Beyond Memorization and Recitation: Evaluating LLMs on Deep Understanding of Ancient Chinese Poetry

Anonymous ACL submission

Abstract

Ancient Chinese Poetry (ACP) stands as a brilliant embodiment of cultural heritage, using concise forms to convey profound emotions. While Large Language Models (LLMs) have made rapid progress in mimicking linguistic styles and reciting verses, whether they truly understand the poets' underlying intent remains an open question. Current works primarily focus on knowledge-driven, surface-level understanding, failing to assess the understanding gap between rote memorization and aesthetic appreciation. To address this, we propose CP-DUE (**C**lassical **P**oem – **D**eep **U**nderstanding **E**valuation), a top-down framework that treats poetry comprehension as a five-level progressive process. CP-DUE systematically evaluates LLMs across dimensions ranging from basic cultural facts to precise word choice (*Tui Qiao*), hidden allusions, and overall aesthetic appreciation. Through extensive experiments comparing LLMs with human experts, we reveal that even advanced models struggle with the artistic nuances that define the soul of ACP. This work provides new insights into bridging the understanding gap and enhancing LLMs' competence in genuine cultural connection.

1 Introduction

> "*Poetry is when an emotion has found its thought and the thought has found words.*" ——***Robert Frost***

Ancient Chinese Poetry (ACP) has emerged as a global cultural phenomenon that captivates international audiences through its minimalist aesthetics and profound emotional depth(Liu, 2022; Owen, 2020). This process of artistic appreciation transcends mere information decoding and involves a complex spiritual resonance between the reader and the poet's inner world(Cai, 2008). Recent studies have shown the proficiency of LLMs in generating classical Chinese poems (He et al., 2012; Yi et al., 2020). Consequently, these models have become popular tools for assisting in ACP composition. However, while LLMs can mimic the structure of ACP, whether they can truly perform high-level artistic appreciation and grasp the soul of this medium remains an open question.

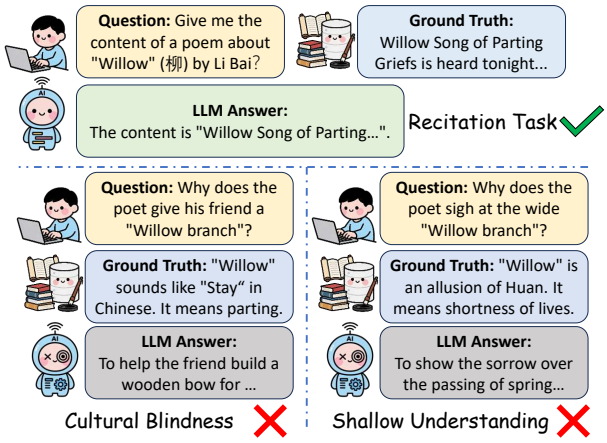

Figure 1: Examples of LLMs' weaknesses in ACP: LLMs can do well in recitation tasks, but often make mistakes in tasks involving deep understanding, leading to vacuous and inaccurate interpretations.

Previous studies on LLMs' abilities in ACP most focus on knowledge-based tasks(Liu et al., 2025; Zhang and Li, 2023; Zhou et al., 2023), which fall short of assessing the core question of whether LLMs can achieve a deep understanding of ACP. However, as shown in Figure 1, the artistic significance of ACP extends far beyond its literal surface. Even subtle contextual variations can imbue a poetic element with vastly different connotations and cultural symbolism(Hightower and Yeh, 2020). Merely retrieving the text of a poem remains a far cry from achieving a profound understanding of its essence. As a cornerstone of Chinese cultural heritage, ACP represents a sophisticated form of literary art that possesses an

aesthetic depth. Therefore, there is an essential need to develop a novel, multidimensional evaluation framework that can more comprehensively and profoundly assess the artistry of LLMs' understanding of ACP.

To overcome the limitations of existing evaluation tasks for ACP, we propose **CP-DUE** (**C**lassical **P**oem –**D**eep **U**nderstanding **E**valuation), a multidimensional and progressive evaluation framework designed to assess the depth of LLMs' understanding of ACP. By adopting this design, our framework measures LLMs' understanding across multiple dimensions, enabling a more thorough and nuanced assessment of whether models attain a profound comprehension of ACP. CP-DUE systematically evaluates LLMs across dimensions ranging from basic cultural facts to precise word choice (*Tui Qiao*), hidden allusions, and overall aesthetic appreciation. Together, these modules form a hierarchically structured evaluation system and ensure a comprehensive and fine-grained measurement of LLMs' true comprehension of ACP, beyond surface-level memorization and imitation.

Next, we conduct experiments on mainstream LLMs and compare their performance with that of humans from different educational backgrounds. By thoroughly analyzing the limitations of LLMs in ACP understanding tasks, we identify some important abilities that can be further improved in this domain. We believe our work provides valuable directions for developing future LLMs with stronger abilities in understanding traditional Chinese culture.

The proposed framework and dataset will be publicly released upon the paper's acceptance.

Our contributions are as follows:

- We propose a novel and more challenging task suite designed to deeply evaluate the understanding abilities of LLMs in the domain of ACP.
- We design and construct **CP-DUE**, a multidimensional and progressive framework that systematically measures LLMs' comprehension across cultural, stylistic, linguistic, and semantic dimensions.
- We conduct comprehensive evaluations on multiple mainstream LLMs and provide in-depth analyses, revealing both their current limitations and potential in achieving deep understanding of ACP, as well as traditional Chinese culture.

