From *Shijing* to English and German: Resources and Evaluation for LLM Translation of Early Chinese Poetry

Ying Jiao
KU Leuven, Leuven.AI
Belgium
ying.jiao@kuleuven.be

Meng Sun
Shanghai International Studies University
P. R. China
sunmeng920@gmail.com

Abstract

While large language models (LLMs) show promise in literary translation, *Shijing* (The Book of Songs) serves as a rigorous yet underexplored testbed for testing their limits, given its linguistic antiquity and complex poetic constraints. Automated evaluation in this domain is currently hindered by a scarcity of multilingual resources and the inadequacy of existing metrics in capturing both semantic fidelity and aesthetic quality. In this paper, we bridge these gaps by curating a *Shijing* parallel corpus with line-by-line Chinese-English-German alignments, together with a fine-grained lexical knowledge base (KB) for archaic expressions. Based on these resources, we propose a hybrid evaluation framework that integrates knowledge-driven, rule-based, and LLM-as-judge metrics. Experimental results show that our framework achieves significantly higher human correlation than traditional metrics and demonstrates high statistical stability. By applying this framework to evaluate representative LLMs, we reveal that while top-tier models like Gemini-2.5-Pro and DeepSeek-3.1 show potential, achieving semantic precision and aesthetic sophistication—particularly in lower-resource directions like German—remains a persistent challenge. Our code, lexical KB, and corpus reconstruction protocols are available at https://github.com/ML-KULeuven/ShijingLLMTrans.

1 Introduction

Poetry translation is uniquely challenging, requiring the preservation of semantic meaning alongside stylistic and cultural characteristics. Large language models (LLMs) have recently demonstrated strong performance in literary translation, including poetry, often outperforming traditional machine translation systems in capturing content, tone, and aesthetic qualities (Wang et al., 2024; Chen et al., 2025). While promising results have been reported for modern and Tang-Song poetry, early Chinese poetry, as exemplified by *Shijing* (The Book of Songs), remains largely unexplored, leaving open questions about LLMs' ability to handle archaic lexemes and formulaic structures.

As the earliest collection of Chinese poetry, *Shijing* dates from the Western Zhou Dynasty to the Spring and Autumn periods (eleventh to sixth centuries BCE). It stands as the fountainhead of the Chinese literary tradition and provides an indispensable record of the pre-Qin socio-cultural landscape (McNaughton, 1963; Granet, 2015). Translating this canon poses unique linguistic and cultural challenges that distinguish it from later poetic forms (see Figure 1 for an example). These challenges are demonstrated in its sophisticated syntax and imagery design (Yu, 1983; Zhi, 2007), as well as in expressions deeply embedded in historical contexts that complicate interpretation through modern knowledge alone. Consequently, these multifaceted characteristics render *Shijing* a compelling case for examining the capacity of LLMs to perform context-sensitive semantic inference under linguistic ambiguity, and to achieve genre-sensitive generation when engaging with archaic poetic forms.

Beyond the intrinsic linguistic and interpretive challenges of *Shijing*, systematic research is hindered by a severe lack of evaluation resources. Most authoritative translations are restricted to English, with even fewer in languages like German (Song, 2025). Crucially, to our knowledge, fine-grained lexical resources for cross-lingual adequacy evaluation remain largely absent. Furthermore, existing metrics like BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) fall short of capturing the rigid formal constraints and nuanced literariness essential to poetry, as they focus on surface or semantic similarity (Chakrabarty et al., 2021; Thai et al., 2022; Chen et al., 2025).

To address these gaps, we curate a multilingual

Findings of the Association for Computational Linguistics: ACL 2026, pages 11143–11162

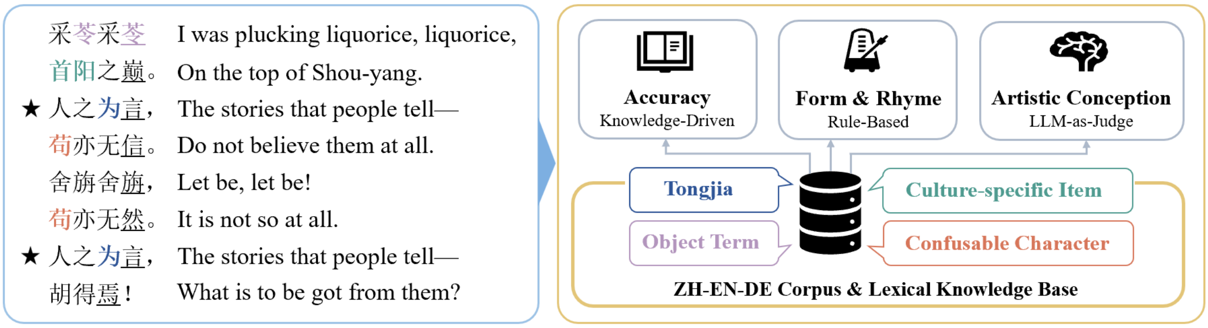

Figure 1: Illustrative example and framework overview. Left: *Shijing* example with translation from Waley (1937). ⋆, underlines, and different colors denote incremental repetitions, rhymes, and different lexical categories, respectively. Right: Multilingual resources and the hybrid evaluation framework.

parallel corpus for *Shijing* [1], featuring line-by-line Chinese-English-German alignments across multiple authoritative translations. This is complemented by a lexical knowledge base (KB) that maps linguistically complex and culturally-loaded archaic expressions to their English and German equivalents. Leveraging these resources, we propose a hybrid evaluation framework that integrates knowledge-driven metrics for semantic accuracy, rule-based metrics for form and rhyme, and selective LLM-as-judge scoring for artistic conception.

Our experimental results validate the effectiveness of the proposed framework. Compared to existing metrics, our method achieves significantly higher correlations with human judgments and robustly consistent system-level rankings across both languages. Applying this framework to a benchmark of seven representative LLMs, we find that while top-tier models such as Gemini-2.5-Pro (Google DeepMind, 2025) and DeepSeek-3.1 (DeepSeek, 2025) show promising potential, achieving high-quality translation remains a persistent challenge for all models. Furthermore, our analysis reveals a notable performance disparity between languages, with models consistently underperforming in the lower-resource Chinese-to-German direction compared to Chinese-to-English.

Our work makes three main contributions: (i) Multilingual Resources: We curate a *Shijing* Chinese-English-German parallel corpus and construct a lexical KB, providing a vital foundation for cross-lingual research on early Chinese poetry; (ii) Hybrid Evaluation Framework: We introduce a multi-dimensional evaluation approach that combines knowledge-driven, rule-based, and LLM-based metrics, significantly outperforming traditional baselines in human correlation; (iii) LLM Evaluation: We evaluate seven representative LLMs to reveal their strengths and limitations in archaic poetry translation, establishing a benchmark for future cross-lingual research in this domain.

2 Related Work

2.1 Classical Chinese Poetry Resources

Progress in classical Chinese poetry translation has benefited from parallel corpora, annotated datasets, and knowledge bases. Chen et al. (2019) released a corpus of Chinese quatrains annotated with fine-grained emotion labels, while Li et al. (2021) developed a parallel dataset of classical Chinese poems and their modern translations for evaluating semantic understanding. Liu et al. (2020) and Chen et al. (2025) introduced classical Chinese poetry datasets enriched with metadata such as authorship, topics, and historical context. Most of these resources, however, focus on Tang and Song poetry, leaving *Shijing*, a foundational corpus of early Chinese poetry, largely underrepresented. We address this gap by curating a parallel Chinese-English-German corpus for *Shijing*, together with a multilingual lexical knowledge base.

2.2 Poetry Translation

Translating poetry requires preserving semantic meaning, aesthetic style, and cultural resonance, which poses challenges for conventional machine translation (Chakrabarty et al., 2021). Early studies showed that statistical and neural machine translation models struggled with poetic form, such as rhyme and meter (Genzel et al., 2010), motivating style-aware or domain-adapted approaches (Yang

[1]Due to copyright restrictions on the specific *Shijing* edition and several modern translations, we will release a corpus reconstruction protocol and representative samples to ensure reproducibility. Please refer to Appendix A for details.

et al., 2018; Song et al., 2023). Recent work demonstrates that LLMs achieve promising results in poetry translation (Wang et al., 2024; Chen et al., 2025). Wang et al. (2024) studied modern English poetry translation into Chinese, while Chen et al. (2025) focused on Tang and Song poetry into English. However, translating *Shijing*, a foundational corpus of early Chinese poetry, presents distinct challenges due to its archaic language, dense cultural references, and repetitive structures, making it an useful testbed for LLM translation abilities.

2.3 Poetry Translation Quality Evaluation

Standard machine translation evaluation metrics such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020) developed for general-domain translation correlate poorly with human judgments on literal (Thai et al., 2022; Zhang et al., 2025) and poetry (Chakrabarty et al., 2021) translation tasks. In the poetic translation domain, Genzel et al. (2010) emphasized the importance of rhyme and meter, using rule-based computations to optimize these aspects in statistical machine translation outputs. More recently, Chen et al. (2025) employed knowledge-driven metrics for accuracy of classical Chinese poem translation along with LLM-based judgment for sound, form, and meaning. Their results indicate that LLM-based assessment of aesthetic quality aligns most closely with human evaluations, while sound and form scores show larger discrepancies. Our work builds on this line and proposes a hybrid evaluation approach that integrate knowledge-driven, rule-based, and selective LLM-based metrics, combining interpretability with the advantages of LLMs for capturing poetic qualities.

3 Dataset and Resources

3.1 Multilingual Parallel Corpus

Shijing presents a comprehensive picture of social life in China during the Zhou Dynasty and is traditionally divided into three sections: *Feng* (Songs), *Ya* (Odes), and *Song* (Hymns). In this work, we focus on *Feng* as a representative example, while noting that our resource construction and evaluation framework can be readily applied to the remaining sections as well as to other poetic corpora. *Feng* accounts for more than half of the collection (160 out of 305 poems) and exhibits a rich diversity of themes and forms. Its vivid depiction of the emotional experiences and sociocultural realities of commoners has also attracted comparatively more attention in human translation studies, making it particularly suitable for both resource development and systematic evaluation (Wagner, 2007; Kim, 2016).

