Decoding Emotion in Ancient Poetry: Leveraging Generative Models for Classical Chinese Sentiment Analysis

Quanqi Du, Loic De Langhe, Els Lefever and Véronique Hoste
LT3, Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

This study explores the use of generative language models for sentiment analysis of classical Chinese poetry, aiming to better understand emotional expression in literary texts. Using the FSPC dataset, we evaluate two models, Qwen-2.5 and LLaMA-3.1, under various prompting strategies. Initial experiments show that base models struggle with task-specific instructions. By applying different instruction tuning strategies with Low-Rank Adaptation (LoRA), we significantly enhance the models' ability to follow task instructions and capture poetic sentiment, with LLaMA-3.1 achieving the best results (67.10% accuracy, 65.42% macro F1), demonstrate competitive performance against data-intensive, domain-adapted baselines. We further examine the effects of prompt language and multi-task learning, finding that English prompts outperform Chinese ones. These results highlight the promise of instruction-tuned generative models in sentiment analysis of classical Chinese poetry, and underscore the importance of prompt formulation in literary understanding tasks.

1 Introduction

Sentiment analysis is a core task in natural language processing (NLP) that focuses on identifying affective states within text, typically categorizing them into negative, neutral and positive sentiments (Liu, 2020). While the role of emotion in literary comprehension has long been recognized in literary theory and hermeneutics (Hogan, 2011), computational literary studies have only recently begun to systematically incorporate sentiment analysis, especially since the mid-2010s (Rebora, 2023). This shift has opened up new possibilities for analyzing emotional expression in literature at scale, offering novel insights into patterns of affect, authorial voice, and reader reception.

Classical Chinese poetry, an ancient literary form dating back to the 11th century BC, is deeply rooted in emotional expression and aesthetic subtlety (Tian et al., 2024). Investigating how emotions are conveyed in classical Chinese poetry not only helps reveal the poets' inner world, but also reflects broader sociopolitical and philosophical concerns, thus offering deeper insight into traditional Chinese values and cultural paradigms (Zhang et al., 2023c). However, several linguistic and stylistic features – such as strict tonal and syntactic constraints, the use of parallelism and antithesis, and the frequent reliance on indirect emotional expression – render standard sentiment analysis techniques less effective. Compared to modern texts, classical poetry often communicates affect through allusion and symbolic imagery rather than through explicit evaluative language (Meng et al., 2024). Consequently, expert knowledge is often indispensable for accurate interpretation and annotation (Chen et al., 2019).

Previous work on sentiment analysis in classical Chinese poetry has primarily leveraged large pre-trained language models, such as the domain-adapted BERT- and RoBERTa-based architectures (Du and Hoste, 2024). While these models have shown promise, they typically require extensive supervised training on domain-specific labeled corpora – resources that are scarce in the realm of classical literature. Annotating such texts is both time-consuming and cognitively demanding, often requiring interdisciplinary expertise that spans NLP, literary studies, and classical Chinese philology.

In contrast, recent advances in generative models, particularly when combined with instruction tuning and prompt engineering, have the capacity to learn efficiently from only a few labeled examples (Li et al., 2006), showing potential as an alternative to data-hungry approaches (Song et al., 2023). Previous studies have demonstrated the effectiveness

Proceedings of Recent Advances in Natural Language Processing,pages 306–315
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-038

of generative models for sentiment analysis across various domains, including movie and restaurant reviews (Zhong et al., 2021), and financial texts (Fatemi et al., 2025). In the case of classical Chinese poetry, where emotional nuance is often implicit and culturally embedded, sentiment analysis poses unique challenges that go beyond standard classification tasks. Investigating how generative models interpret poetic emotion thus opens new directions for computational literary studies.

In this paper, we explore how generative language models can be applied to sentiment analysis of classical Chinese poetry. We first evaluate zero- and few-shot prompting as a lightweight alternative to fine-tuning. Observing that prompt adherence limits performance, we further examine instruction tuning to enhance output quality. Our findings show that combining prompting with instruction tuning improves accuracy and consistency, advancing sentiment analysis for historical texts and supporting the integration of NLP in literary studies.