2 Related Works

2.1 Datasets of classical Chinese

Several datasets have been introduced to evaluate models on tasks related to Ancient Chinese Culture. Zhou et al. (2023) defines several ancient text tasks, with its ACP-related tasks including line-by-line sentiment (positive/negative) labeling and machine translation. Zhang and Li (2023) also proposes a set of ancient culture tasks, focusing on ACP tasks such as knowledge-based questions, next-line prediction, appreciation (analysis), quality assessment (evaluating informativeness, fluency, and coherence), and sentiment classification (positive, implicit positive, neutral, implicit negative, negative).The Wemind Ancient Culture QA dataset includes ACP tasks like machine translation, recitation/memorization-based question answering, and Theme identification. (Liu et al., 2025) primarily focuses on ACP retrieval, often used in Retrieval-Augmented Generation (RAG).However, current evaluations using these datasets generally remain superficial, failing to deeply assess the models' comprehensive understanding of ACP.

2.2 Researches on Classical Chinese Poetry

The advancement of Large Language Models (LLMs) has spurred increasing research into ACP-related capabilities. Current studies on LLMs' capacity in ACP primarily fall into two categories:Generative Tasks: Works such as Wang et al. (2016); Yi et al. (2018); Zhang and Lapata (2014); Chen et al. (2019) have proposed various methods that successfully address the task of ACP generation. Furthermore, (Zhipeng et al., 2019)introduced datasets suitable for fine-tuning models for poetry generation.Traditional Text Analysis Tasks: THU-FSPC is an early dataset for ACP sentiment analysis, providing positive/negative sentiment labels for each line of quatrains. Studies such as Liu et al. (2019); Li et al. (2021) have explored translation and matching tasks, respectively. Building on this, Chen et al. (2024) conducts detailed machine translation evaluations.While LLMs have shown achievements in both generative and traditional text analysis tasks, it remains unclear whether these models possess a deep, task-independent understanding of ACP.

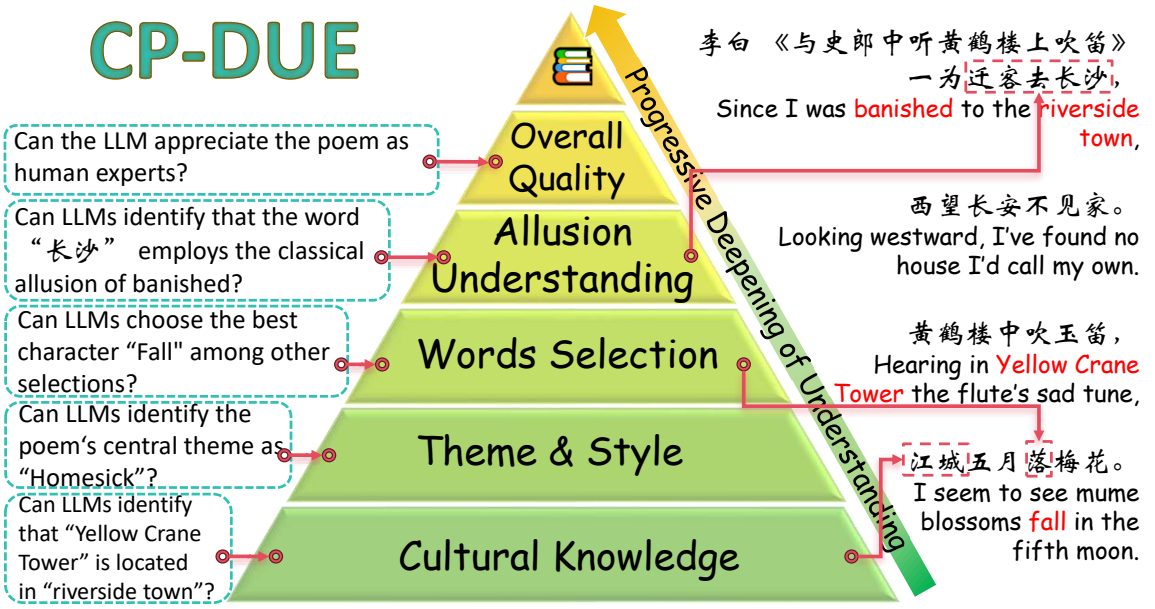

Figure 2: This figure illustrates the evaluation content of the **CP-DUE** framework using a Tang poem (translated by Xu(Xu, 2021)) as an example. Through five progressively deeper layers of assessment, we thoroughly evaluate the LLM's deep comprehension of ACP.

3 CP-DUE: A Framework Evaluating LLMs' Deep Understanding of ACP

In this section, we introduce the CP-DUE framework, designed for the in-depth evaluation of ACP comprehension abilities.

3.1 Framework Design

Inspired by (Yuan, 2005), we design a suite of layered, progressively challenging evaluation tasks organized into five core modules: Cultural Knowledge (**CK**),Theme and Style (**TH&ST**), Words Selection (**WS**), Allusion Understanding (**AU**), and Overall Quality (**OQ**). CP-DUE comprises five core modules rooted in monographs on ACP(Yuan, 2005, 2009) and Chinese Educational Curriculum Guidelines, which aim to comprehensively examine the understanding abilities of LLMs regarding ACP. As shown in Figure 2, this process of understanding the poem involves mapping the spatial relationships between ancient landmarks at a factual level, identifying thematic elements such as the author's nostalgia, performing a fine-grained lexical analysis of specific word choices, grasping the allusive connotations embedded within the text and judging the quality of the poem as a whole.

Cultural Knowledge (CK) tests how well LLMs master the cultural background specific to ACP. This foundation allows LLMs to interpret poetry with the perspective of someone truly rooted in Chinese tradition.

Theme and Style (TH&ST) tests the ability of LLMs to identify the theme and style of a poem. This ensures LLMs can grasp the emotional tone and artistic spirit from a holistic perspective.

Word Selection (WS) tests whether LLMs can recognize the most effective words and sentences within a context. This measures their ability to capture the fine-grained beauty and the precise language that define poetic art(Liu, 1933).

Allusion Understanding (AU) tests if LLMs understand the historical allusions and hidden meanings in the text. This effectively measures whether the model can look beyond the literal words to reach the deeper soul of the poem.

Overall Quality (OQ) tests whether the model's overall appreciation of poetry aligns with human experts. This determines if LLMs can share our aesthetic values and judge the quality of art in a way that humans do.