	Chinese	English	German
Unique Tokens	1676	10882	7826
Avg. Tokens / Line	4.1	6.2	5.3
Avg. Lines / Poem	16.3	16.3	16.3
Avg. Tokens / Poem	66.6	100.4	87.1

Table 1: Statistics of the parallel corpus.

We curate a multilingual parallel corpus based on *Feng* (Cheng and Jiang, 2017), in which all source poems are in simplified Chinese, to enable cross-lingual analysis. The corpus contains 160 classical Chinese poems, each aligned line-by-line with multiple high-quality human translations by recognized translators. Specifically, each Chinese poem is aligned with four independent English translations (Legge, 1871; Jennings, 1891; Waley, 1937; Xu, 1993) and two independent German translations (von Strauß, 1880; Simon, 2015), capturing both the inherent ambiguity and stylistic diversity of poetic translation. Table 1 presents statistics of the multilingual parallel corpus. Chinese token counts are computed at the character level, while English and German tokens are whitespace-based. The target-language translations have substantially larger vocabularies than the Chinese originals, highlighting the broader lexical choice space and the inherent complexity of poetic translation.

3.2 Lexical Knowledge Base

The lexical knowledge base (KB) is constructed by an expert [2] consulting multiple classical commentaries (Karlgren, 1950; Gao, 2009; Zhu, 2011; Zhou, 2013; Granet, 2015; Lv, 2015; Cheng and Jiang, 2017). Each entry corresponds to a word or multi-character expression in classical Chinese and includes its associated poem and line, as well as multiple English and German renderings. The target-language renderings are provided by the expert based on source-side semantic interpretations drawn from these authoritative commentaries. This design ensures high-quality semantic annotations that are directly usable for translation evaluation.

The resulting KB contains 1195 entries in total, averaging 7.5 entries per poem. Entries are selected based on expert judgment, and only words

[2] A researcher specialized in the translation of *Shijing* with 6-year experience.

or expressions that are linguistically challenging, prone to confusion, or of particular cultural significance are included, making the overall coverage comparable to that of classical commentaries. This ensures that the KB emphasizes high-value items that are most informative for evaluating translation quality. Each entry is classified into one of four types: tongjia (phonetic loan character) [3], term for object, culture-specific item, and confusable character [4]. Table 2 provides representative examples of each category; additional samples are included in Appendix F.1.

Beyond its role in translation evaluation, the KB constitutes a curated inventory of archaic lexical-semantic conventions grounded in authoritative classical commentaries. Given that early classical texts share a substantial cultural and conceptual substrate, the KB may also support tasks involving related pre-Qin corpora and later poetic traditions that reuse and reinterpret canonical expressions. More broadly, it holds potential value as an independent linguistic resource for diachronic lexical analysis, culturally grounded term interpretation, and systematic error diagnosis in cross-lingual generation involving early Chinese texts. While the current version is expert-curated to ensure quality, we plan to investigate semi-automatic methods for scaling the KB in future work.

4 Evaluation Framework

Following prior work that characterizes poetry translation quality in terms of accuracy and literariness (Raffel, 1988; Robinson, 2010), we model the overall quality score of a translation t as a weighted combination of these two dimensions:

$$\mathrm{Ovl}(t) = \alpha \cdot \mathrm{Acc}(t) + (1-\alpha) \cdot \mathrm{Lit}(t),$$

where $\alpha \in [0, 1]$ controls the relative importance of the two components. In this work, we set $\alpha = 0.5$ to treat accuracy and literariness equally, but the weighting can be adjusted depending on evaluation priorities. Notably, we apply a zero-score penalty to all metrics if the translation fails to be rendered in the target language. Further details on fail translation detection are provided in Appendix B. In the following subsections, we describe in detail how $\mathrm{Acc}(t)$ and $\mathrm{Lit}(t)$ are computed.

[3]Tongjia is homophonous or near-homophonous character employed to convey the semantic meaning of the word it phonetically represent.

[4]Confusable character refers to archaic lexeme that is orthographically identical to modern Chinese but possess distinct, often deceptive, meanings.

4.1 Accuracy

We define poem-level accuracy based on the coverage of lexical knowledge base (KB) entries (see Section 3.2). For a given translation t of source poem s, accuracy is computed as

$$\mathrm{Acc}(t) = \frac{\#_{\text{matched}}(t)}{\#_{\text{total}}(s)},$$

where $\#_{\text{total}}(s)$ is the number of KB entries associated with s, and $\#_{\text{matched}}(t)$ is the number of matched entries in the translation.

To determine whether an entry is matched in a translation, we first align its source line to a corresponding target line. Let a source poem contain L_s lines and its translation contain L_t lines. For a source line indexed by $i \in \{0, \dots, L_s - 1\}$, we approximate its aligned target line index as

$$\hat{j} = \mathrm{round}\left(\frac{i}{L_s} \cdot L_t\right),$$

where $\mathrm{round}(\cdot)$ denotes standard rounding to the nearest integer. Given the creative nature of poetic translation, we define a local search window $\mathcal{W}(\hat{j}) = \{\hat{j}-1, \hat{j}, \hat{j}+1\} \cap \{0, \dots, L_t - 1\}$, to allow limited restructuring of content in the translation.

An entry is considered matched if the search window $\mathcal{W}(\hat{j})$ contains any of the provided renderings associated with the specific target language. We adopt exact string matching to prioritize precision. This approach may produce false negatives and potentially leading to an underestimation of accuracy scores. The impact on recall is partially mitigated by applying stemming to both KB renderings and target translations, and by including multiple human-provided renderings per KB entry to broaden lexical coverage. More flexible matching strategies remain a direction for future work.

4.2 Literariness

Building on prior analyses of the aesthetic characteristics of *Shijing* (Saussy, 1997; Smith, 2015), we define the literariness of a translation $Lit(t)$ as the average of three dimensions: rhyme, form, and artistic conception.

4.2.1 Rhyme Score

We quantify the rhyme quality of a translation t as the proportion of lines that participate in rhyming patterns, defined as

$$\mathrm{Rhyme}(t) = \frac{\#_{\text{rhymed}}(t)}{\#_{\text{total}}(t)}.$$

Category	Lexicon	Poem Source	English	German
Tongjia	养	中心养养(二子乘舟)	restless, unrest, unease	unruhig, Bangen, Sorge
Object Term	蒲	有蒲与荷(泽陂)	cattail, rush	Binse, Rohrkolben
Culture-specific Item	归	之子于归(桃夭)	wed, marriage, bride	Ehe, Braut, heiraten
Confusable Character	微	式微式微(式微)	dusk, darkness	dunkeln, dämmern, dunkel

Table 2: Examples from the lexical knowledge base. "Poem Source" represents original line (title).

Higher values indicate stronger and more consistent rhyming throughout the poem. Our computational framework maintains consistent criteria for English and German translations, enabling a comparable cross-linguistic assessment of rhyme quality.

To determine $\#_{\text{rhymed}}(t)$, we follow the rhyme definition of Greene et al. (2012) as the phonetic linkage between line endings achieved through the identity of stressed vowels and all succeeding phonemes. Let $R = [r_0, r_1, \ldots, r_{L-1}]$ denote the sequence of rhymes of a translation t with L lines. We scan R sequentially to identify lines that participate in rhyming sequences according to two criteria inspired by Baxter (1992) and Cai (2008):

- Consecutive Rhyme: any two or more consecutive lines with identical rhyme endings are considered rhymed.
- Pattern-based Rhyme: for segments of common stanza length n, we normalize the rhymes to a canonical pattern (mapping the first unique rhyme to "A", the next to "B", etc.). If the normalized pattern matches any predefined canonical pattern of the same length, all lines in the segment are considered rhymed.

The pseudocode of rhyme score computation is shown in Algorithm 1.

Algorithm 1: Rhyme Score Computation

```
Input  : List of extracted rhymes R of translation t,
         Length of translation L, Canonical pattern
         dictionary P
Output : Rhyme score Rhyme(t)
rhymed_lines ← 0;
i ← 0;
while i < L do
    // Check consecutive rhymes (e.g., AA)
    if lines starting at i share the same rhyme ending
      then
        k ← count of consecutive lines;
        rhymed_lines ← rhymed_lines + k;
        i ← i + k;
    else
        match_found ← false;
        // Check specific patterns by
           length (e.g., ABAB, ABCABC)
        foreach n in descending order of pattern
          lengths do
            if i + n < L then
                S ← R[i : i + n];
                S_norm ← Normalize(S) ;
                  // Canonical form
                if S_norm ∈ P[n] then
                    rhymed_lines ←
                      rhymed_lines + n;
                    i ← i + n;
                    match_found ← true;
                    break;

        // No pattern matched, skip line
        if not match_found then
            i ← i + 1;

return rhymed_lines/L;
```

4.2.2 Formal Score

A defining formal characteristic of *Shijing* is incremental repetition, where identical or near-identical lines recur across stanzas to reinforce rhythm and meaning (Shaughnessy, 1992; Yen, 2021). To evaluate whether translations preserve this repetitive structure, we measure the extent to which line-level repetition in the source poem is reflected in the translation. We define the formal score as

$$\text{Form}(t) = \min\left(1, \frac{\#_{\text{repeated}}(t)}{\#_{\text{repeated}}(s)}\right).$$

This formulation assigns a maximum score of 1 when the translation fully preserves the degree of repetition observed in the source, while avoiding rewarding over-repetition beyond the original structure. If the source poem contains no repeated lines (i.e., $\#_{\text{repeated}}(s) = 0$), we define $\text{Form}(t) = 0$.

4.2.3 Artistic Conception Score

We employ LLMs to assess the artistic conception of translations, motivated by recent findings that LLM-based evaluations demonstrate high alignment with human aesthetic judgment in poetry and other creative domains (Chiang and Lee, 2023; Sawicki et al., 2025).

