Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Paradox of poetic intent in back-translation: evaluating the quality of large language models in Chinese translation*#

Li WEIGANG‡[1], Pedro Carvalho BROM[2]

[1]*Department of Computer Science, University of Brasilia, Brasilia 70919-900, Brazil*

[2]*Department of Mathematics, Federal Institute of Brasilia, Brasilia 71200-020, Brazil*

E-mail: weigang@unb.br; pedro.brom@ifb.edu.br

Received May 8, 2025; Revision accepted Sept. 4, 2025; Crosschecked Nov. 10, 2025

Abstract: Large language models (LLMs) excel in multilingual translation tasks, yet often struggle with culturally and semantically rich Chinese texts. This study introduces the framework of back-translation (BT) powered by LLMs, or LLM-BT, to evaluate Chinese → intermediate language → Chinese translation quality across five LLMs and three traditional systems. We construct a diverse corpus containing scientific abstracts, historical paradoxes, and literary metaphors, reflecting the complexity of Chinese at the lexical and semantic levels. Using our modular NLPMetrics system, including bilingual evaluation understudy (BLEU), character F-score (CHRF), translation edit rate (TER), and semantic similarity (SS), we find that LLMs outperform traditional tools in cultural and literary tasks. However, the results of this study uncover a high-dimensional behavioral phenomenon, the paradox of poetic intent, where surface fluency is preserved, but metaphorical or emotional depth is lost. Additionally, some models exhibit verbatim BT, suggesting a form of data-driven quasi-self-awareness, particularly under repeated or cross-model evaluation. To address BLEU's limitations for Chinese, we propose a Jieba-segmentation BLEU variant that incorporates word-frequency and n-gram weighting, improving sensitivity to lexical segmentation and term consistency. Supplementary tests show that in certain semantic dimensions, LLM outputs approach the fidelity of human poetic translations, despite lacking a deeper metaphorical intent. Overall, this study reframes traditional fidelity vs. fluency evaluation into a richer, multi-layered analysis of LLM behavior, offering a transparent framework that contributes to explainable artificial intelligence and identifies new research pathways in cultural natural language processing and multilingual LLM alignment.

Key words: Back-translation; Chinese natural language processing; Large language model-based back-translation (LLM-BT); Paradox of poetic intent; Quasi-self-awareness; Verbatim back-translation

https://doi.org/10.1631/FITEE.2500298 **CLC number:** TP391.1

1 Introduction

The synergistic advancement of large language models (LLMs) and artificial intelligence (AI) has ushered in a paradigm shift in natural language processing (NLP) (Vaswani et al., 2017). Transformer-based architectures have enabled remarkable progress in text generation and cross-lingual translation. Nevertheless, applying LLMs

‡ Corresponding author

* Project supported by the Brazilian National Council for Scientific and Technological Development (CNPq) (No. 309545/2021-8)

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2500298) contains supplementary materials, which are available to authorized users

ORCID: Li WEIGANG, https://orcid.org/0000-0003-1826-1850; Pedro Carvalho BROM, https://orcid.org/0000-0002-1288-7695

to non-English languages, particularly those that are structurally and culturally distinct, such as Chinese, remains an enduring challenge. Key difficulties include preserving poetic intent, cultural nuance, and domain-specific terminology (Zhang XE, 2021; Zhong et al., 2024).

Spoken by more than 1.6 billion people worldwide (Eberhard et al., 2022), Chinese exhibits significant linguistic complexity. Its syntactic ambiguity, idiomatic richness, and specialized terminology (e.g., heterocyclic compounds like 噻唑 (thiazole) and 吡啶 (pyridine)) have long hindered machine understanding. Chinese⇔English translation, in particular, continues to pose a major challenge for multilingual AI systems (Cao et al., 2020; Sun et al., 2021; Li YH et al., 2025). While much previous work has focused on optimizing three-dimensional trade-offs, such as lexical fidelity (信), surface fluency (达), and stylistic elegance (雅), in Chinese translation studies (Chen et al., 2024a), we argue that LLM behavior must now be understood in higher dimensions, where semantic recoverability, cultural alignment, and terminology consistency interact.

