

Seeing the Poem: Image-Semantic Detection of AI-Generated Modern Chinese Poetry with MLLMs

Shanshan Wang¹ Fengying Ye¹ Hanjia Lyu² Caiwen Gou³ Junchao Wu¹
Jingming Yao¹ Chengzhong Xu¹ Jiebo Luo^{4,*} Derek F. Wong^{1,*}

¹University of Macau

²Singapore Management University

³Sichuan University

⁴University of Rochester

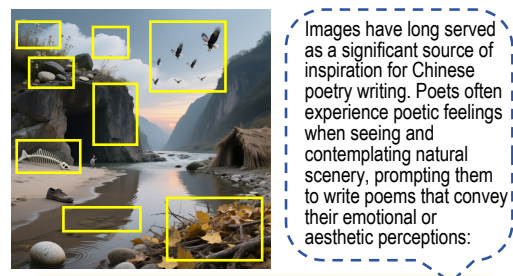
nlp2ct.shanshan@gmail.com, jluc@cs.rochester.edu, derekfw@um.edu.mo

Abstract

Previous detection studies have shown that LLMs cannot be effectively used as detectors, but these studies have not addressed modern Chinese poetry. Moreover, no relevant research has explored the performance of LLMs in detecting modern Chinese poetry. This paper evaluates and enhances the performance of LLMs as detectors for modern Chinese poetry, and proposes an image-semantic guided poetry detection method. Compared with traditional detection approaches, our method innovatively incorporates images that reflect the content of the poetry. Through example-driven approaches, our method effectively integrates information such as meaning, imagery, and feeling from the image, then forms a complementary judgment with the poem text. Experimental results demonstrate that the LLM detectors based on our method outperform baseline detectors based on plain text, and even surpass the best-performing traditional detector, RoBERTa. The Gemini detector using our method achieves a Macro-F1 score of 85.65%, reaching the state-of-the-art level. The performance improvements of different LLM detectors on multiple LLMs-generated data prove the effectiveness of our method.

1 Introduction

Large Language Models (LLMs) are increasingly being applied to a wide range of tasks, including poetry (Antar, 2023; Wang et al., 2024; Zhang and Eger, 2024; Wang et al., 2026a; Lan et al., 2026; Wang et al., 2026b). The rapid advancement of LLMs has significantly improved the quality of AI-generated text, making it increasingly similar to human-written content (Lai et al., 2024). While this progress demonstrates notable technical achievements, it has also raised concerns about the growing difficulty in distinguishing AI-generated text from genuine human-written text (Jakesch



Empty Valley

When the river waters dips low
They pull the sky to the lowest
Stones and wildflowers at the top of the cliff
Are the scattered stars falling to the earth
Pebbles echo the eagles on the white clouds
In the open space of the beach propped up by fish bones
Running water writes down concession and forbearance
Yellow leaves and withered branches...

Figure 1: The relationship between human poetry and images. In this example, key nouns from the poem are underlined, and their corresponding visual elements are highlighted with yellow bounding boxes in the image.

et al., 2023; Porter and Machery, 2024; Hayawi et al., 2024; Najjar et al., 2025).

The blurring boundary between human and AI-generated text now poses a particularly pressing challenge, extending even into the domain of poetry (Wang et al., 2025). Poetry, as a precious part of human cultural heritage, carries deep historical and cultural significance, permeating various aspects of our lives comprehensively (Ren et al., 2023; Zhang et al., 2019). With a history spanning thousands of years, it has profoundly influenced the development of different nations and cultures (Yi et al., 2018; Zhao et al., 2022). Alarmingly, the total amount of AI-generated poetry has been predicted to exceed all human-composed poetry throughout history (Huo, 2020).

In China, since the release of DeepSeek-R1 (Guo et al., 2025), the quantity of LLMs-generated poetry has grown rapidly. Some individuals now

*Corresponding author.

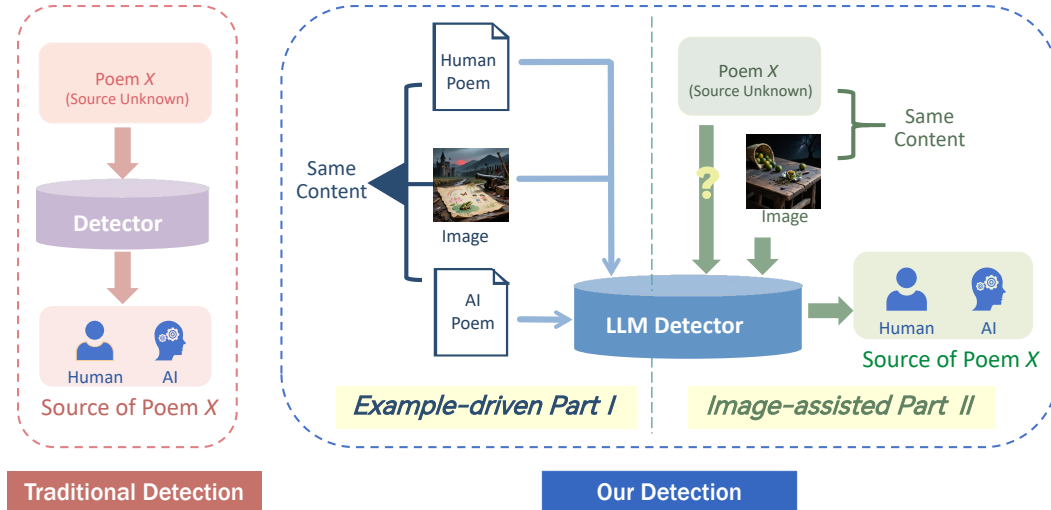


Figure 2: Comparison between the framework of the traditional detection method and our proposed IMAGINE method.

use LLMs to imitate the style of famous poets to generate modern Chinese poems and submit them to regular literary journals. This phenomenon creates considerable confusion and disruption for editors, readers, and even professional poets, as humans often struggle to distinguish AI-generated poems from those written by humans (Köbis and Mossink, 2021; Jakesch et al., 2023; Darda et al., 2023). Therefore, the development of effective automatic detection methods has become not only necessary but also urgent (Wu et al., 2025b).

However, a recent benchmark study (Wang et al., 2025) has revealed that traditional text detectors such as Fast-DetectGPT, Binoculars, LRR, Log-Rank, Log-Likelihood, and RoBERTa-based classifiers, which perform well in general text domains, are *not* reliable for detecting AI-generated modern Chinese poetry. In fact, most detectors other than RoBERTa achieve less than 70% accuracy in detecting LLM-generated poems.