2 Related Studies

2.1 Prompt-based learning for sentiment analysis

With the advent of generative models, prompt-based learning has emerged as a key approach in few-shot learning for NLP (Colombo et al., 2023). Unlike traditional fine-tuning, which requires updating a model's parameters, prompt-based learning enables models to generate responses by leveraging their pre-trained knowledge (Liu et al., 2023b). This method utilizes natural language instructions or templates to effectively elicit knowledge from pre-trained language models (PLMs) for downstream tasks such as named entity recognition (NER) (Huang et al., 2022) and question answering (Chen et al., 2024).

Prompt-based learning has also demonstrated strong performance in various sentiment analysis tasks. In aspect-based sentiment analysis (ABSA), which focuses on identifying sentiment towards specific aspects within a text, prompt-based techniques have facilitated multi-task learning and syntactic enhancement. Gao et al. (2022) introduced a unified generative multi-task framework, which constructs task prompts by combining multiple element prompts to handle various ABSA tasks. Similarly, Yin et al. (2024) proposed a syntax-aware enhanced prompt method to effectively extract essential syntactic information related to aspect words.

Prompt-based learning has also facilitated sentiment analysis in low-resource languages. Šmíd and Přibáň (2023) found that prompting yields significantly better results than traditional fine-tuning when applied to Czech sentiment analysis with limited training data. Similarly, Debess et al. (2024) explored different prompt configurations and found that clear task instructions improved sentiment classification performance using GPT-4 on Faroese news texts. Additionally, Nešić et al. (2024) applied prompt-based learning to sentiment analysis of Serbian novels from the 1840-1920 period, achieving an accuracy of 68.2%.

2.2 Sentiment analysis for classical Chinese poetry

Classical Chinese poetry, as a form of ancient literature, shares similarities with low-resource languages in that it lacks large-scale annotated datasets. One key challenge in sentiment analysis is the necessity of expert knowledge (Chen et al., 2019), for example, understanding historical and cultural contexts that influence the meaning and emotional nuances of the text.

To address this limitation, various strategies have been explored to maximize the utility of existing annotated datasets. Some researchers have proposed methods to enhance knowledge transfer and improve model performance. For instance, Hong et al. (2023) aligned classical Chinese poetry with its modern Chinese translations, facilitating effective knowledge transfer from pre-trained models. Similarly, Li et al. (2022) constructed a multimodal knowledge graph that integrates visual information into the semantic learning of classical Chinese poetry.

Beyond incorporating external resources, scholars have also sought to extract richer information directly from the poems themselves. Zhang et al. (2023a) introduced a multi-layer feature extraction approach, while Du and Hoste (2024) leveraged line-level information to refine overall sentiment analysis. Additionally, domain-specific pre-trained models for ancient Chinese, such as BERT-ancient-Chinese (Wang and Ren, 2022), GuwenBERT[1], GujiBERT (Wang et al., 2023), and BERT_CCPoem [2], have been developed to support research on classical Chinese poetry.

[1]https://github.com/ethan-yt/guwenbert
[2]https://github.com/THUNLP-AIPoet/BERT-CCPoem

2.3 Enhancing Prompt-Based Learning with Instruction Tuning

The prompt-based use of generative models holds significant potential for tackling the challenge of sentiment analysis in classical Chinese poetry. However, a major issue with this approach is the inherent misalignment between the objectives of model training and the expectations embedded in user prompts. Specifically, while model training typically aims to minimize response errors across a wide range of general-purpose tasks, user prompting seeks responses that align closely with the given instructions (Fedus et al., 2022).

To bridge this gap, instruction tuning has emerged as a promising strategy for improving both the capabilities and controllability of generative models (Ouyang et al., 2022). This technique involves fine-tuning large language models on curated datasets comprising input-output pairs based on explicit human instructions (Liu et al., 2023a). By learning to follow natural language commands more precisely, instruction-tuned models exhibit enhanced alignment with user intent across a variety of domains (Zhang et al., 2023b). Moreover, recent studies have demonstrated that instruction-tuned models not only generalize better to unseen instructions but also exhibit improved sample efficiency (Luo et al., 2023; Jiang et al., 2024), making them well-suited for domains with limited annotated data, i.e., the classical Chinese poetry.

In the following sections, we explore the feasibility of using prompt-based learning and instruction tuning to enable generative models to support sentiment analysis in classical Chinese poetry.

3 Experiment

Our first experiment consisted in evaluating the performance of prompting techniques alone, as an accessible approach to bootstrap the sentiment analysis process. Subsequently, we incorporated instruction tuning to enhance the models' ability to adhere to sentiment analysis guidelines. The following sections provide a detailed account of our methodology and findings.