Chinese characters also encode layers of meaning beyond surface tokens. Rooted in the ancient *Liu Shu* (六书, Six-Writings) framework and exemplified in Xu Shen's *Shuowen Jiezi* (Zhou, 2014), Chinese combines ideographic and phonosemantic structures. Although our experiments focus on the lexical level, these multimodal features graphic (形), phonetic (声), and semantic (意) help explain why literal fidelity often comes at the cost of expressive depth. Recent studies suggest that LLMs trained predominantly on alphabetic languages may struggle to represent such orthographic complexity (Weigang et al., 2024).

In today's globalized landscape, cross-lingual translation serves as an infrastructure across international trade, scientific communication, and cultural exchange. In 2024, China's foreign trade volume approached US$ 6 trillion, reported by *China Daily*, where translation quality directly affects contract execution, product compliances, and supply chain performance. Scientifically, more than 25% of the world's science citation index publications in 2023 originated from China, with more than 40% requiring English translation (Jiang and Liu, 2020; Bahji et al., 2023). Culturally, the 2025 global release of "Nezha: Rebirth of the Demon Child 2" earned more than US$ 200 million at the North American box office (https://www.globaltimes.cn/page/202502/1328145.shtml), highlighting the critical role of dubbing and subtitling in international media.

Despite their success in English-dominant tasks, LLMs such as GPT-4.5 remain challenged when applied to Chinese texts. Transformer models trained primarily on space-delimited languages must contend with Chinese's unique properties: (1) the lack of explicit word boundaries and (2) a more uniform information distribution at the character level (Wong et al., 2010; Weigang et al., 2025b). Such challenges necessitate a shift from token-level to concept-level evaluation, especially in poetic, historical, and scientific contexts.

However, modeling Chinese purely at the character level is suboptimal. While (one, two)-character words (e.g., 技术(technology), 翻译(translation)) constitute approximately 95% of high-frequency items (Liu and Liang, 1986), many semantically dense (three, four)-character expressions (e.g., 机器人(robot), 人工智能(AI), and idioms like 刻舟求剑(a futile act of marking the boat to retrieve a lost sword)) carry disproportionate contextual weight. Treating characters as isolated units often leads LLMs to lose the semantic integrity of these word-level constructs, resulting in unnatural phrasing or cultural misalignment. This phenomenon helps explain the comparative underperformance of LLMs on Chinese tasks, particularly those involving metaphor, poetry, or specialized jargon.

Empirical evaluations highlight this gap: mainstream machine translation (MT) systems typically report bilingual evaluation understudy (BLEU) scores (Papineni et al., 2002) of not greater than 0.45 for English-to-Chinese tasks and under 0.27 for Chinese-to-English ones. These scores are substantially lower than those for many other language pairs—for example, Russian-to-English reaches 0.44 (Zhu et al., 2025). This discrepancy is especially evident in sensitive domains such as science, law, and healthcare. Although hybrid architectures, such as sparse mixture-of-experts (MoE) models integrated into GPT-4.0 and LLaMA, show promise (Zhu et al., 2025), emerging platforms (e.g., Claude 3.7, DeepSeek V3, Gemini 2.0, and Grok 3) lack rigorous benchmarking, especially for content involving cultural depth, historical references, or scientific

terminology.

In this study, we present a structured back-translation (BT) framework, LLM-based back-translation (LLM-BT), to assess bidirectional semantic preservation. Central to this concept is the paradox of poetic intent, a phenomenon in which surface fluency is maintained at the expense of cultural or poetic nuance (Fig. 1). This observation extends beyond traditional fidelity–fluency trade-offs and introduces a high-dimensional framework for evaluating translation quality.

