Gender Analysis of Ancient Chinese Texts based on Large Language Models

Anonymous ACL submission

Abstract

In ancient China, women's literary expression was constrained by their low social status, established literary conventions, and cultural norms, which led many female poets to adopt pseudonyms and resulted in the loss or fragmentation of a large number of their works. There were significant differences in themes and emotional expressions between women's and men's poems; however, most current gender-related studies focus on modern texts, leaving a noticeable gap in the gender analysis of ancient Chinese poetry. Therefore, this study proposes a new task named *Gender Analysis of Ancient Chinese Texts*, which focuses on ancient Chinese poetry with two core objectives: analyzing the gender of the authors and exploring the themes and emotions conveyed in their works. To support this task, we construct and annotate a dataset of 30,000 ancient Chinese poems from the Ming and Qing dynasties with gender, theme, and emotion labels. Considering the challenge posed by the implicit associations between gender, themes, and emotions in ancient poetry, an *Implicit Multi-Step Reasoning* framework is constructed to identify and utilize these associations. Experimental results demonstrate the effectiveness of the proposed framework.

1 Introduction

In ancient China, women's literary expression was constrained by both their relatively low social status and established literary conventions that prescribed appropriate themes and emotional tones (Fong and Widmer, 2017; Nanxiu and Fong, 2019). As a result, many female-authored works were lost or fragmented over time (Chang and Saussy, 1999). Rediscovering these works is therefore not only of scholarly value but also essential for recovering and respecting the lived experiences and inner worlds of these women.

Moreover, women's poetry was subject to thematic and emotional restrictions. For instance, married women were often discouraged from expressing homesickness or strong negative emotions. These social constraints led to systematic differences between women's and men's poetry in both thematic focus and emotional expression (Fong and Widmer, 2017). Classical Chinese poetry also reflects gendered experiences and social norms within historical contexts.

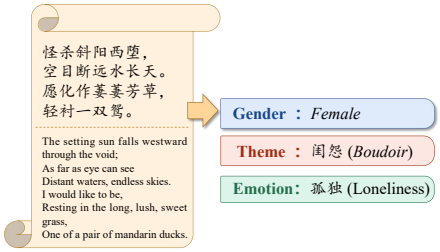

Figure 1: Example of gender analysis of ancient Chinese texts.

To address these challenges, we propose a new task, *Gender Analysis of Ancient Chinese Texts*, which jointly models author gender, themes, and emotions in classical poetry. As shown in Figure 1, this task not only fills a gap in computational analysis of ancient Chinese texts, but also provides a new lens for examining gender culture, as poetry encodes both individual expression and broader cultural expectations (Nanxiu and Fong, 2019).

Specifically, we construct and annotate a 30,000-poem dataset from the Ming and Qing dynasties to support this task. The dataset includes works by both male and female poets; despite their smaller number, female-authored poems cover diverse themes, social backgrounds, and regions. Each poem is annotated with theme and emotion, capturing its semantic focus and affective tone.

However, gender analysis of ancient Chinese poetry faces inherent challenges due to its unique linguistic and cultural features. It lacks explicit gender indicators, relying instead on subtle, poem-

wide stylistic cues like imagery and diction. Determining gender also requires integrating multiple semantic clues (e.g., poetic intent, emotion), while independent prediction of gender, theme, and emotion leads to inconsistencies. To address these, we propose the *Implicit Multi-Step Reasoning* (IMSR) framework, which first extracts low-level cues (e.g., imagery) and progressively infers high-level attributes, capturing latent semantic interactions and enhancing prediction coherence.

Experimental results demonstrate the framework's effectiveness. Our method outperforms baseline and state-of-the-art models, achieving the best F1 scores on triple prediction, theme, and emotion. While competitive on gender prediction, these results highlight the advantage of modeling the task as implicit multi-step reasoning.

2 Related Work

2.1 Gender and Stylistic Analysis

Gender has long been studied in literary and linguistic research. Feminist criticism suggests that women's writing often exhibits distinctive thematic and emotional patterns (Woolf, 1929; Showalter, 1977, 1985), and similar observations have been made in Chinese literary studies (Wei, 1998; Hu, 2004; Mann, 1997). However, these analyses are largely qualitative.

In NLP, gender classification has been widely explored in author profiling (Schler et al., 2006; Rangel and Rosso, 2013), with contextualized models such as BERT (Devlin et al., 2019) significantly improving performance (Gonen and Goldberg, 2019). Meanwhile, substantial work has examined gender bias in language representations (Bolukbasi et al., 2016; Bordia and Bowman, 2019; Nadeem et al., 2021). Recent studies further investigate bias in large language models from structural and reasoning perspectives (Qiu et al., 2025; Zhang et al., 2025; Dubey, 2025).

Despite these advances, most work focuses on modern texts and bias mitigation. Modeling gender as a stylistic attribute in highly implicit literary genres such as ancient Chinese poetry remains underexplored.

2.2 Information Extraction on Ancient Chinese Texts

Recent work has made progress in NLP for ancient Chinese, initially focusing on basic tasks such as word segmentation and syntactic analysis (Cui and Che, 2021; Guo et al., 2020). Subsequent studies extend to information extraction tasks, including named entity recognition, relation extraction, and historical event extraction (Yamada et al., 2020; Wang et al., 2021; Han et al., 2021).

Efforts have also been made to construct structured knowledge from historical texts, such as historical knowledge graphs and person–relation extraction (Zhang et al., 2021; Liu et al., 2022). More recent work introduces datasets and benchmarks for ancient Chinese IE (Tang et al., 2024; Blouin et al., 2024), and explores advanced modeling approaches, including multimodal frameworks and large language models (Bao et al., 2025a, 2024; AI@Alibaba, 2024; Dong, 2024).