Following the protocol of Chen et al. (2025), we use the same 5-point Likert scale (Likert, 1932) for LLM evaluation. Specifically, a score of 1 indicates poor preservation of imagery and artistic effect, 3 denotes an adequate but unremarkable rendering, and 5 corresponds to an excellent translation that strongly preserves the original imagery, atmosphere, and overall artistic conception (see Appendix C.4 for the full prompts). Both rhyme and formal scores naturally fall in the range $[0, 1]$. To ensure consistency in scale, we normalize the artistic conception score to $[0, 1]$ using $\mathrm{AC}_{\mathrm{norm}}(t) = \frac{\mathrm{AC}(t)-1}{4}$.

5 Experiments

5.1 Research Questions

Our experiments aim to answer the following research questions:

- **Evaluation Validity:** Is the proposed evaluation framework more consistent with human judgments than existing automatic metrics for assessing translations of poems in the *Feng* section of *Shijing*?
- **LLM Translation Performance:** How well do LLMs translate the *Feng* section of *Shijing*, and how does translation quality vary between English and German?

5.2 Experimental Setup

5.2.1 Human Evaluation

Human evaluation is performed by three expert annotators on a stratified subset of 60 poems. The stratification is based on translation difficulty, captured by poem length and the density of linguistically complex and culturally-loaded archaic expressions (see Appendix C.1 for details). This subset is used to validate the correlation between automatic metrics and human judgments. Each translation of the poems in this subset, produced by the LLMs in English and German, is assessed along three dimensions: accuracy, literariness, and overall quality, all scored on a 1–5 scale. The annotation protocol and scoring guidelines are described in Appendix D.1. Human judgments show strong reliability across evaluation dimensions for both English and German, as evidenced by high Krippendorff's α (Krippendorff, 2019) and excellent Intra-class Correlation Coefficient (Shrout and Fleiss, 1979) for aggregated scores (Appendix D.2).

5.2.2 Translation Systems

We evaluate seven LLMs for translating poems from the *Feng* section of *Shijing* into English and German. The closed-source models are GPT-5 (OpenAI, 2025), Gemini-2.5-Pro (Google DeepMind, 2025), ERNIE-4.5 (Baidu Inc., 2025), and Spark-4.0 (iFLYTEK, 2025), while the open-source models are DeepSeek-3.1 (DeepSeek, 2025), Qwen-3-14B (Yang et al., 2025), and LLaMA-3.1-8B (Meta AI, 2024). Among these, ERNIE, Spark, DeepSeek, and Qwen are Chinese-developed models trained extensively on Chinese corpora. All models are evaluated in a zero-shot setting. The full prompts and all hyperparameter configurations are provided in Appendix C.2.

5.2.3 Baseline Evaluation Metrics

We compare our evaluation framework against widely used automatic metrics in machine translation. Surface-level metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), which measure n-gram overlap between translation and reference. We also include chrF++ (Popović, 2017), a character-level F-score metric suitable for short texts such as poetry. Embedding- and model-based semantic metrics include BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020), capturing semantic similarity beyond surface forms. Implementation details are provided in Appendix C.3.

5.2.4 Our Evaluation Framework Implementation

For evaluating artistic conception, we employ Gemini-2.5-Pro, as it achieved the highest overall and literariness ranking among the seven evaluated LLMs on our human-scored 60-poem subset. All prompts, technical details, and hardware specifications are provided in Appendix C.4.

5.2.5 Data Contamination Analysis

To discern whether LLM performance stems from genuine capability or memorization, we evaluate potential contamination using CoDeC (Zawalski et al., 2025) on Qwen3-14B and LLaMA-3.1-8B. We construct seven evaluation corpora: one Shijing-source-poem-only and six paired with specific translation mentioned in Section 3.1. Pythia-1.4B (Biderman et al., 2023), trained on the documented Pile corpus (Gao et al., 2020), serves as the uncontaminated baseline.

Pythia produces near-zero scores across all seven datasets. Both Qwen and LLaMA achieve comparable scores on the six poem–translation paired corpora, providing no evidence of direct contamination with the translation datasets. For the Chinese-only source corpus, LLaMA scores near zero while

Qwen yields a relatively higher score (0.27), suggesting potential exposure to classical Chinese text distributions during pretraining, even without direct access to *Shijing* itself. Nonetheless, this value remains well below the 0.80 threshold indicative of direct contamination (Zawalski et al., 2025).

We note that contamination detection methods including CoDeC are not fully reliable indicators (Samuel et al., 2025). These results should therefore be interpreted as evidence against direct memorization rather than a definitive guarantee of contamination-free evaluation.

Nonetheless, our evaluation framework is by design robust to potential memorization: accuracy is assessed against a semantically grounded lexical knowledge base constructed independently of any of the six translations, and literariness is evaluated through rule-based structural criteria and LLM-based judgment of poetic atmosphere rather than comparison to specific reference editions. Consequently, reproducing a historically established translation would not automatically yield a high score, as the scoring function rewards semantic precision and structural quality rather than surface overlap with any reference.

5.3 RQ1: Evaluation Validity

To evaluate how well automatic metrics align with human judgments, we compute correlations at both poem and system levels. We report Spearman correlation (Spearman, 1961) as the primary measure, as it reflects ranking consistency, while Pearson (Pearson, 1895) and Kendall correlations (Kendall, 1938) are included in Appendix E.1.

On the human-evaluated 60-poem subset, we evaluate four aspects: poem-level accuracy, literariness, and overall quality, as well as system-level overall quality. Poem-level correlations are computed by comparing automatic and human scores over all individual poem translations produced by the LLMs (60×7 instances per language), while system-level correlations are computed by comparing average scores over the 60 poems for each LLM. Our evaluation framework explicitly produces separate scores for accuracy and literariness. In contrast, baseline metrics output a single overall score; for these metrics, we compute correlations between the metric scores and the corresponding human ratings for each aspect. For reference-based metrics, scores are averaged over all references for each language (four for English and two for German). We report ROUGE-L as a representative ROUGE variant, with results for ROUGE-1 and ROUGE-2 provided in Appendix E.2.

Table 3 compares our framework with baseline metrics across all aspects and target languages. Across both English and German, our evaluation framework consistently achieves substantially higher correlations than all baseline metrics at both poem and system levels. At the poem level, our method exhibits strong alignment with human judgments for accuracy, literariness, and overall quality, whereas baseline metrics show substantially weaker correlations across aspects. At the system level, our framework produces system rankings that are fully consistent with human judgments on this subset, while baseline metrics exhibit ranking discrepancies. The correlations achieved by our framework are stable across all evaluation aspects, with narrow bootstrap confidence intervals (CIs), indicating robustness to sampling variation (see Appendix E.3 for complete CI results).

We further conduct ablation studies to assess the robustness of our framework to the LLM evaluator and the validity of its design choices (see Appendix E.4). First, replacing the artistic conception evaluator with Qwen-3-14B—a smaller, locally deployable model—yields correlations that remain substantially higher than baseline metrics, confirming that our framework's effectiveness does not rely on a specific proprietary model. Second, ablating individual dimensions (accuracy vs. literariness) and literary sub-dimensions leads to decreased performance, demonstrating that each component contributes meaningfully to the overall assessment.

5.4 RQ2: LLM Translation Performance

We apply our evaluation framework to assess English and German translations of seven LLMs on the full multilingual corpus of 160 poems. Table 4 summarizes system-level scores, which are computed by averaging the evaluation results over the 160 translations produced by each LLM for each target language.

The results suggest that current LLMs exhibit encouraging capability in translating early Chinese poetry from the *Feng* section of *Shijing*, but still fall short of high-quality literary translation. When mapped to the 1–5 human evaluation scale, overall scores range from approximately 2.0 to 3.6. According to our human rating criteria, this corresponds to translations spanning from basic to the upper end of average quality. Notably, even the strongest model does not consistently reach

	English				German			
Metric	Acc-P	Lit-P	Ovl-P	Ovl-S	Acc-P	Lit-P	Ovl-P	Ovl-S
BLEU	0.219	0.058	0.164	0.643	0.234	0.075	0.182	0.821
ROUGE-L	0.350	0.091	0.269	0.607	0.288	0.130	0.237	0.857
METEOR	0.256	0.082	0.214	0.857	0.239	0.107	0.200	0.857
chrF++	0.261	0.059	0.184	0.929	0.252	0.113	0.215	0.893
BERTScore	0.342	0.132	0.272	0.786	0.325	0.186	0.289	0.929
BLEURT	0.014	0.070	0.085	0.429	0.147	0.286	0.220	0.286
COMET	0.183	0.097	0.181	0.786	0.230	0.172	0.223	0.857
Ours	**0.818**	**0.813**	**0.810**	**1.000**	**0.824**	**0.807**	**0.814**	**1.000**

Table 3: Spearman correlation between automatic metrics and human judgments on the 60-poem subset. Acc-P, Lit-P, and Ovl-P denote poem-level accuracy, literariness, and overall quality, respectively; Ovl-S denotes system-level overall quality.

	English			German		
Model	Overall	Accuracy	Literariness	Overall	Accuracy	Literariness
Closed-source models						
GPT-5	0.568 (3.272)	0.480	0.656	0.546 (3.185)	0.453	0.640
Gemini-2.5 Pro	**0.655** (3.619)	**0.610**	**0.700**	**0.617** (3.468)	**0.549**	**0.685**
ERNIE-4.5	0.558 (3.232)	0.475	0.640	0.479 (2.914)	0.411	0.546
Spark-4.0	0.505 (3.022)	0.455	0.556	0.336 (2.343)	0.302	0.370
Open-source models						
DeepSeek-3.1	<u>0.566</u> (3.263)	<u>0.506</u>	<u>0.625</u>	<u>0.483</u> (2.934)	<u>0.386</u>	<u>0.581</u>
Qwen3-14B	0.499 (2.994)	0.407	0.590	0.363 (2.452)	0.267	0.459
LLaMA-3.1-8B	0.316 (2.265)	0.233	0.400	0.247 (1.987)	0.113	0.380

Table 4: System-level evaluation results on translations of the full 160-poem corpus using our evaluation framework. Scores lie in $[0, 1]$ (higher is better); Overall scores are additionally mapped to a 1–5 scale for interpretability (shown in parentheses).

the "good" level, indicating that while current LLMs show promising potential, high-quality poetry translation remains a challenging task. Descriptions of the 1–5 rating criteria are provided in Appendix D.1. We emphasize that the mapping to the 1–5 scale is introduced solely for interpretability and does not affect the quantitative analysis.