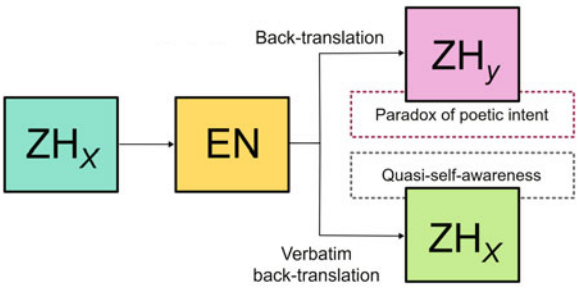

Fig. 1 Conceptual diagram: the paradox of poetic intent in back-translation (BT) vs. emergent quasi-self-awareness in LLM verbatim BT, where $\mathbf{ZH}_x$ is the original Chinese text, EN is the translated English text, and $\mathbf{ZH}_y$ is the back-translated Chinese text

Verbatim BTs further reveal model-specific behaviors. GPT-4.5 and DeepSeek V3, for instance, often regenerate the original input even without prompting. Building on memory-related research (Brown et al., 2020; Zhang ZY et al., 2024), we describe this as a form of quasi-self-awareness, an emergent, non-memorized, but stable recovery behavior across model boundaries and time delays.

We have constructed a multi-domain corpus covering (1) scientific naming paradoxes (e.g., Xue Dejiong's dilemma (He, 2019; Weigang et al., 2025a)), (2) metaphorical and poetic expressions (e.g., Dao Lang's lyrics), and (3) terminology-rich abstracts from CNKI. We evaluated five LLMs (Claude 3.7, DeepSeek V3, Gemini 2.0, GPT-4.5, and Grok 3) and three commercial translation tools (Google, Baidu, and Sogou) using multi-sample BT and statistical analysis. Our contributions are:

1. The LLM-BT framework is proposed to analyze bidirectional translation, together with NLP-Metrics, a modular system for multi-dimensional evaluation including BLEU, character F-score (CHRF), translation edit rate (TER), and semantic similarity (SS). Results reveal the paradox of poetic intent, especially in metaphorical and poetic texts.

2. BT is shown to enhance semantic clarity in scientific abstracts when compared to direct translation. In contrast, traditional tools retain domain-specific fluency advantages, as illustrated by Sogou's BLEU score of 0.57 for scientific content.

3. The concept of quasi-self-awareness is formalized to capture highly deterministic LLM behavior across sessions and models, quantified via the LLM-BT consistency index (LBCI).

4. BLEU is adapted for Chinese through Jieba-based segmentation and frequency-weighted n-gram scoring. This adjustment better reflects the structural and statistical properties of Chinese word formation.

5. Supplementary benchmarks against human poetic translations (Chen et al., 2024a) demonstrate that GPT-4.5, DeepSeek V3, and Claude 3.7 approach human-level BLEU scores, underscoring emerging capabilities in semantic preservation and creative paraphrasing.

2 Corpus design and linguistic properties

This section presents corpora that span scientific terminology dilemmas, academic abstracts, and metaphor-rich lyrics. These datasets form a multi-level benchmark to assess LLM-BT performance.

2.1 Corpus of "Xue Dejiong's dilemma"

Historically, scholars have debated whether the Chinese naming of heterocyclic organic compounds should follow semantic principles or phonetic transcription. He (2019) conducted a retrospective review of this issue. Building on that, the present study compiles statements by chemist Xue Dejiong into what is now referred to as the "Xue Dejiong's dilemma" corpus (Weigang et al., 2025a). The original Chinese passage is provided in Section 4.2 and the English translation by GPT-4.5 is listed in the supplementary materials.

This passage weaves together technical critique, cultural metaphor, and rhetorical flourish, offering an ideal benchmark for evaluating MT of terminology-rich and stylistically complex texts.

Beyond its surface linguistic difficulty, "Xue Dejiong's dilemma" embodies a historical turning

point in Chinese scientific terminology development. During the mid-20th century, Chinese chemists faced growing pressure to transliterate complex organic compounds from the Western literature (He, 2019). While proponents of phonetic transliteration ultimately prevailed, favoring ease of standardization, Xue Dejiong strongly advocated for semantically meaningful and morphologically expressive translations rooted in the structural features of Chinese characters, especially using phonosemantic compounds. Despite his efforts, phonetic naming conventions for heterocyclic compounds ultimately prevailed, resulting in widespread transliterations that lack semantic transparency and complicate downstream NLP tasks such as terminology extraction and MT.