Although these studies improve extraction performance, they typically treat tasks independently. The interaction between author attributes (e.g., gender) and literary semantics, especially in poetry, remains largely unexplored.

3 Data

3.1 Data Collection

The dataset consists of poems written by 495 poets from the Ming and Qing dynasties (1368–1912), whose gender information is clearly documented in historical biographical records.

Poems were collected from the online classical poetry archive *SouYun*[1]. We first crawled the list of poets from the Ming and Qing periods available on the website, and then retrieved all poems associated with each poet.

During preprocessing, we retained only complete poetic texts. Prefaces, prose commentaries, biographical introductions, duplicated entries, and incomplete records were removed. All poems were kept in their original ancient Chinese form without modern rewriting.

After cleaning, we obtained a large-scale corpus containing approximately 30,000 poems written by 495 poets from the Ming and Qing dynasties. However, the corpus exhibits a severe gender imbalance, as male-authored poems substantially outnumber female-authored poems.

To alleviate data imbalance in downstream evaluation, we further construct a relatively balanced experimental subset consisting of 3,275 poems, including 1,735 male-authored and 1,540 female-authored poems. All poems in the experimental

[1] https://www.sou-yun.cn

subset are manually annotated with theme and emotion labels.

3.2 Annotation

Each poem is annotated with three attributes: **gender**, **theme**, and **emotion**.

Gender labels are obtained directly from the documented biographical information of each poet in historical records.

Theme describes the dominant semantic focus of the poem. The theme taxonomy is inspired by traditional classifications in ancient Chinese literary studies (i.e., *landscape poetry*, *nostalgic poetry*, *historical reflection*, *travel and frontier poetry*, *boudoir poetry*, and *social interaction*) (, 2003; , 1994). Each poem is assigned one primary theme based on its central imagery and narrative focus.

Emotion captures the overall affective tone of the poem. Following common emotion categories used in literary analysis and computational poetry studies (Mohammad and Turney, 2013; Khattak et al., 2022), we define several major emotional orientations, including *melancholy*, *nostalgic longing*, *aspirational or heroic sentiment*, *tranquility*, *joy or celebration*, and *solitude*.

We provide an example to illustrate the annotation process, while the complete annotation guidelines and detailed procedures are presented in Appendix A:

> **Poem:** 乘鸾夜下蓬莱岛，闲辗麟车踏瑶草。海风吹折碧桃花，玉盘满摘安期枣。 (*gender* = *female*, *theme* = 咏物言志, *emotion* = *tranquility*)

To improve efficiency and consistency, we adopt a two-stage annotation procedure. First, ChatGPT (Gilardi et al., 2023) is used to produce preliminary labels for theme and emotion. These initial labels are then manually reviewed and corrected by the authors. When ambiguous cases arise, three annotators jointly discuss the poem and determine the final label through consensus. The average Cohen's κ agreement score among the annotators is 0.82 for theme and 0.76 for emotion, indicating substantial inter-annotator agreement (Landis and Koch, 1977).

3.3 Overall Statistics

The final dataset contains 30000 poems, written by 495 poets from the Ming and Qing dynasties, spanning approximately 500 years (1368–1912).

Total poems	30,000
Time span	1368–1912
Total poets	495
Female poets	45
Male poets	450
Female poems	1,540
Male poems	28,460
Avg. poems per poet	60.61
Avg. poems per female poet	34.22
Avg. poems per male poet	63.24

Table 1: Overall statistics of the dataset

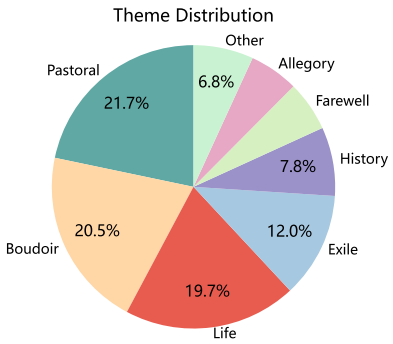

Figure 2: Distribution of theme

Among the poets, 45 are female and 450 are male. Female poets contribute 1,540 poems, while male poets contribute 28460 poems.

On average, each poet contributes 60.61 poems. Female poets contribute 34.22 poems per author on average, while male poets contribute 63.24 poems per author. Table 1 summarizes the overall dataset characteristics.

To better understand the composition of the dataset, we present the distributions of gender, themes, and emotions in Figure 2 and Figure 3.

3.4 Correlations Among Gender, Theme, and Emotion

Gender and Theme. We observe notable differences in thematic preferences between male and female poets. As shown in Figure 4, Historical reflection and frontier or travel-related themes appear more frequently in male-authored poems, accounting for approximately 27.5% of male texts compared with 14.0% of female texts. In contrast, boudoir-related themes are substantially more common in female-authored poems, representing 32.4% of female texts but only 8.2% of male-authored poems. For example, the poem 恼人最是西风，断肠听尽归鸿…… *(Most troubling is the west-*

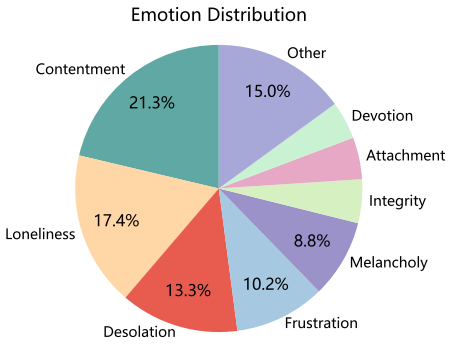

Figure 3: Distribution of emotion

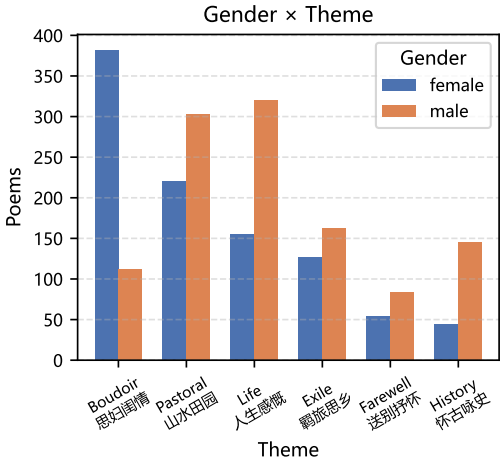

Figure 4: Distribution of theme categories across male and female poets

ern wind; heartbroken, one listens as the returning wild geese pass by.....) expresses delicate emotions associated with **boudoir sentiments**, and the author is identified as female in our dataset.