3.1 Models

In our experiments, we employed widely used open-source models such as Qwen-2.5 (Yang et al., 2024) and LLaMA-3.1[3]. To ensure a fair comparison, we selected models with similar parameter sizes.

Qwen-2.5 represents the latest iteration of the Qwen large language model series. Compared to its predecessors, it exhibits notable improvements in structured data comprehension and generation[4].

LLaMA-3.1, developed by Meta, is a state-of-the-art open-source model that competes with leading foundation models such as GPT-4, GPT-4o, and Claude 3.5 Sonnet across a range of NLP tasks[5].

We selected Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct for our experiments and did not include DeepSeek-R1-Distill-Llama-8B[6]. While it is the only comparable model in its family, its focus on general reasoning and efficiency, along with potential capacity loss from distillation, makes it less suitable for our experiment.

3.2 Dataset

We utilized FSPC (Chen et al., 2019), a fine-grained sentiment-annotated poetry corpus comprising 5,000 classical Chinese poems as experimental data. Each poem, as shown in Figure 2, is manually annotated by experts both at the line level and poem level with sentiment labels, including *negative*, *implicit negative*, *neutral*, *implicit positive*, and *positive*.

As shown in Figure 1 and exemplified in Figure 2, the negative and positive labels are significantly outnumbered by the implicit sentiment labels. In order to achieve a more balanced distribution of sentiment labels and enable a fair comparison with prior work, we adopt the same preprocessing strategy as the baselines in Section 3.5, merging the *implicit negative* and *negative* class and also the *implicit positive* and *positive* class, resulting in a three-class sentiment scheme: negative (1,756), neutral (1,328), and positive (1,916).

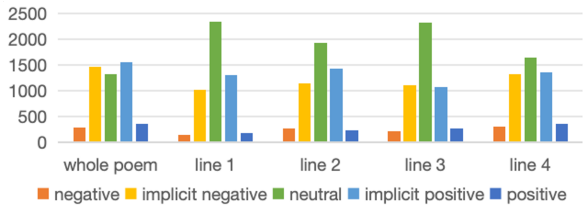

Figure 1: Sentiment label distribution of FSPC dataset.

For the main experiments, we randomly selected a test set of 1,000 poems for performance evalu-

[3]https://huggingface.co/meta-llama

[4]https://github.com/QwenLM/Qwen2.5

[5]https://ai.meta.com/blog/meta-llama-3-1

[6]https://huggingface.co/deepseek-ai/DeepSeek-R1

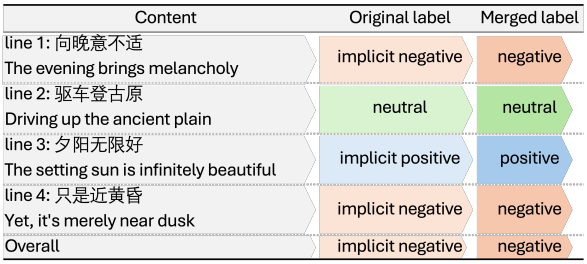

Figure 2: An annotated classical Chinese poem from the FSPC dataset.

ation with the following distribution of sentiment labels at the overall poem level: 372 negative, 258 neutral, and 370 positive. For the prompting experiments, we randomly selected 3 poems from FSPC as exemplars in the few-shot prompt for testing. These three poems are labeled as negative, neutral, and positive, respectively, at the overall sentiment level. For prompt optimization, we used 200 poems to refine the prompt, ensuring better instruction compliance across models.

3.3 Prompt

We designed multiple types of prompts to investigate different aspects of sentiment analysis in classical Chinese poetry. The first type of prompting differentiates between the zero-shot and few-shot approaches, with the latter including three randomly selected poems representing different overall sentiment labels. The second type of approach examines the impact of prompt language, comparing model performance using English and Chinese prompts. The third type of prompting explored the effectiveness of single-task versus multi-task prompting. In the single-task setting, the model predicts a sentiment label for the entire poem, whereas in the multi-task setting, it first assigns sentiment labels to individual lines before determining the overall sentiment based on both the line-level predictions and the poem as a whole. An example template of the prompts is shown in Figure 3.

3.4 Evaluation Matrix

Accuracy and Macro F1 scores were used to evaluate the model's performance. Accuracy provides an overall measure of the model's correctness, particularly useful for balanced class distributions. However, since the sentiment labels in the classical Chinese poetry corpus are somewhat imbalanced, Macro F1 was also used to ensure equal importance is given to all classes, regardless of their frequency.