We argue that this compromise, driven by practical constraints and limited computational tools at the time, contributed to the long-term bottleneck in Chinese terminology processing. The historical loss of systematic form-meaning mappings makes term disambiguation, MT, and domain adaptation significantly more difficult today. This dilemma provides the philosophical and linguistic foundation for our recent LLM-BT-Terms framework (Weigang and Brom, 2025), which seeks to recover and standardize term semantics across multilingual BT cycles using LLMs.

2.2 CNKI abstract corpus of the scientific literature

Scientific abstracts were collected from the international portal of the China National Knowledge Infrastructure (CNKI: oversea.cnki.net), including journal names, article titles, authors, keywords, abstracts, volumes, and page numbers. A domain-general corpus was then constructed based on 10 scientific and technological themes: chemistry, biotechnology, nanotechnology, telemedicine, AI, data science, digital economy, linguistics, sociology, and distance education.

For each domain, 295 Chinese-language scientific papers were selected to ensure statistical robustness, and all metadata have been released via GitHub (https://github.com/pcbrom/bt-conference) for reproducibility. Due to space constraints, this paper focuses on a chemistry-specific subcorpus comprising 295 abstracts, of which 89 entries were randomly sampled to form the CNKI-CHE-89 subset.

One example, labeled CNKI-CHE-89-18 (Feng, 2024), is shown in Section 4.3, and the English translation by Sogou is listed in the supplementary materials.

2.3 Corpus from Dao Lang's *Hua Yao* lyrics

Dao Lang is a Chinese pop musician known for incorporating folklore and literary aesthetics into his lyrics. His song *Hua Yao* (花妖, flower demon) explores themes of love, reincarnation, and spiritual yearning. The lyrics function as a lyrical epic, interweaving poetic sentiment, historical geography, Buddhist cosmology, and nonlinear temporality.

The text relies on culturally unique semiotic systems, such as I Ching (易经) and Taoist geomancy (罗盘经), and features complex rhetorical structures, including tonal layering and phonetic ambiguity. These characteristics present a triple challenge for MT:

1. symbolic deconstruction of cultural metaphors (e.g., reducing "compass classic" to a mere navigational guide);

2. disconnection from historical spatiotemporal context (e.g., rendering place names without semantic depth);

3. flattening of poetic rhythm and prosody into literal or unstructured output.

For these reasons, *Hua Yao* was chosen as the centerpiece of the literary BT task. The original Chinese is shown in Section 4.4; the English translation was generated using Google Translate and is included in the supplementary materials.

Recent works such as Chen et al. (2024a) and Zhao et al. (2025) have evaluated LLMs in translating classical Chinese texts, emphasizing adequacy, fluency, and stylistic fidelity. Our study complements this by probing cognitive alignment and semantic recoverability during BT of semantically rich and culturally embedded texts. Human-translated references from Chen et al. (2024a) are included as parallel data for comparative evaluation.

In sum, the corpora described here, terminologically dense historical texts, scientifically structured abstracts, and metaphor-laden lyrics, cover a wide spectrum of linguistic complexity. Each presents unique challenges in tokenization, cultural interpretation, and semantic recovery. Together, they constitute a robust benchmark for evaluating LLMs in

cross-lingual BT.

3 Methodology: LLM-BT framework and evaluation metrics

This section presents the LLM-BT framework, covering its conceptual basis, system design, and evaluation metrics. Section 4.5 defines key terms, and Section 7 reviews the evolution of BT in neural MT (NMT).

3.1 Overall framework of LLM-BT

BT, also known as round-trip translation (RTT), is a widely used method for evaluating translation accuracy (Tu and Li, 2017; Artetxe et al., 2018; Ozolins et al., 2020). It has proven effective in data augmentation, translation quality assessment, and cross-lingual alignment, particularly in complex linguistic systems such as Chinese and Japanese. The core process involves translating a source language (e.g., Chinese, ZH_x) into a target language (e.g., English, EN) and then translating it back into the source language (ZH_y). By comparing ZH_x and ZH_y, one can assess how well the semantic structure is preserved.