In the real world, poetry often originates from the poet’s interaction with visual scenes, a principle summarized in the famous concept of “expressing emotions through scenery.” In other words, the poet sees the image first, and then creates the poem. Specifically, poetry is the artistic presentation of subjective emotions inspired by objective external objects (Hu, 2003). Images have long served as a significant source of inspiration for Chinese poetry writing (Cheng et al., 2018; Li et al., 2021). Expressing emotions through scenery is a common method in poetry writing (Wang, 2003; Xiong,

2001). Poets often experience poetic feelings when seeing and contemplating natural scenery (Li et al., 2021), prompting them to write poems that express their emotional or aesthetic perceptions (Xu et al., 2018; Cheng et al., 2018). However, the essence of poetry lies not in describing the objective facts in an image, but in condensing the objects, scenes, and emotions within them to reveal the deeper meaning and poetic essence behind them (Liu et al., 2018). Images, in this context, carry rich information, extending beyond immediate visual details to encompass higher-level semantic themes (Xu et al., 2018). Take Figure 1 as an example, imagery often emerges across scattered lines and through seemingly simple nouns such as river, cliff, eagles, fish bones, and yellow leaves. They collectively form a coherent emotional landscape or artistic conception. This emotional core is rarely stated explicitly, but rather constructed through the subtle interplay of visuals, themes, and affect. Recent work indicates that LLMs can exhibit semantic drift and rely on surface lexical cues in metaphor processing (Ye et al., 2026), further motivating auxiliary semantic signals beyond plain text.

Building on this poetically grounded generation setting where visual and emotional imagery plays a central role, rather than relying solely on textual analysis, models can benefit from visual semantics to better grasp the implicit meanings and emotional depth encoded in poetry. A similar benefit of auxiliary semantic guidance has been observed in poetry translation, where explanation-

assisted prompting improves the performance of LLMs (Wang et al., 2024). Therefore, in this paper, we propose a novel image-semantic guided poetry detection framework based on Multi-modal Large Language Models (MLLMs), termed **IMAGINE**, which integrates multimodal information to further refine LLM-based detection.

Specifically, we first provide the model with paired examples of AI- and human-written poems along with corresponding images reflecting the content of the poem. The model then receives both the poem under examination and a related image to guide its classification. This framework leverages image-semantic guidance to deepen the model’s understanding of poetic nuance and origin.

Experimental results demonstrate that our image-semantic guided method yields substantial improvements in poetry detection task. Gemini, when enhanced with our approach, achieves a state-of-the-art Macro-F1 score of 85.65% across all LLM-generated and human-written poems. Furthermore, different categories of LLM-based detectors exhibit varying performance when distinguishing between LLM-generated and human-written poems, but our method significantly narrows this performance gap. The consistent improvements observed across both overall and fine-grained metrics demonstrate the robustness and effectiveness of our proposed method. Our approach transforms LLM-based detectors from an unreliable state into viable tools, providing a new paradigm for poetry detection.

Main contributions of our work are as follows:

- We construct a dataset for detecting AI-generated modern Chinese poetry that is grounded in poetic creation practices, comprising 800 human-written poems and 3,200 LLM-generated poems. Complementary to prior datasets, our dataset requires LLMs to retain comprehensive features of human poetry, including theme, meaning, imagery, and emotion/feeling, thus enabling a semantically grounded evaluation setting.
- We propose IMAGINE, a novel image-semantic guided detection framework, which achieves state-of-the-art performance on poetry detection and surpassing all baselines.
- We determine the initial performance of LLMs as a poetry detector and effectively enhance their detection capability through an example-driven approach.

	Poems	Stanzas	Lines	Words
Human	800	2.1K	15.0K	137.5K
GPT-4.1	800	2.3K	15.5K	173.5K
GLM-4	800	2.2K	15.1K	132.7K
DeepSeek-V3	800	2.8K	15.3K	119.3K
DeepSeek-R1	800	3.1K	16.6K	172.2K

Table 1: The statistics of our poem data.

2 Related Work

Previous research has achieved notable progress in detecting LLM-generated text (Chen et al., 2025), but these approaches face limitations when applied to modern Chinese poetry. Wu et al. (2024) proposed a detection method based on grammatical error correction scores, achieving effective results across various real-world text sources. However, this method is not applicable to modern Chinese poetry because modern poetry does not require adherence to grammatical rules (Blohm et al., 2018; Chen, 2012; Deng, 2007). Köbis and Mossink (2021) attempted a novel Turing test by increasing judges’ motivation to improve the accuracy of their judgments. They assessed human ability to distinguish rhymed, rhythmic English poems generated by GPT-2, including human preferences for model-generated and human-written poems. However, the results demonstrated that humans could not accurately identify AI-generated poems. Nguyen et al. (2023) proposed a machine learning classification method based on handcrafted feature engineering and a feature detection method based on text similarity to detect human-written and synthetically generated content. The first method extracts multi-dimensional features such as lexical, syntactic, semantic, and readability features, achieving an F1 score as high as 99.93% using models such as Random Forest (Kam, 1995) and XGBoost (Chen and Guestrin, 2016). The second method, through inverse generation and cosine similarity calculation, can effectively distinguish between human-generated and ChatGPT-generated text.

3 Dataset

Our dataset consists of 800 human-written poems, 3,200 LLMs-generated poems, and 800 images corresponding to the human-written poems. Unless otherwise stated, all poems mentioned in this paper refer to modern Chinese poetry.

Human-written Poems The human-written poems employed in our work are sourced directly

from the AIGenPoetry dataset (Wang et al., 2025), which comprises 800 modern Chinese poems authored by professional poets. These poems feature well-documented provenance and exhibit diversity in style, form, and theme, ensuring both high quality and representativeness for our study.

LLMs-generated Poems We follow Wang et al. (2025) and employ four LLMs including GPT-4.1 (OpenAI) (Achiam et al., 2023), GLM-4 (Zhipu AI) (ZHIPU, 2024), DeepSeek-V3 (Liu et al., 2024), and DeepSeek-R1 (Guo et al., 2025) to generate one corresponding poem for each human-written poem. For poem generation, we use the recommended hyperparameters from DeepSeek-R1 (Liu et al., 2024): temperature set to 1.5 and top_p to 0.95. This results in a total of 3,200 LLMs-generated poems. The specific prompt used is provided in Table 12. Table 1 presents the summary statistics of our poem dataset.

Corresponding Images To support image-semantic guided detection, we associate each human-written poem with a representative image. We first select 30 poems from the dataset, covering diverse styles. Next, six mainstream image generation models are prompted according to our designed prompt (shown in Table 13) to generate one corresponding image per poem. Thus, each model generates a total of 30 images. These models include DALL-E 3 (Betker et al., 2023), GPT-Image-1 (OpenAI, 2025), GPT-4o-image (Hurst et al., 2024), Doubao, Qwen-image (Wu et al., 2025a), and Hidream (Cai et al., 2025).

To determine the best model for large-scale generation, we invite professional poets to evaluate the images produced by different models. Based on their assessments, we select Qwen-Image as the image generation model and prompt it to generate one image for each of the 800 human-written poems. We also invite **professional poets** to review all the images to ensure their semantic alignment with the corresponding text. This guarantees that the images are of high quality and faithfully reflect the content of the poems.

4 Text-Only Detection

In this section, we aim to empirically demonstrate the limitations of text-only detectors in reliably distinguishing between human-written and LLM-generated modern Chinese poetry.

Gen. →	GPT	GLM	V3.1	R1	Avg.
Fast-Detec.	33.33	51.96	62.73	33.28	45.33
LRR	33.83	39.31	48.18	33.33	38.66
Log-Likeli.	33.61	45.03	48.93	33.33	40.23
Log-Rank	33.61	43.56	47.17	33.33	39.42
Binoculars	34.32	61.73	69.08	33.83	49.74
RoBERTa	89.44	73.31	78.44	96.74	84.48
Avg.	43.02	52.48	59.09	43.97	49.64

Table 2: The performance (Macro-F1, %) of various traditional detectors. Det. and Gen. represent Detector and Generator respectively. GPT, GLM, V3.1, and R1 represent GPT-4.1, GLM-4, DeepSeek-V3.1, and DeepSeek-R1, respectively. In the first column, Fast-Detec. and Log-Likeli. represent Fast-DetectGPT and Log-Likelihood, respectively.

