection

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9533–9552
November 4-9, 2025

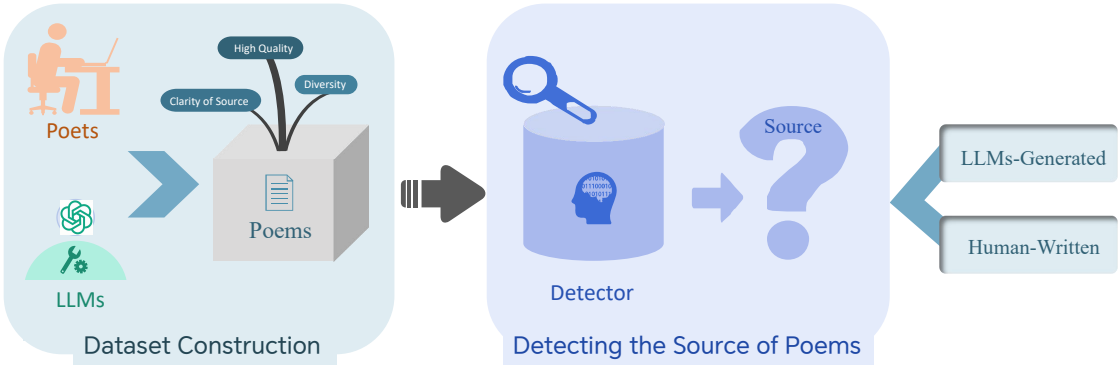


Figure 1: The framework of our proposed benchmark.

instance, Uchendu et al. (2023) proposed that incoherence in text can serve as an indicator of LLM generation, but this metric fails to generalize to modern Chinese poetry because coherence and fluency are not necessary factors for modern Chinese poetry. In recent work, Wu et al. (2024a) introduced GECSCORE, a detection method based on grammatical error correction scores, to distinguish between LLM-generated and human-written texts. Their core idea relies on the disparity in grammatical errors as a discriminative feature, because they found that human-written texts typically contain more grammatical errors than those produced by LLMs. This method performs well across diverse real-world text sources. However, GECSCORE is still not applicable to modern Chinese poetry because one of the most unique intrinsic features of modern poetry is its deliberate violation of grammatical conventions (Deng, 2007; Blohm et al., 2018). Modern poets achieve maximum rhetorical tension and novel aesthetics by breaking grammatical rules (Chen, 2012b). Therefore, the seemingly incorrect grammar in modern poetry is actually the result of careful design, rendering grammar-based detection methods ineffective.

Recently, there are authors who submit AI-generated poems to journals or publish them on online platforms without indicating that they come from LLMs. The proliferation of AI-generated modern Chinese poetry has significantly disrupted the poetry ecosystem: (1) it deceives both readers and journal editors, and (2) misleads poetry novices. Therefore, developing reliable text source identification techniques is critically urgent and practically significant for distinguishing human-written poetry from AI-generated poetry. However, no detection research has focused on modern Chinese poetry.

Based on the urgency of identifying AI-generated poetry in the real Chinese world, in this paper, we propose a novel benchmark for detecting LLMs-generated modern Chinese poetry. Our goal is to determine the recognition capabilities of detectors on LLMs-generated modern Chinese poetry. We first constructed a high-quality modern Chinese poetry dataset named AIGenPoetry, which contains 800 poems written by six professional poets and 41,600 poems generated by four mainstream LLMs. This dataset focuses on different aspects of poetry, including intrinsic qualities, external structures, and emotions. Then, we conducted detection experiments on this dataset using six detectors. The experimental results demonstrate that current detectors cannot be used as reliable tools to detect modern Chinese poems generated by LLMs. Among them, the most difficult poetic features to detect are intrinsic qualities, especially style. When conducting in-domain detection experiments on poems with different features generated by different models, RoBERTa-based detector has the best comprehensive detection performance. For data with different characteristics, poems with the same style as human poems are the most difficult to detect, while poems that literally express specific emotions, especially fear, are the easiest to detect. These detection results verify the effectiveness and necessity of our proposed benchmark.

Main contributions of our work are as follows:

- We constructed the first modern Chinese poetry dataset that includes both high-quality poems written by professional poets and those generated by several mainstream LLMs.
- We propose the first benchmark for detecting LLMs-generated modern Chinese poetry. Both in-domain and out-of-domain experimental results from six detectors verify the

effectiveness and necessity of our work.

- Our work reveals the vulnerabilities of existing detectors in this task, and lays a foundation for future detection of AI-generated poetry.

2 Related Work

With the improvement in the quality of AI-generated poetry, some studies have made efforts to identify AI-generated poetry, but primarily focused on English poetry. Köbis and Mossink (2021) evaluated humans' ability to discriminate English poems generated by GPT-2 (Radford et al., 2019) that meet appearance criteria such as rhyme and repetition. They introduced a new Turing test by incentivizing the accuracy of human judgments (i.e., increasing judges' motivation). Specifically, they evaluated humans' preference for model-generated poems and human-written poems. The results showed that participants preferred human-written poems to model-generated poems, regardless of whether they were told the source of the poems. Furthermore, although humans were motivated to participate in the experiment because of monetary rewards, they were still unable to accurately identify AI-generated poems. Porter and Machery (2024) explored whether non-professional readers can distinguish between English poems of the same style generated by GPT-3.5 and poems written by famous poets. Their experiments proved that poems generated by general AI models are more likely to be misidentified as human-written because these poems are considered simpler and easier to understand. For example, AI-generated poems have obvious themes and emotions. In addition, humans' ability to discern AI-generated poems does not improve with more poetry experience. Hayawi et al. (2024) used four traditional models to classify AI-generated texts and human-written texts, including random forest (RF) (Liaw et al., 2002), support vector machine (SVM) (Burges, 1998), logistic regression (LR) (Kleinbaum et al., 2002) and long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). Among them, the AI-generated English poems are derived from GPT-3.5 (Radford et al., 2019) and BARD (Thoppilan et al., 2022). Their experimental results show that SVM is the most reliable model, which can correctly detect the poems generated by BARD.

Different from these studies, our work focuses on modern Chinese poetry, which is significantly different from the poetry in previous work.

	Poems	Stanzas	Lines	Words
Human	800	2.1K	15K	137K
GPT-4.1	10.4K	41.8K	254K	2284K
DeepSeek-V3	10.4K	36.4K	205K	1446K
DeepSeek-R1	10.4K	39.8K	230K	1816K
GLM-4	10.4K	56.2K	340K	2764K

Table 1: The statistics of AIGenPoetry dataset.

3 Task

In our work, the detection task of LLM-generated poetry is defined as a binary classification task, formulated as:

$$Y(P) = \begin{cases} 1, & \text{if } P \text{ is generated by LLMs,} \\ 0, & \text{if } P \text{ is written by human,} \end{cases} \quad (1)$$

where P denotes the input poem to be detected, and $Y(P) \in \{0, 1\}$ represents the binary output of the detector (1 for LLMs-generated, 0 for human-written). Figure 1 presents the framework of our proposed benchmark.

4 AIGenPoetry Dataset

A high-quality benchmark dataset is a prerequisite for advancing research on detecting poetry generated by LLMs (Wu et al., 2025). Currently, there is no publicly available dataset specifically designed for identifying LLMs-generated modern Chinese poetry. To address this gap, we constructed a novel modern Chinese poetry dataset named AIGenPoetry. AIGenPoetry includes both 800 high-quality poems written by six professional human poets and 41,600 poems generated by four state-of-the-art LLMs. To the best of our knowledge, this constitutes the first modern Chinese poetry dataset comprising data pairs of both poems written by professional poets and those generated by AI.

Human-Written We collected 800 high-quality modern Chinese poems written by six professional modern Chinese poets, which are particularly valuable due to three distinctive characteristics:

- **Clarity of Source**: The misclassification of LLM-generated texts as human-written ones and their subsequent use as training data has led to an unresolved data ambiguity problem (Wu et al., 2025). To ensure the reliability and clarity of data sources in our work, all human-written poems in our dataset are derived from

poems provided by trustworthy professional modern Chinese poets themselves. This effectively avoids the data ambiguity issues highlighted in the previous detection tasks (Alemohammad et al., 2023; Cardenuto et al., 2023), where existing training datasets contain content generated by LLMs.

- **High Quality**: Human-written poems in our dataset are directly provided by six professional modern Chinese poets with rich experience in poetry creation, including: senior editors of famous poetry journals, university professors, members of the Chinese Writers Association[2], etc. They have published excellent poetry collections, published influential research papers on poetry, and won many high-quality poetry awards. They are not only professional poetry creators, but also senior researchers of poetry theory. Therefore, the poetry data from them in our work are precious and valuable for promoting the development of related domain.
- **Diversity**: Human-written poems in our dataset encompass a diverse range of styles, themes, and forms in real-world modern Chinese poetry. For example, the dataset includes poems in diverse forms, such as single-stanza poems and multi-stanza poems with uniform or varying line counts. These poems cover a wide range of topics, including emotions, daily life, social issues, philosophy, and so on.

LLMs-Generated Integrating text generated by multiple LLMs as training data can significantly enhance the detector's performance in cross-model detection (Wu et al., 2025). Therefore, the AI-generated poems in our work are produced by various mainstream LLMs, including GPT-4.1 (OpenAI) (Achiam et al., 2023), DeepSeek-V3 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025), and GLM-4 (Zhipu AI) (ZHIPU, 2024).