Together, these modules form a comprehensive framework that shifts the focus from simple text matching to deep artistic appreciation. By evaluating these diverse dimensions, we can better understand whether LLMs are merely processing lin-

guistic data or are beginning to grasp the unique spirit of ACP. Ultimately, this systematic approach provides a new way to measure the gap between machine intelligence and the human soul in the realm of literary art.

3.2 Data Construction

To ensure objectivity in our evaluation, all assessment datasets are presented in the form of multiple-choice questions rather than generative tasks. The relevant data statistics for the dataset is shown in Table 1.

Task	Questions	Options
CK (Module 1)	505	3
TH (Module 2)	162	10
ST (Module 2)	808	8
WS (Module 3)	404	4
AU (Module 4)	500*3	4
OQ (Module 5)	690*2	3

Table 1: Data statistics of CPDUE. This includes the data volume for the five modules and six tasks, as well as the number of options for each question. (* means a single data point is associated with multiple questions.)

The datasets for our CP-DUE framework are curated from several authoritative platforms to ensure a reliable foundation for evaluation. Data for the **CK** Module is derived from the official question pools of renowned TV competitions, such as *Chinese Poetry Congress*[1] and *Classical Chinese Poetry Challenge*[2]. These materials are vetted by expert committees and further screened by our team to serve as high-standard benchmarks for assessing ACP literacy. For the **TH**&**ST** and **WS** Modules, we utilize *Souyun*[3] as our primary source, adopting classical academic classifications and professional literary criticism as references(Zhan, 1980; Wu, 2010). To ensure the absolute correctness of the data, these samples are independently labeled by annotators with at least a bachelor's degree in Classical Chinese, maintaining an inter-annotator agreement above 90%. The **AU** Module is also built upon *Souyun*, where we incorporate distractors based on textual similarity to increase the rigor of the evaluation. Finally, the **OQ** Module features modern works from the *shici52* community[4]. Each Anonymous poem has undergone a rigorous review process by three to ten senior judges, whose expert scores serve as a definitive gold standard for human aesthetic judgment. This process guarantees the professional quality and correctness of this module.

With such a meticulously curated and verified corpus, we provide a solid point of support for Large Language Models to move beyond superficial patterns and truly engage with the deep artistic essence of ACP. This corpus enables models to be evaluated against professional-grade standards, ensuring that our benchmark serves as an effective tool for measuring high-level literary comprehension.

Cultural Knowledge. We selected 500 multiple-choice questions from TV poetry competitions, excluding those relying on simple rote memorization, as well as questions with exceptionally high accuracy rates in the original programs. Covering dozens of themes, these questions focus on core ACP features to measure the depth of LLMs' cultural mastery.

Theme and Style. Tang poetry is categorized into ten themes, while Song Ci is divided into eight unique styles based on the classification systems. We curated a balanced dataset of representative works from the Souyun website, manually screened by experts to evaluate LLMs' ability to discern refined poetic themes and textual styles (see Appendix C).

Words Selection. Using regulated poems from various periods, experts developed a set of discriminative multiple-choice questions through a manual fill-in-the-blank design process. After removing key words from the original texts, experts manually crafted plausible distractors to challenge the model. These tasks evaluate an LLM's grasp of verbs, reduplicated words, and imagery, as well as its ability to satisfy strict constraints of parallelism and prosody(Wang, 1977).

Allusion Understanding. We sourced over a thousand allusions from the Souyun library. To test deep comprehension, questions include distractors with similar meanings and a "no allusion" option. This evaluates whether LLMs can accurately identify the figures, sources, and deep-seated meanings of allusions within poetic lines.

Overall Quality. We selected three poems of the same theme from the 52shici competition—one winning, one median, and one zero-scored work. LLMs must identify the best and worst poems, measuring whether their aesthetic judgment

[1]https://tv.cctv.com
[2]https://www.hebtv.com/
[3]https://sou-yun.cn/
[4]https://www.52shici.com/

Models	CK	TH/ST	WS	AU(Q1/Q2/Q3/all)	OQ(best/worst/all)
GPT-4o	75.45	51.85 / 18.07	50.25	64.80 / 42.40 / 37.40 / 14.60	45.01 / 45.73 / 24.75
DeepSeek-chat	84.55	53.70 / **30.32**	**58.32**	67.60 / 57.60 / 40.40 / 21.00	50.36 / 39.22 / 22.00
Doubao-seed-1.6	**90.30**	58.64 / 29.70	54.60	**79.20 / 78.40 / 81.80 / 64.60**	**51.81 / 53.98 / 31.11**
Qwen-Max	85.15	**59.88** / 17.70	55.35	67.00 / 52.80 / 48.60 / 27.40	50.22 / 49.20 / 29.81
Qwen3-235B	**85.94**	**54.94 / 28.22**	**55.60**	**81.60 / 54.80** / 43.20 / **20.80**	**49.35** / 48.05 / **27.79**
Qwen3-32B	78.81	47.53 / 26.17	47.43	64.00 / 37.40 / **43.60** / 7.80	47.76 / 43.99 / 23.73
Qwen3-14B	76.44	41.98 / 19.80	42.23	42.60 / 34.80 / 29.20 / 1.80	46.02 / **49.20** / 26.20
Random	33.33	10.00 / 12.50	25.00	25.00 / 25.00 / 25.00 / 1.56	33.33 / 33.33 / 11.11

Table 2: Comparison of performance across various LLMs. Models are categorized into general commercial models (top) and the Qwen3 series with different parameter scales (bottom). **Bold** values indicate the best performance within each category.

aligns with human experts.

4 Experiments

4.1 Experiment Setup

In our experiments, we select the following LLMs for evaluation:

- Open-source LLMs : Qwen3-235B, Qwen3-32B, Qwen3-14B(Bai et al., 2023).
- Closed-source LLMs : GPT-4o(Achiam et al., 2023), Doubao-seed-1.6, Deepseek-Chat(Liu et al., 2024), Qwen-Max(Bai et al., 2023).