Among all models, Gemini-2.5-Pro achieves the strongest performance across accuracy, literariness, and overall quality in both English and German, indicating a balanced advantage rather than strength in a single dimension. While closed-source models generally exhibit higher performance than open-source ones, some open-source models approach or even surpass certain closed-source counterparts. In particular, DeepSeek-3.1 stands out among open-source models, achieving competitive performance across languages, which may be attributed to its strong Chinese-language modeling and training on large-scale bilingual data.

Across all evaluated models, we observe a consistent performance disparity, with German translations systematically underperforming their English counterparts. This gap potentially stems from the relative paucity of Chinese-German training data,

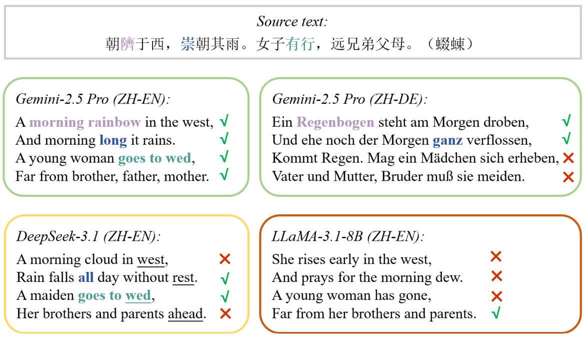

Figure 2: Translation comparison for a *Shijing* stanza. Color spans and underlines indicate challenging expressions and rhymes, respectively.

which limits the models' ability to capture the complex semantic and formal mappings required for archaic Chinese-to-German translation. Furthermore, the morphological complexity of German and the structural differences in its poetic tradition may pose additional challenges for models trained primarily on more abundant English-centric data.

5.5 Case Study

We present a case study of a stanza from the *Shijing* poem *Didong* (Rainbow) to illustrate the varying capabilities of LLMs in translating early Chinese poetry. Gemini-2.5-Pro demonstrates the most robust semantic fidelity in its English translation. While it offers vivid imagery and accurate content, the translation lacks a formal rhyme scheme. DeepSeek-3.1 represents the most capable open-source model. It achieves a consistent rhyme scheme at the cost of accuracy: it fails to recognize the term for "rainbow" and the final line "ahead" loses the original logic of the woman's separation from her family. LLaMA-3.1-8B yields a substandard translation. Its deficiency in both semantic precision and aesthetic grace reflects the inherent limitations of smaller open-source models when tasked with the dense semantics of archaic Chinese poetry.

While Gemini achieves near-perfect accuracy in English, its German translation shows a marked decline in quality. Although it correctly identifies the rainbow ("Regenbogen"), it misses the central theme of marriage and the motif of the woman leaving her family. This discrepancy supports our conclusion that the paucity of Chinese-German parallel corpora hinders models from achieving the same level of cross-linguistic mapping as seen in the Chinese-English task.

6 Conclusion

In this work, we present a comprehensive evaluation ecosystem for the multilingual translation of *Shijing*. By curating a Chinese-English-German parallel corpus and constructing a fine-grained lexical knowledge base, we effectively bridge the resource gap for archaic Chinese poetry. Our proposed hybrid evaluation framework demonstrates superiority over existing metrics, achieving significantly higher correlations with human judgment and providing a reliable benchmark for this nuanced domain.

Our evaluation reveals that while current LLMs show promising potential, there remains substantial room for improvement in both semantic precision and aesthetic sophistication. Furthermore, the performance disparity in the Chinese-to-German direction underscores the challenges of literary translation in lower-resource scenarios. By establishing a robust benchmark, we hope to catalyze future advancements in cross-lingual literary research, moving closer to models that can truly capture the cultural resonance and aesthetic integrity of archaic poetry across linguistic boundaries.

Limitations

Despite the contributions of our framework, several limitations remain.

First, while our human-evaluated 60-poem subset demonstrates high statistical stability, it represents only a fraction of the entire *Shijing* corpus. Expanding the scale of expert validation would further strengthen the benchmark.

Second, our work is currently limited to English and German. The framework's generalizability to linguistically distant families remains to be explored.

Third, the use of LLMs as evaluators introduces potential self-preference bias and the risk of cultural homogenization. Models might prioritize modern, standardized linguistic fluency over the unique, archaic aesthetic nuances of the original text. More broadly, translation quality is inherently subjective: different translators and readers may prioritize semantic fidelity, aesthetic form, or cultural resonance differently, and no single evaluation framework can fully capture this diversity of interpretive stances. This underscores the risk of misinterpretation if automated metrics are used in isolation. We thus emphasize that our framework serves as a complementary tool, intended to assist rather than replace human philological expertise.

Fourth, while our CoDeC analysis provides no evidence of direct memorization of the evaluated translations, the memorization–generalization distinction cannot be fully resolved for closed-source models, and residual ambiguity remains an inherent limitation of LLM benchmark evaluation.

Acknowledgments

Ying Jiao acknowledges support from the EU Framework Program for Research and Innovation Horizon under the Grant Agreement No 101073307 (MSCA-DN LeMuR).

References

Baidu Inc. 2025. Ernie 4.5 technical report. Accessed: 2026-01-02.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

William H Baxter. 1992. *A Handbook of Old Chinese Phonology*. Mouton de Gruyter, Berlin.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.

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

Carnegie Mellon University. 2014. The CMU pronouncing dictionary.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036.

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

Junying Cheng and Jianyuan Jiang. 2017. *Commentary on Shijing (诗经注析)*. Zhonghua Book Company, Beijing. In Chinese; ISBN: 9787101126914.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

DeepSeek. 2025. Deepseek v3.1 release. Accessed: 2026-01-02.

Reece Dunn, Alexander Epaneshnikov, and others. 2024. eSpeak-NG: Compact open source speech synthesizer. Open source text-to-speech software, GPL-3.0 licensed.

Heng Gao. 2009. *Modern Annotations on the Book of Songs (诗经今注)*. Shanghai Classics Publishing House, Shanghai. In Chinese; ISBN: 9787532553136;.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Josef Och. 2010. "poetic" statistical machine translation: rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166.

Google DeepMind. 2025. Gemini 2.5: Our newest gemini model with thinking. Accessed: 2026-01-02.

Marcel Granet. 2015. *Festivals and songs of ancient China*. Routledge, New York.

Roland Greene, Stephen Cushman, Clare Cavanagh, Jahan Ramazani, and Paul Rouzer. 2012. *The Princeton Encyclopedia of Poetry and Poetics*. Princeton University Press, Princeton.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and others. 2020. spacy: Industrial-strength natural language processing in python.

iFLYTEK. 2025. iflytek spark large model series (including 4.0). Accessed: 2026-01-02.

William Jennings. 1891. *The Shi King. The old "Poetry Classic" of the Chinese*. George Routledge and Sons, London.

Bernhard Karlgren. 1950. *The Book of Odes*. Museum of Far Eastern Antiquities, Stockholm.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Ha Poong Kim. 2016. *Joy and Sorrow Songs of Ancient China: A New Translation of Shi Jing Guo Feng*. Liverpool University Press, Liverpool.

Klaus Krippendorff. 2019. *Content analysis: An introduction to its methodology*. Sage publications, Thousand Oaks.

James Legge. 1871. *The Chinese Classics. Vol. IV. The She-King, or the Book of Poetry*. Lane, Crawford Co, Hong Kong.

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

Rensis Likert. 1932. *A technique for the measurement of attitudes*. Archives of Psychology, New York.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yutong Liu, Bin Wu, and Ting Bai. 2020. The construction and analysis of classical chinese poetry knowledge graph. *Journal of Computer Research and Development*, 57(6):1252–1268.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Hualiang Lv. 2015. *Shijing: An Analysis of Objects - Feng (诗经名物注析国风篇)*. Huangshan Book Company, Hefei. In Chinese; ISBN: 9787546150451;.

William McNaughton. 1963. The composite image: Shy jing poetics. *Journal of the American Oriental Society*, 83(1):92–106.

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. Accessed: 2026-01-02.

OpenAI. 2025. Gpt-5 is here. Accessed: 2026-01-02.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762.

Burton Raffel. 1988. *The art of translating poetry*. Pennsylvania State University Press, University Park, London.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Peter Robinson. 2010. *Poetry & Translation: The Art of the Impossible*. Liverpool University Press, Liverpool.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. 2025. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5058–5070.

Haun Saussy. 1997. Repetition, rhyme, and exchange in the book of odes. *Harvard Journal of Asiatic Studies*, 57(2):519–542.

Piotr Sawicki, Marek Grześ, Dan Brown, and Fabrício Góes. 2025. Can large language models outperform non-experts in poetry evaluation? a comparative study using the consensual assessment technique. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 31901–31918.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Edward L Shaughnessy. 1992. Marriage, divorce, and revolution: Reading between the lines of the book of changes. *The Journal of Asian Studies*, 51(3):587–599.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428.

Rainald Simon. 2015. *Shijing: Das altchinesische Buch der Lieder*. Reclam, Ditzingen. In German; ISBN: 9783150108659;.

Jonathan Smith. 2015. Sound symbolism in the reduplicative vocabulary of the shijing. *Journal of Chinese Literature and Culture*, 2(2):258–285.

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

Yunhong Song. 2025. The research trends and prospect of translation studies of the book of songs — based on citespace literature visualization analysis (《诗经》翻译研究动态与展望-基于CiteSpace的文献可视化分析). *Modern Linguistics*, 13:284–292.

Charles Spearman. 1961. The proof and measurement of association between two things. In J. J. Jenkins and D. G. Paterson, editors, *Studies in Individual Differences: The Search for Intelligence*, pages 45–58. Appleton-Century-Crofts, New York. Reprinted from The American Journal of Psychology, 15, 72–101 (1904).

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902.