<System>: You are a sentiment analysis assistant specializing in analyzing the sentiment of Chinese poetry (Positive, Negative, Neutral). The possible answers are "Positive.", "Negative.", and "Neutral.".
(Single-task prompt)
<User>: What is the overall sentiment of the poem? Text: 几度秋深未到家|满湖开遍水红花|逶人日日频中酒|明月扁舟索买瓜
<Assistant>: Label: Positive.
(Multi-task prompt)
<User>: Please analyze this Chinese poem by first determining the sentiment of each line, then provide the overall sentiment. Text: 几度秋深未到家|满湖开遍水红花|逶人日日频中酒|明月扁舟索买瓜
<Assistant>:
Line analysis:
Line 1:几度秋深未到家 sentiment: Neutral.
Line 2:满湖开遍水红花 sentiment: Positive.
Line 3:逶人日日频中酒 sentiment: Positive.
Line 4:明月扁舟索买瓜 sentiment: Positive.
Overall sentiment: Positive.

Figure 3: Examples of prompts.

3.5 Baseline

The results from recent studies on sentiment analysis in classical Chinese poetry using the FSPC dataset, as shown in Table 1. These studies employed BERT- or RoBERTa-based models that were pre-trained on domain-specific corpora related to classical Chinese literature, leveraging domain-specific linguistic patterns to achieve high performance (70.96% accuracy and 68.51% F1). In contrast, our work explores the use of generative models without domain-specific pretraining, highlighting their potential to generalize across literary tasks with minimal supervision.

	Accuracy	F1
Zhang et al. (2023a)	NA	64.38
Hong et al. (2023)	67.10	64.80
Du and Hoste (2024)	**70.96**	**68.51**

Table 1: Benchmark results of the FSPC dataset from domain-specific data-intensive BERT- or RoBERTa-based models pretrained on large-scale classical Chinese texts.

4 Preliminary Results

We first evaluated the performance of the original base models without any tuning by applying zero-shot and few-shot prompting. As shown in Table 2, under the zero-shot setting, Qwen hardly produced 118 valid responses, i.e. responses in which

Model	Prompt	Acc	F1	Valid
Qwen	zero-shot	7.20	9.22	118
	few-shot	29.00	28.31	481
LLaMA	zero-shot	61.20	49.01	1,000
	few-shot	52.00	30.58	946

Table 2: Overall sentiment prediction results of the base models on the evaluation set, which contains 1,000 poems. *Acc* refers to accuracy, *F1* represents the macro-averaged F1-score. *Valid* refers to the number of valid answers generated.

the poem was assigned one of the three polarity labels, while LLaMA generated 1000. In the few-shot prompting setting, Qwen's valid responses increased to 481, whereas LLaMA yielded 946. One possible explanation for LLaMA's reduced number of valid outputs in the few-shot setting is the introduction of biases by the example prompts and the increased input length, both of which may hinder the model's ability to generalize or accurately interpret the task. The substantially higher number of valid outputs from LLaMA in both settings suggests a stronger ability to capture task intent compared to Qwen.

In terms of accuracy and Macro F1 score, LLaMA also consistently outperforms Qwen. Specifically, under the zero-shot prompt, LLaMA achieves an accuracy of 61.20% and a macro F1 of 49.01%, while under the few-shot setting, it obtains 52.00% and 30.58%, respectively. In comparison, Qwen exhibited relatively low performance in the zero-shot configuration, likely due to its high incidence of invalid or off-target responses.

5 Further Experiment with Instruction Tuning

In the preliminary study, it was observed that both Qwen-2.5 and LLaMA-3.1 did not consistently adhere to instructions or generate the desired responses. In some cases, the models' outputs even lacked a sentiment label altogether. This behavior is a known challenge when prompting generative models (Zhang et al., 2023b), and it can significantly affect the accuracy of output evaluation.

To further investigate the models' ability to generate valid responses and potentially improve the accuracy of sentiment analysis, we also applied instruction tuning to both Qwen and LLaMA. To improve a model's ability to understand user intent and follow instructions more precisely, a common approach is to fine-tune it for the specific task. However, given the large number of model parameters and the limited availability of annotated data for supervised fine-tuning, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022), a prominent parameter-efficient tuning technique. LoRA preserves the original model weights while updating only the low-rank matrices, thereby reducing computational overhead. In our experiment, we set the LoRA alpha to 16 and the rank parameter r to 128.