Gen. →	GPT	GLM	V3.1	R1	Avg.
Detector →	Gemini				
$\mathcal{TP}_1$	81.58	74.92	69.70	54.90	70.28
$\mathcal{TP}_2$	86.86	77.46	69.31	84.35	79.50
Det. →	Qwen				
$\mathcal{TP}_1$	33.44	34.80	33.16	33.44	33.71
$\mathcal{TP}_2$	44.15	47.55	33.73	42.64	42.02
Detector →	GPT-5				
$\mathcal{TP}_1$	74.33	69.23	46.07	48.08	59.43
$\mathcal{TP}_2$	74.75	70.70	49.54	76.89	67.97

Table 3: The performance (Macro-F1, %) of LLM-based plain-text detectors (Gemini, Qwen, GPT-5) in distinguishing between poems generated by different LLMs and human-written poems.

4.1 Traditional Text-Only Detection

Following Wang et al. (2025), we evaluate a total of six traditional detectors, including Fast-DetectGPT (Bao et al., 2023), LRR (Su et al., 2023), Log-Likelihood (Solaiman et al., 2019), Log-Rank (Gehrmann et al., 2019), Binoculars (Hans et al., 2024), and RoBERTa-based classifier (Liu et al., 2019). The implementation details are provided in Appendix A.4.

We apply these detectors to the dataset described in Table 1. Both the testing and training sets consist of 400 pairs of poems, with each pair containing one LLM-generated poem and one human-written poem. Table 2 presents the performance (Macro F1-scores) of traditional detectors while Tables 18 through 23 provide detailed results of different detectors, respectively, including the Macro-F1 and F1 scores for both AI-generated (F1_AI) and human-written poems (F1_Human).

Evidently, traditional detectors remain inca-

You are a professional expert in analyzing modern Chinese poetry, skilled at distinguishing between human-written and AI-generated poems.

The following is an image and two modern Chinese poems, where the AI poem was generated by AI based on the human-written poem. The AI poem shares identical title, theme, meaning, imagery, mood/emotion, and other content with the human poem, and retains all the nouns from the human poem, differing only in its writing style. The image was also generated by AI in a realistic style based on the human poem, with particular attention paid to the poem’s title, theme, meaning, imagery, and mood/emotion. It embodies all the noun elements present in the human poem.

Image: *{{Example of Image}}*

Human Poem: *{{Example of Human Poem}}*

AI Poem: *{{Example of AI Poem}}*

The following is a new image and a new modern Chinese poem generated by AI in a realistic style based on a human poem, with particular attention paid to the poem’s title, theme, meaning, imagery, and emotion/feeling, and it embodies all the noun elements from that human poem. Now, please determine whether the following modern Chinese poem was written by human or generated by AI. Please note that the object of your judgment is the “Poem Text”, not the image. Output only the final judgment as either “Human” or “AI.” The output format should be: Answer: [Human or AI]

Image: *{{Image corresponding to the Poem Text to be judged}}*

Poem Text: *{{Poem text to be judged}}*

Table 4: The image-semantic guided prompt ($\mathcal{IP}_3$) we designed for poetry detection.

pable of reliably identifying LLMs-generated poems. Even the best-performing detector, RoBERTa, achieves a Macro-F1 score of only 73.31% for GLM-4-generated poems.

4.2 LLMs-based Text Detection

We further evaluate the performance of LLM-based text-only detectors under two prompting strategies: vanilla prompting and example-driven prompting.

Initial Detection To evaluate the initial ability of LLMs in distinguishing between human-written and different LLMs-generated poetry, we design a baseline prompt, denoted as $\mathcal{TP}_1$. The complete $\mathcal{TP}_1$ is presented in Table 15. Subsequently, mainstream LLMs including qwen-vl-max-2025-08-13, GPT-5, and gemini-3-flash-preview are prompted to judge whether the poem texts in the dataset are created by AI or humans. Table 3 presents the Macro-F1 scores. We also report the F1-scores for AI poems as positive examples (denoted as F1_AI), and the F1-scores for human poems as positive examples (denoted as F1_Human), as shown in the third row of Tables 24, 25, and 26.

The results show that, except for Gemini, which scores higher than RoBERTa in Macro-F1 (74.92%) for poems generated by GLM-4, other LLM detectors score lower than RoBERTa in F1 for poems generated by different LLMs. This demonstrates that LLM-based plain-text detectors are ineffective at distinguishing between AI-generated poems and human-written poems.

Enhanced Detection We enhance the detection capability of LLMs for poetry texts through an example-driven approach. Specifically, we design a new plain text prompt, denoted as $\mathcal{TP}_2$, as shown in Table 16. Based on the vanilla baseline prompt $\mathcal{TP}_1$, we add a pair of examples containing AI-generated and human-written poems to $\mathcal{TP}_2$, specifying the relationship between the two poems in the examples that their titles, themes, meanings, imagery, and moods/emotions are identical. Next, different LLMs are prompted to determine whether the source of the poem to be judged was AI or human. Table 3 presents the results (Macro-F1) of different LLM-based plain-text detector using $\mathcal{TP}_2$ to judge poems generated by different LLMs and human-written poems. Tables 24, 25, and 26 respectively present the detailed performance, including F1_AI, F1_Human, and Macro-F1.

Experimental results show that our designed example-driven method improves the overall Macro-F1 score to 79.50% (Table 3). When employing $\mathcal{TP}_2$, Qwen, GPT-5, and Gemini achieve varying degrees of improvement in Macro-F1 scores for poetry generated by LLMs other than DeepSeek-V3.1. Particularly for poems generated by DeepSeek-R1, Gemini using $\mathcal{TP}_2$ achieves a 29.45% improvement in Macro-F1 compared to $\mathcal{TP}_1$ (Table 26), while GPT-5 achieves a 28.81% improvement (Table 25). However, LLM detectors still remain unreliable in judging poetry generated by LLMs other than GPT-4.1 and DeepSeek-R1.

5 Image-Semantic Guided Poetry Detection

Images are the significant source of inspiration for Chinese poetry writing (Cheng et al., 2018; Li et al., 2021). Expressing emotion and feeling through scenery is a common method in poetry writing (Wang, 2003; Xiong, 2001). Furthermore, images carry extremely rich information, transcending mere visual details and even delving into semantic themes (Xu et al., 2018).