Victor von Strauß. 1880. *Schī-kīng: Das kanonische Liederbuch der Chinesen*. Carl Winter's Universitätsbuchhandlung, Heidelberg. In German; https://books.google.de/books/about/Schi_king.html?id=NhF4em_jMZ8C&redir_esc=y.

Hans-Günter Wagner. 2007. *Hell ein Vogelruf ertönt: Altchinesische Volkslyrik aus dem Buch der Lieder und Gedichte (Shijing–Guofeng)*. YinYang Media Verlag, Kelkheim. In German; ISBN: 9783935727129;.

Arthur Waley. 1937. *The Book of Songs*. Houghton Mifflin Company, Boston.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yuanchong Xu. 1993. *Book of poetry (诗经)*. Hunan Publishing House, Changsha. In Chinese and English; ISBN: 7543806878;.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3960–3969.

Shih-hsuan Yen. 2021. A tentative discussion of some phenomena concerning early texts of the shi jing. *Bamboo and Silk*, 4(1):45–93.

Pauline R Yu. 1983. Allegory, allegoresis, and the classic of poetry. *Harvard Journal of Asiatic Studies*, 43(2):377–412.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. 2025. Detecting data contamination in llms via in-context learning. *arXiv preprint arXiv:2510.27055*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are llms for literary translation, really? literary translation evaluation with humans and llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. https://arxiv.org/abs/1904.09675.

Chen Zhi. 2007. *The Shaping of the Book of Songs: From Ritualization to Secularization*. Steyler Verlag, Nettetal.

Zhenfu Zhou. 2013. *Translation and Annotation of the Book of Songs (诗经译注)*. Zhonghua Book Company, Beijing. In Chinese; ISBN: 9787101093414;.

Xi Zhu. 2011. *Collected Commentaries on the Book of Songs (诗集传)*. Zhonghua Book Company, Beijing. In Chinese; ISBN: 9787101077162;.

A Copyright and Data Availability Statement

A.1 Source Material and Copyright Status

The *Shijing* corpus utilized in this study is compiled from multiple sources with varying copyright statuses. We categorize these materials into three groups to ensure full compliance with intellectual property rights:

1. Original Chinese Text: We adopt the edition by Cheng and Jiang (2017), which is protected by its own copyright. While the ancient text of *Shijing* is in the public domain, this specific modern collation and arrangement are proprietary.

2. Open-Domain Translations: The translations by Legge (1871), Jennings (1891), and von Strauß (1880) are in the public domain globally.

3. Copyrighted Translations: The modern translations by Xu (1993), Waley (1937), and Simon (2015) remain under active copyright protection.

A.2 Data Distribution Policy

To facilitate research while respecting intellectual property, we will implement the following distribution strategy upon the publication of this work. We will provide the full text for the open-domain translations. For the original Chinese text and the copyrighted translations, we will provide representative snippets for illustrative purposes under "fair use" guidelines for academic research. For the copyrighted portions not included in the release, we will provide SHA-256 hashes for each line. This allows researchers who independently acquire the texts to verify that their data matches our experimental setup.

A.3 Reproducibility and Reconstruction

To ensure our corpus can be fully reconstructed by other researchers, we will provide detailed ISBNs, editions, and stable URLs for all source books. For poems where we performed manual line-merging or splitting to achieve line-by-line alignment, we provide a structured mapping file. This file includes the poem title, the original line numbers, and the specific coordinates of our modifications (e.g., "Line [line index]-[line index] merged," "Split at [character index]").

A.4 Licensing and Terms of Use

Our release will be under an open-source license (e.g., MIT License) to support future research. All data and resources released through this project are provided strictly for non-commercial research and educational purposes.

B Failure Translations

A generation is flagged as a failure if it meets any of the following three criteria:

1. Empty or Null Responses: Instances where the model fails to generate any content or explicitly returns a "None" string.
2. Untranslated Source Retention: We identify cases where the model fails to perform the translation task and instead outputs exclusively Chinese text. Specifically, a response is categorized as a failure if it contains at least one CJK Unified Ideograph but contains no Latin characters (including German-specific umlauts and eszett).
3. Lexical Invalidity (Garbled Text): To filter out gibberish or non-lexical "hallucinations," we calculate a validity ratio for each generation. A word is considered "valid" if its cleaned form (stripping punctuation) exists in the target language's frequency dictionary (English or German). If the ratio of valid words to total words falls below a threshold of 0.5, the output is marked as garbled.

C Experimental Setup Details

All experiments were conducted on a workstation with 2 $\times$ NVIDIA RTX A5000 (24GB VRAM each), 128-thread processor architecture, 256 GB RAM. Note that all tasks were configured to utilize a single GPU.

C.1 Data Subset Selection

To ensure that the 60-poem subset used for human evaluation is representative of the entire 160-poem corpus, we implement a stratified sampling strategy based on text complexity.

For each poem i in the 160-poem corpus, we define a Composite Complexity Score (S_i) based on poem length (L_i) and the density of challenging expressions (D_i). L_i is calculated as the total Chinese character number in the poem. D_i is calculated as the ratio of entries in our lexical knowledge base corresponding to the poem relative to its total character length. Both metrics are normalized to the range $[0, 1]$. The final score for each poem is calculated as $S_i = \text{Norm}(L_i) + \text{Norm}(D_i)$.

The 160 poems were ranked in ascending order according to their composite scores S_i. We divide the ranked list into five equal strata and randomly select 12 poems to be included in the evaluation subset. To ensure the reproducibility of the selection, we utilized a fixed random seed (Seed=42) for all sampling operations. The *Feng* section contains 15 parts, each represents poems from a specific region. We guaranteed the 60 subset includes at least one poem from each of the 15 regions. This approach ensures that the subset covers a balanced spectrum of translation difficulties while maintaining geographical and cultural diversity.

C.2 Translation Models

We evaluate seven LLMs: GPT-5 (OpenAI, 2025) via OpenAI gpt-5 API, Gemini-2.5-Pro (Google DeepMind, 2025) via Google gemini-2.5-pro API, ERNIE-4.5 (Baidu Inc., 2025) via Baidu QianFan ernie-4.5-turbo-128k API, Spark-4.0 (iFLYTEK, 2025) via iFlytek 4.0Ultra API, DeepSeek-3.1 (DeepSeek, 2025) via Baidu QianFan deepseek-v3.1-250821 API rather than locally deployed due to computational constraints, as well as Qwen-3-14B (Yang et al., 2025)[5] and LLaMA-3.1-8B (Meta

[5]https://huggingface.co/Qwen/Qwen3-14B

AI, 2024)[6] via HuggingFace (Wolf et al., 2020).

Gemini is evaluated using temperature=0, top_p=1, top_k=0, and 10000 maximum output tokens for both target languages. GPT is used with its default setting (temperature=1.0) as the API restricts temperature adjustment with 10000 maximum completion tokens for both target languages. Spark is evaluated using temperature=0, top_p=1, top_k=0 with 1200 and 1600 maximum output tokens for English and German, respectively. For ERNIE and DeepSeek accessed via the Qianfan API, which does not support zero temperature, we apply a near-greedy setting (temperature=0.01, top_p=1, top_k=0, seed=42) with 1200 and 1600 maximum output tokens for English and German, respectively. LLaMA and Qwen are evaluated using greedy decoding (do_sample=False) with 1200 and 1600 maximum new tokens for English and German, respectively. The translation generation of open-source models completed within 3 GPU hours. The prompts used for all LLMs are shown in Table 5.

English Prompt
You are an expert in classical Chinese poetry and translation studies. Translate the following poem from the Book of Songs (Shijing) into English. Ensure that the translation is faithful to the original meaning while preserving its poetic beauty and rhythm. Only output the translation. Do not add any explanations or notes. `{poem_text}`
German Prompt
Du bist ein Experte für klassische chinesische Poesie und Übersetzungswissenschaften. Übersetze das folgende Gedicht aus dem Buch der Lieder (Shijing), einer klassischen Sammlung altchinesischer Dichtung, ins Deutsche. Stelle sicher, dass die Übersetzung dem ursprünglichen Sinn treu bleibt und gleichzeitig die poetische Schönheit und den Rhythmus bewahrt. Geben Sie nur die Übersetzung aus. Füge keine Erklärungen oder Anmerkungen hinzu. `{poem_text}`

Table 5: Prompts used for zero-shot translation of *Shijing* into English and German.

C.3 Baseline Evaluation Metrics

The specific implementations and configurations of the baseline evaluation metrics are detailed as follow. We report BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) using the default signatures of the sacreBLEU (Post, 2018) package. METEOR (Banerjee and Lavie, 2005) is computed using the NLTK implementation (Loper and Bird, 2002). We report ROUGE scores (Lin, 2004) using the rouge-score implementation. For semantic similarity, we utilize bert_score (Zhang et al., 2020) with the default roberta-large model. We utilize the comet package with the Unbabel/wmt22-comet-da model (Rei et al., 2022) and the bleurt package with the BLEURT-20 checkpoint (Pu et al., 2021).

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

C.4 Our Implementation Details

We utilize the spaCy library (Honnibal et al., 2020) with "en_core_web_sm" for English and "de_core_news_sm" for German to perform stemming. All tokens are converted to lowercase during the matching process.

We extract phonetic representations using the CMU Pronouncing Dictionary (Carnegie Mellon University, 2014) via NLTK (Loper and Bird, 2002) for English and espeak-ng (Dunn et al., 2024) for German.

Following the rhyme analysis of *Shijing* by Cai (2008) and Baxter (1992), we define a set of target rhyme schemes based on stanza length. For 4-line stanzas, we consider `abab`, `abba`, and `abcb`. For 6-line stanzas, we detect `abcabc` and `ababab` patterns.

For the evaluation of artistic conception, we employ Gemini-2.5-Pro (Google DeepMind, 2025) and Qwen-3-14B (Yang et al., 2025). Gemini with temperature=0, top_p=1, top_k=0, and 10000 maximum output tokens. Qwen with do_sample=False and 10000 maximum new tokens. The prompts following the design of Chen et al. (2025) are provided in Table 6.