The performance of LLMs is highly sensitive to input prompts (Antar, 2023; Gao et al., 2023). Based on the unique characteristics of modern Chinese poetry and the way people often use LLMs to generate poetry in the real Chinese world, we carefully designed 13 prompts, recorded as $\mathcal{P}_{i\in\{1...13\}}$. Table 7 in in Appendix A.1 presents the detailed prompts used for poetry generation. These prompts focus on different aspects and characteristics of modern Chinese poetry, including intrinsic qualities, external structures, and emotions:

- $\mathcal{P}_1$: In the real world, people tend to write poems with the same title. Therefore, $\mathcal{P}_1$ requires LLMs to generate poems with the same title as human poems, but without any other restrictions. The data generated by $\mathcal{P}_1$ is used as the baseline.
- $\mathcal{P}_2$ ~ $\mathcal{P}_5$: Both intrinsic quality and external form are essential components of modern poetry (Huo, 2020). So we designed $\mathcal{P}_2$ to $\mathcal{P}_5$ based on the style (Yuan, 1991; Luo, 2000), thought and sentiment (Xi, 2019; Mo, 2009), and theme (Zhong, 2003) of the poem.
- $\mathcal{P}_6$ ~ $\mathcal{P}_8$: For external structure, the most intuitive external features of modern poetry include stanza-division and line-breaking (Xue, 2010). Line is the most fundamental formal unit of modern Chinese poetry (Yang and Liu, 1985; Shen, 1999), constituting a key distinction from classical Chinese poetry and prose (Wang, 2007; Sun, 2011). Line-breaking not only manifests the structural aesthetics of poetry but also carries its inherent tension (Chen, 2012a; Xue, 2016). A line is connected to another adjacent line or lines to form a higher-level structure called a stanza. Stanza is a section of text in modern poetry that has a complete meaning (Song et al., 2023). Therefore, we designed $\mathcal{P}_6$ to $\mathcal{P}_8$ to focus mainly on the external structure of the poem, namely stanzas and lines.
- $\mathcal{P}_9$ ~ $\mathcal{P}_{13}$: Poets also express their emotions through poetry, so $\mathcal{P}_9$ to $\mathcal{P}_{13}$ require LLMs to generate poems with different emotions, but the titles are the same as human poems.

Following DeepSeek's recommendations for the poetry writing task (Liu et al., 2024), we set temperature to `1.5` and top_p to `0.95` for all models. Then, each LLM generated 800 poems per prompt, matching the number of original human-written poems. This ensures a one-to-one correspondence between human-written and LLM-generated poems under each prompt. Human-written poems and LLMs-generated poems through $\mathcal{P}_{i\in\{1...13\}}$ constitute a pair of data (HD_i, LD_i), denoted as $\mathcal{D}_{i\in\{1...13\}}$.

Table 1 presents the statistics of AIGenPoetry dataset. Examples from the dataset are provided in Appendix A.2. Appendix A.3 shows the detailed

[2] https://www.chinawriter.com.cn/

Generator → / Detector ↓	GPT-4.1		DeepSeek-V3		DeepSeek-R1		GLM-4		Avg.	
	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1
Fast-DetectGPT	43.24	46.97	54.67	51.99	94.89	85.19	96.59	92.25	72.35	69.10
LRR	68.84	62.35	55.71	54.32	55.08	55.74	99.79	98.37	69.86	67.70
Log-Likelihood	73.44	65.19	50.08	47.18	64.78	61.11	99.66	98.50	71.99	68.00
Log-Rank	73.23	65.89	47.97	47.80	62.94	57.46	99.83	**98.87**	70.99	67.51
Binoculars	47.64	44.02	68.63	65.17	93.89	87.34	83.97	78.43	73.53	68.74
RoBERTa	82.00	**81.45**	96.25	**96.25**	99.75	**99.75**	87.38	87.21	91.35	**91.17**
Avg.	64.73	60.98	62.22	60.45	78.56	74.43	94.54	92.27	75.01	72.03

Table 2: The performance (%) of various detectors on D_1 data pairs (in-domain).

distribution of dataset, the average number of stanzas, lines, and words per poem (Table 8), as well as the statistics of word frequency statistics (Table 9).

5 Experiment

Detector We employed a variety of detectors to evaluate AIGenPoetry. For statistics-based methods, we utilized Fast-DetectGPT (Bao et al., 2023), LRR (Su et al., 2023), Log-Likelihood (Solaiman et al., 2019), Log-Rank (Gehrmann et al., 2019), and Binoculars (Hans et al., 2024). For fine-tuning-based approaches, we adopted the RoBERTa (Liu et al., 2019) based classifier, which has demonstrated superior performance on DetectRL benchmark (Wu et al., 2024b).

Since previous work has mainly been validated on English data, migrating to Chinese requires certain configuration modifications. For the Log-Likelihood, Log-Rank, and LRR methods, we employed Qwen2.5-3B[3] as the scoring model to optimize detector efficiency while maintaining a model size comparable to the widely used English scoring model GPT-Neo-2.7B[4]. For Fast-DetectGPT, we used Qwen2.5-3B as the reference model and Qwen2.5-7B[5] as the scoring model. For Binoculars, we used Qwen2.5-7B as the observer, and Qwen2.5-7B-Instruct[6] as the performer. For the RoBERTa-based classifier, we fine-tuned the Chinese RoBERTa model[7] with `3` epochs, a learning rate of `1e-6`, and a batch size of `16`.

We used the above detectors to conduct both in-domain and generalization (out-of-domain) experiments on AIGenPoetry dataset. In-domain experiments were trained and tested on datasets with the same characteristics generated by individual LLMs. Both the training set and the test set contain 400 pairs of poems. Each pair consists of a human-written poem and a LLM-generated poem. The data for generalization experiments merged poems generated by multiple LLMs, but the test set had different features from the training set. We balanced the dataset by up-sampling the human-written poems to match the size of the LLM-generated poems across all prompts.

Temperature Experiments To explore the effects of different temperatures, we conducted additional experiments with temperatures of `0.7` and `0.0`. The experiment is divided into two steps:

1. *Four LLMs generate new poetry data*. Based on $\mathcal{P}_1$, each LLM generates 800 new poems at temperatures of `0.7` and `0.0` respectively.
2. *Detectors detect LLMs-generated poems*. 1) Overall performance of the detector: the detectors detect the data generated by the 4 LLMs. 2) Performance of the detector for a single LLM: the detectors detect the data generated by a single LLM. 3) Generalization of the detector: the generalization ability of the detectors on the datasets generated under different temperature settings.

6 Analysis

Tables 2 and 3 present the performance of various detectors on the D_1 data pairs and the D_{2-13} data pairs (in-domain) with different characteristics, respectively. Table 4 shows the generalization performance of various detectors across out-domain data pairs with different features from all LLMs. The results (F1-scores) of the temperature experiments are presented in Table 5 and 6.

[3] https://huggingface.co/Qwen/Qwen2.5-3B

[4] https://huggingface.co/EleutherAI/gpt-neo-2.7B

[5] https://huggingface.co/Qwen/Qwen2.5-7B

[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[7] https://huggingface.co/hfl/chinese-RoBERTa-wwm-ext

Generator →	GPT4.1	DS-V3	DS-R1	GLM4	Avg.	GPT4.1	DS-V3	DS-R1	GLM4	Avg.
Detector ↓	**Focus on the intrinsic qualities of poetry (D_{2-5})**									
	D_2 (Style)					D_3 (Style)				
Fast-DetectGPT	33.33	52.21	75.82	54.58	53.99	43.28	57.50	80.99	50.89	58.17
LRR	47.00	53.17	51.54	58.07	52.45	52.28	63.97	52.00	54.89	55.79
Log-Likelihood	43.58	51.81	56.82	54.34	51.64	48.19	67.61	60.01	50.26	56.52
Log-Rank	44.10	50.63	52.47	57.15	51.09	50.21	66.50	56.50	50.16	55.84
Binoculars	38.00	59.87	79.12	68.25	61.31	52.47	65.79	82.36	64.12	66.19
RoBERTa	**86.35**	**84.85**	**97.00**	**82.08**	**87.57**	**79.09**	**73.94**	**96.62**	**64.62**	**78.57**
Avg.	48.73	58.76	68.80	62.41	59.68	49.29	64.27	66.37	54.06	58.50
	D_4 (Thought and sentiment)					D_5 (Theme)				
Fast-DetectGPT	33.33	54.54	79.45	69.73	59.26	33.22	61.78	83.09	71.12	62.30
LRR	52.20	57.72	57.15	77.12	61.05	50.14	63.40	56.74	77.11	61.85
Log-Likelihood	49.35	59.99	63.91	75.16	62.10	50.47	65.22	59.20	71.05	61.49
Log-Rank	50.25	62.32	62.89	76.72	63.05	51.60	63.82	61.07	75.85	63.09
Binoculars	33.33	61.41	81.54	74.50	62.70	34.32	72.24	83.47	76.56	66.65
RoBERTa	**84.05**	**87.87**	**98.12**	**82.90**	**88.24**	**86.18**	**91.12**	**99.00**	**81.05**	**89.34**
Avg.	50.42	63.98	73.84	76.02	66.07	50.99	69.60	73.76	75.46	67.45
	Focus on the external structures of poetry (D_{6-8})									
	D_6 (Stanza)					D_7 (Line)				
Fast-DetectGPT	40.27	46.98	71.89	80.00	59.78	33.33	65.79	87.62	77.75	66.12
LRR	58.84	49.81	48.07	94.62	62.84	58.31	54.99	54.58	94.00	65.47
Log-Likelihood	68.08	41.97	50.16	93.87	63.52	68.94	53.91	61.83	**94.12**	69.70
Log-Rank	66.42	45.69	50.37	**94.75**	64.31	68.36	55.79	58.81	94.00	69.24
Binoculars	36.65	60.89	77.08	80.28	63.72	33.33	74.23	87.37	70.37	66.32
RoBERTa	**84.87**	**94.36**	**98.87**	87.35	**91.36**	**83.36**	**97.62**	**99.62**	83.26	**90.97**
Avg.	59.19	56.62	66.07	88.48	67.59	57.60	67.05	74.97	85.58	71.30
	D_8 (Stanza and line)					D_9 (No emotion)				
Fast-DetectGPT	33.72	60.57	87.37	71.08	63.18	33.38	36.62	78.85	74.25	55.78
LRR	56.49	56.34	55.64	90.48	64.74	44.13	43.91	45.74	86.74	55.13
Log-Likelihood	68.55	51.61	59.98	**91.09**	67.81	52.79	40.16	50.83	80.74	56.13
Log-Rank	66.89	54.02	62.06	90.96	68.48	50.52	40.44	49.23	83.40	55.90
Binoculars	33.33	69.98	85.32	77.22	66.46	45.86	59.35	82.43	81.59	67.31
RoBERTa	**86.19**	**97.62**	**99.37**	84.73	**91.98**	**98.37**	**98.37**	**100.00**	**95.75**	**98.12**
Avg.	57.53	65.02	74.96	84.26	70.44	54.18	53.14	67.85	83.75	64.73
	Focus on emotions in poetry (D_{10-13})									
	D_{10} (Happiness and joy)					D_{11} (Sadness and despair)				
Fast-DetectGPT	33.38	49.45	84.22	77.96	61.25	39.42	51.48	87.87	91.11	67.47
LRR	71.85	60.20	58.35	97.87	72.07	62.06	56.04	56.93	98.00	68.26
Log-Likelihood	74.48	49.96	66.39	98.50	72.33	68.80	51.90	65.47	98.37	71.14
Log-Rank	77.71	56.45	66.06	**98.87**	74.77	68.33	54.54	65.91	**98.87**	71.91
Binoculars	33.33	60.49	85.35	63.47	60.66	41.15	66.11	86.85	75.02	67.28
RoBERTa	**86.57**	**97.25**	**99.12**	94.86	**94.45**	**91.05**	**97.87**	**100.00**	95.37	**96.07**
Avg.	62.89	62.30	76.58	88.59	72.59	61.80	62.99	77.17	92.79	73.69
	D_{12} (Anger)					D_{13} (Fear)				
Fast-DetectGPT	50.76	54.91	86.66	93.87	71.55	46.80	55.60	85.81	92.87	70.27
LRR	58.48	54.90	56.42	97.00	66.70	59.84	55.38	54.99	98.12	67.08
Log-Likelihood	60.32	50.77	63.39	98.12	68.15	65.88	50.60	64.66	98.25	69.85
Log-Rank	61.76	52.19	62.86	98.25	68.77	63.13	53.78	62.42	**98.50**	69.46
Binoculars	53.41	66.04	87.87	80.76	72.02	58.40	72.28	87.75	76.54	73.74
RoBERTa	**97.12**	**97.87**	**100.00**	**98.37**	**98.34**	**94.49**	**98.37**	**100.00**	98.25	**97.78**
Avg.	63.64	62.78	76.20	94.40	74.26	64.76	64.33	75.94	93.76	74.70