To more accurately determine the source of po-

Gen. →	GPT	GLM	V3.1	R1	Avg.
Det. →	Gemini				
$\mathcal{IP}_2$	88.61	81.08	68.99	83.86	80.64
$\mathcal{IP}_3$	91.87	81.48	74.51	94.75	85.65
Det. →	Qwen				
$\mathcal{IP}_3$	63.25	61.98	46.46	74.96	61.66
Det. →	GPT-5				
$\mathcal{IP}_3$	76.06	73.13	54.93	77.40	70.38

Table 5: The performance (Macro-F1, %) of LLM-based detectors (Gemini, Qwen, GPT-5) applying our proposed method.

etry, we propose a novel image-semantic guided poetry detection method named **IMAGINE** (IMAge-semaNtic GuIded poEtry detection). Compared to text-only poetry detection, IMAGINE innovatively introduces images that reflect the content of the poem text. IMAGINE consists of two parts:

- In the first part, we provide the model with a pair of examples comprising an AI-generated poem and a human-written poem, along with an image reflecting the content of the poems in the examples. We also inform the model of the relationship between the image and the two poems, the relationship between the two poems themselves, and the specific source of the image and each poem in the examples.
- In the second part, we provide the model with an image reflecting the content of the poem to be judged, informing it of the relationship between the image and this poem, and the specific source of the image. The model is then asked to determine whether the poem to be judged originates from AI or human.

Both the AI-generated poem and the image are generated from the same human-written poem. Figure 2 presents the framework of our proposed method. The complete prompt ($\mathcal{IP}_3$) of IMAGINE is presented in Table 4. Based on $\mathcal{TP}_2$, we also design another image-semantic guided prompt, denoted as $\mathcal{IP}_2$, which is presented in Table 17 of the Appendix A.5. The different focuses and characteristics of all prompts are presented in Table 14. The AI-generated poem share the same title, theme, meaning, imagery, and emotion/feeling as the human poem. At the beginning of each prompt, the model is asked to act as a professional expert in analyzing modern Chinese poetry.

6 Experiment

Section 4.2 shows that Gemini is the best-performing model for distinguishing the source of poem text, while the performance of Qwen is less satisfactory. Therefore, in subsequent main experiments, we employ Gemini as the primary detector, and GPT-5 and Qwen as the secondary detector to implement our proposed IMAGINE method. Specifically, Gemini is utilized to determine the source of poems generated by different LLMs and human written poems through $\mathcal{IP}_2$ and $\mathcal{IP}_3$, while GPT-5 and Qwen are used to judge the source of poems through $\mathcal{IP}_3$. Due to cost constraints, the text data for GPT-5 as a detector consisted of 100 poems randomly sampled from the dataset described in Section 3. The text data used by other LLMs as detectors is exactly the same as that in Section 4.2.

Table 5 presents the performance (Macro-F1) of LLM-based detectors (Gemini, Qwen, GPT-5) applying our proposed method in distinguishing between poems generated by different LLMs and human-written poems. Tables 6, 27, and 28 respectively present the detailed performance (%) of LLM-based detectors (Gemini, Qwen, and GPT), including F1_AI, F1_Human, and Macro-F1.

IMAGINE for classical Chinese poetry To demonstrate that our IMAGINE method is not inherently limited to modern poetry, we conducted additional experiments on classical Chinese poetry. Specifically, we randomly selected 50 classical Chinese poems from the recently published ChangAn dataset (Li et al., 2026). Using Gemini as the detector under both the text-only baseline ($\mathcal{TP}_1$) and our IMAGINE method ($\mathcal{TP}_3$), we obtained the results presented in Tables 7 and 29.

Experimenting with images generated from AI poems We conducted experiments where images were generated from the AI poems instead of the human-written poems. Due to cost constraints, we randomly sampled 50 poems and used Gemini as the detector under $\mathcal{IP}_3$. The results are shown in Tables 8 and 30. Note that the images used in all other experiments except this one were generated from human-written poetry.

CLIP-based Experiment We also conducted the CLIP-based experiment using the following setup: the text encoder is paraphrase-multilingual-MiniLM-L12-v2 and the image encoder is openai/clip-vit-base-patch32. The text

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{IP}_2$	89.00	88.23	88.61	80.21	81.96	81.08
$\mathcal{IP}_3$	91.68	92.06	91.87	78.77	84.19	81.48
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{IP}_2$	64.53	73.44	68.99	83.44	84.29	83.86
$\mathcal{IP}_3$	69.47	79.54	74.51	94.59	94.90	94.75

Table 6: The detailed performance (%) of LLM-based detector (Gemini) applying our proposed method on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

Gen. →	GPT	GLM	V3.1	R1	Avg.
$\mathcal{TP}_1$	57.42	61.56	50.83	56.60	56.60
$\mathcal{IP}_3$	79.87	88.87	76.14	77.00	80.47

Table 7: The performance (Macro-F1, %) of LLM-based detector (Gemini) applying plain-text method $\mathcal{TP}_1$ and our method $\mathcal{IP}_3$ to classical Chinese poetry.

Gen. →	GPT	GLM	V3.1	R1	Avg.
	89.94	80.67	73.13	95.99	84.93

Table 8: The performance (Macro-F1, %) of LLM-based detector (Gemini) applying our method $\mathcal{IP}_3$ (images from AI poems) to modern Chinese poetry.

and image embeddings were then concatenated together and connected to a logistic regression model as the classifier. The training and test sets are identical to those used for the traditional detectors, with the only difference being the addition of images. The experimental results are shown in Table 9.

Experimenting with images unrelated to the poem text We used GPT-5 as the detector ($\mathcal{IP}_3$) and conducted additional experiments with images that were unrelated to the poem text. The experimental results (%) on poems generated by all LLMs are as shown in Tables 10 and 31.

7 Analysis

7.1 Effectiveness of Our Method

Our proposed IMAGINE method ($\mathcal{IP}_3$) effectively improves the detection performance of LLMs for poetry by integrating images.

Gemini Detector Table 5 shows that the overall performance of the LLM detector based on our method ($\mathcal{IP}_3$) is superior to the baseline LLM detector based on plain text ($\mathcal{TP}_1$) and the enhanced plain text detector ($\mathcal{TP}_2$), and even surpasses the best-performing traditional detector,

Gen. →	GPT	GLM	V3.1	R1	Avg.
	85.12	71.24	78.32	96.37	82.76

Table 9: The experimental results (Macro-F1, %) of CLIP-based experiment.

Gen. →	GPT	GLM	V3.1	R1	Avg.
	70.24	70.89	46.34	67.84	63.83

Table 10: The performance (Macro-F1, %) of LLM-based detector (GPT-5) applying our method $\mathcal{IP}_3$ with images that were unrelated to the poem text.

RoBERTa. Specifically, the Gemini detector using our method ($\mathcal{IP}_3$) achieves an overall Macro-F1 score of 85.65% for all poems generated by LLMs and poems written by humans, reaching the state-of-the-art level. This represents a 15.37% improvement over the baseline ($\mathcal{TP}_1$, 70.28%) and a 6.15% improvement over the enhanced plain text detector ($\mathcal{TP}_2$, 79.50%) under the same settings but without added images. Furthermore, it is noteworthy that the overall Macro-F1 score of our method ($\mathcal{IP}_3$) is also superior to RoBERTa (84.48%) by 1.17%, and surpasses all other detectors (Binoculars) except RoBERTa by at least 35.91%. These improvements demonstrate that the rich information contained in the images is beneficial for the model’s judgment of poetic text, strongly proving the effectiveness of our proposed IMAGINE method. Similarly, Gemini also achieves an overall improvement in detection performance compared to the baseline ($\mathcal{TP}_1$) after using $\mathcal{IP}_2$ with added images corresponding to the poems, with the overall Macro-F1 increasing from 70.28% to 80.64%.