During the instruction tuning phase, two types of instructions were designed, analogous to zero-shot and few-shot prompts. Instruction I contains no annotation examples, whereas instruction II includes three annotated examples – one randomly selected instance from each sentiment category.

For the instruction tuning experiments, the test set partition of 1,000 poems and the prompt data remained unchanged. The remaining corpus of 3,797 poems was used for training, including 1,231 poems with negative labels, 1,018 with neutral labels and 1,458 with positive labels at the overall sentiment level.

6 Results and Discussion

After applying instruction tuning, as shown in Table 3, both Qwen and LLaMA demonstrated improved understanding of prompts, generating 100% valid responses in all cases – except for one instance where Qwen was tuned with instruction II and evaluated with the few-shot prompt. This particular setting yielded the lowest Macro F1 score (34.51%), despite achieving the highest accuracy (57.20%) among Qwen's tuning scenarios. A closer examination revealed that Qwen produced 731 negative responses out of 1,000 in this configuration, resulting in a severe class imbalance. This imbalance plausibly accounts for the notably low F1 score (34.51%).

6.1 Effectiveness of Instruction Tuning

Comparing the results before and after instruction tuning (Table 2 and 3), we observe significant improvements in both accuracy and F1 scores, underscoring the effectiveness of instruction tuning.

Further analysis of the two instruction types reveals consistent performance differences. As shown in Table 3, instruction II, which includes annotated examples, consistently outperforms instruction I. While the performance gap between the two instruction types is relatively small under few-shot prompting, it becomes substantial under zero-shot prompting. For instance, Qwen's accuracy

Model	Tuning	Prompt	Acc	F1	Valid
Qwen	I	zero-shot	50.90	50.90	1,000
	I	few-shot	52.10	51.98	1,000
	II	zero-shot	**57.20**	34.51	999
	II	few-shot	52.50	**53.33**	1,000
LLaMA	I	zero-shot	61.53	57.94	1,000
	I	few-shot	58.60	47.61	1,000
	II	zero-shot	**67.10**	**65.42**	1,000
	II	few-shot	58.90	54.20	1,000

Table 3: Evaluation results of generated overall poem sentiment answers from individual models tuned with different instructions and prompted with different numbers of shots. Instruction type I is without annotation examples, while Instruction type I contains three randomly selected annotation examples, one for each sentiment category. *Acc* refers to accuracy, and *F1* represents F1-Macro.

improves from 50.90% to 57.20%, and LLaMA's accuracy increases from 61.53% to 72.00%, highlighting the advantage of incorporating annotation examples in the instruction during tuning. A plausible explanation is that these examples help the models better distinguish between the three sentiment categories.

The models' performance under different prompting strategies also varies depending on the instruction type. For instruction I, Qwen performs slightly better with few-shot prompting, while LLaMA performs better with zero-shot prompting. The difference is relatively minor in both cases. However, under instruction II, both models exhibit significantly better performance with zero-shot prompting. Qwen's accuracy increases from 52.50% to 57.20%, and LLaMA's from 58.90% to 72.00%. It suggests that the annotation examples in instruction II play a more important role in enhancing zero-shot prompting performance.

When considering the effect of instruction type in the context of few-shot prompting, both models show similar performance regardless of the tuning instruction used. In contrast, under zero-shot prompting, switching from instruction I to instruction II yields a substantial performance boost. This indicates that while annotation examples in the instruction during tuning have limited impact in few-shot prompting, they significantly enhance the models' ability to generalize in zero-shot scenarios.

This discrepancy may be attributed to an interference effect, as "examples don't always help" (Reynolds and McDonell, 2021). In the few-shot prompting scenario, the model tends to rely more on the examples provided in the prompt itself, rather than leveraging the knowledge acquired from instruction tuning. In contrast, under the zero-shot prompting condition, the model depends solely on its pre-trained and instruction-tuned knowledge, as no additional examples are provided at inference time. During instruction tuning, the model likely learns useful task-related representations from the annotated examples included in the instructions. However, in the prompting stage, the few-shot examples may introduce conflicting signals or redundant information, which in turn hampers the model's ability to generalize, thereby diminishing the benefits gained from instruction tuning.