Table 3: The performance (F1-score) of various detectors on D_{2-13} data pairs (in-domain) with different characteristics. DS-V3 and DS-R1 represent DeepSeek-V3 and DeepSeek-R1 respectively.

	Train on D_{2+4+5}, test on different data						Train on D_{6-8}, test on different data				
Test on →	D_{2+4+5}	D_1	D_{6-8}	D_9	D_{10-13}	Avg.	D_{6-8}	D_1	D_{10-13}	D_{2+4+5}	Avg.
Fast-DetectGPT	60.50	67.81	63.80	59.11	67.03	64.44	64.68	68.91	67.88	60.86	65.88
LRR	56.57	59.44	57.93	50.15	60.46	57.00	59.55	60.97	62.39	57.55	60.30
Log-Likelihood	56.57	58.32	57.85	51.25	59.85	56.82	64.24	63.61	66.23	60.20	63.35
Log-Rank	58.11	60.50	59.89	52.17	62.24	58.70	63.02	62.69	65.38	59.53	62.53
Binoculars	64.24	69.61	67.69	68.43	68.02	68.44	67.69	69.61	68.02	64.24	67.29
RoBERTa	**84.19**	**84.32**	**84.31**	**83.53**	**83.87**	**84.01**	**89.12**	**89.16**	**89.04**	**87.12**	**88.44**
Avg.	63.36	66.67	65.25	60.77	66.91	64.90	68.05	69.16	69.82	64.92	67.97

	Train on D_{10-13}, test on different data									
Test on →	D_{10-13}	D_1	D_{2+4+5}	D_2	D_3	D_{6-8}	D_6	D_9	D_{10}	Avg.
Fast-DetectGPT	67.88	68.91	60.85	55.71	60.09	64.68	61.04	59.22	63.39	61.74
LRR	63.33	61.97	58.17	53.94	56.97	60.51	58.89	52.12	66.05	58.58
Log-Likelihood	65.11	62.87	59.51	54.13	59.05	63.04	59.66	54.19	67.10	59.94
Log-Rank	65.36	62.69	59.52	53.77	58.73	62.99	59.85	53.39	68.00	59.87
Binoculars	68.02	69.61	64.24	60.86	63.13	67.69	66.06	68.43	61.66	65.21
RoBERTa	**93.59**	**93.07**	**85.90**	**81.88**	**83.25**	**92.80**	**92.51**	**88.62**	**93.41**	**88.93**
Avg.	70.55	69.85	64.70	60.05	63.54	68.62	66.34	62.66	69.93	65.71

Table 4: The generalization performance (F1-score) of various detectors on different feature data pairs from all LLMs.

Detector →	**Fast-Det.**	**LRR**	**Log-Likelihood**	**Log-Rank**	**Binoculars**	**RoBERTa**	**Avg.**
Train on `1.5`, test on `1.5`	68.88	62.07	62.76	61.72	69.61	92.44	69.58
Train on `0.7`, test on `0.7`	76.15	71.22	67.81	70.52	74.28	93.10	75.51
Train on `0.0`, test on `0.0`	**79.34**	**74.34**	**70.36**	**74.03**	**78.43**	**93.29**	**78.30**
Train on `1.5`, test on `0.7`	76.15	66.11	66.90	65.97	74.28	92.84	73.71
Train on `1.5`, test on `0.0`	**79.34**	69.28	68.65	67.86	**78.43**	92.75	76.05

Table 5: The results (F1-scores) of the temperature experiments on poetry data generated by 4 LLMs. The `1.5`, `0.7`, and `0.0` in the first column represent the different temperatures of the LLMs used to generate the poetry data. Fast-Det. in the first row represent Fast-DetectGPT.

6.1 In-domain

Baselines As shown in Table 2, current detectors cannot serve as reliable tools for detecting LLMs-generated modern Chinese poems, despite their unexpected performance on certain individual LLM-generated poems. Specifically, when the evaluated LLMs-generated poems share identical titles with human-written poems, most detectors exhibit unsatisfactory performance. Among them, RoBERTa-based detector demonstrates markedly distinct detection capabilities compared to other detectors when distinguishing between data pairs composed of poems from different LLMs and human-written poems. RoBERTa-based detector achieves the best overall detection performance, significantly outperforming other detectors. Its average F1-score is 91.17%, surpassing the second-best detector, Fast-DetectGPT, by 22.07%. However, the average detection performance of the remaining five detectors is comparable, with F1-scores clustered between 67.51% and 68.74%. Obviously, RoBERTa-based detector outperforms other detectors in detecting poems generated by DeepSeek-R1, DeepSeek-V3, and GPT-4.1. Especially for the detection of DeepSeek-R1-generated poems, RoBERTa-based detector achieves an F1-score of 99.75%. Intriguingly, for GLM-4-generated data, RoBERTa-based detector attains only an 87.38% F1-score, while Fast-DetectGPT, LRR, Log-Likelihood, and Log-Rank all achieve F1-scores above 92%. However, GPT-4.1-generated poems pose challenges to all detectors, with the highest detection performance being merely 81.45% (RoBERTa-based detector). In contrast, among individual LLMs, GLM-4-generated poems are the most detectable, with an average F1-score of 92.27%. We infer that this is related to the length of GLM-4-generated poems.

External Structures Under identical prompts, GLM-4 generates longer poems than both human and other LLMs (Table 8). Specifically, the poems generated by GLM-4 contain on average 2.73 more stanzas, 13.83 more lines, and 93.95 more words

Detector →	Fast-Det.	LRR	Log-Likelihood	Log-Rank	Binoculars	RoBERTa	Avg.
	Experiments on poetry data generated by GPT-4.1						
Train on `1.5`, test on `1.5`	46.97	62.35	65.19	65.89	44.02	81.45	60.98
Train on `0.7`, test on `0.7`	69.74	95.25	95.87	96.12	61.84	80.84	83.28
	Experiments on poetry data generated by DeepSeek-V3						
Train on `1.5`, test on `1.5`	51.99	54.32	47.18	47.80	65.17	96.25	60.45
Train on `0.7`, test on `0.7`	61.50	64.32	58.10	61.95	74.59	97.37	69.64
	Experiments on poetry data generated by DeepSeek-R1						
Train on `1.5`, test on `1.5`	85.19	55.74	61.11	57.46	87.34	99.75	74.43
Train on `0.0`, test on `0.0`	87.15	60.64	65.16	63.97	**91.87**	**99.87**	78.11
	Experiments on poetry data generated by GLM-4						
Train on `1.5`, test on `1.5`	92.25	98.37	98.50	98.87	78.43	87.21	92.27
Train on `0.0`, test on `0.0`	**93.87**	**98.87**	**99.12**	**99.12**	78.83	88.78	**93.10**

Table 6: The results (F1-scores) of the temperature experiments on poetry data generated by a single LLM.

than those written by human, and exceed GPT-4.1-generated poems (the second longest) by 1.38 stanzas, 8.21 lines, and 46.16 words, respectively. In the process of constructing the dataset, we found that GLM-4 can effectively capture the meaning of $\mathcal{P}_6$, that is, the generated poems have exactly the same number of stanzas as human poems. However, when generating poems with identical line counts (D_7), all models generate poems containing 1–3 additional lines per poem compared to human counterparts, though remaining structurally proximate. Overall, the structure of GLM-4-generated poems is similar to that of human poems when the number of stanzas is restricted. However, compared to other poems without restrictions on the structure, the number of stanzas and lines generated by GLM-4 is far more than that of human poems.