For individual LLM-generated poems and human-written poems, the Gemini detector using our method also achieves surprising results. For example, Gemini as a detector achieves a Macro-F1 of 94.75% for detecting DeepSeek-R1-generated

poems and human-written poems, exceeding the baseline ($\mathcal{TP}_1$, 54.90%) by 39.85%. Furthermore, Gemini using $\mathcal{IP}_3$ outperforms RoBERTa in detecting poems generated by GPT-4.1 and GLM-4, as well as human-generated poems. Especially for the judgment of GLM-4 and human-generated poems, Macro-F1 score of $\mathcal{IP}_3$ (81.48%) is 8.17% higher than RoBERTa.

Qwen Detector While Qwen’s detection performance for LLM-generated and human-written poems remains unreliable, with an average Macro-F1 score of only 61.66%, its overall detection performance and its performance for individual LLM-generated and human-written poems are significantly better than traditional text-based ($\mathcal{TP}_1$) methods (Table 3).

GPT Detector Compared to its text-only performance ($\mathcal{TP}_1$, average Macro-F1 of 59.43%), $\mathcal{IP}_3$ brings an improvement of 10.95%. The gains are particularly notable on GPT-4.1 and DeepSeek-R1 generated poems, with Macro-F1 reaching 76.06% and 77.40% respectively.

The strongest (Gemini), weakest (Qwen), and intermediate (GPT-5) LLM detectors based on text detection methods all show varying degrees of performance improvement after using our proposed IMAGINE method. This again highlights the effectiveness of images in poetry detection, suggesting that deeper information beyond visual details plays a crucial role. Similar to poetry translation, where style injection helps preserve poeticity (Song et al., 2023), image semantics can serve as auxiliary information to help the model capture the implicit meanings of poetry. Using our method, the model effectively integrates information such as theme, meaning, imagery, and emotion/feeling from the image, forming complementary judgments with the text. For example, nouns like river, sunset, fallen leaves, thatch, and worn-out shoes in a poem might be distributed across different lines, and the poet uses the imagery represented by these nouns to construct a artistic conception. This artistic conception itself contains the poet’s feelings and emotions, but is not explicitly presented in the text. However, when these nouns or images appear together in a single scene or image, they naturally become a unified whole and are more visually impactful, evoking emotional responses or deep reflection. This is the most important reason why we require AI-generated poetry and images to retain or reflect the nouns found in human poetry. This also aligns

with the role of visual knowledge in multimodal reasoning (Jiang et al., 2025).

CLIP-based Method The CLIP-based method achieves an average Macro-F1 of 82.76%, which is below both RoBERTa (84.48%) and our proposed IMAGINE method (85.65%). The gap is especially pronounced on the most challenging generator, GLM-4 (71.24% vs. 81.48%). This result is consistent with our previous conclusions, further demonstrating the effectiveness of our method.

7.2 Meaningful Findings

Performance Differences Among LLM Detectors. Different LLMs exhibit varying detection performance for poetry. Table 3 shows that Qwen and GPT-5 have limited initial detection capabilities, especially when processing poetry from DeepSeek-V3.1 and DeepSeek-R1. This indicates that not all LLMs are inherently capable of reliable poetry text identification, and therefore not all LLMs are suitable for use as poetry detectors.

Detection Differences between AI and Human Poetry. Different types of detectors exhibit differences in their performance when dealing with poems generated by LLMs and human-written poems. Specifically, traditional detectors, including RoBERTa (Table 18), LRR (Table 20), Log-Likelihood (Table 21), Log-Rank (Table 22), and Binoculars (Table 23), are better at identifying poems generated by LLMs because their average F1_AI score is higher than their F1_Human score. Conversely, Fast-DetectGPT (Table 19) is better at detecting human-written poems. However, there are extreme cases: Fast-DetectGPT scores 0 on F1_AI for poems generated by GPT-4.1 and DeepSeek-R1, while LRR, Log-Likelihood, and Log-Rank score 0 on F1_AI for poems generated by DeepSeek-R1, but achieve an F1_Human score of 66.66%. This indicates that traditional detectors, except for RoBERTa, tend to classify poems generated by DeepSeek-R1 as human when they cannot identify them. The misclassification of a large number of LLM poems as human poems reflects the severe inadequacy of traditional detectors in capturing the features of AI-generated poetry.

Unlike traditional detectors, LLMs exhibit completely opposite performance. As shown in Tables 26, 24, and 25, Gemini, Qwen, and GPT-5 are all better at identifying human-written poems, with significantly higher F1_Human scores than F1_AI scores ($\mathcal{TP}_1$).

However, promisingly, after applying our method (Table 27, $\mathcal{IP}_3$), Qwen’s F1_AI scores ($\mathcal{TP}_1$) for poems generated by GPT-4.1, DeepSeek-V3, and DeepSeek-R1 improved to 63.43%, 38.17%, and 76.99%, respectively. Furthermore, it is worth noting that the LLM detector using our method ($\mathcal{IP}_3$) significantly reduces the gap between the F1_AI score and the F1_Human score compared to the baseline ($\mathcal{TP}_1$).

The improvements in both overall and detailed performance across different LLM detectors repeatedly underscore the effectiveness of our method. In real-world deployment, detectors may face distribution shifts across generators and styles, and test-time adaptation methods like CAP-TTA (Shen et al., 2026) could further enhance robustness.

7.3 Failure Case

Taking the poem titled "Shadows on the Stone Wall (青石墙上的影子)" generated by DeepSeek-V3.1 as an example (Table 11), Gemini incorrectly identified it as human-written. Our inspection confirms that the corresponding image is correct, so the error does not stem from image quality or image-text mismatch. Instead, the issue appears to lie in Gemini’s tendency to associate anthropomorphic expressions with human authorship, while overlooking signs such as the poem’s weak internal logic and overly unfamiliar phrase combinations. For example, the phrase "using withered branches to correct the shadow’s deviation (用枯枝校对影子的偏差)" feels overly abrupt. Although human poets may also use unconventional vocabulary, such expressions are usually integrated in a way that strengthens the poem’s aesthetic and emotional coherence. These failures are consistent with evidence that LLMs struggle with context-sensitive literary interpretation (Zhang et al., 2026).

7.4 Practicality

Experimental results (Tables 7) applying our method to classical poetry demonstrate three critical findings: 1) The IMAGINE method achieves a remarkable average Macro-F1 of 80.47% on classical poetry, an improvement of 23.87% over the text-only baseline. 2) The magnitude of improvement is even larger than what we observed on modern poetry (15.37% improvement for Gemini), suggesting that the image-semantic guidance principle may be particularly powerful for the condensed and imagery-dense nature of classical Chinese poetry. 3) The highest score reaches 88.87% for GLM-

Poem Title: 青石墙上的影子
Poem Content:
青苔在石缝里签下暗绿的名字
紫荆用枯枝校对影子的偏差
月光和夕阳轮流调色
把空白的墙面涂成谜面
十年才挪动一寸的谜底
工笔的承诺碎在墨水瓶底
紫荆突然抖落所有笔画
风清点着散佚的偏旁
光斑在墙头不断修改遗嘱
青石的沉默里藏着拓印的年轮

Table 11: The poem analyzed in failure case.

generated poems, further establishing state-of-the-art performance across poetic genres. These findings demonstrate that our IMAGINE approach is not limited to modern poetry but can be effectively extended to classical poetry as well.