Finally, when comparing the two models, LLaMA consistently outperforms Qwen across all configurations. One exception occurs under instruction I with few-shot prompting, where Qwen achieves a higher F1 score (51.98%) than LLaMA (47.61%). Nevertheless, LLaMA still leads in accuracy (58.60% vs. 52.10%). Overall, LLaMA demonstrates stronger performance than Qwen in the task of sentiment analysis for classical Chinese poetry. Therefore, in the following experiments, we focus solely on LLaMA in zero-shot setting for further analysis.

6.2 English Prompt vs. Chinese Prompt

Since the focus of our research was on Chinese Poetry, we also investigated the viability of using Chinese prompts instead of English prompts. Given that LLaMA demonstrated better performance with the zero-shot prompting approach (Table 3), we took this approach as a starting point and switched the prompt language from English to Chinese. However, as shown in Table 4, the Chinese prompt does not outperform the English prompt, achieving only 61.0% and 63.40% accuracy, with F1-macro scores of 50.20% and 63.52%, respectively. This discrepancy may be attributed to cultural differences between languages and potential language bias present in the pretrained models, as previous experiments also suggest that prompting models in different languages often result in varying performance (Behzad et al., 2024; Agarwal et al., 2024; Vida et al., 2024).

6.3 Single-task vs. Multi-task

Previous work (Du and Hoste, 2024) suggests that a multi-task framework and the inclusion of additional line-level information in the RoBERTa-based model can enhance overall sentiment labeling performance for classical Chinese poems. We initially

Tuning	Prompt	Acc	F1
I	En	61.53	57.94
I	Cn	61.00	50.20
II	En	67.10	65.42
II	Cn	63.40	63.52

Table 4: Comparison of model performance with zero-shot prompts in English and Chinese. *No. in Tuning* means the number of example poems provided in instruction tuning. *Acc* refers to accuracy, and *F1* represents F1-Macro.

designed a configuration that enabled LLaMA to predict sentiment labels for individual lines as well as for the entire poem, aiming to compare its performance with the best-performing model in Table 3. However, this approach did not yield reliable results. Consequently, we adopted an alternative configuration and compared it with the second-best performing model (bold in Table 5). As shown in Table 5, the multi-task approach achieves an accuracy of 60.10% and an F1-macro of 53.26%, which is a decrease in performance compared to the set-up focused solely on overall sentiment prediction (accuracy = 61.53%, F1-macro = 57.94%). This result suggests that the multi-task approach may introduce bias into the model. To gain deeper insights from this experiment, we conduct a thorough error analysis in the following section.

	Acc	F1
Single	67.10	65.42
Single	**61.53**	**57.94**
Multi	60.10	53.26

Table 5: Comparison of LLaMA-3.1 performance in single- and multi-task zero-shot contexts. *Acc* refers to accuracy, and *F1* represents macro-averaged F1.

6.4 Error Analysis

To investigate the reasons behind the incorrect sentiment labels assigned by the model to classical Chinese poems, we perform an error analysis on the results of LLaMA in the instruction II and zero-shot prompt set-up, which achieved the best performance in the experiments.

As shown in Figure 4, the models generally perform better on non-neutral poems than on neutral ones. Among the true positive poems, 293 were correctly identified as positive, accounting for approximately 79.19%. Similarly, 249 of the true negative poems were correctly predicted, yielding an accuracy of 66.94%. In contrast, only 129 out of 258 neutral poems were correctly classified, corresponding to an accuracy of 50%. These results suggest that neutral poems pose greater challenges for sentiment analysis, possibly due to their subtle or context-dependent emotional cues. This difficulty aligns with observations from previous studies using RoBERTa-based approaches (Du and Hoste, 2024). Furthermore, the model often confuses neutral poems with positive ones, suggesting an oversensitivity to emotionally suggestive language even when the overall tone is balanced or ambivalent.

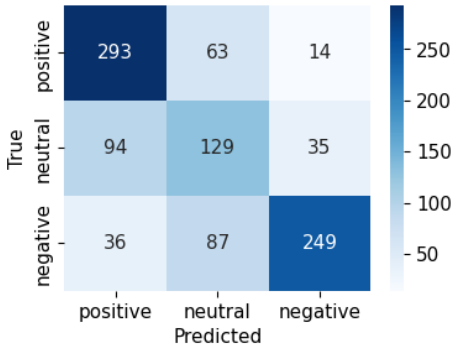

Figure 4: Confusion matrix of the true and predicted overall sentiment labels produced by LLaMA under the instruction II set-up with zero-shot prompting.