For every model and each task mentioned in Chapter 3, we employ an identical Chinese prompt to guide the models in answering the single-choice questions. We set the temprature to zero to ensure the stability. The models' capability is then measured by the correct answer rate for the single-choice questions for each task. Detailed prompts are provided in Appendix A.

4.2 Main Results

The experiment results for LLMs are shown in Table 2. From the results, we can come to some interesting conclusions.

The Knowledge-Reasoning Gap. A significant gap exists between knowledge acquisition and semantic understanding in LLMs. Specifically, their capacity for factual mastery far exceeds their ability for semantic reasoning. Experimental results reveal a performance fracture between tasks requiring factual recall and those demanding deep semantic inference. Nearly all evaluated closed-source models excel in the **CK** Module, with leading models such as Doubao-seed-1.6 and Qwen-Max achieving accuracies above 85%. This success merely indicates that LLMs have successfully internalized ACP corpora during pre-training. However, performance declines sharply in the **WS** and **AU** Modules. This suggests that while models can recite cultural facts, they struggle to select the literarily optimal word during the polishing process, which requires balancing context, prosodic constraints, imagery, and emotion. Furthermore, they fail to accurately identify the nuanced interpretations of classical allusions, as this necessitates deep understanding grounded in specific contexts. Consequently, LLMs exhibit a shallow understanding of ACP: they possess the encyclopedic knowledge of a polymath but lack the deep interpretive insight of a literary critic.

The Parameter Scaling Bottleneck. Model capability exhibits diminishing marginal returns relative to parameter scale, as increased capacity does not directly translate into artistic intuition. Comparing models of varying scales within the same series (Qwen3-14B, 32B, and 235B) reveals distinct performance evolution patterns. As the parameter count increases from 14B to 235B, performance improves significantly in fact-based tasks such as the **CK** Task and **AU**-Q1 Task (as an example, **CK** accuracy rises from 76.44% to 85.94%). However, improvements in the **ST** Task and **OQ** Task are marginal or even fluctuate. While expanding parameter scale effectively enhances memory capacity which allows LLMs to store more ACP knowledge, the high-order semantic features do not follow a linear learning curve relative to scale. In the domain of ACP, scaling parameters is effective for addressing founda-

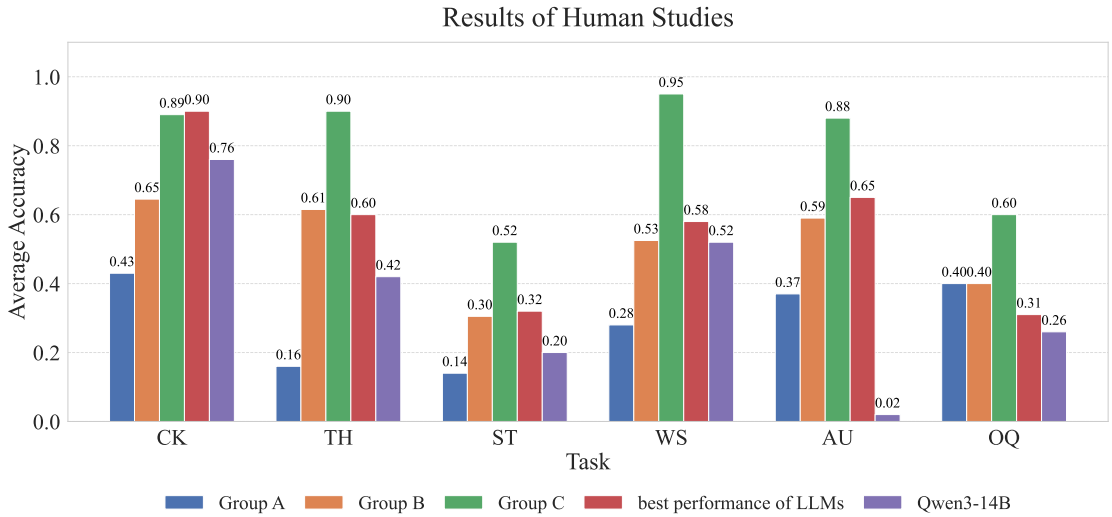

Figure 3: Experiment results of the human studies. The figure shows the average accuracy of volunteers from each group and compares it with the best performance achieved by tested LLMs, as well as a Qwen3-14B Model. The best performance of LLMs reach the C-group volunteers' level only on the basic **CK** task. On the **TH** and **ST** tasks they only match group B, corresponding to 67% and 62% of the C-group's performance, respectively. For the **WS** and **AU** tasks they only slightly exceed group B, amounting to 61% and 74% of the C-group's ability. In the **OQ** task, LLMs reach only 75% of group B and 52% of group C.

tional knowledge coverage but yields diminishing marginal returns in achieving high-level conceptual or aesthetic understanding.

The Abstract Cognitive Ceiling. The CP-DUE framework reveals a significant bottleneck for LLMs in processing high-order abstract concepts, particularly regarding artistic style and aesthetic quality. A sharp contrast exists between the identification of subject matter and artistic style. While models achieve acceptable accuracy in identifying themes (approximately 50-60%), their performance drops to near-random levels (16-30%) when identifying more abstract styles. Subject matter is generally explicit and often triggered by specific keywords. In contrast, style is implicit and defined by the synergy of sentence structure, tone, rhythm, and imagery. These features are highly abstract and remain difficult to master through simple word frequency statistics or pattern matching. Furthermore, within the **OQ** Module, even the most advanced models demonstrate poor alignment with the aesthetic preferences of human experts. This indicates that LLMs have not yet acquired the artistic judgment criteria that are inherent in human specialists. While understanding at the content layer is reaching saturation, LLMs encounter an abstraction bottleneck at the aesthetic level. Current models excel at analyzing the literal content of ACP but fail to perceive the underlying artistic tension within the creative execution.