Furthermore, the results of experiments using images generated from AI-generated poems (Table 8) are highly consistent with those using images generated from human poems (Table 5). This confirms that our method does not depend on privileged access to human-poem-derived images. In practical applications, images can be generated directly from the poem (whether human or AI-generated) to be evaluated, and the detector will still benefit from multimodal integration.

These findings enhance the practical applicability of our proposed IMAGINE method. The core significance of IMAGINE lies in transforming LLM-based detectors from largely unreliable into viable tools. Its purpose is not to replace traditional detectors, but to provide a new solution with more accurate detection for AI-generated poetry.

8 Conclusion

In this paper, we evaluate and enhance the performance of LLMs as detectors for modern Chinese poetry and propose an image-semantic guided poetry detection method (IMAGINE) based on LLMs. Compared with traditional detection methods, our method innovatively introduces images reflecting the content of the poems. Experimental results demonstrate that the LLMs detector based on our method outperforms plain text-based LLMs detectors and even surpasses the best-performing traditional detector, RoBERTa. The Gemini detector using our method achieves a comprehensive Macro-F1 score of 85.65% for all LLMs-generated poems and human-written poems, reaching the state-of-the-art level. Our work provides a new paradigm for LLM-based poetry detection.

Limitations

Our method draws inspiration from the way humans write poetry, where poets often express their emotions and thoughts by writing poems based on images or scenes they see in reality. Moreover, poetry typically contains numerous images represented by nouns. Therefore, incorporating images during detection can effectively enhance model performance. However, our method may not be fully applicable to other types of text, such as news articles or academic papers.

Ethics Statement

The human poems used in this study are derived from existing datasets, and all authors of these poems have authorized our detection work to use the data.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), 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 Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- D Antar. 2023. [The effectiveness of using chatgpt4 in creative writing in arabic: Poetry and short story as a model](#).
- 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*.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. [Improving image generation with better captions](#). *Computer Science*, 2(3):8.
- 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.
- Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, and 1 others. 2025. [Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer](#). *arXiv preprint arXiv:2505.22705*.
- Tianqi Chen and Carlos Guestrin. 2016. [Xgboost: A scalable tree boosting system](#). In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- 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. 2012. [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.
- Wen-Feng Cheng, Chao-Chung Wu, Ruihua Song, Jianlong Fu, Xing Xie, and Jian-Yun Nie. 2018. [Image inspired poetry generation in xiaoice](#). *arXiv preprint arXiv:1808.03090*.
- 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.
- 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, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- 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.
- Min Hu. 2003. On the imagery beauty of scene-emotion blending in classical chinese poetry. *Jiangxi Social Sciences*, (4):3.
- Junming Huo. 2020. Clone li bai and 100 trillion poems - the "quasi-text" production and possible prospects of ai poetry. *Southern Literature*, (4):5.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- 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.
- Tianxiang Jiang, Sheng Xia, Yicheng Xu, Linqun Wu, Xiangyu Zeng, Limin Wang, Yu Qiao, and Yi Wang. 2025. Vknowu: Evaluating visual knowledge understanding in multimodal llms. *arXiv preprint arXiv:2511.20272*.
- H Tin Kam. 1995. Random decision forests. *proceedings of 3rd international conference on document analysis and recognition*.
- 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.
- Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.
- Tian Lan, Shanshan Wang, Zehua Duo, Jiang Li, Guanglai Gao, Derek F Wong, and Xiangdong Su. 2026. Do large language models perform well on comprehending poetic logic in modern chinese poetry? *arXiv preprint arXiv:2608.21827*.
- Haotian Li, Jiatao Zhu, Sichen Cao, Xiangyu Li, Jiajun Zeng, and Peng Wang. 2021. Poetic expression through scenery: sentimental chinese classical poetry generation from images. In *International Conference on Database Systems for Advanced Applications*, pages 629–637. Springer.
- Jiang Li, Tian Lan, Shanshan Wang, Dongxing Zhang, Dianqing Lin, Guanglai Gao, Derek F. Wong, and Xiangdong Su. 2026. Who wrote this line? evaluating the detection of LLM-generated classical Chinese poetry. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5395–5413. Association for Computational Linguistics.
- 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*.
- Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 783–791.
- 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.
- 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*.
- Trung T Nguyen, Amartya Hatua, and Andrew H Sung. 2023. How to detect ai-generated texts? In *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0464–0471. IEEE.
- OpenAI. 2025. Gpt-image-1 model documentation.
- 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.
- Xuelian Ren, Xiaolong Chai, and Mingzhi Mao. 2023. Generating chinese poetry from images based on deep learning. In *Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management*, pages 134–138.
- Hanwen Shen, Ting Ying, Jiajie Lu, and Shanshan Wang. 2026. Preconditioned test-time adaptation for out-of-distribution debiasing in narrative generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20071–20104. Association for Computational Linguistics.
- 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. Asia-Pacific Association for Machine Translation.

- 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*.
- Deming Wang. 2003. [The three stages in the formation of the technique of expressing emotions through describing scenes in early ancient chinese poetry creation](#). *Journal of Baise University*, (5):42–46.
- 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. Association for Computational Linguistics.
- Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2026a. [Can ChatGPT really understand Modern Chinese poetry?](#) In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4152–4162. Association for Computational Linguistics.
- Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2026b. [Poller: Are LLMs suitable for evaluating poetry understanding task?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9720–9735. Association for Computational Linguistics.
- Shanshan Wang, Junchao Wu, Fengying Ye, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2025. [Benchmarking the detection of LLMs-generated Modern Chinese poetry](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9533–9552. Association for Computational Linguistics.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. [Qwen-image technical report](#). *arXiv preprint arXiv:2508.02324*.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025b. [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. 2024. [Who wrote this? the key to zero-shot llm-generated text detection is gecscore](#). *arXiv preprint arXiv:2405.04286*.
- Zhengrong Xiong. 2001. [Why so many spring sorrows and autumn regrets: A preliminary exploration of the creative mentality of ancient chinese poets](#). *Journal of Huainan Normal University*, (4):2.
- Linli Xu, Liang Jiang, Chuan Qin, Zhe Wang, and Dongfang Du. 2018. [How images inspire poems: Generating classical chinese poetry from images with memory networks](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Fengying Ye, Shanshan Wang, Lidia S. Chao, and Derek F. Wong. 2026. [Probing semantic alignment, lexical invariance, and syntactic influence in LLM metaphor processing](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27914–27932. Association for Computational Linguistics.
- Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018. [Chinese poetry generation with a working memory model](#).
- Daniel Zhang, Bo Ni, Qiyu Zhi, Thomas Plummer, Qi Li, Hao Zheng, Qingkai Zeng, Yang Zhang, and Dong Wang. 2019. [Through the eyes of a poet: Classical poetry recommendation with visual input on social media](#). In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 333–340.
- Ran Zhang and Steffen Eger. 2024. [Llm-based multi-agent poetry generation in non-cooperative environments](#). *arXiv preprint arXiv:2409.03659*.
- Ran Zhang, Steffen Eger, Arda Tezcan, Wei Zhao, Simone Paolo Ponzetto, and Lieve Macken. 2026. [Beyond reproduction: A paired-task framework for assessing LLM comprehension and creativity in literary translation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40862–40884. Association for Computational Linguistics.
- Jiaqi Zhao, Ting Bai, Yuting Wei, and Bin Wu. 2022. [Poetrybert: Pre-training with sememe knowledge for classical chinese poetry](#). In *International conference on data mining and big data*, pages 369–384. Springer.
- ZHIPU. 2024. [Zhipu ai devday glm-4](#). Accessed: 2024-05-01.