In addition to more closely investigating the output of the best-performing single-task model, we also examined the classifications provided in the multi-task setup, which showed improved performance in previous RoBERTa-based studies but did not outperform in our generative model configuration. To further investigate, we conducted an additional error analysis on poems predicted by LLaMA under the multi-task setup.

As shown in Figure 5, the performance of LLaMA under the multi-task setup has similar results to LLaMA in the best performance set-up. However, the model's performance on neutral sentiment is significantly weaker, with only 42 neutral poems correctly identified, accounting for about 16.28%. One possible explanation is that neutral sentiment in classical Chinese poetry may be more ambiguous and context-dependent, making accurate classification challenging. Neutral poems may contain mildly positive elements, such as descriptions of nature, philosophical reflections, or balanced emotional tones, which the model may misinterpret as positive, as illustrated in Figure 6. This confusion between the neutral and positive class also seems to confirm earlier findings in Figure 4.

Another possible explanation is error propaga-

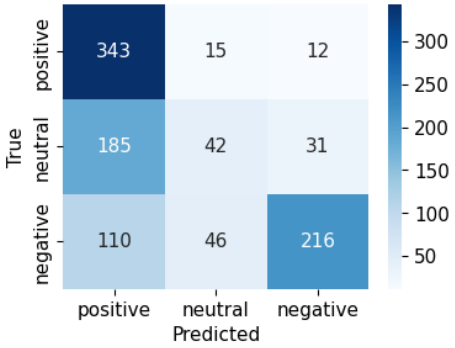

Figure 5: Confusion matrix of the true and predicted overall sentiment labels produced by LLaMA under the multi-task set-up.

	Accuracy	F1-Macro
Line 1	55.80	56.79
Line 2	59.70	58.23
Line 3	60.80	61.01
Line 4	55.10	53.93
Overall	60.10	53.06

Table 6: Evaluation results of generated line and overall sentiment labels from LLaMA-3.1-8B-Instruct.

tion from misclassifications at the line level. If the model makes small errors when classifying individual lines, as shown in Table 6, these errors can accumulate and mislead the overall sentiment predictions. Additionally, sentiment in classical Chinese poetry is often expressed through complex interplays of emotion, imagery, and historical allusions. Analyzing sentiment at the line level may cause the model to overlook these complexities, resulting in less accurate overall predictions. Furthermore, the relationship between line-level and overall sentiment predictions is multifaceted, as different lines of a poem may convey contrasting emotions that contribute to a more complex overarching sentiment, as demonstrated in Figure 6.

Content	True label	Predicted label
line 1: 万顷烟波百尺丝 Vast misty waves, a hundred-foot line	neutral	positive
line 2: 禅家宗旨有谁知 Who truly grasps the Zen school's mind	neutral	neutral
line 3: 自嫌固陋如高叟 I mock myself, stubborn as an old recluse	negative	neutral
line 4: 却为僧笺把钓诗 Yet pen fishing verses for monks to find	neutral	positive
Overall	neutral	positive

Figure 6: An example of true and predicted lines and overall sentiment labels of a classical Chinese poem. English translations are obtained from ChatGPT.

7 Conclusion

This study investigates the potential of generative models for analyzing the sentiment of classical Chinese poetry, focusing on their ability to capture poetic emotion and navigate the linguistic subtlety of this historically rich literary genre. Initial experiments with basic zero- and few-shot prompting revealed that limited instruction adherence significantly constrained model performance. Subsequent instruction tuning significantly improved the models' ability to generate consistent and meaningful sentiment predictions, underscoring their potential for sentiment analysis of classical Chinese poetry. Additionally, our results suggest that prompt language plays a substantial role in performance, with English prompts generally yielding better outcomes than Chinese ones. Multi-task strategies that incorporate line-level analysis introduced biases that degraded overall prediction accuracy, suggesting that more focused modeling may be preferable.

These findings suggest that prompt-based and instruction-tuned generative models offer promising tools for computational literary studies. From a sentiment analysis perspective, our results highlight both the strengths and limitations of using large language models for interpreting affect in poetic texts. While out-of-the-box LLMs like LLaMA-3.1 can achieve over 60% accuracy and near 50% macro F1 in zero- and few-shot settings, instruction tuning clearly enhances their ability to make contextually grounded predictions, achieving an accuracy of 67.10% and an F1 score of 65.42%. Although our results slightly trail domain-specific baselines, those rely on models pre-trained on large-scale classical Chinese texts. In contrast, our approach uses general-purpose models, yet still achieves competitive performance. This underscores the generalization potential of LLMs for literary sentiment analysis.