D_{6-8} in Table 3 validate our assertion that GLM-4's detectability stems from generating poems exceeding human-written lengths. Specifically, when required to match human stanza counts (D_6), detectors' average F1-score on GLM-4 decreases from 92.27% (Table 2) to 88.48% (Table 3). Similarly, with line count constraints (D_7), detectors' performance drops to 85.58%, and further to 84.26% under combined stanza-line constraints (D_8). This result also verifies the necessity of constructing poetry data with different structures (D_{6-8}).

Intrinsic Qualities Modern Chinese poetry transcends mere line-breaking techniques, prioritizing intrinsic qualities over external structures. D_{2-5} in Table 3 present detector performance on LLMs-generated poems same as human-written poems in style, thought & sentiment, and themes. When detecting style-matched poems with distinct titles and content (D_2), all detectors exhibit significant performance declines. For instance, average detection performance drops from 72.03% (Table 2) to 59.68% (Table 3), with the F1-score of RoBERTa-based detector decreasing from 91.17% (D_1) to 87.57% (D_2). Other detectors show reductions of at least 7.43% (Binoculars), with four remaining detectors declining by at least 15.11% (Fast-DetectGPT). For GLM-4, detection performance plummets from 92.27% to 62.41%. Notably, GPT-4.1-generated poems with the same style as human poems are the most difficult to detect among all the data, with detectors achieving only an average F1-score of 48.73% (Table 3). These results demonstrate that LLMs can effectively evade the detection of detectors by imitating the style of poetry, and generating poetry in the same style is currently one of the most commonly used method for AI-generated poetry in real-world scenarios. Disappointingly, except for the detection performance of RoBERTa-based detector (97.00%) on DeepSeek-R1, other detectors performed poorly in detecting different LLMs-generated poems with the same style as human poems. Similarly, detection performance on D_{4-5} decreases compared to D_1, confirming the difficulty of detecting LLMs-generated poems matching human intrinsic qualities.

The analysis of emotions is shown in Appendix A.4. In summary, when conducting in-domain detection experiments on poems with different features generated by different models, RoBERTa-based detector has the best comprehensive detection performance. For data with different characteristics, poems with the same style as human poems are the most difficult to detect, while poems that

literally express specific emotions, especially fear, are the easiest to detect.

6.2 Generalization

As shown in Table 4, detectors trained on poems focusing on intrinsic qualities (D_{2+4+5}) have the ability to generalize to poems with other features besides D_9. Specifically, detectors trained on poems focusing on intrinsic qualities (D_{2+4+5}) can generalize to the baseline (D_1), poems with restricted structures (D_{6-8}), and poems with obvious emotions in the surface meaning of the text (D_{10-13}). For example, Fast-DetectGPT improves its F1-score from 60.50% to 67.81% when detecting a baseline (D_1) where only the title is the same as human poems. Similarly, it also improves its F1-score to 67.03% when detecting poems focusing on different emotions (D_{10-13}). This shows that intrinsic qualities including style, theme, thought and sentiment are the core issues that need to be solved to accurately detect poems generated by LLMs. The F1-scores of detectors trained on datasets D_2, D_3, D_4, and D_5 in Table 3 when detecting in-domain poems are all lower than those of detectors trained on other data (except D_9), which further reinforces this conclusion.

In contrast, detectors trained on poems that focus on structure (D_{6-8}) cannot generalize to poems that focus on intrinsic qualities (D_{2+4+5}). Specifically, the average F1-score of all detectors trained on data D_{6-8} decreased from 68.05% to 64.92% when detecting data D_{2+4+5}. And the performance of each detector decreased. Similarly, detectors trained on poems that literally express different emotions (D_{10-13}) also have difficulty generalizing to poems with other characteristics. Among them, the detectors have the most difficulty generalizing to poems with the same style as human poems but different titles and contents, and the average F1-score of all detectors decreased from 70.55% to 60.05% (D_2).

The results in Table 3 and 4 prove that the most difficult poetry features to detect are intrinsic qualities, especially style. However, in real scenarios, imitating style is one of the most common way to generate poetry using LLMs. Unfortunately, current detectors are still inaccurate for detecting modern Chinese poems with multiple features and cannot be used as reliable poetry detectors. This once again highlights the effectiveness and necessity of building high-quality poetry datasets.

6.3 Analysis of Temperature Experiments

Detectors except RoBERTa-based Table 5 and 6 show that when LLM's temperature is set to `1.5`, the generated poems are the most difficult to detect, and the detector performs the worst under this temperature setting. As the temperature decreases, the detector's performance improves (compared to a temperature of `1.5`, the average improvement of the detector at temperatures of `0` and `0.7` is 8.72% and 5.93%, respectively). This indicates that poems generated by LLMs at lower temperature settings are easier to detect. More specifically, taking GPT-4.1 generator as an example, when the temperature decreases from `1.5` to `0.7`, the F1-score of LRR improves by 32.90% (Table 6). Furthermore, when the test set comes from a lower temperature setting, the performance of all detectors is improved. This shows that the detectors trained on the dataset with a temperature of `1.5` can be effectively generalized to datasets with lower temperatures.

RoBERTa-based detector The temperature of LLMs generating poetry data has an extremely minimal, negligible impact on the performance of RoBERTa-based detector. In addition, GLM-4 remains relatively unaffected by variations in temperature settings compared to the other three LLMs.

These results prove that setting the temperature to `1.5` for the poetry generation task is reasonable, which facilitates the construction of a challenging benchmark to reveal the difficulties posed by LLM-generated poetry for existing detection systems.

7 Conclusion

In this paper, we propose the first benchmark for detecting LLMs-generated modern Chinese poetry. Experimental results demonstrate that current detectors cannot be used as reliable tools to detect LLMs-generated modern Chinese poems. The most difficult poetic features to detect are intrinsic qualities, especially style, while poems that literally express specific emotions, especially fear, are the easiest to detect. GPT-4.1-generated poems with the same style as human poems are the most difficult to detect among all the data. The detection results verify the effectiveness of our proposed benchmark. Our work reveals the vulnerabilities of existing detectors in this task, and lays a foundation for future detection of AI-generated poetry. Meanwhile, we call on the research community to pay attention to the oversight and protection of modern Chinese poetry and other forms of artistic creation.

Limitations

Our research focuses on modern Chinese poetry, which is characterized by a highly flexible form. Our method may not be applicable to classical poetry or rhymed modern poetry. Therefore, the results of our study cannot represent or cover all types of poetry.

Ethics Statement

The dataset we built consists of high-quality poems. Poetry may contain negative emotions, but there is no information harmful to society.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/0007/2024/AKP), the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/0070/2022/AMJ, China Strategic Scientific and Technological Innovation Cooperation Project Grant No. 2022YFE0204900), the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/060/2022/AFJ, National Natural Science Foundation of China Grant No. 62261160648), the UM and UMDF (Grant Nos. MYRG-GRG2023-00006-FST-UMDF, MYRG-GRG2024-00165-FST-UMDF, EF2024-00185-FST), and the National Natural Science Foundation of China (Grant No. 62266013).

References

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

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 4:14.

D. Antar. 2023. The effectiveness of using chatgpt4 in creative writing in arabic: Poetry and short story as a model. In *Information Sciences Letters*.

Abdul Ghafoor Awan and Yasmin Khalida. 2015. New trends in modern poetry. *Journal of Literature, Languages and Linguistics*, 13:63–72.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Brendan Bena and Jugal Kalita. 2020. Introducing aspects of creativity in automatic poetry generation. *arXiv preprint arXiv:2002.02511*.

Stefan Blohm, Valentin Wagner, Matthias Schlesewsky, and Winfried Menninghaus. 2018. Sentence judgments and the grammar of poetry: Linking linguistic structure and poetic effect. *Poetics*, 69:41–56.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation. *Center for Security and Emerging technology*, 1(1):2.

Christopher JC Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda Andaló, Sébastien Marcel, Anderson Rocha, and 1 others. 2023. The age of synthetic realities: Challenges and opportunities. *APSIPA Transactions on Signal and Information Processing*, 12(1).

Soma Chaudhuri, Maura Dooley, Dan Johnson, Roger Beaty, and Joydeep Bhattacharya. 2024. Evaluation of poetic creativity: Predictors and the role of expertise—a multilevel approach. *Psychology of Aesthetics, Creativity, and the Arts*.

Xin Chen, Junchao Wu, Shu Yang, Runzhe Zhan, Zeyu Wu, Ziyang Luo, Di Wang, Min Yang, Lidia S Chao, and Derek F Wong. 2025. Repreguard: Detecting llm-generated text by revealing hidden representation patterns. *arXiv preprint arXiv:2508.13152*.

Zhongyi Chen. 2012a. Branch cross-industry, the trivial skills of external form? - research on modern poetic rhetoric v. *Journal of Nanjing University of Science and Technology: Social Science Edition*, 25(1):7.

Zhongyi Chen. 2012b. Why "twist the neck of grammar" - a study of modern poetic rhetoric. *Journal of Central China Normal University: Humanities and Social Sciences Edition*, 51(1):7.

Nicole A Cooke. 2018. *Fake news and alternative facts: Information literacy in a post-truth era*. American Library Association.

Kohinoor Darda, Marion Carre, and Emily Cross. 2023. Value attributed to text-based archives generated by artificial intelligence. *Royal Society Open Science*, 10(2):220915.