4.3 Human Studies

To further validate the ACP comprehension evaluation system proposed in this paper and to explore the relationship between LLMs and humans regarding poetry understanding capability, we also invite some volunteers to participate in a human experiment. All volunteers are from China and share the same cultural background. They are divided into three groups, A, B, and C, based on their education backgrounds and familiarity with ACP:

- Group A: All members have an elementary or junior high school diploma.
- Group B: All members have a high school diploma or a bachelor's degree in science and engineering, and they have little exposure to ACP after graduating from high school.
- Group C: This group includes volunteers with senior undergraduate status or a bachelor's degree in Chinese Language and Literature, as well as active poetry enthusiasts who have created over 200 works, at least 20 of which are featured on ACP forums.

Each group of participants is required to answer the same balanced questions sampled from each

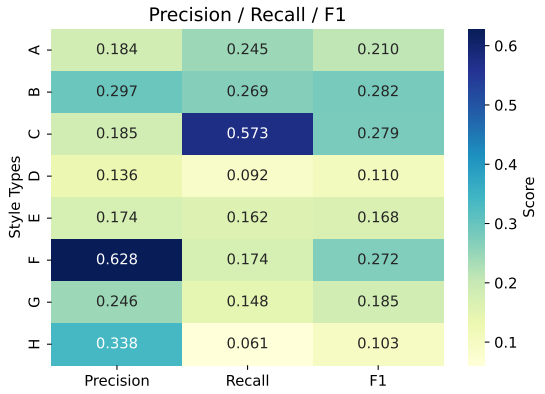

Figure 4: Detailed LLMs' accuracy, recall and F1 score for different Styles.LLMs reach their peak Recall in Category C, while their Precision is highest in Category F.

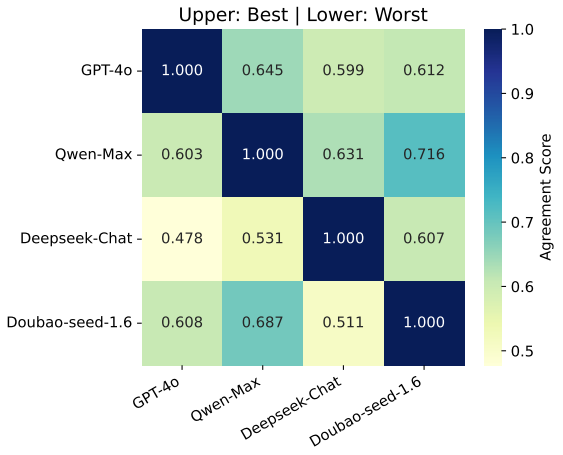

Figure 5: The consistency of tested models in the **OQ** module. The upper matric shows the consistency of the *which is the best* Task. The lower matric shows the consistency of the *which is the worst* Task.

dataset. The accuracy of human responses is then evaluated. The experimental results for each task across the three volunteer groups (A, B, and C) are shown in Figure 3. Meanwhile, we include the best performance of tested LLMs across all tasks(red bars) and a small-parameter model that can be deployed locally, Qwen3-14B(purple bars), for comparison. The detailed experiment results and extra analysis are shown in Appendix B.

Comparing the average scores from the human experiments with LLMs' results, we can draw the conclusion that LLMs have an expert-Level memory but a layman-level Intuition. Despite the fact that the best performance of models achieve parity with Group C experts in **CK** (0.90 vs. 0.89), demonstrating expert-level factual recall, their performance in deeper comprehension tasks (**TH**, **WS**, **AU**) regresses to the level of Intermediate Group B, and even falls behind Novice Group A in **OQ** (0.31 vs. 0.40). While human experts utilize deterministic structural rules such as prosody and allusion logic, LLMs appear to mirror the linguistic intuition of non-experts—relying on probabilistic pattern matching rather than the rigorous artistic constraints essential for a deep understanding of ACP. Smaller models perform substantially worse than groups B and C and fall below their performance on **CK**, **WS**, **AU**, and **OQ**. In everyday use, we should recognize the value LLMs provide as a repository of cultural knowledge. However, when LLMs assist with poetic composition or interpretation their ability may be inferior to that of an ordinary university student; therefore practitioners should exercise caution when relying on their suggestions and outputs.

4.4 Analysis of Each Module

In this section we analyze the items that show high error rates in LLMs and examine them further.

Cultural Knowledge. Although LLMs achieve high accuracy here, errors frequently occur in items requiring secondary understanding. Models often rely on surface-level cues rather than deep reasoning regarding implicit cultural contexts. Detailed Examples are shown in appendix D.

Theme & Style. We observed a significant performance imbalance across categories. Figure 4 shows the accuracy of different styles. This likely stems from training data bias and a lack of fine-grained stylistic understanding. Models often default to broad classifications and struggle to distinguish the subtle lexical and structural nuances of specific poets.

Words Selection. Models exhibit widespread errors in this module, particularly regarding advanced poetic techniques. While they possess factual knowledge, they lack the sensitivity to diction required to serve as competent assistants for poetic composition or comprehension.

Allusion Understanding. Performance drops significantly when tasks shift from rote memorization to contextual application. Most models can recall the content of an allusion but fail to interpret its specific meaning within the constraints of a verse.

Overall Quality. The consistency of tested closed-source models is shown in Figure 5. Both humans and LLMs show low accuracy in aesthetic judgment. High internal inconsistency between models suggests that LLMs cannot replicate human evaluative judgment for complex aesthetic tasks.

5 Limitations of LLMs and Implications for LLM Improvement

Through a detailed analysis of our experimental results, we identify significant limitations in the capabilities of current LLMs regarding the comprehension and appreciation of ACP. Consequently, the gap between the artistic nature of poetry and the statistical nature of current models manifests in three primary aspects:

Predominance of Surface-level Matching over Deep Cultural Inference. While LLMs demonstrate robust performance in retrieving factual cultural knowledge, they lack the capacity for deep reasoning driven by cultural context. Model failures are concentrated in tasks requiring second-order reasoning that combines literal information with implicit cultural backgrounds and the profound meanings of allusions.