Future work will include both open-source and closed-source models, with a focus on investigating reasoning mechanisms and chain-of-thought prompting in the experiment of sentiment analysis for classical Chinese poetry.

Acknowledgments

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (FAIR) (174K02325). We extend our thanks the anonymous reviewers for their insightful and constructive feedback.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Cheng Yang, Wenhao Li, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China.

Xiusi Chen, Jyun-Yu Jiang, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Wei Wang. 2024. MinPrompt: Graph-based minimal prompt data augmentation for few-shot question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 254–266, Bangkok, Thailand. Association for Computational Linguistics.

Pierre Colombo, Victor Pellegrain, Malik Boudiaf, Myriam Tami, Victor Storchan, Ismail Ayed, and Pablo Piantanida. 2023. Transductive learning for textual few-shot classification in API-based embedding models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4214–4231, Singapore. Association for Computational Linguistics.

Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2024. Good or bad news? Exploring GPT-4 for sentiment analysis for faroese on a public news corpora. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7814–7824.

Quanqi Du and Veronique Hoste. 2024. A multi-task framework with enhanced hierarchical attention for sentiment analysis on classical Chinese poetry: Utilizing information from short lines. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 113–122, Miami, USA. Association for Computational Linguistics.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. 2025. A comparative analysis of instruction fine-tuning large language models for financial text classification. *ACM Trans. Manage. Inf. Syst.*, 16(1).

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Patrick Colm Hogan. 2011. *What literature teaches us about emotion*. Cambridge University Press.

Jie Hong, Tingting He, Jie Mei, Ming Dong, Zheming Zhang, and Xinhui Tu. 2023. A hybrid corpus based fine-grained semantic alignment method for pre-trained language model of ancient chinese poetry. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4794–4801. IEEE.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. COP-NER: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5421–5434, Bangkok, Thailand. Association for Computational Linguistics.

Fei-Fei Li, Fergus Robert, and Perona Pietro. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Yuqing Li, Yuxin Zhang, Bin Wu, Ji-Rong Wen, Ruihua Song, and Ting Bai. 2022. A multi-modal knowledge graph for classical Chinese poetry. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2318–2326, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36:29615–29627.

Yingying Meng, Yuwei Wan, and Chunyu Kit. 2024. Du fu's conspicuous negativity and li bai's hidden positivity: a sentiment comparison and exploration. *Digital Scholarship in the Humanities*, 39(1):280–295.

Milica Ikonić Nešić, Saša Petalinkar, Mihailo Škorić, Ranka Stanković, and Biljana Rujević. 2024. Advancing sentiment analysis in Serbian literature: A zero and few–shot learning approach using the mistral model. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 58–70, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Simone Rebora. 2023. Sentiment analysis in literary studies. a critical survey. *Digital Humanities Quarterly*, 17(3).

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

Jakub Šmíd and Pavel Přibáň. 2023. Prompt-based approach for Czech sentiment analysis. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1110–1120, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40.

Mengmeng Tian, Qi Jia, Cong Wang, Juwang Yang, and Xin Liu. 2024. Tang Chang'an poetry automatic classification: A practical application of deep learning methods. *Digital Scholarship in the Humanities*, 39(2):756–764.

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. 2023. GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.

Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Wen Yin, Cencen Liu, Yi Xu, Ahmad Raza Wahla, Huang Yiting, and Dezhang Zheng. 2024. SynPrompt: Syntax-aware enhanced prompt engineering for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15469–15479, Torino, Italia. ELRA and ICCL.

Lingli Zhang, Yadong Wu, Qikai Chu, Pan Li, Guijuan Wang, Weihan Zhang, Yu Qiu, and Yi Li. 2023a. SA-Model: Multi-feature fusion poetic sentiment analysis based on a hybrid word vector model. *CMES-Computer Modeling in Engineering & Sciences*, 137(1).

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Wei Zhang, Hao Wang, Min Song, and Sanhong Deng. 2023c. A method of constructing a fine-grained sentiment lexicon for the humanities computing of classical Chinese poetry. *Neural Computing and Applications*, 35(3):2325–2346.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. UserAdapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488, Online. Association for Computational Linguistics.