Cheng Deng. 2007. Dilemma and solution: Reflections on current new poetry. *Literary Review*, (3):4.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Tao Fang, Tianyu Zhang, Derek F. Wong, Keyan Jin, Lusheng Zhang, Qiang Zhang, Tianjiao Li, Jinlong Hou, and Lidia S. Chao. 2025. Llmcl-gec: Advancing grammatical error correction with llm-driven curriculum learning. *Expert Systems with Applications*, 279:127397.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

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

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*.

Moruo Guo. 1957. Talking about literary translation work. In *Complete Works of Guo Moruo*.

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

Kadhim Hayawi, Sakib Shahriar, and Sujith Samuel Mathew. 2024. The imitation game: Detecting human and ai-generated texts in the era of chatgpt and bard. *Journal of Information Science*, page 01655515241227531.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Junming Huo. 2020. Clone li bai and 100 trillion poems - the "quasi-text" production and possible prospects of ai poetry. *Southern Literature*, (4):5.

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

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

Nils Köbis and Luca D Mossink. 2021. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in human behavior*, 114:106553.

Kaixin Lan, Tao Fang, Derek Wong, Yabo Xu, Lidia Chao, and Cecilia Zhao. 2024. FOCUS: Forging originality through contrastive use in self-plagiarism for language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14432–14447.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. McBE: A multi-task Chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056.

Andy Liaw, Matthew Wiener, and 1 others. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Christina Linardaki. 2022. Poetry at the first steps of artificial intelligence. *Humanist Studies & the Digital Age*, 7(1).

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. *CoRR*, abs/1907.11692.

Zhenya Luo. 2000. "anti-traditional" singing - the new artistic quality of bian zhilin's poetry. *Literary Review*, 000(002):84–91.

Hanjia Lyu, Jiebo Luo, Jian Kang, and Allison Koenecke. 2025. Characterizing bias: Benchmarking large language models in simplified versus traditional chinese. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2815–2846. Association for Computing Machinery.

Qingwei Mo. 2009. Emotional classification of new poetry and its cultural roots. *Management Observation*, (13):1.

Ayat Najjar, Huthaifa I Ashqar, Omar Darwish, and Eman Hammad. 2025. Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llms-generated text. *arXiv preprint arXiv:2501.03212*.

Brian Porter and Edouard Machery. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133.

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

Qi Shen. 1999. Poetic, poetic form and non-poetic. *Contemporary Writers Review*, (6):3.

Brian Phillips Skerratt. 2013. *Form and Transformation in Modern Chinese Poetry and Poetics*. Ph.D. thesis, Harvard University.

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*.

Wai Lei Song, Haoyun Xu, Derek F Wong, Runzhe Zhan, Lidia S Chao, and Shanshan Wang. 2023. Towards zero-shot multilingual poetry translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 324–335.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Liyao Sun. 2011. The art of "travel": A new exploration of modern poetry form. *Academic Monthly*, (1):14.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and 1 others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, and 1 others. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Jian Wang. 2007. The rationale and function of the construction of line divisions in modern poetry. *Journal of Mianyang Normal University*, 26(7):6.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043.

Yuchun Wang. 2006. Moruo guo's view on literary translation. *Moruo Guo Journal*, page 5.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. 2024a. Who wrote this? the key to zero-shot llm-generated text detection is gecscore. *arXiv preprint arXiv:2405.04286*.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024b. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yunshu Xi. 2019. Two characteristics of new poetry language—talking about the creation of new poetry part three. *Masterpiece Appreciation: Appreciation Edition (First Decade)*, (8):6.

Jiang Xu. 2015. "modern poetry" and "new poetry". *Xingxing Monthly*.

Shichang Xue. 2010. Modern poetry formal processing methods - talk about "leaving blanks", "raising lines" and "sectioning". *Reading and Writing*, (4):3.

Shichang Xue. 2016. Something deeper than poetry itself – on the artistic intention of the line arrangement of poetry. *Write*, page 6.

Jianhao Yan, Yun Luo, and Yue Zhang. 2024. RefuteBench: Evaluating refuting instruction-following for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13775–13791.

Kuanghan Yang and Fuchun Liu. 1985. *Discussion on Chinese Modern Poetry, Part 1*. Discussion on Chinese Modern Poetry, Part 1.

Kejia Yuan. 1991. A moving quartet - the trajectory of feng zhi's poetic style. *Literary Review*, (4):7.

ZHIPU. 2024. Zhipu ai devday glm-4. Accessed: 2024-05-01.

Ling Zhong. 2003. *American Poetry and Chinese Dream: Chinese Cultural Model in Modern American Poetry*. American Poetry and Chinese Dream: Chinese Cultural Model in Modern American Poetry.

A Appendix

A.1 Prompts for Poetry Generation

Table 7 presents the detailed prompts we designed for poetry generation.

The prompts in our work were carefully crafted and validated to meet specific goals. Below, we address the key considerations:

- **Diversity:** Previous benchmark studies [1-3] typically used simplistic prompts, often limited to a single title for text generation. In

Prompts	Specific Content
$\mathcal{P}_1$	Please create a modern Chinese poem titled T_i.
$\mathcal{P}_2$	Please create a new modern Chinese poem, requiring the style to imitate the following poem, but with a completely different title and content.
$\mathcal{P}_3$	Please create a new modern Chinese poem, requiring the style to imitate the following poem, with the same title as the following poem, but with completely different content.
$\mathcal{P}_4$	Please create a new modern Chinese poem, requiring the expressed thought and sentiment to be the same as the following poem, but with a completely different title and content.
$\mathcal{P}_5$	Please create a new modern Chinese poem, requiring the theme to be the same as the following poem, but with a completely different title and content.
$\mathcal{P}_6$	Please create a modern Chinese poem titled T_i, containing S_i stanzas.
$\mathcal{P}_7$	Please create a modern Chinese poem titled T_i, with a total of L_i lines.
$\mathcal{P}_8$	Please create a modern Chinese poem titled T_i, containing S_i stanzas and a total of L_i lines.
$\mathcal{P}_9$	Please create a modern Chinese poem titled T_i, ensuring that the content does not convey any emotion.
$\mathcal{P}_{10}$	Please create a modern Chinese poem titled T_i, expressing emotions of happiness and joy.
$\mathcal{P}_{11}$	Please create a modern Chinese poem titled T_i, expressing emotions of sadness and despair.
$\mathcal{P}_{12}$	Please create a modern Chinese poem titled T_i, expressing emotions of anger.
$\mathcal{P}_{13}$	Please create a modern Chinese poem titled T_i, expressing emotions of fear.

Table 7: The prompts we designed for poetry generation.

contrast, our study goes beyond this baseline ($\mathcal{P}_1$) by designing and verifying 12 additional diverse prompts. We designed these prompts to reflect the unique characteristics of modern Chinese poetry and the ways in which users typically interact with large language models (LLMs) for poetic generation in practical settings. Our 13 prompts address different aspects of modern Chinese poetry, ranging from intrinsic qualities (e.g., emotions, themes, and styles) to external structures (e.g., number of stanzas or lines). The complexity of these prompts varies depending on their focus. For instance, $\mathcal{P}_3$ requires the model to generate a new modern Chinese poem by imitating the style of a human-written poem while creating entirely different content under the same title (Table 7). This demonstrates the diversity and sophistication of the prompts we developed.

- **Applicability:** Our prompts were specifically designed for generating modern Chinese poetry, reflecting its unique characteristics. Importantly, these prompts were developed based on feedback from professional modern Chinese poets. Their expertise as both practitioners and researchers of poetry theory has been invaluable in ensuring the relevance and applicability of our prompts to the domain.

- **Practicality:** The prompts we designed reflect how people commonly use LLMs in real-world scenarios. For example, users often generate poems with the same title ($\mathcal{P}_1$), the same style ($\mathcal{P}_2$ and $\mathcal{P}_3$), or specific structures ($\mathcal{P}_{6-8}$). These practical use cases informed our prompt designs.

- **Effectiveness:** The goal of our study is to evaluate the ability of detectors to identify the source of poems. The poems generated using our prompts exhibit high quality and are comparable to human-written poems, posing significant challenges for detection models. This underscores the effectiveness of our prompt designs in achieving the objectives of our study.

A.2 Examples

Here are examples of a human-written poem and LLMs-generated poems under $\mathcal{P}_2$.

- Human Poem:

捣蒜
(dǎo suàn)

一层一层，剥脱自己
(yì céng yì céng, bāo tuō zì jǐ)
一瓣一瓣，掰碎自己
(yí bàn yí bàn, bāi suì zì jǐ)
在彻底坦白前
(zài chè dǐ tǎn bái qián)
一切都是羁绊
(yī qiè dōu shì jī bàn)

在一个容器底部，感受弧度的力量
(zài yí gè róng qì dǐ bù, gǎn shòu hú dù de lì liàng)

直到，交出体内全部的辽阔
(zhí dào, jiāo chū tǐ nèi quán bù de liáo kuò)
直到，所有绿色倒塌在白色里
(zhí dào, suǒ yǒu l ǜ sè dǎo tā zài bái sè lǐ)
直到变成液体
(zhí dào biàn chéng yè tǐ)

把所有光芒都说出来
(bǎ suǒ yǒu guāng máng dōu shuō chū lái)

- GPT-4.1 Generated Poem:

煮茶
(zhǔ chá)

一片一片，舒展自己
(yí piàn yí piàn, shū zhǎn zì jǐ)
一缕一缕，漂浮自己
(yì l ǚ yì l ǚ, piāo fú zì jǐ)
在沸水之前
(zài fèi shuǐ zhī qián)
一切都还是安静
(yī qiè dōu hái shì ān jìng)

在透明的杯壁，感受温度的环绕
(zài tòu míng de bēi bì, gǎn shòu wēn dù de huán rào)

直到，叶脉写尽内心的苦与甘
(zhí dào, yè mài xiě jìn nèi xīn de kǔ yǔ gān)
直到，所有青涩褪成澄澈
(zhí dào, suǒ yǒu qīng sè tuì chéng chéng chè)
直到溶进琥珀
(zhí dào róng jìn hǔ pò)
让余香与回声缓缓升起
(ràng yú xiāng yǔ huí shēng huǎn huǎn shēng qǐ)