Gender and Emotion. Emotional tendencies likewise exhibit gender-related patterns. Figure 5 shows that Male-authored poems show a relatively higher proportion of aspirational or outward-looking sentiments (26.1%), often associated with themes of historical reflection or frontier experience. Female-authored poems, on the other hand, display higher proportions of nostalgic longing and melancholic emotions (40.1%), reflecting more introspective or personal modes of expression. Nevertheless, substantial overlap exists between the two groups, suggesting that gender differences should be interpreted as stylistic tendencies rather than rigid categorical boundaries. For example, the poem 黄衣冠子翠云裘，四海三山挟弹游..... *(Clad in yellow robes and a jade-green fur coat, he roams across the seas and sacred mountains*

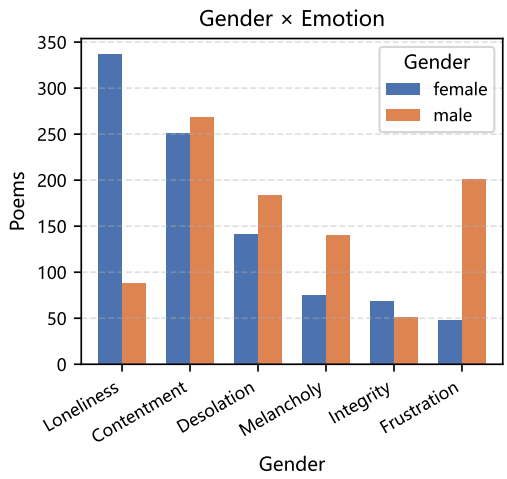

Figure 5: Distribution of emotion categories across male and female poets

with carefree spirit.....) conveys a sense of **heroism and impassioned sentiment**, characterized by an outward-looking and unrestrained tone, and the author is identified as male in our dataset.

Triangular interaction. Beyond pairwise relationships, gender, theme, and emotion form a structured interaction pattern. For example, historical reflection themes are frequently associated with aspirational emotional tones and are predominantly authored by male poets. Conversely, boudoir-related themes often co-occur with nostalgic or melancholic emotions and are more commonly written by female poets. These patterns suggest that gender signals in classical Chinese poetry are often implicitly expressed through thematic preference and emotional orientation rather than explicit lexical indicators.

4 Problem Definition

4.1 Task Definition

Given a ancient Chinese poem x written by an author, our objective is to jointly predict three interrelated attributes: author gender g, thematic category t, and emotional category e.

Formally, let $\mathcal{X}$ denote the space of ancient Chinese poems. We define three discrete label spaces: $\mathcal{G}$ for gender, $\mathcal{T}$ for theme, and $\mathcal{E}$ for emotion. For an input poem $x \in \mathcal{X}$, the task is to learn a mapping function:

$$f : \mathcal{X} \rightarrow \mathcal{G} \times \mathcal{T} \times \mathcal{E} \quad (1)$$

that predicts the attribute triplet (g, t, e).

Unlike independent multi-label classification, we assume that these attributes exhibit latent de-

pendencies. In particular, thematic preference may correlate with gender-related stylistic tendencies, and emotional tone may further interact with thematic selection. Therefore, instead of modeling $P(g \mid x)$, $P(t \mid x)$, and $P(e \mid x)$ independently, we aim to model the joint distribution:

$$P(g, t, e \mid x), \tag{2}$$

allowing structured dependency modeling among attributes.

4.2 Challenges

Ancient Chinese poetry gender analysis is challenging due to implicit stylistic signals, the need for multi-step semantic reasoning, and strong dependencies among gender, theme, and emotion. These characteristics motivate structured reasoning-aware modeling instead of independent classification.

5 Implicit Multi-Step Reasoning for Gender Analysis

Gender analysis of ancient Chinese poetry is challenging because it involves multiple correlated factors such as themes, emotions, narrative perspective, and imagery. However, these associations are often implicit rather than explicitly expressed. Therefore, we construct an *Implicit Multi-Step Reasoning* framework to model and utilize such latent semantic relationships.

As illustrated in Figure 6, the framework formulates gender analysis as a hierarchical reasoning process within large language models (LLMs) through Chain-of-Thought (CoT) prompting. The reasoning process begins with *Semantic Understanding* (s), where the model captures contextual semantics, symbolic imagery, and latent textual implications.

Based on these semantic representations, the framework performs *Perspective and Gender Cue Extraction* (p), identifying implicit gender-related contextual cues embedded in narrative voice and stylistic expression. These cues then guide *Gender Inference* (g), enabling the model to infer latent gender orientation through semantic reasoning rather than explicit lexical matching.

Building upon the inferred gender representation, the framework further conducts *Emotion Recognition* (e) and *Theme Identification* (t), allowing emotional and thematic information to interact with semantic and gender representations during reasoning. The inferred gender, emotion, and theme representations are then aggregated into a unified *Informed Representation* for final *Triple Prediction*.

By structuring the inference process into interconnected reasoning stages, the framework models the dependencies among semantic understanding, gender cues, emotion, and theme, thereby improving prediction consistency and interpretability.