D Human Evaluation

D.1 Human Evaluation Protocol

Human evaluation plays a central role in this work and serves as the reference for validating the proposed automatic evaluation framework. We invited three expert annotators, all PhD researchers specializing in translation studies. All annotators are proficient in Chinese, English, and German, and their research focuses on *Shijing* and its translation. They were fully informed about the research's objectives. We obtained their explicit consent before the evaluation process began, and they were aware that their anonymized feedback would be used for academic analysis and publication. The evaluators participated on a voluntary basis as members of the research group.

Human evaluation is conducted on a stratified

English Prompt
/* Task */ Evaluate the translation of classical Chinese poetry from the Book of Songs (Shijing) for the beauty of meaning, focusing on whether the translation effectively conveys the themes, emotions, and messages of the original. This includes the use of concise and precise language to create vivid imagery and a rich atmosphere. 1 point: Poor translation, fails to convey the depth and richness of the original poetry. 2 point: Basic translation with significant shortcomings in capturing themes, emotions, and messages. 3 point: Average translation, conveys basic themes and emotions but lacks refinement or depth. 4 point: Good translation, effectively captures most themes, emotions, and messages with minor imperfections. 5 point: Excellent translation, accurately conveys the depth, richness, and atmosphere of the original poetry with full thematic and emotional resonance. /* Input */: Original Chinese poem: `{source}` English translation: `{translation}` /* Evaluation (score only) */:
German Prompt
/* Aufgabenbeschreibung */ Bewerten Sie die Übersetzung eines klassischen chinesischen Gedichts aus dem Buch der Lieder (Shijing) im Hinblick auf die Bedeutungsschönheit, also darauf, ob die Übersetzung die Themen, Emotionen und zentralen Aussagen des Originals angemessen vermittelt. Dazu zählt auch der Einsatz knapper und präziser Sprache, die eindrucksvolle Bilder und eine stimmige Atmosphäre entstehen lässt. 1 Punkt: Schlechte Übersetzung, gibt Tiefe und Reichtum des Originals kaum wieder. 2 Punkte: Einfache Übersetzung mit deutlichen Mängeln bei der Vermittlung von Themen, Emotionen und Aussagen. 3 Punkte: Durchschnittliche Übersetzung, vermittelt grundlegende Themen und Emotionen, bleibt jedoch in Feinheit und Tiefe zurück. 4 Punkte: Gute Übersetzung, erfasst die meisten Themen, Emotionen und Aussagen überzeugend, mit nur geringfügigen Schwächen. 5 Punkte: Hervorragende Übersetzung, gibt Tiefe, Reichtum und Atmosphäre des Originals präzise wieder und erreicht eine hohe thematische und emotionale Resonanz. /* Eingabedaten */: Originaltext des chinesischen Gedichts: `{source}` Deutsche Übersetzung: `{translation}` /* Bewertung (nur Punktzahl) */:

Table 6: Prompts used for artistic conception evaluation.

subset of 60 poems. For each poem in the subset, translations produced by the seven LLMs (see Section 5.2.2) in both English and German are evaluated. Each translation is assessed along three dimensions: accuracy, literariness, and overall quality, using a 1–5 Likert scale (Likert, 1932).

Before the annotation process, we developed detailed scoring guidelines for each evaluation dimension. The guidelines were established in consultation with two senior scholars specializing in *Shijing* studies and classical Chinese poetry translation. To calibrate the annotators' understanding of the criteria, they first independently evaluated five English translations and five German translations based on the guidelines, followed by a discussion to resolve discrepancies and achieve a shared interpretation of the scoring standards. After this calibration phase, the remaining translations were annotated independently.

For each translation, the final human evaluation score for each dimension is computed as the median of the scores assigned by the three annotators. The median was chosen to ensure a more robust measure of central tendency, thereby mitigating the impact of potential outliers or a single divergent rater that might otherwise skew the final results in our 3-rater setup.

The guidelines for each dimension are:

Dimension 1: Accuracy

5 (Excellent) The translation is perfectly accurate. All challenging elements—including culture-specific items, terms for object, phonetic loan characters, and confusable characters—are precisely conveyed.

4 (Good) The translation is mostly accurate, with the core meaning correctly preserved. Most challenging terms and cultural nuances are handled appropriately.

3 (Average) The general meaning is correct, but there are notable inaccuracies in translating challenging terms.

2 (Basic) Only part of the meaning is correctly conveyed. There are significant mistranslations or omissions that distort the original intent.

1 (Poor) The translation completely deviates from the original meaning or is entirely incoherent.

Dimension 2: Literariness

5 (Excellent) Exceptional literary quality with a strong sense of rhyme. The translation perfectly reproduces the formal beauty and aesthetic imagery of the original poem.

4 (Good) Good literary quality. Most lines follow rhyme schemes, successfully capturing most of the original form, imagery, and mood.

3 (Average) Average literary quality. The presentation of rhyme schemes as well as the reproduction of poetic form and aesthetic imagery is inconsistent throughout the translation.

2 (Basic) Basic literary quality. The translation feels prosaic, with significant flaws in rhyme, formal structure, or the evocation of imagery.

1 (Poor) No recognizable poetic structure or rhyme. The translation fails to convey any aesthetic or formal beauty of the original poem.

Dimension 3: Overall Quality

5 (Excellent) A superior translation that excels in both accuracy and poetic elegance, achieving a perfect balance among meaning, rhyme, form, and imagery.

4 (Good) A high-quality translation. The meaning is largely correct, and the translation possesses satisfying rhyming, formal and aesthetic beauty.

3 (Average) An acceptable translation. The meaning is basically correct, showing a noticeable but inconsistent effort to maintain poetic rhyme, form and imagery.

2 (Basic) A subpar translation with significant deficiencies in either semantic accuracy or literary expression.

1 (Poor) Fails to convey the original meaning and lacks any recognizable poetic or formal structure.

D.2 Inter-Annotator Agreement (IAA)

To validate the reliability of our human evaluation, we calculate inter-annotator agreement (IAA) across the three human evaluation dimensions: accuracy, literariness, and overall quality. We report two complementary metrics: Krippendorff's α with an ordinal metric (Krippendorff, 2019) and the Intra-class Correlation Coefficient (ICC) (Shrout and Fleiss, 1979). Specifically, for ICC, we utilize the two-way random-effects model. We report both the single rater absolute agreement (ICC2) and the average rater absolute agreement (ICC2k). The latter is particularly relevant as our final gold-standard scores are derived from the median of the three annotators' ratings. The agreement coefficients are calculated by concatenating all annotations across all seven LLMs and 60 poems per language, resulting in $N = 420$ translations per annotator for each language. The results for English and German translations are summarized in Table 7.

The IAA results demonstrate high reliability across all dimensions. According to the criteria suggested by Cicchetti (1994), the ICC2 and ICC2k scores indicate excellent agreement. All ICC2k values exceed 0.92, demonstrating that the pooled judgment used in our experiments is highly robust and reliable.

Furthermore, all Krippendorff's α coefficients are above 0.72, which satisfies the threshold for reliable conclusions in content analysis (Krippendorff, 2019). Notably, the accuracy dimension achieves the highest consistency, while literariness remains firmly within the substantial agreement category. This slight variation is theoretically expected, given the inherent subjectivity of aesthetic appreciation compared to the more objective nature of semantic fidelity. Overall, these high agreement scores validate that our proposed rubrics (see Appendix D.1) provided precise and consistent guidance to the annotators.

E Extended Results and Ablation Studies

E.1 Pearson and Kendall Correlation Results

We report Pearson's r (Pearson, 1895) and Kendall's τ (Kendall, 1938) correlation coefficients in Table 8. These additional results consistently align with the Spearman's ρ (Spearman, 1961) reported in the main paper (Table 3), reinforcing the observations that our framework consistently outperforms all baselines across all evaluation aspects and achieves perfect alignment with human rankings at the system level.

E.2 ROUGE-1 and ROUGE-2 Spearman Correlation Results

We provide the Spearman correlation results for ROUGE-1 and ROUGE-2 in Table 9. The results show that the performance of ROUGE-1 and ROUGE-2 is generally comparable to that of ROUGE-L (Table 3), while both remain significantly weaker than our proposed method across all evaluation aspects.

E.3 Confidence Intervals

To evaluate the statistical stability of our framework, we compute 95% confidence intervals (CIs) for all correlation measures across English and German. Following standard practices, we perform $B = 10,000$ bootstrap iterations with replacement from the human-evaluated samples. The 2.5th and 97.5th percentiles of the resulting bootstrap distri-

Language	Dimension	Krippendorff's α	ICC2 (Single)	ICC2k (Average)
English	Accuracy	0.81	0.85 [0.82, 0.87]	0.94 [0.93, 0.95]
	Literariness	0.74	0.81 [0.78, 0.83]	0.93 [0.91, 0.94]
	Overall	0.77	0.81 [0.78, 0.83]	0.93 [0.91, 0.94]
German	Accuracy	0.83	0.87 [0.85, 0.89]	0.95 [0.95, 0.96]
	Literariness	0.73	0.83 [0.81, 0.86]	0.94 [0.93, 0.95]
	Overall	0.78	0.85 [0.82, 0.87]	0.94 [0.93, 0.95]

Table 7: IAA Results for English and German Poem Evaluations. ICC values are reported with 95% Confidence Intervals (CI) in brackets. All p-values < 0.001.