- DeepSeek-V3 Generated Poem:

折信
(zhé xìn)

一点一点，拆开自己
(yì diǎn yì diǎn, chāi kāi zì jǐ)
一字一句，摊平自己
(yí zì yí jù, tān píng zì jǐ)
在抵达纸心前
(zài dǐ dá zhǐ xīn qián)
笔画都是墙壁

(bǐ huà dōu shì qiáng bì)
在信封的折痕里，触碰往事的锋利

(zài xìn fēng de zhé hén lǐ, chù pèng wǎng shì de fēng lì)
直到，墨迹洇成云朵的形状
(zhí dào, mò jì yīn chéng yún duǒ de xíng zhuàng)
直到，所有暗语坠落在空白里
(zhí dào, suǒ yǒu àn yǔ zhuì luò zài kòng bái lǐ)
直到变成羽毛
(zhí dào biàn chéng yǔ máo)
把整个秋天都轻轻举起
(bǎ zhěng gè qiū tiān dōu qīng qīng jǔ qǐ)

- DeepSeek-R1 Generated Poem:

晾衣绳
(liàng yī shéng)

一褶一褶，展平蜷缩的潮声
(yì zhě yì zhě, zhǎn píng quán suō de cháo shēng)
在风干之前
(zài fēng gān zhī qián)
所有褶皱都是未愈合的伤口
(suǒ yǒu zhě zhòu dōu shì wèi yù hé de shāng kǒu)

金属丝震颤着承受天空的重量
(jīn shǔ sī chàn zhàn zhe chéng shòu tiān kōng de zhòng liàng)

直到，布纹里渗出的咸涩结晶
(zhí dào, bù wén lǐ shèn chū de xián sè jié jīng)
直到，纤维绷紧成透明的琴弦
(zhí dào, xiān wéi bēng jǐn chéng tòu míng de qín xián)
直到某夜失踪
(zhí dào mǒu yè shī zōng)
月光里只留下锈蚀的勒痕
(yuè guāng lǐ zhǐ liú xià xiù shí de lēi hén)

- GLM-4 Generated Poem:

打磨鹅卵石
(dǎ mó é luǎn shí)

一遍一遍，水流打磨棱角
(yí biàn yí biàn, shuǐ liú dǎ mó léng jiǎo)
一寸一寸，月光削减粗糙
(yí cùn yí cùn, yuè guāng xuē jiǎn cū cāo)
在真正圆润之前
(zài zhēn zhèng yuán rùn zhī qián)
每道伤痕都是印记
(měi dào shāng hén dōu shì yìn jì)

在一道河床深处，感受冲刷的耐心
(zài yí dào hé chuáng shēn chù, gǎn shòu chōng shuā de nài xīn)

直到，所有尖锐都变得柔和
(zhí dào, suǒ yǒu jiān ruì dōu biàn dé róu hé)
直到，深色内里泛起温润的光泽
(zhí dào, shēn sè nèi lǐ fàn qǐ wēn rùn de guāng zé)
直到化作浑圆

(zhí dào huà zuò hún yuán)
把所有过往都沉淀为静默
(bǎ suǒ yǒu guò wǎng dōu chén diàn wéi jìng mò)

The example demonstrates that the LLM-generated poem adheres to the prompt's requirements, imitating the style of the human-written poem while creating distinct titles and content. These examples highlight the challenges of detecting AI-generated poetry and emphasize the significance of our study.

A.3 Data Distribution

Table 8 presents the average number of stanzas, lines, and words per poem for both human-written and LLM-generated poetry, along with the proportion of poems that exceed or fall below these averages relative to their respective total counts (human-written or LLMs-generated). Table 9 shows the frequency of nouns in human-written poems and LLM-generated poems. In addition, we provide the detailed distributions of all human-written and LLMs-generated poems, covering three categories: (1) the number of stanzas per poem and the corresponding count of poems, (2) the number of lines per poem and the corresponding count of poems, and (3) the number of lines per stanza and the corresponding count of stanzas:

- Figures 2 to 6 present the distribution of all human-written and LLMs-generated poems, focusing on the number of stanzas per poem and the corresponding count of poems. The horizontal axis represents the number of stanzas in a single poem, while the vertical axis indicates the number of poems corresponding to a specific stanza count.
- Figures 7 to 11 show the distribution of all human-written and LLMs-generated poems, focusing on the number of lines per poem and the corresponding count of poems. The horizontal axis denotes the total number of lines in a single poem, and the vertical axis represents the number of poems with a specific line count. Among them, the number of lines in a single poem includes the title and the blank lines within the poem.
- Figures 12 to 16 illustrate the distribution of human-written and LLMs-generated poems, focusing on the number of lines per stanza and the corresponding count of stanzas. The horizontal axis indicates the number of lines in a single stanza, while the vertical axis shows the number of stanzas with a specific line count.

A.4 The Analysis for Emotions

Emotions Unlike poems focusing on intrinsic qualities or structures, LLMs-generated poems expressing specific emotions (D_{10-13}) prove more detectable than poems that do not express any emotions (D_9). For instance, the detectors' F1-score improved from 64.73% to over 72.59% when detecting poems expressing emotions. Remarkably, RoBERTa-based detector achieves 100.00% detection accuracy on emotion-related poems (D_9, D_{11-13}) generated by DeepSeek-R1.

This result aligns with real-world poetic practices. Human poets typically write to express emotions, resulting in frequent emotional content in human-written poems. Thus, poems that do not contain any emotions (D_9) can be used as a detection feature. However, Chinese poetry aesthetics emphasize conveying infinite meaning through finite words, which means poets embed emotions implicitly rather than through explicit lexical markers. Human poems rarely contain direct emotional terms (e.g., happiness, sadness, anger, fear, etc). Therefore, the poems literally containing these emotions (D_{10-13}) are easier to be detected than other types of poems.

	Stanzas			Lines			Words		
	Avg.	$P_{>Avg.}$ (%)	$P_{<Avg.}$ (%)	Avg.	$P_{>Avg.}$ (%)	$P_{<Avg.}$ (%)	Avg.	$P_{>Avg.}$ (%)	$P_{<Avg.}$ (%)
Human	2.68	54.50	45.50	18.86	53.50	46.50	171.95	43.38	56.63
GPT-4.1	4.03	37.38	62.62	24.48	52.96	47.04	219.74	49.22	50.78
DeepSeek-V3	3.51	46.07	53.93	19.73	47.93	52.07	139.12	42.91	57.09
DeepSeek-R1	3.83	69.75	30.25	22.15	43.06	56.94	174.77	45.54	54.46
GLM-4	5.41	45.60	54.40	32.69	43.25	56.75	265.90	44.18	55.82

Table 8: The average number of stanzas, lines, and words per poem for both human-written and LLM-generated poetry, along with the proportion of poems that exceed or fall below these averages relative to their respective total counts (human-written or LLMs-generated).

	Human							
Noun	人 (rén)	时 (shí)	风 (fēng)	天空 (tiān kōng)	梦 (mèng)	雨水 (yǔ shuǐ)	城市 (chéng shì)	石头 (shí tou)
Frequency	265	130	125	99	82	80	68	65
	DeepSeek-R1							
Noun	时 (shí)	指纹 (zhǐ wén)	年轮 (nián lún)	褶皱 (zhě zhòu)	月光 (yuè guāng)	玻璃 (bō lí)	青铜 (qīng tóng)	影子 (yǐng zi)
Frequency	8571	4453	4167	3904	3554	2984	2619	2478
	DeepSeek-V3							
Noun	时 (shí)	风 (fēng)	光 (guāng)	月光 (yuè guāng)	雪 (xuě)	信 (xìn)	人 (rén)	月亮 (yuè liang)
Frequency	3471	2155	1989	1867	911	843	824	802
	GLM-4							
Noun	风 (fēng)	世界 (shì jiè)	时间 (shí jiān)	影子 (yǐng zi)	人 (rén)	天空 (tiān kōng)	声音 (shēng yīn)	心 (xīn)
Frequency	4424	3584	2468	2381	2261	2144	2142	1557
	GPT-4.1							
Noun	风 (fēng)	影子 (yǐng zi)	世界 (shì jiè)	人 (rén)	夜色 (yè sè)	梦 (mèng)	光 (guāng)	时间 (shí jiān)
Frequency	5122	3892	2900	2413	2339	2227	2203	2199

Table 9: Frequency of nouns in human-written poems and LLM-generated poems.

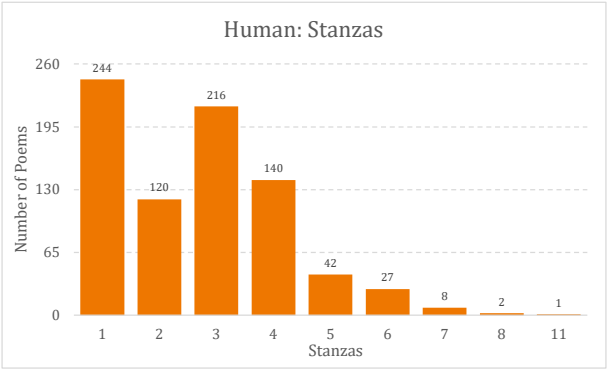


Figure 2: Stanza distribution of human-written poems.