5.1 Text Representation

Given an input poem x, we first employ a pretrained language model encoder to transform the raw text into contextualized semantic representations. Unlike traditional static word embeddings, the encoder captures bidirectional contextual dependencies across the entire poem, allowing the model to encode not only lexical meanings but also implicit semantic associations, stylistic characteristics, and long-range contextual interactions embedded within poetic expressions. Such contextual modeling is particularly important for poetry analysis, where emotional implication, symbolic imagery, and narrative perspective are often conveyed implicitly rather than through explicit surface patterns.

Formally, the poem is tokenized into a sequence of tokens and fed into the encoder, which produces contextualized hidden representations for each token. These hidden states dynamically integrate information from surrounding textual contexts, enabling the representation space to preserve both local semantic structures and global discourse-level semantics. The resulting token-level representations are subsequently aggregated into a unified sequence-level embedding, which serves as the shared semantic foundation for downstream reasoning and multi-attribute prediction tasks, including gender inference, emotion recognition, and theme identification:

$$h = \text{Encoder}(x). \tag{3}$$

The obtained representation h encapsulates high-level semantic and contextual information of the poem, providing a robust latent feature space for the subsequent implicit multi-step reasoning process.

5.2 Implicit Multi-Step Reasoning Framework

In classical Chinese poetry, gender is rarely expressed through explicit lexical indicators, but is often reflected through thematic preference, emotional tendency, narrative perspective, and imagery

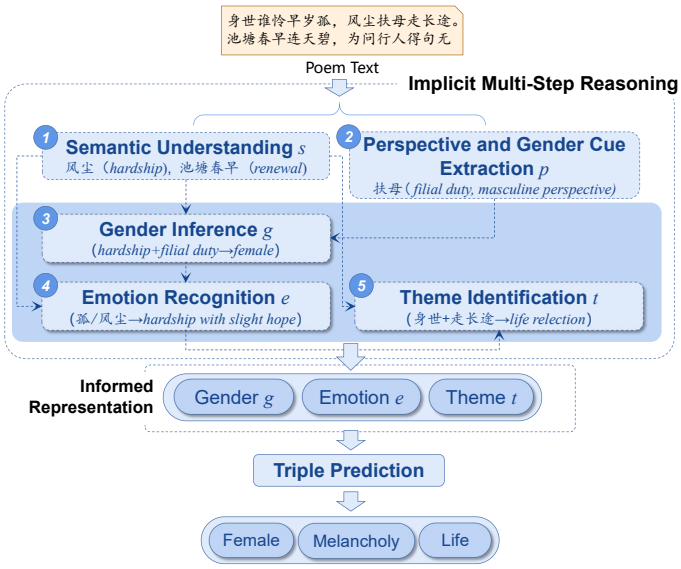

Figure 6: An example of the CoT Reasoning prompt construction

selection. Therefore, gender, emotion, and theme should not be treated as independent prediction targets, since these attributes exhibit strong semantic correlations.

To address this issue, we propose an **Implicit Multi-Step Reasoning** framework, which formulates triplet prediction as a progressive semantic reasoning process. The framework first performs semantic understanding to capture contextual information such as imagery and emotional implication, and then extracts perspective-related and gender-sensitive cues for gender inference. The inferred gender representation further interacts with emotion recognition and theme identification, enabling joint reasoning among semantic, emotional, and thematic information.

Through this reasoning paradigm, the framework models the dependencies among gender, emotion, and theme, thereby improving prediction consistency and interpretability.

Given an input poem x, the model aims to predict a triplet:

$$y = (g, e, t),$$

where g, e, and t denote the gender, emotion, and theme labels, respectively. The overall objective is to maximize the joint probability:

$$P(g, e, t \mid x).$$

From a reasoning perspective, this process can be abstracted into the following stages:

- **Semantic Understanding:** Identify key imagery, scenes, and tonal characteristics in the poem;
- **Perspective and Gender Cue Extraction:** Extract potential gender-related signals based on lexical choices (e.g., narrative perspective and activity descriptions);
- **Gender Inference:** Predict the author's gender based on the extracted cues;
- **Emotion Recognition:** Infer the emotional category by integrating semantic and stylistic information;
- **Theme Identification:** Determine the thematic category based on emotion and contextual semantics.

Following the implicit multi-step reasoning structure, the joint probability can be approximately factorized as:

$$\begin{aligned} P(g, e, t \mid x) = P(s \mid x)\, P(p \mid s, x) \\ P(g \mid s, p, x) \\ P(e \mid s, g, x) \\ P(t \mid s, e, x). \end{aligned} \tag{4}$$

Notably, the reasoning process is not executed in an explicit step-by-step manner, but is implicitly encoded in the model's high-dimensional representations. Although the model outputs a triplet (g, e, t), it learns a latent mapping from low-level semantics to high-level attributes during training.

This framework approximates structured reasoning without explicit intermediate rationales, while leveraging theme and emotion as auxiliary constraints to improve semantic consistency and maintain efficiency, achieving a better balance between expressiveness and computational cost.

5.3 Decoder

A decoder then generates the output sequence y, which represents the structured prediction of the three attributes. For example, the output may follow a template such as:

> Gender: Female; Theme: Landscape; Emotion: Melancholy

Formally, the decoder models the conditional probability of the output sequence as:

$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid y_{<i}, x), \tag{5}$$

where y_i denotes the i-th generated token.

5.4 Training

During training, the model parameters are optimized by minimizing the negative log-likelihood of the gold output sequence:

$$\mathcal{L} = -\sum_{i=1}^{|y|} \log P(y_i \mid y_{<i}, x). \tag{6}$$

This formulation allows the model to jointly generate the three attributes in a unified decoding process, enabling potential interactions among gender, theme, and emotion during prediction.