	English				German			
Metric	Acc-P	Lit-P	Ovl-P	Ovl-S	Acc-P	Lit-P	Ovl-P	Ovl-S
BLEU	.19 / .16	.07 / .04	.15 / .13	.61 / .52	.20 / .18	.13 / .06	.16 / .14	.74 / .62
ROUGE-1	.28 / .21	.10 / .04	.20 / .14	.73 / .62	.33 / .19	.22 / .05	.31 / .14	.87 / .71
ROUGE-2	.26 / .20	.10 / .06	.19 / .16	.69 / .62	.21 / .18	.11 / .06	.17 / .14	.83 / .71
ROUGE-L	.34 / .26	.13 / .07	.28 / .21	.72 / .52	.36 / .22	.28 / .10	.35 / .18	.82 / .71
METEOR	.26 / .19	.13 / .06	.23 / .17	.72 / .71	.32 / .18	.27 / .08	.33 / .15	.82 / .71
chrF++	.26 / .20	.11 / .05	.21 / .14	.86 / .81	.37 / .19	.37 / .09	.42 / .18	.81 / .81
BERTScore	.35 / .26	.18 / .10	.30 / .21	.76 / .71	.39 / .25	.35 / .14	.41 / .23	.88 / .81
BLEURT	.04 / .01	.11 / .05	.13 / .07	.26 / .33	.21 / .11	.38 / .22	.32 / .17	.46 / .14
COMET	.18 / .14	.14 / .08	.21 / .14	.82 / .71	.31 / .18	.36 / .13	.39 / .17	.85 / .71
Ours	**.81 / .72**	**.85 / .70**	**.82 / .68**	**.97 / 1.0**	**.84 / .73**	**.85 / .70**	**.83 / .70**	**.97 / 1.0**

Table 8: Pearson / Kendall correlation between automatic metrics and human judgments on the 60-poem subset. Acc-P, Lit-P, and Ovl-P denote poem-level accuracy, literariness, and overall quality; Ovl-S denotes system-level overall quality.

butions are reported as the lower and upper bounds of the CIs in Table 10.

As shown in the results, the CIs across all four evaluation aspects and both languages are consistently narrow. These concentrated intervals demonstrate that our framework's high correlation with human judgment is not a byproduct of sampling bias but reflects a robust and reliable evaluation capability for *Shijing* translations.

E.4 Ablation Studies

E.4.1 Artistic Conception Evaluator

Table 11 presents the evaluation performance when replacing the artistic conception evaluator with Qwen-3-14B (Yang et al., 2025), a smaller, locally deployable model. The results demonstrate that our framework maintains high correlations with human judgments at both the poem and system levels, consistently outperforming the baselines reported in Table 3. Notably, the system-level correlations (Ovl-S) consistently reach or approach 1.000, indicating near-perfect alignment with human rankings. Furthermore, the narrow CIs across all dimensions confirm that our framework's reliability remains robust regardless of the evaluator's parameter scale, ensuring high-quality evaluation even in resource-constrained environments.

E.4.2 Core Evaluation Dimensions

The ablation results in Table 12 demonstrate that the full evaluation framework consistently outperforms individual components across the evaluation aspects for overall quality. This synergy suggests that both accuracy and literariness are indispensable for human-like poetry evaluation. While both branches maintain strong correlations for system-level evaluation, the full version achieves perfect alignment by integrating complementary information of semantic accuracy and literary quality. The consistent performance patterns observed in both English and German further validate the robustness of our framework across different languages.

E.4.3 Literary Sub-components

The ablation results in Table 13 evaluate the contribution of each component—Rhyme, Form, and Artistic Conception (AC)—to the overall assessment of literariness. The full (Rhyme+Form+AC) configuration consistently achieves the highest correlation with human judgments in both English and German, confirming that literariness is a multi-dimensional construct that requires a holistic evaluation of both structural and semantic fea-

Metric	English				German			
	Acc-P	Lit-P	Ovl-P	Ovl-S	Acc-P	Lit-P	Ovl-P	Ovl-S
ROUGE-1	0.278	0.056	0.186	0.821	0.256	0.062	0.184	0.857
ROUGE-2	0.262	0.081	0.203	0.714	0.241	0.073	0.185	0.893

Table 9: Spearman correlation between ROUGE-1, ROUGE-2 and human judgments on the 60-poem subset.

Correlation	English				German			
	Acc-P	Lit-P	Ovl-P	Ovl-S	Acc-P	Lit-P	Ovl-P	Ovl-S
Spearman	[.77, .86]	[.77, .85]	[.77, .85]	[1.0, 1.0]	[.77, .87]	[.77, .84]	[.77, .85]	[1.0, 1.0]
Pearson	[.77, .85]	[.81, .88]	[.79, .85]	[.78, 1.0]	[.79, .87]	[.82, .87]	[.80, .86]	[.96, 1.0]
Kendall	[.67, .80]	[.66, .74]	[.65, .72]	[1.0, 1.0]	[.68, .77]	[.66, .74]	[.65, .74]	[1.0, 1.0]

Table 10: 95% confidence intervals (2.5th and 97.5th percentiles) for our framework on the 60-poem subset, estimated via 10,000 bootstrap iterations.

tures.Specifically, the removal of any single component leads to a performance degradation. In the English dataset, the Form+AC variant (omitting Rhyme) shows the most significant drop in correlation, suggesting that rhyme plays a dominant role in human perceptions of English poetic quality. Conversely, in German, the dependencies appear more balanced. These patterns underscore the necessity of the integrated framework in capturing the complex, synergistic nature of literary translation quality.

F Supplementary Examples

F.1 Lexical Knowledge Base Examples

Table 14 provides additional representative examples for each category of the lexical knowledge base, supplementing those discussed in the main text.

F.2 LLM Translation Examples

Figure 3 and Figure 4 present example translations across the LLMs we evaluated.

	English			German		
Correlation	Lit-P	Ovl-P	Ovl-S	Lit-P	Ovl-P	Ovl-S
Spearman	.80 [.75, .83]	.78 [.74, .82]	1.0 [1.0, 1.0]	.73 [.67, .77]	.78 [.73, .83]	.96 [.70, 1.0]
Pearson	.80 [.77, .84]	.78 [.74, .82]	.98 [.93, 1.0]	.82 [.78, .86]	.83 [.79, .86]	.98 [.97, 1.0]
Kendall	.68 [.64, .72]	.65 [.61, .70]	1.0 [1.0, 1.0]	.63 [.58, .68]	.67 [.62, .71]	.90 [.56, 1.0]

Table 11: Ablation results with Qwen-3-14B as the backbone evaluator. Cells report the mean correlation coefficients and the corresponding 95% confidence intervals.

	English		German	
Variant	Ovl-P	Ovl-S	Ovl-P	Ovl-S
Full (Acc+Lit)	0.81 / 0.82 / 0.68	1.00 / 0.97 / 1.00	0.81 / 0.83 / 0.70	1.00 / 0.97 / 1.00
Acc Only	0.67 / 0.65 / 0.56	0.96 / 0.96 / 0.90	0.68 / 0.67 / 0.58	0.89 / 0.95 / 0.81
Lit Only	0.60 / 0.67 / 0.50	0.93 / 0.97 / 0.81	0.65 / 0.71 / 0.54	0.96 / 0.93 / 0.90

Table 12: Ablation results comparing the full evaluation framework with variants using only accuracy or only literariness. Spearman / Pearson / Kendall correlations are reported on the 60-poem subset.

Variant	English	German
Full (Rhyme+Form+AC)	0.81 / 0.85 / 0.70	0.81 / 0.85 / 0.70
Rhyme+Form	0.78 / 0.75 / 0.67	0.66 / 0.67 / 0.56
Rhyme+AC	0.69 / 0.71 / 0.58	0.67 / 0.68 / 0.58
Form+AC	0.36 / 0.50 / 0.30	0.64 / 0.70 / 0.56

Table 13: Ablation study on the literariness evaluation component. We compare the full framework (Rhyme + Form + Artistic Conception) against variants using only two components. Spearman / Pearson / Kendall correlations are reported on the 60-poem subset.

Category	Lexicon	Poem Source	English	German
Tongjia	流	左右流之(关雎)	pick	pflücken, sammeln
Tongjia	说	我心则说 (草虫)	joy, happy	froh, freudig, Freude
Tongjia	畜	畜我不卒(日月)	love	lieben
Tongjia	骄	维莠骄骄 (甫田)	tall, high	hoch, stolz
Object Term	薪	翘翘错薪 (汉广)	firewood, fuel wood	Reisig
Object Term	鹊	维鹊有巢(鹊巢)	magpie	Elster
Object Term	荼	谁谓荼苦(谷风)	sowthistle, sow-thistle	Gänsedistel
Object Term	桐	椅桐梓漆(定之方中)	wutong, parasol tree	Parasolbaum
Culture-specific Item	狱	何以速我狱 (行露)	law, court	Gericht
Culture-specific Item	浚	在浚之下(凯风)	Jun, town of Jun	Jun, Stadt von Jun
Culture-specific Item	东宫	东宫之妹(硕人)	crown prince	Kronprinz, Thronfolger
Culture-specific Item	德音	德音不忘(有女同车)	gracious virtue	tugendhafter Ruf
Confusable Character	微	微我无酒(柏舟)	not, nor	nicht
Confusable Character	终	终鲜兄弟(扬之水)	already	schon
Confusable Character	斯	斧以斯之(墓门)	hew, cut, chop	hauen, fällen, schlagen
Confusable Character	伤	伤如之何(泽陂)	I, me	ich, mich

Table 14: Supplementary examples from the lexical knowledge base.