Poetry serves as a carrier of historical and cultural aesthetics, yet models tend to rely on explicit lexical cues and surface-level patterns instead of comprehending the historical context and cultural consensus underpinning the verses. This suggests that current LLMs have not yet developed a robust cultural reasoning capability. To adapt to the complexity of poetic language, which seeks to evoke aesthetic experience rather than provide simple answers, models must evolve from pattern matching toward acquiring deeper internal mechanisms for utilizing cultural knowledge.

Coarse-grained Modeling of Artistic Style and Rhetorical Nuance. The modeling of poetic style and artistic expression in current LLMs remains highly coarse-grained and lacks sensitivity to linguistic and rhetorical details. LLMs exhibit prevalent predictive collapse and category bias in style and theme classification tasks. Furthermore, their accuracy in perceiving the quality of specific diction is notably low.

This reflects that the representation of poetic art in these models remains at a macro-label level. Current models fail to capture the fine-grained features or the intrinsic connections between character elements, rendering them unable to distinguish between masterpieces and works that are merely formally similar but artistically hollow. We hypothesize this stems from unbalanced stylistic representation in training corpora and a lack of fine-grained annotation. Future work should introduce more granular representations of rhetoric and style to bridge this perceptual gap.

Instability in Aesthetic Evaluation and the Irreplaceability of Human Judgment. In comprehensive appreciation tasks, consistency across different models and between models and humans is universally low, with significant divergences observed across models of different scales.

Results across various task levels indicate that aesthetic evaluation remains a complex cognitive task heavily dependent on humanistic experience and sensibility. Current LLMs do not possess the stable evaluative capability or the empathetic understanding required to replace human judgment, a limitation that is expected to persist for the foreseeable future. Therefore, users should regard LLMs as auxiliary analytical tools rather than authoritative critics. We must remain cautious of their output, recognizing that while they can process the text, they cannot yet fully comprehend the comprehensive conceptual framework or the unique creative spirit embedded within the art of poetry.

6 Conclusions

We introduce CP-DUE, a framework designed for the evaluation of LLMs' abilities in deep ACP understanding. Through a top-down, progressively deepening task design and a meticulously annotated expert dataset, we evaluate the ACP understanding capabilities of existing LLMs across five dimensions. The experimental results reveal that LLMs still face significant challenges in deeply comprehending ACP and cannot be fully trusted as tools for guided learning or poetry creation. In addition, we further analyze the performance and error patterns of LLMs across all tasks and provide experiment-based recommendations for enhancing their understanding of ACP in future developments. We believe that our work makes a substantial contribution to advancing LLM capabilities in ACP and promoting the dissemination of Chinese traditional culture.

Limitations

Lack of the linguistic scope of the poetry studied.

In fact, poetic forms in other languages (such as English metrical poetry or Japanese haiku) also embody unique cultural meanings and rhythmic aesthetics. Experimental results suggest that an LLM's performance on poetry is highly correlated with its proficiency in the corresponding language. Nevertheless, we believe that our dataset construction and evaluation methodology can be transferred to other languages to assess models' understanding of poetry across different cultural contexts.

The absence of generative tasks.

We believe that LLMs are capable of composing poems under relaxed conditions given a title. However, within the modular framework of this study, it is difficult to evaluate generated outputs objectively—manual evaluation is impractical due to scale. For the sake of fairness, we therefore exclude generative tasks. While we acknowledge their importance, we argue that expert-controlled multiple-choice questions can more precisely convey the models' level of understanding.

Acknowledgments

References

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zong-qi Cai. 2008. *How to read Chinese poetry: A guided anthology*. Columbia University Press.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. *arXiv e-prints*, pages arXiv–2408.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *IJCAI*, pages 4925–4931.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 1650–1656.

James R Hightower and Florence Chia-ying Yeh. 2020. *Studies in Chinese poetry*, volume 47. BRILL.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

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*.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.

James JY Liu. 2022. *The art of Chinese poetry*. Routledge.

Xie Liu. 1933. *The Literary Mind and the Carving of Dragons (Wen Xin Diao Long)*. Art China Network.

Yang Liu, Lan Lan, Jiahuan Cao, Hiuyi Cheng, Kai Ding, and Lianwen Jin. 2025. Large-scale corpus construction and retrieval-augmented generation for ancient chinese poetry: New method and data insights. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 779–817.

Stephen Owen. 2020. *Readings in Chinese literary thought*, volume 30. Brill.

Li Wang. 1977. *The Prosody of Classical Chinese Poetry*. Zhonghua Book Company.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*.

Xionghe Wu. 2010. *A Comprehensive Study of Tang and Song Ci*. Zhejiang Ancient Books Publishing House.

Yuanchong Xu. 2021. *Selected Peoms od LiBai*. China Translation and Publishing House.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9450–9457.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018. Chinese poetry generation with a working memory model. *arXiv preprint arXiv:1809.04306*.

Xingpei Yuan. 2005. *The History of Chinese Literature*. Higher Education Press.

Xingpei Yuan. 2009. *Studies on the Art of Chinese Poetry*. Peking University Press.

Antai Zhan. 1980. *Essays on Song Peotry*. Guangdong People's Publishing House.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 670–680. Association for Computational Linguistics.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend ancient chinese? a preliminary test on aclue. *arXiv preprint arXiv:2310.09550*.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative chinese classical poetry generation system. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pages 25–30.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese. *arXiv preprint arXiv:2305.14150*.

A Prompts Used in Experiments

The original prompts are shown in Figure 6.