6 Experiments

6.1 Data and Setting

For experiments, we use the balanced subset introduced in Section 3.1, which contains 3,275 poems, including 1,735 male-authored poems and 1,540 female-authored poems. Each poem was manually annotated with both a thematic category and an emotional category. The dataset is split into training, validation, and test sets with a ratio of 8:1:1. Macro-F1 is used to address class imbalance.

6.2 Implementation Details and Evaluation

We optimize all models using AdamW (Loshchilov and Hutter, 2019) with a batch size of 1 and a maximum sequence length of 256. We tune learning rates $\{1 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$ and epochs $\{2, 4, 6\}$, selecting 5×10^{-4} and 4 epochs based on validation performance with early stopping.

We report Macro-F1 for gender, theme, and emotion classification, together with a joint triple metric (Manning et al., 2008).

6.3 Comparison with Baselines

We compare our proposed implicit multi-step reasoning framework with multiple categories of baselines:

- **Pre-trained Language Models.** We fine-tune Chinese BERT-wwm (Cui et al., 2019), RoBERTa-wwm-ext (Cui et al., 2020a), MacBERT (Cui et al., 2020b), T5 (Raffel et al., 2020), and LLaMA (Touvron et al., 2023) under the same training setting.
- **API-based Large Language Models.** GPT (Brown et al., 2020), DeepSeek (DeepSeek-AI, 2024), and Qwen (Bai et al., 2023) are evaluated via direct prompting.
- **Task-related baselines.** We additionally include a classical Chinese event extraction framework (Bao et al., 2025b) and a transformer-based author profiling model (Shahnazari and Ayyoubzadeh, 2025).

As shown in Table 2, pre-trained Chinese language models and API-based large language models achieve acceptable performance, while task-specific baselines do not outperform general pre-trained language models, likely because they were designed for different tasks.

Our Implicit Multi-Step Reasoning framework significantly outperforms all baselines ($p < 0.05$), demonstrating the effectiveness of explicit reasoning over implicit gender, thematic, and emotional cues in classical Chinese poetry. For example, expressions like 人晚倦梳裹，恨海茫茫*(In late years, weary of dressing; boundless sorrow stretches like the sea.)* reflect a delicate and introspective perspective often associated with female authorship, while historical allusions such as 孟尝*(Lord Mengchang, a historical figure)* indicate history-reflection themes, which are better captured by our framework.

6.4 Effectiveness of Implicit Multi-Step Reasoning

We conduct an ablation study with six variants to evaluate the impact of different reasoning compo-

Method	Gender			Theme			Emotion			Triple		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT	0.592	0.402	0.479	0.243	0.103	0.145	0.108	0.074	0.088	0.213	0.184	0.197
DeepSeek	0.698	0.563	0.623	0.424	0.255	0.318	0.102	0.196	0.134	0.163	0.139	0.150
Qwen	0.469	0.335	0.391	0.446	0.508	0.475	0.303	0.269	0.285	0.224	0.243	0.233
T5	0.563	0.781	0.655	0.327	0.262	0.291	0.167	0.150	0.158	0.183	0.166	0.174
LLaMA	0.772	0.847	**0.808**	0.433	0.675	0.527	0.514	0.477	0.495	0.291	0.347	0.317
Roberta	0.782	0.828	0.804	0.531	0.443	0.483	0.314	0.351	0.332	0.365	0.381	0.373
Bert-wwm	0.786	0.829	0.807	0.518	0.498	0.508	0.317	0.319	0.318	0.286	0.432	0.344
Macbert	0.784	0.826	0.804	0.438	0.485	0.461	0.306	0.287	0.296	0.363	0.352	0.357
Bao25	0.835	0.812	0.823	0.438	0.456	0.447	0.312	0.326	0.319	0.360	0.348	0.354
Shahnazari25	0.852	0.834	0.843	0.444	0.457	0.450	0.264	0.293	0.278	0.325	0.352	0.349
Ours	0.764	0.803	0.783	0.622	0.598	**0.610**	0.616	0.711	**0.660**	0.591	0.516	**0.551**

Table 2: Overall performance of baseline models measured by Precision (P), Recall (R), and F1

Method	Gender	Theme	Emotion	Triple
Baseline	0.808	0.527	0.495	0.317
+Struct	0.766	0.537	0.509	0.390
+Shallow	0.772	0.567	0.520	0.477
-Understanding	0.765	0.537	0.509	0.426
-Cue Extra	0.771	0.550	0.527	0.459
Full Reasoning	0.781	**0.610**	**0.660**	**0.551**

Table 3: Ablation study on different chain-of-thought reasoning strategies.

Model	Gender	Theme	Emotion	Joint
Full Model	**0.781**	**0.610**	**0.660**	**0.551**
-Gender	–	0.503	0.401	0.398
-Theme	0.748	–	0.400	0.379
-Emotion	0.774	0.540	–	0.503

Table 4: Ablation study on attribute interdependency. Each variant removes one attribute from the joint prediction framework.

nents (Table 3).

As can be seen from the table, the *Baseline* model achieves the highest gender score but performs poorly on theme, emotion, and joint prediction, indicating that direct prediction favors surface-level gender cues while lacking semantic consistency. Compared with the Baseline, *+Struct* slightly improves most metrics, while *+Shallow* further improves overall performance, indicating that even simple reasoning steps help capture richer contextual information.

For the variants with key reasoning steps removed, both *-Understanding* (removing the Semantic Understanding step) and *-Cue Extra* (removing the Perspective and Gender Cue Extraction step) show lower performance than the Full Reasoning variant in all metrics. This indicates that both gender inference and imagery analysis steps are crucial components of the Implicit Multi-Step Reasoning framework. For instance, imagery like 花*(flower)* and 酒*(drink)* in 月下花会我留酌*(Under the moon, among flowers, I linger to drink)* conveys subtle emotional tones that are difficult to capture without explicit reasoning.