The LLM-BT procedure ($ZH_x \rightarrow EN \rightarrow ZH_y$) consists of three stages (Fig. 1):

1. Forward translation ($ZH_x \rightarrow EN$): The original Chinese text is translated using a system (e.g., neural MT or LLM) to generate an English version.

2. Back translation ($EN \rightarrow ZH_y$): The English translation is then translated back into Chinese.

3. Comparative analysis: ZH_x and ZH_y are evaluated using both automatic metrics (e.g., BLEU and TER) and qualitative analysis (e.g., semantic drift, fluency, and cultural retention).

The method is grounded in the translation equivalence theory (Nida, 1964) and the cyclic consistency hypothesis (Artetxe et al., 2018), which assert that high-quality translations should preserve meaning bidirectionally. Significant divergence between ZH_x and ZH_y (e.g., BLEU$\leq$0.30 or human adequacy score$\leq$3/5) indicates semantic loss or distortion.

In the era of large-scale generative models, BT is better understood as more than a symmetric language translation tool. Within our LLM-BT framework, BT functions as a probing mechanism for emergent behaviors such as terminological drift, semantic flattening, and stylistic memory. This redefinition allows BT to evaluate not just accuracy, but interpretability, robustness, and potential alignment in generative multilingual systems. Recent advances in MT and LLMs have enabled the following applications:

1. system evaluation: using round-trip tests ($ZH_x \rightarrow EN \rightarrow ZH_y$) to compare translation performance across platforms (e.g., Google Translate, ChatGPT, and DeepSeek).

2. terminology retention: assessing fidelity in transliteration (e.g., "比特" for "bit") vs. semantic translation (e.g., "人工智能" for "artificial intelligence").

3. linguistic representation: evaluating semantic mappings of Chinese expressions through embedding similarity and tokenization schemes, including phonosemantic encodings derived from the Six-Writings framework (Weigang et al., 2024).

More nuanced behavioral phenomena such as verbatim back-translation, poetic drift, and others are introduced and formally defined in Section 4.

3.2 Implementation of LLM-BT

The LLM-BT workflow is divided into three stages: data preparation, model selection, and multi-dimensional evaluation.

3.2.1 Data preparation in three categories

We organize the source materials into three categories to cover diverse text types and linguistic styles:

1. Translation paradox cases: Texts exhibiting semantic asymmetry in RRT ($ZH_x \rightarrow EN \rightarrow ZH_y$), such as "Xue Dejiong's Dilemma," are used to evaluate model sensitivity to linguistic ambiguity and cultural nuance.

2. Scientific abstracts: A curated sample of 295 Chinese abstracts from CNKI, covering science and engineering domains, is used to assess terminological alignment and logical consistency.

3. Literary texts: Representative works such as the lyrics to *Hua Yao* by Dao Lang are selected to test each model's ability to handle metaphor, rhetoric, and imagery.

3.2.2 Model selection with two categories

We evaluated two categories of systems: (1) NMT baselines (Google Translate, Baidu Translate,

and Sogou Translate) and (2) LLMs (GPT-4, Grok 3, Mistral, Claude 3.7, DeepSeek V3, Gemini 2.0, GPT-4.5, and Grok Beta). All models were accessed via public application programming interfaces (APIs) or official interfaces with default parameters, ensuring comparability across platforms.

3.2.3 Multi-dimensional evaluation

Evaluation spans four key dimensions (Table 1): semantic consistency, fluency, cultural fidelity, and terminological consistency (Chen et al., 2024a).

To enhance objectivity, both single-sample qualitative analyses and multi-sample statistical tests were conducted. The Friedman test (Demšar, 2006), a non-parametric method, was used to identify statistically significant differences across model outputs. Supporting visualizations (e.g., heatmaps) aid interpretation. Although human-referenced metrics such as COMET-Kiwi are valuable in downstream applications, they were intentionally omitted here to maintain interpretability: (1) to ensure comparability with BLEU, CHRF, and TER-based studies; (2) to keep analytical focus on the quasi-self-awareness phenomenon, which was already captured via our proposed LBCI. Future work may incorporate COMET without affecting our core conclusions.