Prompts(Cultural Knowledge)
请根据以下单项选择题，仅返回A、B、C中的一个字母作为答案，不要包含任何解释、标点或多余的文字。
题目: {question} A. {opt_a} B. {opt_b} C. {opt_c}

Prompts(Theme/Style)
请你根据以下指示选择所给词作的语言风格。仅返回A、B、C、D、E、F、G、H中的一个字母作为答案，不要包含任何解释、标点或多余的文字。
风格解释:{instruction}
词作: {title} {content}

Prompts(Word Selection)
请根据以下在*处空缺的七律，结合对仗、格律、连贯性等角度综合考虑，选择最合适填入空缺的选项。返回A、B、C、D中的一个字母作为答案，不要包含任何解释、标点或多余的文字。
被挖空的诗句: {question} A. {opt_a} B. {opt_b} C. {opt_c} D. {opt_d}

Prompts(Allusion Understanding)
请你阅读以下诗句。并判断其涉及的典故人物、典故出处、典故含义。依次返回A、B、C、D中的字母作为三道题的答案（如:CAB)，中间不要包含任何解释、标点或多余的文字.
诗句:{question}
典故人物:A. {opt_1a} B. {opt_1b} C. {opt_1c} D. 没有使用典故或使用的典故不在以上三个选项里。
典故出处:A. {opt_2a} B. {opt_2b} C. {opt_2c} D. 没有使用典故或使用的典故不在以上三个选项里。
典故含义:A. {opt_3a} B. {opt_3b} C. {opt_3c} D. 没有使用典故或使用的典故不在以上三个选项里。

Prompts(Overall Quality)
你是一个诗歌擂台的评委。以下三首都是相同题材的擂台作品，请你从从中选出整体创作最好(坏)的一项古诗。返回A、B、C中的一个字母作为答案，不要包含任何解释、标点或多余的文字。
A. {opt_a} B. {opt_b} C. {opt_c}

Figure 6: Prompts used in each module.

B Detailed Data and Analysis of Human Experiments

40 volunteers grouped by their educational background participated in our human experiments. There are 10 questions in each task. The volunteers' number of correct answers are shown in Table 3. The web interface for the human study is shown in Figure 7. Informed consent for the use of experimental results has been obtained from all participants (or their guardians), and they were compensated at a rate no less than the local average wage.

UserID	CK	TH	ST	WS	AU	OQ	UserID	CK	TH	ST	WS	AU	OQ
A01	4	1	2	4	4	3	B01	4	6	3	5	6	3
A02	8	1	0	4	2	6	B02	3	6	2	3	5	5
A03	7	2	4	3	8	2	B03	9	9	5	9	8	5
A04	3	4	1	2	4	4	B04	6	6	2	6	5	4
A05	6	0	2	3	2	3	B05	4	3	2	5	7	2
A06	3	1	1	2	3	2	B06	7	6	3	4	6	4
A07	3	1	1	3	4	6	B07	6	5	2	5	6	2
A08	2	1	2	2	2	4	B08	7	5	3	7	6	6
A09	4	2	0	2	3	5	B09	6	4	6	5	5	5
A10	3	3	1	3	5	5	B10	5	5	3	4	8	6
C01	9	10	9	10	10	8	B11	5	4	5	5	6	4
C02	8	10	4	10	8	6	B12	8	8	3	7	8	3
C03	10	10	5	10	9	5	B13	6	7	4	6	6	3
C04	10	8	5	9	8	5	B14	7	10	2	4	5	4
C05	9	10	3	10	9	6	B15	7	5	3	5	7	4
C06	8	7	3	8	8	4	B16	7	5	2	3	6	3
C07	10	8	5	10	7	6	B17	7	5	4	5	5	4
C08	9	10	8	9	10	7	B18	8	6	4	6	4	5
C09	9	10	4	9	10	6	B19	8	10	1	6	4	4
C10	7	7	6	10	9	7	B20	9	8	2	5	5	4

Table 3: Human evaluation results for groups A(left)、B(right)、C(left).

Figure 7: The web interface of human study.

From the figure, it is evident that volunteers with different educational backgrounds exhibit significant differences in ACP tasks. Group A (elementary and junior high school level) avolunteers only achieves a certain correct rate in the Cultural Knowledge module, while they find it difficult to identify the correct options in other tasks. Group B volunteers (high school/university science and engineering level) show significantly better performance in the Theme, Words Selection, and Allusion Understanding tasks, but their ability in the most difficult tasks, Overall Quality and Style Judgment, still appears insufficient. Group C volunteers (with ACP backgrounds) generally achieve extremely high accuracy, or even perfect scores, in relatively shallow tasks, and their abilities in Overall Quality and Style Judgment are also much higher than those of the other two groups.

In addition, we conduct interviews with selected volunteers from Groups B and C. We choose ques-

tions with high error rates and ask them why they make those selections. Even for the same questions, Group C (with ACP backgrounds) possesses high-level skills and abilities (such as the harmony of prosody or the specific meaning of allusions) to form clear and explicit judgments on the correct answers, while Group B volunteers merely rely on a vague sense of language to make their choices. This further proves that our dataset can profoundly assess the ability to understand ACP.