6.5 Influence of Attribute Interdependency

We remove gender, theme, and emotion respectively to analyze their interactions, with a particular focus on examining the impact of gender information on the prediction of other attributes. The results in Table 4 show that removing any single attribute leads to a decline in overall performance, indicating strong interdependency among the three.

From the perspective of gender, removing gender information significantly degrades the performance of both theme and emotion prediction, suggesting that gender plays an important role in understanding poetic semantics and emotional expression. In contrast, removing theme or emotion has a relatively smaller effect on gender prediction, indicating that gender is more independent, while still providing valuable prior information for other attributes.

7 Conclusion

This study proposes a new task named Gender Analysis of Ancient Chinese Texts, focusing on ancient Chinese poetry with two core objectives: analyzing author gender and exploring poetic themes and emotions. We construct a corpus of 30,000 ancient Chinese poems together with an annotated balanced subset containing theme and emotion labels. Furthermore, we propose an Implicit Multi-Step Reasoning framework to model latent dependencies among gender, theme, and emotion. Experimental results demonstrate the effectiveness of the proposed framework for joint gender, theme, and emotion prediction in classical Chinese poetry.

Limitations

The limitations of our work can be summarized from several perspectives. First, while our approach focuses on semantic and stylistic features, other potentially informative factors (e.g., phonological or rhythmic characteristics) are not explicitly considered, which may be explored in future work. Second, our experiments are conducted on a dataset mainly drawn from the Ming and Qing dynasties. Although this provides a consistent setting, it remains valuable to examine the effectiveness of our approach across broader historical periods.

References

AI@Alibaba. 2024. Qwen2 technical report. Technical report.

Jinze Bai, Shuai Bai, Yunfei Yang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiaoyi Bao, Jinghang Gu, Zhongqing Wang, and Chu-Ren Huang. 2024. Employing glyphic information for chinese event extraction with vision-language model. In *Findings of EMNLP*.

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, and Chu-Ren Huang. 2025a. Revisiting classical chinese event extraction with ancient literature information. In *Proceedings of ACL*.

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, and Chu-Ren Huang. 2025b. Revisiting classical chinese event extraction with ancient literature information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8440–8451.

Baptiste Blouin, Cécile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in chinese historical newspapers. In *Proceedings of LREC-COLING 2024*, pages 385–394.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *NAACL*.

Tom Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Judith Butler. 1990. *Gender Trouble*. Routledge.

Kang-i Sun Chang and Haun Saussy. 1999. *Women Writers of Traditional China: An Anthology of Poetry and Criticism*. Stanford University Press.

Yiming Cui and Wanxiang Che. 2021. Pre-training with whole word masking for chinese bert. In *ACL*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2020a. Revisiting pre-trained models for chinese natural language processing. *Findings of EMNLP*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2020b. Revisiting pre-trained models for chinese nlp: Macbert. In *Findings of EMNLP*.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Xiaoyi et al. Dong. 2024. Internlm-xcomposer2: Vision-language large model.

Shivam Dubey. 2025. Activation steering for bias mitigation: An interpretable approach to safer llms. *arXiv preprint arXiv:2508.09019*.

Grace S. Fong and Ellen Widmer. 2017. *Nan Nü: Men, Women, and Gender in Early and Imperial China*. Brill.

Fabrizio Gilardi, Lucas Meyers, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *PNAS*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases. In *NAACL*.

Jiaxin Guo, Hao Cheng, and Xu Li. 2020. Revisiting pre-trained models for chinese nlp. In *Findings of EMNLP*, pages 657–668.

Xu Han, Tianyu Gao, and Yankai Lin. 2021. A novel pointer network for named entity recognition. In *Proceedings of ACL*.

Xiaoming Hu. 2004. *Zhongguo Gudai Funü Wenxue Shi Lun*. Fudan University Press. In Chinese.

Asad Khattak, Muhammad Zubair Asghar, Hassan Ali Khalid, and Hussain Ahmad. 2022. Emotion classification in poetry text using deep neural network. *Multimedia Tools and Applications*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

H. Liu and 1 others. 2022. Information extraction for classical chinese historical documents. *Journal of Chinese Information Processing*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Susan Mann. 1997. *Precious Records: Women in China's Long Eighteenth Century*. Stanford University Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. In *Proceedings of the NAACL-HLT 2013 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *ACL*.

Qian Nanxiu and Grace Fong. 2019. *Women and Literary Culture in Late Imperial China*. Brill.

Hongye Qiu, Yue Xu, Meikang Qiu, and Wenjie Wang. 2025. Dr.gap: Mitigating bias in large language models using gender-aware prompting with demonstration and reasoning. *arXiv preprint arXiv:2502.11603*.

Colin Raffel, Noam Shazeer, Adam Roberts, and 1 others. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Francisco Rangel and Paolo Rosso. 2013. Overview of the author profiling task at pan 2013. In *CLEF*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI*.

Kourosh Shahnazari and Seyed Moein Ayyoubzadeh. 2025. Who are you behind the screen? implicit mbti and gender detection using artificial intelligence. *arXiv preprint arXiv:2503.09853*.

Elaine Showalter. 1977. *A Literature of Their Own: British Women Novelists from Brontë to Lessing*. Princeton University Press.

Elaine Showalter. 1985. *The New Feminist Criticism: Essays on Women, Literature, and Theory*. Pantheon Books.

Xuemei Tang, Zekun Deng, Qi Su, Hao Yang, and Jun Wang. 2024. Chisiec: An information extraction corpus for ancient chinese history. In *Proceedings of LREC-COLING 2024*, pages 3391–3400.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xu Wang, Tianyu Gao, and Zhen Li. 2021. Kepler: A unified model for knowledge embedding and pretrained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Shuying Wei. 1998. *Zhongguo Lidai Funü Shige Yanjiu*. Shanghai Ancient Books Publishing House. In Chinese.