A Appendix

A.1 Prompt for Generation

Table 12 presents the prompt we designed for poem generation, and Table 13 presents the prompt we designed for image generation.

A.2 Focuses and Characteristics of Different Prompts

Table 14 presents the focuses and characteristics of different prompts we designed for poetry detection.

A.3 The Role of Image

In the real world, poetry often originates from the poet’s interaction with visual scenes, a principle summarized in the famous concept of “expressing emotions through scenery.” In other words, the poet sees the image first, and then creates the poem. This is the inspiration for our proposed method, whose design principles are as follows:

Please create a new modern Chinese poem. It must share the same title as the poem provided below, and its theme, meaning, imagery, emotion/feeling, etc., must be completely identical to the original. Furthermore, it must retain all the nouns from the provided poem but be composed in a completely different writing style. The output should contain only the poem itself, without any parsing, analysis, or indication of poetic genre.

Poem: $\{\{P_i\}\}$

Table 12: The prompt for poem generation.

1) Image assists models in understanding implicit poetic meanings.

In our method, both the AI-generated poems and images originate from the same human-written poem, explicitly requiring the preservation of the poem’s theme, meaning, imagery, and emotions. Specifically, first, as described in Section 3, we require the AI-generated poems to retain all nouns from the human poem (Table 12), while the images must depict all noun elements present in the human poem (Table 13). Then, we invite professional poets to review the generated AI poems and images. This ensures that the images and AI-generated poems are semantically consistent with the same source (human poem).

Our experimental results and analysis show that the detector captures the deeper, more nuanced meanings between AI and human poetry with the aid of images.

2) Images provide a consistent reference. All test instances in our work (including both human and AI poems) use images derived from the same human poem. In other words, the images don’t provide any privileged information about authorship because they correspond identically to both human and AI poems. Therefore, our successful detection stems from the model’s ability to integrate multi-modal information to better understand the deeper meaning of the poems.

As we discussed in Section 7.1, nouns like “river,” “sunset,” and “fallen leaves” may be scattered across different lines of a poem, but when they appear together in the same image, they form a coherent visual scene that evokes emotional resonance. This allows the model to grasp the implicit emotions and deeper meanings of the poem, which are not explicitly expressed in the text. This deeper understanding derived from the image is key to improving detection performance.

Please generate a realistic-style image for the following modern Chinese poem, paying special attention to its title, theme, meaning, imagery, and emotion/feeling. The image must depict all the noun elements present in the poem.

Important Note: It is absolutely prohibited to include any form of text in the image—including Chinese characters, letters, numbers, symbols, or any text-like elements. Any textual content would severely degrade the image quality. You must avoid generating any text 100%.

Poem: $\{\{P_i\}\}$

Table 13: The prompt for image generation.

A.4 Implementation Details of Traditional Detectors

For traditional statistics-based detection methods, including Log-Likelihood, Log-Rank, and LRR, we utilize Qwen2.5-3B¹ as the scoring model to extract logits-based features. For Fast-DetectGPT, we use Qwen2.5-3B as the reference model and Qwen2.5-7B² as the scoring model. In the case of Binoculars, Qwen2.5-7B serves as the observer model, while Qwen2.5-7B-Instruct³ acts as the performer.

Regarding the RoBERTa-based classifier, we fine-tune a pre-trained Chinese RoBERTa model⁴. The model is fine-tuned for 3 epochs with a batch size of 16 and a learning rate of 1e-6.

A.5 Prompts for Poetry Detection

Tables 15, 16, and 17 respectively present the different prompts we designed for poetry detection.

A.6 Detailed Results

Traditional Detectors Tables 19, 20, 21, 22, and 23 present the detailed performance of different traditional detectors, respectively.

LLM-based Plain-text Detectors Tables 24, 25, and 26 respectively present the detailed performance (%) of LLM-based plain-text detectors (Qwen, GPT-5, Gemini) in distinguishing between poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

LLM-based Detectors Applying Our Method

Tables 6, 27, and 28 respectively present the

¹<https://huggingface.co/Qwen/Qwen2.5-3B>

²<https://huggingface.co/Qwen/Qwen2.5-7B>

³<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

⁴<https://huggingface.co/hfl/chinese-RoBERTa-wm-ext>

Prompts	Characteristics
$\mathcal{TP}_1$	The model is required to determine whether the poem to be judged originates from AI or human. $\mathcal{TP}_1$ is used as the baseline.
$\mathcal{TP}_2$	1) The model is first provided with a pair of examples containing one AI-generated poem and one human-written poem, along with the relationship between the two poems in the example. 2) Then, the model is required to determine whether the poem to be judged originates from AI or human.
$\mathcal{IP}_2$	Based on $\mathcal{TP}_2$. 1) The model is first provided with a pair of examples containing one AI-generated poem and one human-written poem, an image reflecting the content of the poems in the example, the relationship between the image and the two poems , and the relationship between the two poems . 2) Then, the model is provided with an image reflecting the content of the poem to be judged , and the relationship between this image and this poem . Subsequently, the model is required to determine whether the poem to be judged originates from AI or human.
$\mathcal{IP}_3$	Based on $\mathcal{TP}_2$. 1) The model is first provided with a pair of examples containing one AI-generated poem and one human-written poem, an image reflecting the content of the poems in the example, the relationship between the image and the two poems , and the relationship between the two poems . Furthermore, the model is informed of <i>the detailed source of image and each poem in the example</i> . For example, the image is generated by AI in a realistic style based on human-written poem. 2) Then, the model is provided with an image that reflects the text content of the poem to be judged , the relationship between this image and this poem , and <i>the detailed source of the image</i> . Subsequently, the model is required to determine whether the poem to be judged originates from AI or human.

Table 14: The focuses and characteristics of different prompts we designed for poetry detection.

You are a professional expert in analyzing modern Chinese poetry, skilled at distinguishing between human-written and AI-generated poems. Now, please determine whether the following modern Chinese poem was written by human or generated by AI. Output only the final judgment as either "Human" or "AI." The output format should be: Answer: [Human or AI]

Poem Text: *{{Poem text to be judged}}*

Table 15: The text-only prompt ($\mathcal{TP}_1$) we designed for poetry detection.

detailed performance (%) of LLM-based detectors (Gemini, Qwen, and GPT) applying our proposed method on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

Table 29 present the detailed performance (%) of LLM-based detector (Gemini) applying plain-text method $\mathcal{TP}_1$ and our proposed method $\mathcal{IP}_3$ to classical Chinese poetry, including F1_AI, F1_Human, and Macro-F1.

Table 30 present the detailed performance (%) of LLM-based detector (Gemini) applying our proposed method (images from AI poems), including F1_AI, F1_Human, and Macro-F1.