Source text: 南有**樛木**，**葛藟**累之。乐只**君子**，福履绥之。南有**樛木**，**葛藟**荒之。乐只**君子**，福履**将**之。（樛木）			
GPT-5 (EN) ★To the south, a drooping tree, ✗ Kudzu and ivy drape it. ✗ ★Joyous the gentleman, ✓ Blessing and good fortune bring him peace. ✓ ★To the south, a drooping tree, ✗ Kudzu and ivy overrun it. ✗ ★Joyous the gentleman, ✓ Blessing and good fortune attend him. ✗	*Gemini-2.5 Pro (EN)* ★In the south, the drooping tree, ✗ Creeping vines entwine it. ✓ ★Joyful is the noble lord, ✗ Fortune brings him peace. ✓ ★In the south, the drooping tree, ✗ Creeping vines now cover it. ✓ ★Joyful is the noble lord, ✗ Fortune shall uphold him. ✓	*ERNIE-4.5 (EN)* ★In the south stands a twisting tree, ✗ With ivy entwining free. ✓ ★Happy is the noble man, ✗ Blessings come to comfort him. ✓ ★In the south stands a twisting tree, ✗ With ivy thick and green. ✗ ★Happy is the noble man, ✗ Blessings come to guard him keen. ✗	*Spark-4.0 (EN)* The southern tree bent low, ✓ Vines entwined it tight. ✓ ★Joyful is our noble lord, ✗ Blessings ease his plight. ✗ The southern tree bowed down, ✓ Vines spread dark and deep. ✗ ★Joyful is our noble lord, ✗ Fortune guards his keep. ✗
DeepSeek-3.1 (EN) In the south, the bent tree stands, ✓ By clinging vines embraced. ✓ O joyous noble, may your hands ✗ With fortune's grace be blessed. ✗ In the south, the bent tree grows, ✓ Entwined by lush vines' art. ✗ O joyous noble, may life bestow ✗ Prosperity to your heart. ✗	*Qwen3-14B (EN)* ★Southward grows the twisted wood, ✗ Vines of the wild grape entwine it. ✓ ★Happy is the noble man, ✗ May blessings steady his steps. ✗ ★Southward grows the twisted wood, ✗ Vines of the wild grape spread over it. ✗ ★Happy is the noble man, ✗ May blessings guide his steps. ✓	*LLaMA-3.1-8B (EN)* ★In the south, there is a twisted tree, ✗ Vines entwine it. ✓ ★Joy is only for the noble, ✗ Prosperity and good fortune soothe it. ✗ ★In the south, there is a twisted tree, ✗ Vines overgrow it. ✗ ★Joy is only for the noble, ✗ Prosperity and good fortune will take it. ✗	
GPT-5 (DE) ★Im Süden steht ein gebeugter Baum, ✓ Kudzu und wilder Wein behängen ihn. ✗ ★Heiter ist der Edle; ✗ Segen und Glück ruhen auf ihm. ✗ ★Im Süden steht ein gebeugter Baum, ✓ Kudzu und wilder Wein überwuchern ihn. ✗ ★Heiter ist der Edle; ✗ Segen und Glück geleiten ihn. ✗	*Gemini-2.5 Pro (DE)* ★Im Süden der biegende Baum, ✓ die Ranke umschlingt ihn. ✓ ★Glücklich der edle Herr, ✗ Glück und Segen behüten ihn. ✗ ★Im Süden der biegende Baum, ✓ die Ranke umhüllt ihn. ✗ ★Glücklich der edle Herr, ✗ Glück und Segen geleiten ihn. ✗	*ERNIE-4.5 (DE)* ★Im Süden steht ein krummer Baum, ✓ Ranken wachsen um ihn herum. ✗ ★Der glückliche Edle, ✗ sein Glück sei ihm gesichert. ✗ ★Im Süden steht ein krummer Baum, ✓ Ranken bedecken ihn ganz. ✓ ★Der glückliche Edle, ✗ sein Glück sei ihm beschützt. ✗	*Spark-4.0 (DE)* ★Im Süden wachsen krumme Bäume, ✓ Ranken winden sich um sie. ✗ ★Der Edelmann ist heiteren Sinns, ✗ Sein Reichtum besänftigt ihn. ✗ ★Im Süden wachsen krumme Bäume, ✓ Ranken überwuchern sie. ✗ ★Der Edelmann ist heiteren Sinns, ✗ Sein Reichtum unterstützt ihn. ✓
DeepSeek-3.1 (DE) ★Südwärts neigt sich der Baum, ✓ Mit Ranken schwer behangen. ✗ ★Oh, fröhlicher Edler, ✗ Glück und Frieden dir umfangen! ✗ ★Südwärts neigt sich der Baum, ✓ Von Ranken überwuchert. ✗ ★Oh, fröhlicher Edler, ✗ Glück und Heil dir sich verstärken! ✗	*Qwen3-14B (DE)* ★Im Süden wachsen Bäume, ✗ die sich mit Ranken umschlingen. ✓ ★Der fröhliche Gentleman, ✓ wird von Glück und Segen begleitet. ✗ ★Im Süden wachsen Bäume, ✗ die sich mit Ranken umschlingen. ✗ ★Der fröhliche Gentleman, ✓ wird von Glück und Segen begleitet. ✗	*LLaMA-3.1-8B (DE)* ★Im Süden steht ein verworrenes Holz, ✗ Gestrüpp bedeckt es. ✗ ★Freude ist nur für den Edlen, ✗ Glück bringt ihm ein weiches Schicksal. ✗ ★Im Süden steht ein verworrenes Holz, ✗ Gestrüpp verwüstet es. ✗ ★Freude ist nur für den Edlen, ✗ Glück bringt ihm ein nahes Ende. ✗	

Figure 3: Example of *Shijing* translations across seven LLMs. ⋆, underlines, and **boldface** denote incremental repetitions, rhymes, and challenging terms, respectively.

Source text: 毖彼**泉水**，亦流于**淇**。有怀于卫，靡日不思。娈彼**诸姬**，聊与之谋。出宿于**泲**，饮饯于**祢**。（泉水）			
GPT-5 (EN) Brimming is that spring water, ✗ It too runs to the Qi. ✓ I long toward Wei; ✓ There is no day I do not think. ✓ Lovely are those ladies; ✗ Let me confer with them. ✓ I lodge for the night at Ji; ✓ At Ni I drink farewell. ✓	*Gemini-2.5 Pro (EN)* Bubbling forth, that spring water ✗ Flows on to the River Qi. ✓ My heart longs for my homeland Wey; ✓ There is no day I do not think of it. ✓ I think of those gentle ladies, ✓ And with them I would take counsel. ✓ I lodged for the night at Ji, ✓ My farewell feast was held at Ni. ✓	*ERNIE-4.5 (EN)* The spring water flows gently, ✗ And merges with the Qi stream. ✓ My heart yearns for Wei, ✓ Not a day without its dream. ✗ Those lovely ladies fair, ✓ With them I'd share my scheme. ✗ I lodge at Ji's place, ✓ And bid farewell at Ni's stream. ✓	*Spark-4.0 (EN)* Clear springs gush forth, ✗ Flowing into the Qi stream. ✓ Longing for Wei's land, ✓ No moment without thought of home. ✗ Lovely are these maidens, ✗ I seek their counsel alone. ✗ Once lodged at Ji's shore, ✓ Feasted at Mi's plain. ✗
DeepSeek-3.1 (EN) Pure and clear the spring doth flow, ✗ To join the Qi's current below. ✓ My heart to Wei would fain repair— ✓ How can I dismiss its care? ✗ Gracious are those ladies fair; ✓ With them I counsel would share. ✓ Lodging by Ji's stream at night, ✓ At Ni we drank ere daylight. ✗	*Qwen3-14B (EN)* Cool and clear the spring water flows, ✗ Also to Qi it runs. ✓ I long for Wei, ✓ No day I do not think of it. ✓ Lovely are those young girls, ✗ With them I speak and plan. ✗ I spent the night at Ji, ✓ Drank farewell wine at Mi. ✗	*LLaMA-3.1-8B (EN)* I ponder that spring water, ✗ Also flowing to the Qi River. ✓ I have thoughts of the city of Wei, ✗ Never a day without longing. ✓ I have a lovely group of maidens, ✗ Let us just discuss it together. ✗ I leave my lodging at the Li River, ✗ Drink farewell at the Ni River. ✗	
GPT-5 (DE) Rein ist jene Quelle, ✗ sie ergießt sich auch in den Qi. ✓ Mein Herz ist beim Staat Wei, ✓ kein Tag, da ich nicht daran denke. ✓ Hold sind jene Damen, ✗ gern wollt ich mit ihnen raten. ✗ Beim Aufbruch nächtige ich am Ji, ✓ den Abschiedstrunk nehme ich in Ni. ✓	*Gemini-2.5 Pro (DE)* Rauschend die Quelle dort, ✗ fließt hinab zum Qi. ✓ Mein Herz ist in Wey, ✓ und denkt daran ohne Unterlass. ✗ Lieblich die Damen Ji, ✓ mit ihnen will ich beraten. ✓ Aufbrechen, in Ji zu nächtigen, ✓ Abschiedstrunk nehmen in Ni. ✓	*ERNIE-4.5 (DE)* Das Quellwasser strömt, ✗ Fließt auch zum Qi. ✓ An Wei denkt mein Herz, ✓ Nicht einen Tag vergisst es. ✗ Die schönen Damen dort, ✗ Mit ihnen will ich beraten. ✓ Ich übernachte in Ji, ✓ Trinkabschied in Ni. ✓	*Spark-4.0 (DE)* Quellen sprudeln klar, ✗ fließen auch nach Qi. ✗ Ich denke an Wei, mein Land, ✓ kein Tag vergeht ohne Sehnsucht. ✓ Schöne Schwestern hier, ✗ ich berate mich mit ihnen. ✗ Rastete bei Ji, ✗ trank Abschied bei Mi. ✗
DeepSeek-3.1 (DE) Still fließt der Quelle klares Nass, ✗ Dem Qi-Strom eilt es zu. ✓ Weh, meiner Heimat fern ich bin, ✓ Gedenk ich ihrer stets in Ruh. ✗ Die Schwestern hold, sie fraget ich, ✗ Ob Rat sie mir wohl wüssten. ✗ An Dji blieb ich zur ersten Nacht, ✓ An Mi sodann zum Feste. ✗	*Qwen3-14B (DE)* Klarer Quellbach, ✗ fließt auch zum Qi. ✓ Ich sehne mich nach Wei, ✓ kein Tag, an dem ich nicht daran denke. ✓ Schöne sind die jungen Frauen, ✗ mit denen ich mein Herz spreche. ✗ Nacht für Nacht verlasse ich Ji, ✗ trinke Abschiedswein in Mi. ✗	*LLaMA-3.1-8B (DE)* Wer jener Quellensprung ist, ✗ Fließt auch in den Wei. ✗ Wenn ich an Wei denke, ✓ Denke ich nicht an anderes. ✗ Die jungen Frauen sind schön, ✗ Ich rede nur mit ihnen. ✗ Ich verbringe die Nacht bei der Flussbiegung, ✗ Ich trinke Abschiedswein bei der Flussbiegung. ✗	

Figure 4: Example of *Shijing* translations across seven LLMs. Underlines and **boldface** denote rhymes and challenging terms, respectively.