Virginia Woolf. 1929. *A Room of One's Own*. Hogarth Press.

Ikuya Yamada, Hiroyuki Shindo, and Hideaki Takeda. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of EMNLP*, pages 6442–6454.

Renhan Zhang, Lian Lian, Zhen Qi, and Guiran Liu. 2025. Semantic and structural analysis of implicit biases in large language models: An interpretable approach. *arXiv preprint arXiv:2508.06155*.

Y. Zhang and 1 others. 2021. Constructing historical knowledge graphs from classical chinese texts. *Digital Scholarship in the Humanities*.

曹道衡，沈玉成. 2003. 中古文学家丛考. 中华书局出版社, Beijing.

袁行霈. 1994. 中国诗学通论. 新华书店出版社, 北京. In Chinese.

A Annotation Example

We provide an example to illustrate the annotation process:

> **Poem:** 乘鸾夜下蓬莱岛，闲辗麟车踏瑶草。海风吹折碧桃花，玉盘满摘安期枣。

Interpretation: The poem depicts an imaginative and transcendental scene. Imagery such as 鸾车 (phoenix carriage), 瑶草 (jade grass), and 碧桃 (immortal peach blossoms) are commonly associated with celestial or mythical realms, creating an ethereal and refined tone.

Gender inference: The delicate, elegant, and slightly detached stylistic features are more frequently observed in poems written by female poets. Therefore, the gender is inferred as **female**.

Emotion annotation: The poem conveys a sense of calmness and detachment from worldly concerns, corresponding to the category of **tranquility**.

Theme annotation: The use of symbolic imagery to express personal ideals aligns with the category of **object-description with implicit aspiration** (咏物言志).

Final labels: (*gender* = *female*, *theme* = 咏物言志, *emotion* = *tranquility*)

This example demonstrates how annotators integrate semantic interpretation, stylistic cues, and literary conventions to assign consistent labels.

B Discussion

B.1 Gendered Stylistic Patterns and Their Limits

As shown in Table 2, gender prediction in classical Chinese poetry consistently exceeds random baselines across different modeling paradigms, providing quantitative evidence that gender-related stylistic tendencies indeed exist at the corpus level. In particular, most pretrained language models achieve gender F1 scores above 0.80, including LLaMA (0.808), Roberta (0.804), Bert-wwm (0.807), and Bao25 (0.823). These results suggest that gender information is implicitly encoded in poetic expression and can be captured through contextual semantic modeling. Importantly, such signals rarely rely on explicit gender markers. Instead, they emerge from latent stylistic patterns, including imagery preference, thematic orientation, emotional expression, and narrative perspective.

The results further indicate that gender-related stylistic cues are substantially more stable than emotion or theme signals. While several baseline models achieve strong gender performance, their emotion and triple prediction results remain considerably lower. For example, LLaMA obtains an F1 score of 0.808 for gender prediction but only 0.495 for emotion and 0.317 for triple prediction. Similarly, Bert-wwm achieves 0.807 on gender but only 0.318 on emotion. This discrepancy suggests that gender-related patterns are comparatively easier to capture as global stylistic tendencies, whereas emotion and thematic interpretation require deeper semantic understanding and contextual reasoning.

At the same time, gender prediction should not be interpreted deterministically. Poetic style is shaped not only by author gender but also by historical context, literary conventions, genre constraints, and individual authorial intention (Butler, 1990). Male poets may deliberately adopt restrained or delicate lyrical styles, while female poets may engage extensively with historical, political, or landscape-oriented themes. Therefore, the patterns identified in this study should be understood as corpus-level statistical regularities rather than essential or fixed characteristics of individual authors.

B.2 Role of CoT Reasoning in Modeling Implicit Style

As shown in Table 3, CoT reasoning substantially improves the model's ability to capture implicit stylistic semantics, particularly for emotion, theme, and joint prediction tasks. Compared with the Baseline model, the Full Reasoning framework improves theme F1 from 0.527 to 0.610, emotion F1 from 0.495 to 0.660, and triple prediction F1 from 0.317 to 0.551. Notably, the improvement in triple prediction exceeds 23 percentage points, indicating that structured reasoning is especially effective for jointly modeling multiple semantic attributes.

In contrast, the gains for gender prediction alone remain relatively limited. The baseline already achieves 0.808 gender F1, while the full reasoning framework reaches 0.781. This observation suggests that gender-related stylistic patterns can often be partially captured through global semantic representations alone, whereas emotion and theme require more complex interpretive reasoning over contextual interactions. In other words, CoT reasoning contributes less to isolated stylistic classification but becomes increasingly important when the task involves implicit semantic dependency modeling.

The progressive improvements from *+Struct* to *+Shallow* and finally to *Full Reasoning* further support this interpretation. Triple prediction F1 increases from 0.390 under structured prompting to 0.477 with shallow reasoning, eventually reaching 0.551 under the full hierarchical reasoning framework. Moreover, removing key reasoning stages leads to consistent performance degradation. Eliminating the semantic understanding module reduces triple F1 from 0.551 to 0.426, while removing cue extraction decreases it to 0.459. These results indicate that semantic understanding and perspective-related cue extraction are critical intermediate reasoning stages for modeling implicit poetic semantics.

B.3 Advantages of CoT Reasoning

The proposed CoT reasoning framework introduces a structured interpretive mechanism into the joint prediction of gender, theme, and emotion. Rather than directly mapping text representations to out-

put labels, the framework decomposes prediction into semantically connected intermediate reasoning stages, enabling the model to progressively refine its understanding of the poem. Such a design is particularly suitable for classical Chinese poetry, where semantic meaning is often condensed, implicit, and highly context-dependent.