Table 31 present the detailed performance (%) of LLM-based detector (GPT-5) applying our proposed method with images that were unrelated to the poem text, including F1_AI, F1_Human, and Macro-F1.

You are a professional expert in analyzing modern Chinese poetry, skilled at distinguishing between human-written and AI-generated poems.

The following are two modern Chinese poems with identical title, theme, meaning, imagery, and emotion/feeling. One is written by human, and the other is generated by AI.

Human Poem: *{{Example of Human Poem}}*

AI Poem: *{{Example of AI Poem}}*

Now, please determine whether the following modern Chinese poem was written by human or generated by AI. Output only the final judgment as either "Human" or "AI." The output format should be:

Answer: [Human or AI]

Poem Text: *{{Poem text to be judged}}*

Table 16: The text-only prompt ($\mathcal{TP}_2$) we designed for poetry detection.

You are a professional expert in analyzing modern Chinese poetry, skilled at distinguishing between human-written and AI-generated poems.

The following is an image and two modern Chinese poems, with the image reflecting the majority of the textual content of both poems. The two poems share identical title, theme, meaning, imagery, emotion/feeling. One poem was written by human, and the other was generated by AI.

Image: *{{Example of Image}}*

Human Poem: *{{Example of Human Poem}}*

AI Poem: *{{Example of AI Poem}}*

The following is a new image and a new modern Chinese poem, with the image reflecting the majority of the textual content of this poems. Now, please determine whether the following modern Chinese poem was written by human or generated by AI. Please note that the object of your judgment is the "Poem Text", not the image. Output only the final judgment as either "Human" or "AI." The output format should be:

Answer: [Human or AI]

Image: *{{Image corresponding to the Poem Text to be judged}}*

Poem Text: *{{Poem text to be judged}}*

Table 17: The image-semantic guided prompt ($\mathcal{IP}_2$) we designed for poetry detection.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	90.21	88.68	89.44
GLM-4	74.55	72.08	73.31
DeepSeek-V3.1	80.41	76.48	78.44
DeepSeek-R1	96.67	96.83	96.74
Avg.	85.46	83.52	84.48

Table 18: The detailed performance (%) of RoBERTa on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	0.00	66.66	33.33
GLM-4	53.4	50.52	51.96
DeepSeek-V3.1	59.56	65.9	62.73
DeepSeek-R1	0.00	66.56	33.28
Avg.	28.24	62.41	45.33

Table 19: The detailed performance (%) of Fast-DetectGPT on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	66.67	0.99	33.83
GLM-4	66.09	12.53	39.31
DeepSeek-V3.1	51.35	45.01	48.18
DeepSeek-R1	0.00	66.66	33.33
Avg.	46.03	31.30	38.66

Table 20: The detailed performance (%) of LRR on poems generated by different LLMs and human-written poems.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	66.72	0.5	33.61
GLM-4	64.95	25.1	45.03
DeepSeek-V3.1	60.1	37.76	48.93
DeepSeek-R1	0.00	66.66	33.33
Avg.	47.94	32.51	40.23

Table 21: The detailed performance (%) of Log-Likelihood on poems generated by different LLMs and human-written poems.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	66.72	0.5	33.61
GLM-4	65.22	21.91	43.56
DeepSeek-V3.1	60.18	34.16	47.17
DeepSeek-R1	0.00	66.66	33.33
Avg.	48.03	30.81	39.42

Table 22: The detailed performance (%) of Log-Rank on poems generated by different LLMs and human-written poems.

	F1_AI	F1_Human	Macro-F1
GPT-4.1	66.67	1.97	34.32
GLM-4	65.61	57.86	61.73
DeepSeek-V3.1	63.75	74.41	69.08
DeepSeek-R1	66.67	0.99	33.83
Avg.	65.68	33.81	49.74

Table 23: The detailed performance (%) of Binoculars on poems generated by different LLMs and human-written poems.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{TP}_1$	0.50	66.39	33.44	2.93	66.67	34.80
$\mathcal{TP}_2$	20.69	67.61	44.15	27.29	67.81	47.55
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{TP}_1$	0.00	66.33	33.16	0.50	66.39	33.44
$\mathcal{TP}_2$	3.66	63.80	33.73	18.26	67.02	42.64

Table 24: The detailed performance (%) of LLM-based plain-text detectors (Qwen) in distinguishing between poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{TP}_1$	72.21	76.44	74.33	66.30	72.15	69.23
$\mathcal{TP}_2$	77.27	72.22	74.75	74.73	66.67	70.70
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{TP}_1$	29.14	62.99	46.07	32.50	63.65	48.08
$\mathcal{TP}_2$	46.26	52.82	49.54	78.50	75.27	76.89

Table 25: The detailed performance (%) of LLM-based plain-text detectors (GPT-5) in distinguishing between poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{TP}_1$	80.63	82.52	81.58	72.03	77.80	74.92
$\mathcal{TP}_2$	86.49	87.24	86.86	74.93	80.00	77.46
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{TP}_1$	64.78	74.62	69.70	42.16	67.64	54.90
$\mathcal{TP}_2$	63.94	74.68	69.31	83.70	84.99	84.35

Table 26: The detailed performance (%) of LLM-based plain-text detectors (Gemini) in distinguishing between poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{IP}_3$	63.43	63.07	63.25	62.93	61.03	61.98
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{IP}_3$	38.17	54.76	46.46	76.99	72.93	74.96

Table 27: The detailed performance (%) of LLM-based detectors (Qwen) applying our proposed method on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{IP}_3$	79.03	72.83	76.06	77.97	68.29	73.13
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{IP}_3$	53.13	56.73	54.93	78.87	75.94	77.40

Table 28: The detailed performance (%) of LLM-based detectors (GPT-5) applying our proposed method on poems generated by different LLMs and human-written poems, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{TP}_1$	69.77	45.07	57.42	74.63	48.48	61.56
$\mathcal{IP}_3$	81.48	78.26	79.87	90.09	87.64	88.87
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{TP}_1$	61.16	40.51	50.83	68.75	44.44	56.60
$\mathcal{IP}_3$	71.60	80.67	76.14	76.77	77.23	77.00

Table 29: The detailed performance (%) of LLM-based detector (Gemini) applying plain-text method $\mathcal{TP}_1$ and our proposed method $\mathcal{IP}_3$ to classical Chinese poetry, including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{IP}_3$	90.74	89.13	89.94	83.19	78.16	80.67
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{IP}_3$	77.97	68.29	73.13	96.15	95.83	95.99

Table 30: The detailed performance (%) of LLM-based detector (Gemini) applying our proposed method (images from AI poems), including F1_AI, F1_Human, and Macro-F1.

F1-score →	F1_AI	F1_Human	Macro-F1	F1_AI	F1_Human	Macro-F1
Generator →	GPT-4.1			GLM-4		
$\mathcal{IP}_3$	75.00	65.48	70.24	75.11	66.67	70.89
Generator →	DeepSeek-V3.1			DeepSeek-R1		
$\mathcal{IP}_3$	43.39	49.29	46.34	70.09	65.59	67.84

Table 31: The detailed performance (%) of LLM-based detector (GPT-5) applying our proposed method with images that were unrelated to the poem text, including F1_AI, F1_Human, and Macro-F1.