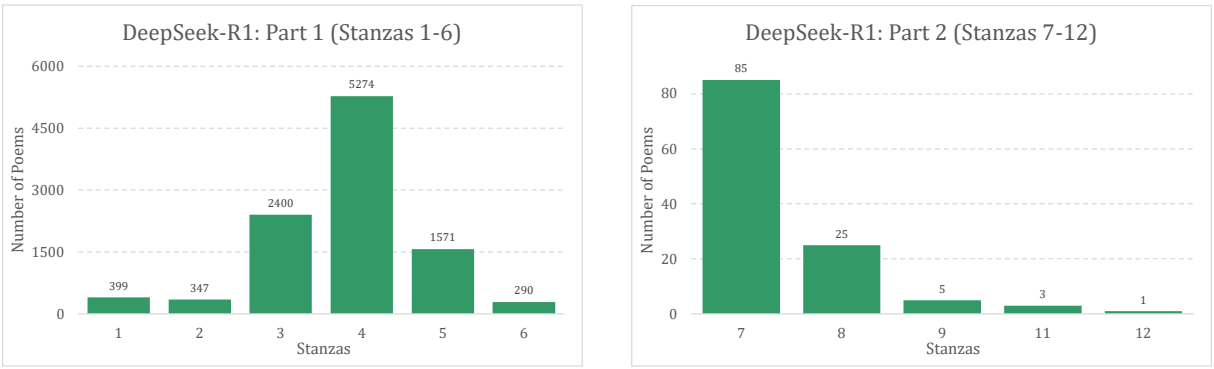


Figure 3: Stanza distribution of DeepSeek-R1-generated poems.

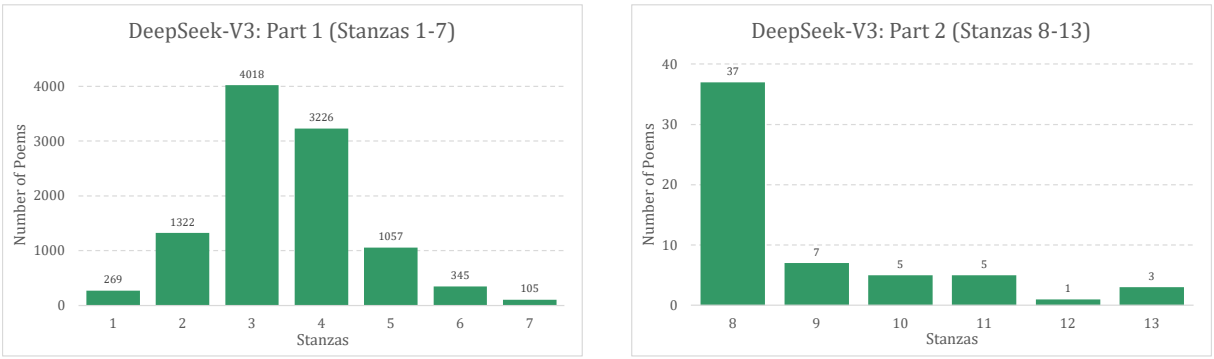


Figure 4: Stanza distribution of DeepSeek-V3-generated poems.

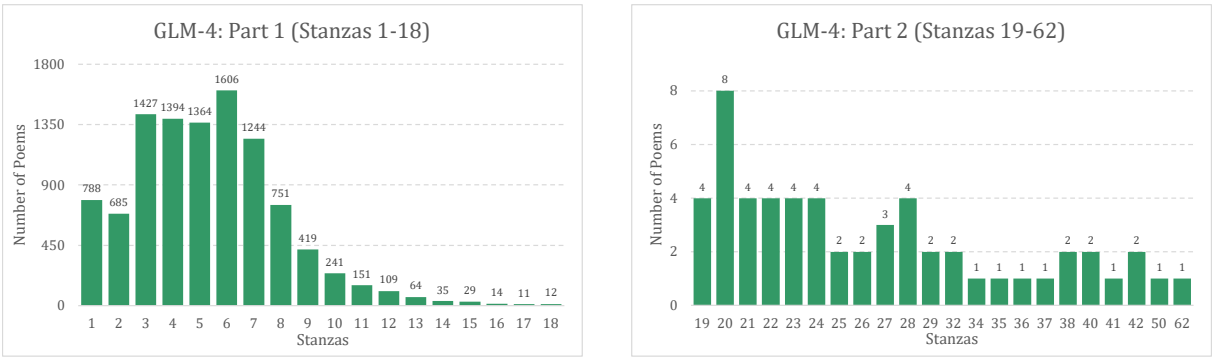


Figure 5: Stanza distribution of GLM-4-generated poems.

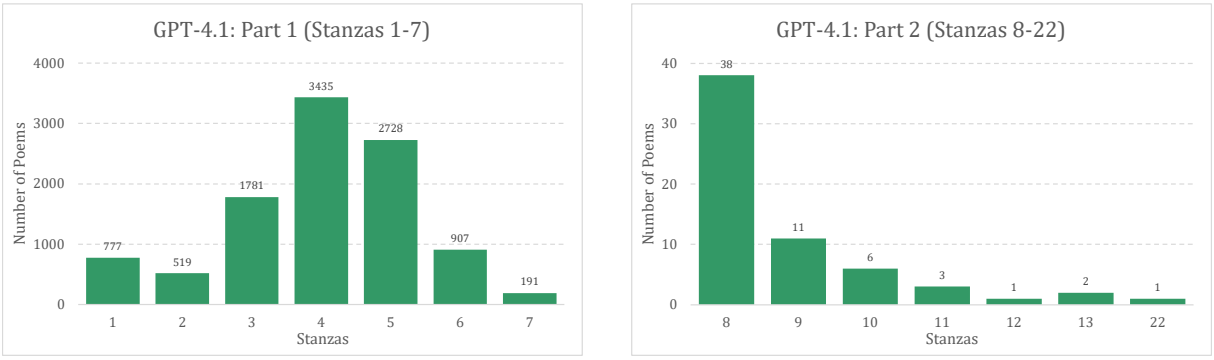


Figure 6: Stanza distribution of GPT-4.1-generated poems.

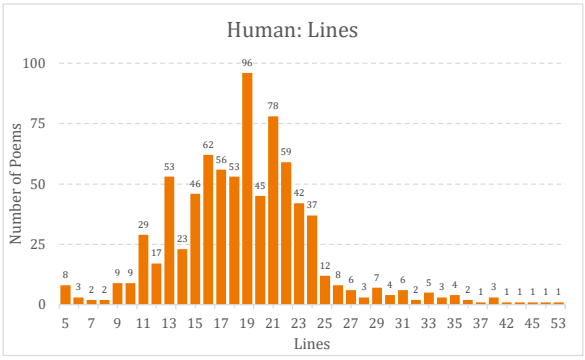
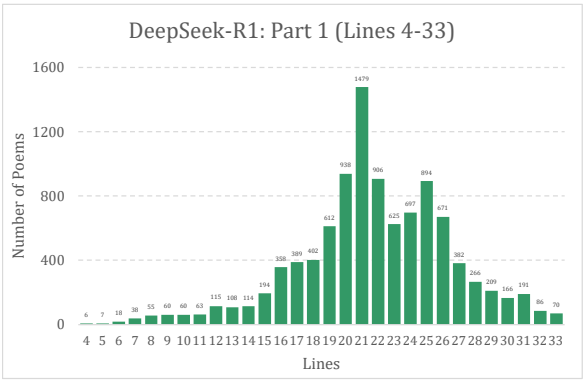


Figure 7: Line distribution of human-written poems.



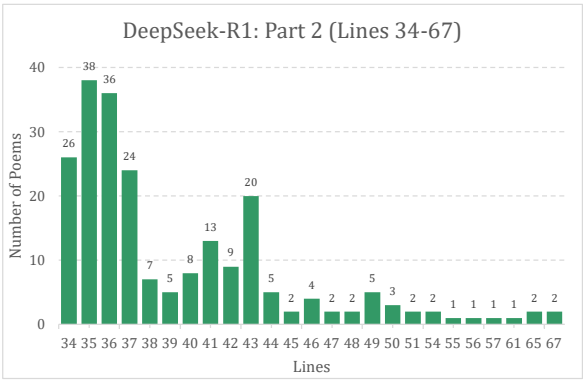
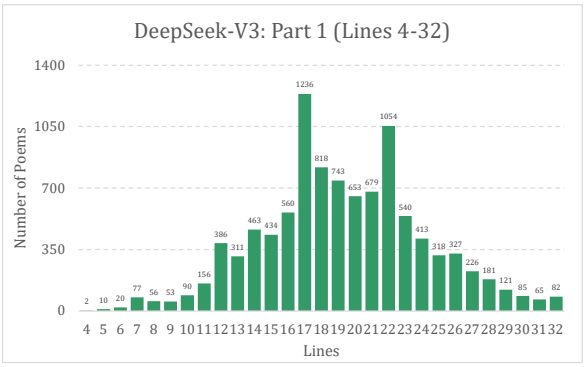


Figure 8: Line distribution of DeepSeek-R1-generated poems.



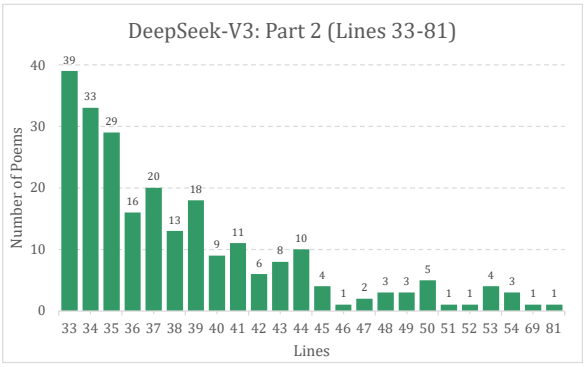
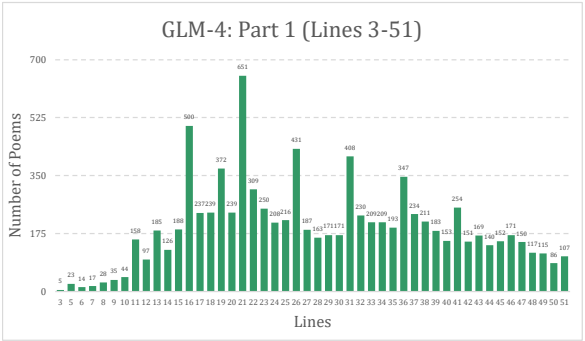


Figure 9: Line distribution of DeepSeek-V3-generated poems.