Results in Table 2 show that CoT reasoning is especially effective for tasks requiring higher-level semantic abstraction. Our method achieves the best performance on theme prediction (0.610), emotion prediction (0.660), and triple prediction (0.551), substantially outperforming both traditional pretrained language models and recent reasoning-based baselines. For example, compared with Shahnazari25, the proposed framework improves emotion F1 from 0.278 to 0.660 and triple prediction F1 from 0.349 to 0.551. These gains indicate that explicit reasoning structures help the model better integrate imagery semantics, contextual implication, and cross-attribute dependencies.

Furthermore, the hierarchical reasoning design improves robustness under challenging interpretive conditions, such as abstract imagery, semantic ambiguity, and weak gender cues. By forcing the model to construct intermediate semantic representations before final prediction, CoT reasoning reduces reliance on superficial lexical correlations and encourages deeper contextual interpretation.

B.4 Implications for Computational Literary Studies

This study demonstrates that NLP methods, particularly large language models augmented with CoT reasoning, can support large-scale stylistic analysis of classical Chinese poetry. The quantitative results reported in Tables 2 and 3 suggest that computational models are capable of capturing latent literary patterns that extend beyond explicit lexical semantics. In particular, the substantial improvements brought by hierarchical reasoning indicate that literary interpretation benefits from modeling structured semantic dependencies rather than treating stylistic attributes independently.

Such approaches complement traditional close reading by enabling large-scale quantitative exploration of implicit stylistic phenomena. Through computational modeling, it becomes possible to identify corpus-level regularities in imagery usage, emotional orientation, thematic preference, and perspectival expression that would be difficult to systematically analyze manually across large poetic collections. Our findings also quantitatively support traditional literary observations regarding gendered stylistic tendencies in classical Chinese poetry. Moreover, the severe imbalance between male- and female-authored poems reflects broader historical asymmetries in literary preservation and authorship visibility.

Nevertheless, computational analysis should not replace humanistic interpretation. Classical Chinese poetry is highly condensed, culturally situated, and historically contextualized, meaning that many interpretive nuances remain difficult to fully encode computationally. Therefore, the results presented in this study should be regarded as supportive and exploratory tools for literary scholarship rather than definitive interpretations of poetic meaning.

C Case Study

To further demonstrate the effectiveness of our method, we present representative examples comparing baseline models and our approach, as shown in Table 5.

In the first example, all methods correctly identify the gender as male, suggesting that certain stylistic patterns related to masculine literary expression can be captured even by general pretrained models. However, substantial differences emerge in theme and emotion prediction. The poem contains strong historical and political allusions, such as references to Sun Quan, Liu Bei, and Tao Yuanming, which implicitly convey dissatisfaction with political reality and frustration over unrealized ambition. Our framework successfully identifies the theme as 怀古咏史 and the emotion as 怀才不遇, which are fully consistent with the gold annotations. In contrast, GPT3 captures the emotional implication of frustration but fails to recognize the historical-reflective theme, while Qwen interprets the poem primarily as symbolic self-expression and predicts 咏物言志. These results indicate that baseline models tend to focus on isolated lexical or imagery-level cues, whereas our framework better captures the deeper semantic interaction between historical reference, symbolic imagery, and emotional implication.

The second example further illustrates the importance of multi-step reasoning for implicit gender inference. The poem describes hardship, loneliness, and familial responsibility through expressions such as 扶母走长途 ("supporting one's mother on a long journey"), which reflects a deli-

Original Text	Results
Title: 辛丑大计闻之哑然 *(Hearing of the grand plans of the Xinchou year, one is left speechless.)*	
Poem: 孙刘要使不三公，点滓微云混太空。比似陶家栽五柳，便无槐棘也春风。 *(Sun Quan and Liu Bei, denied the rank of the Three Excellencies, are like drifting specks clouding the vast sky. Were it like Tao Yuanming planting his five willows, even without thorny growth, all would rest in the grace of spring wind.)*	**Golden:** male; 怀古咏史; 怀才不遇 **GPT3:** male; 怀才不遇; 悲凉惆怅 **Qwen:** male; 咏物言志; 高洁孤傲 **Ours:** male; 怀古咏史; 怀才不遇
Title: 怀竹孙弟 *(Thinking of my younger brother Sun Zhu.)*	
Poem: 身世谁怜早岁孤，风尘扶母走长途。池塘春早连天碧，为问行人得句无。 *(Who would pity a life lonely from early years? Through dust and hardship, I support my mother on a long journey. By the pond, spring arrives early, its green stretching to the horizon—may I ask a passerby whether a verse has yet been formed?)*	**Golden:** female; 羁旅思乡; 羁旅愁思 **GPT3:** male; 羁旅思乡; 羁旅愁思 **Qwen:** male; 羁旅思乡; 羁旅愁思 **Ours:** female; 羁旅思乡; 羁旅愁思

Table 5: Case study comparing GPT3, Qwen, and our method.

cate and restrained emotional perspective characteristic of the annotated female writing style in the dataset. Although GPT3 and Qwen correctly identify the theme and emotion as 羁旅思乡 and 羁旅愁思, both incorrectly predict the gender as male, suggesting that they rely excessively on superficial semantic patterns associated with travel imagery and wandering narratives. By contrast, our framework successfully infers the correct female gender while simultaneously predicting the correct theme and emotion. This result demonstrates that the proposed reasoning framework can better integrate narrative perspective, emotional tone, and contextual semantics during inference, thereby reducing ambiguity in gender prediction.

Overall, the case studies suggest that the proposed Implicit Multi-Step Reasoning framework is more effective at modeling latent semantic dependencies among gender, theme, and emotion. Rather than making isolated predictions based on surface-level textual patterns, the framework progressively constructs structured semantic interpretations, enabling more coherent and contextually grounded prediction results.