C Detailed Classification of Song Ci in The Style Task

ID	Style	Stylistic features	Example poem
A	真率明朗	不事假借，极少粉饰，有甚说甚，而委曲详尽，妥帖谐叶，既明朗，也深切。 It avoids artificial embellishment and employs minimal ornamentation, expressing ideas directly yet with detailed, precise, and harmonious wording that achieves both clarity and depth.	柳永《八声甘州》：对潇潇暮雨洒江天，一番洗清秋。渐霜风凄紧，关河冷落，残照当楼。是处红衰翠减，苒苒物华休。唯有长江水，无语东流……争知我，倚栏杆处，正恁凝愁！
B	高旷清雄	胸次旷远，眼界高绝，笔致清刚，气格雄迈，而兴象超迈，意韵深醇，既高旷，又清雄。 It conveys a broad mind and lofty vision, characterized by a clear and vigorous style, a bold and powerful spirit, and imagery that is both elevated and profound, achieving a balance of grandeur and clarity.	苏轼《西江月》：三过平山堂下，半生弹指声中。十年不见老仙翁，壁上龙蛇飞动。欲吊文章太守，仍歌杨柳春风。休言万事转头空，未转头时皆梦。
C	婉约清新	工细精刻而不露着力，和雅浑融而不陷纤巧，极为含蓄凝整。 It is refined and precise without appearing labored, harmonious and elegant without falling into delicacy, conveying restraint and cohesion.	李清照《声声慢》：寻寻觅觅，冷冷清清，凄凄惨惨戚戚。乍暖还寒时候，最难将息。三杯两盏淡酒，怎敌他、晚来风急！雁过也，正伤心，却是旧时相识……这次第，怎一个愁字了得！
D	奇艳俊秀	抒写平凡景物情事而富韵味，意境奇横、语言精警，推陈出新。 It depicts ordinary scenes and emotions with rich resonance, featuring imaginative conception, concise language, and innovative expression.	张先《倾杯》：飞云过尽，明河浅、天无畔。草色栖萤，霜华清暑，轻🄴弄袂，澄澜拍岸。宴玉尘谈宾，倚琼枝、秀挹雕觞满。午夜中秋，十分圆月，香槽拨凤，朱弦轧雁……烟江艇子归来晚。
E	典丽精工	遣辞造句矜慎，辞语精炼，结构严密，思力深透，音律谐叶。 It employs words and syntax with precision and care, featuring concise diction, rigorous structure, profound thought, and harmonious rhythm.	周邦彦《夜飞鹊》：河桥送人处，凉夜何其。斜月远堕余辉。铜盘烛泪已流尽，霏霏凉露沾衣。相将散离会，探风前津鼓，树杪参旗。华骢会意，纵扬鞭、亦自行迟……极望天涯。
F	豪迈奔放	风格雄奇跌宕、豪迈奔放，经史子集任意驱遣，自然合度。 Its style is bold and dynamic, marked by grandeur and spontaneity, freely drawing on classical texts while maintaining a natural sense of balance.	辛弃疾《永遇乐》：千古江山，英雄无觅，孙仲谋处。舞榭歌台，风流总被，雨打风吹去。斜阳草树，寻常巷陌，人道寄奴曾住。想当年，金戈铁马，气吞万里如虎……凭谁问：廉颇老矣，尚能饭否？
G	骚雅清劲	骚情隐蕴，格调雅正，笔致清峭，气骨劲健。而用典幽微，意境澄冷。扫浮艳、以健笔写柔情。 It embodies restrained emotion and elegant tone, with a clear and vigorous style and strong compositional force. Its allusions are subtle, its imagery pure and serene, rejecting ornamentation while expressing delicate feelings with firm, refined language.	姜夔《扬州慢》：淮左名都，竹西佳处，解鞍少驻初程。过春风十里，尽荠麦青青。自胡马窥江去后，废池乔木，犹厌言兵。渐黄昏，清角吹寒，都在空城……念桥边红药，年年知为谁生。
H	密丽险涩	字面讲究、句法雕琢，工巧丽密，往往险涩而近李贺、李商隐。 It emphasizes careful diction and crafted syntax, exhibiting intricate and refined workmanship, often presenting complexity and subtlety reminiscent of Li He and Li Shangyin.	吴文英《解连环》：思和云结。断江楼望睫，雁飞无极。正岸柳、衰不堪攀，忍持赠故人，送秋行色。岁晚来时，暗香乱、石桥南北。又长亭暮雪，点点泪痕，总成相忆……叹沧波、路长梦短，甚时到得。

Table 4: Detailed classification of Song Ci

D Examples of Error Analysis in Cultral Knowledge Module

We select 3 questions with a high error rate, which are shown in Figure 8-Figure 10. In the first question, the LLMs identifie it as "Havoc in Heaven" based only on the phrase "jade sky clears," but in fact this poem celebrates the "Three Battles with the White Bone Demon." The LLM lacks sufficient cultural knowledge of the work's background. In the second question, The LLMs only know that this is a work

by Du Fu, but they cannot understand that Du Fu's poem is actually expressing his longing for Li Bai. In the third question, the LLMs can only judge whether it relates to someone skilled in music based on the surface presence of "music," but they cannot understand techniques such as analogy applied to music, nor can they grasp the cultural background behind the poetry.

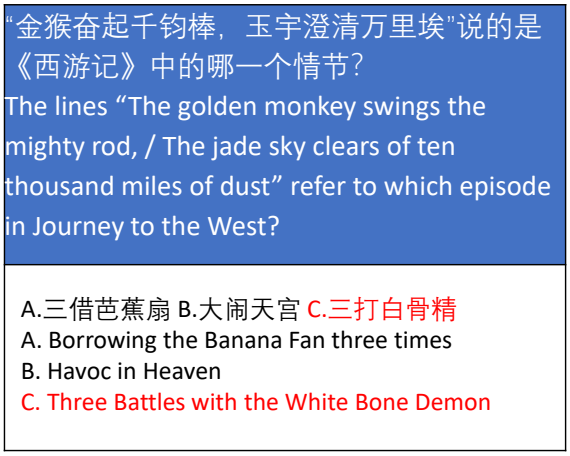

Figure 8: An example of a question with a high error rate.

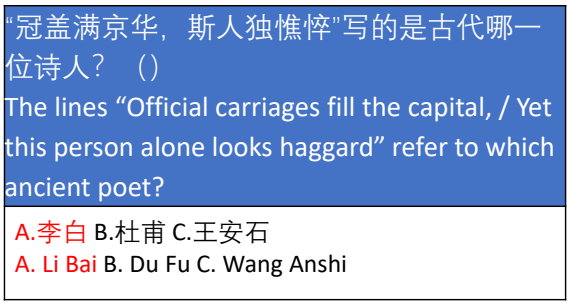

Figure 9: An example of a question with a high error rate.

Figure 10: An example of a question with a high error rate.