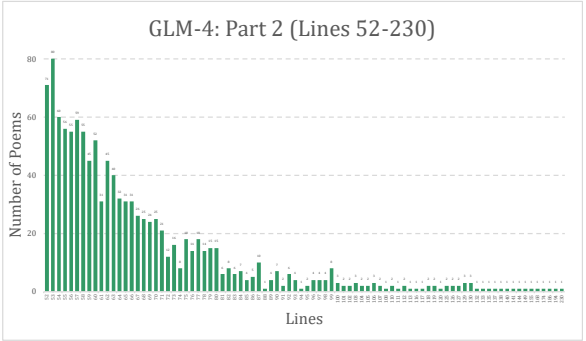
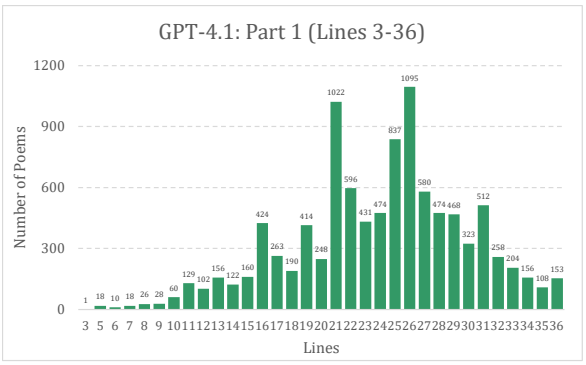


Figure 10: Line distribution of GLM-4-generated poems.



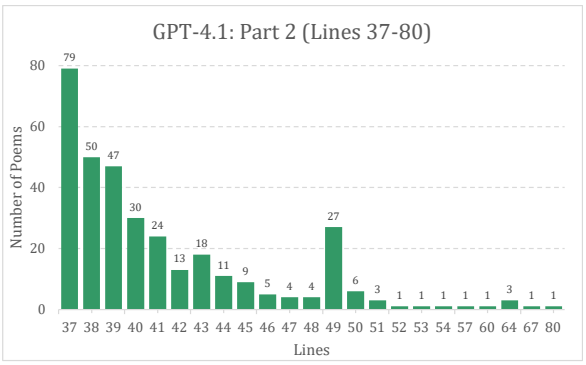
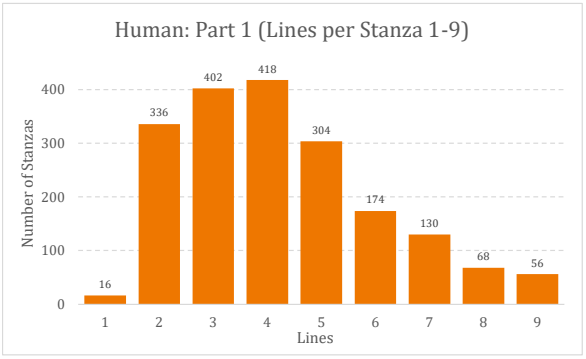


Figure 11: Line distribution of GPT-4.1-generated poems.



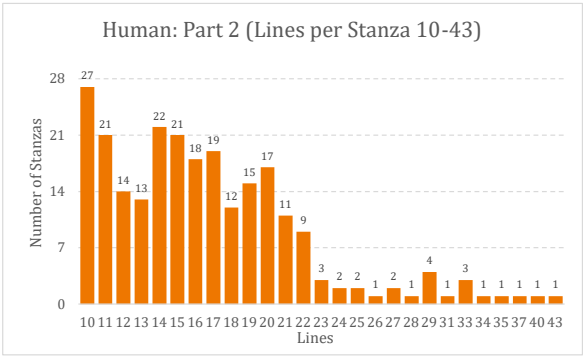
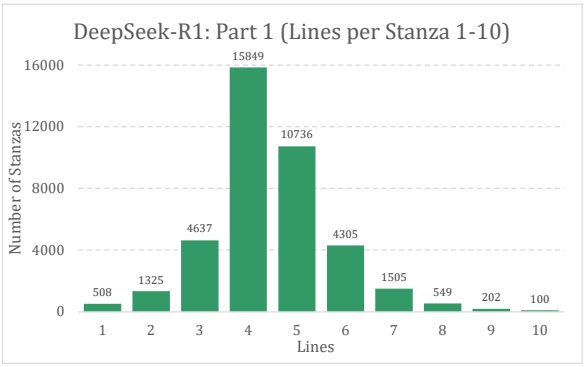


Figure 12: The number of lines per stanza and the corresponding count of stanzas in human-written poems.



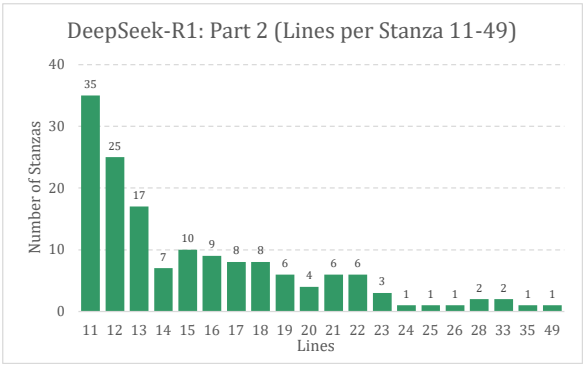
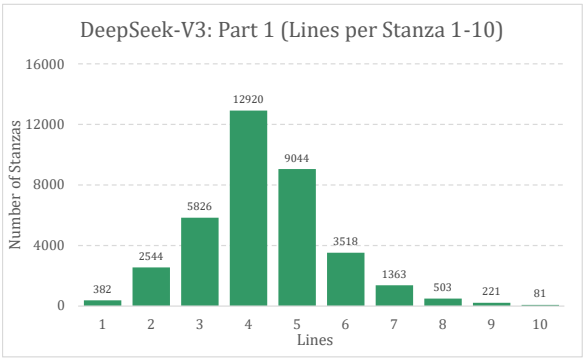


Figure 13: The number of lines per stanza and the corresponding count of stanzas in DeepSeek-R1-generated poems.



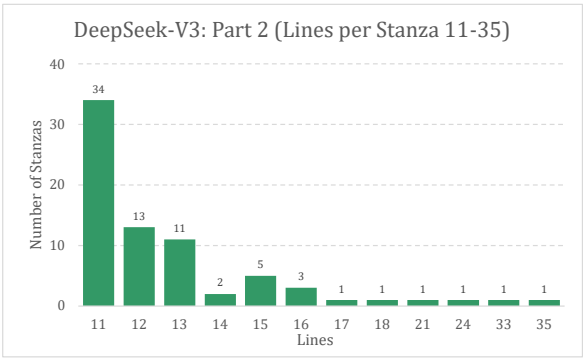
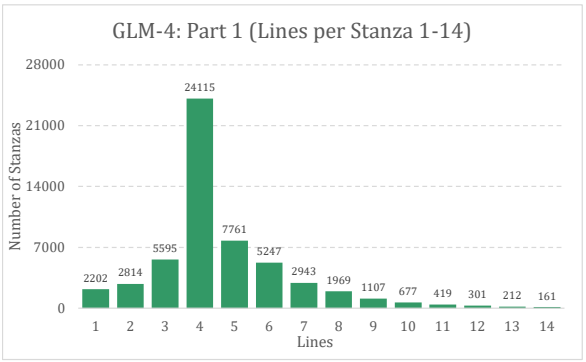


Figure 14: The number of lines per stanza and the corresponding count of stanzas in DeepSeek-V3-generated poems.



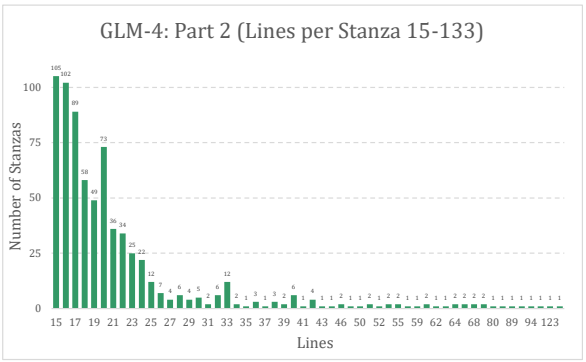
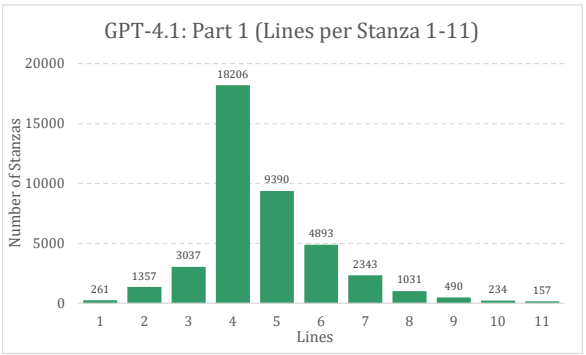


Figure 15: The number of lines per stanza and the corresponding count of stanzas in GLM-4-generated poems.



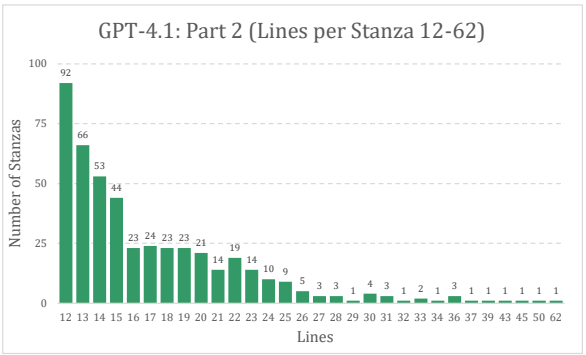


Figure 16: The number of lines per stanza and the corresponding count of stanzas in GPT-4.1-generated poems.