3.3 Assessing BLEU score variability in LLMs with the Friedman test

This subsection outlines the statistical methodology used to evaluate BT quality across five leading LLMs and three commercial translation systems. The analysis focuses on BLEU (Papineni et al., 2002) as the core evaluation metric, complemented by CHRF, TER, and SS.

Each source text is treated as an independent experimental block, with n denoting the total number of texts (blocks). Every block is translated by all $k = 5$ LLMs (Claude 3.7, DeepSeek V3, Gemini 2.0, GPT-4.5, and Grok Beta), enabling model-wise comparisons across a consistent sample space.

During translation, each LLM executes the forward and backward processes independently, translating from source ZH_x to intermediate EN and then back to target ZH_y, using prompt variants and randomized seeds to avoid memory leakage or alignment bias.

The BLEU score for each ZH_y serves as a proxy for semantic and syntactic preservation. We denote the observed BLEU score for the j^{th} text (block) processed by the i^{th} model as Y_{ij}, arranged into a data matrix of size $k \times n$. To account for intra-model variability, each text is translated $r = 3$ times per model using independent sampling runs.

The analysis uses the following linear model:

$$Y_{ij} = \eta + \tau_i + \beta_j + \epsilon_{ij}, \tag{1}$$

where Y_{ij} represents the observed BLEU score for model i on text j, η represents the grand mean of all BLEU scores, τ_i represents the effect of the i^{th} translation model, β_j represents the effect of the j^{th} text block, and ϵ_{ij} represents the random error term for model i on text j.

Given that BLEU score distributions across LLMs may violate parametric assumptions such as normality and homoscedasticity, particularly under repeated-measures settings, we employ the Friedman test as a non-parametric alternative to repeated-measures Analysis of Variance (ANOVA) to detect statistically significant differences among models. For post-hoc analysis, Dunn's test with Bonferroni

Table 1 Evaluation dimensions of Chinese BT quality

Dimension	Evaluation focus	Key metric(s)
Semantic consistency	Alignment of core meaning between original and back-translated texts	SS
Fluency	Grammatical correctness and natural expression in back-translated texts	TER
Cultural fidelity	Accurate transmission of culture-laden elements (e.g., idioms, imagery, and metaphor)	Human expert evaluation (e.g., literary translators)
Terminological consistency	Consistent alignment of scientific and technical terms between Chinese and English	BLEU, CHRF, and term-alignment checks

BLEU: bilingual evaluation understudy; BT: back-translation; CHRF: character F-score; SS: semantic similarity; TER: translation edit rate

correction is applied to control for type I error (Sheldon et al., 1996; Nam and Park, 2015; Arruda-Vasconcelos et al., 2021; Yousufi and Erdely, 2024).

The sample size is determined for power=0.8 and α=0.05, targeting a medium effect size f=0.3. Based on this, we select n=89 distinct texts, each evaluated by k=5 models and r=3 samples, resulting in a total of 1335 translation outputs. The dataset corresponds to the CNKI-CHE-89 corpus (chemistry abstracts), selected to minimize topic-induced variance. Normality and variance assumptions are tested using the Shapiro–Wilk and Levene tests, respectively, justifying the use of non-parametric analysis.

3.3.1 BLEU adaptation for Chinese

To better evaluate LLMs for Chinese text, we define an adapted BLEU score:

$$\mathrm{BLEU_J} = \sum_{n=1}^{4} w_n \ln p_{n\text{-gram}}, \tag{2}$$

where w_n is weight assigned to the n-gram, $p_{n\text{-gram}}$ is precision of n-gram matches after Jieba-based segmentation, and $\mathrm{BLEU_J}$ thus reflects word frequency-aware n-gram alignment, not just token overlap. Two BLEU variants are compared:

1. BLEU: w=(0.5, 0.5, 0, 0), consistent with Chinese word length statistics (Liu and Liang, 1986), where one-character words account for 56.7%, two-character words account for 39.65%, and others account for 3.65%.

2. BLEU-Unif: w=(0.25, 0.25, 0.25, 0.25), aligning with default LLM decoding practices (e.g., GPT-4.5, Gemini, and Grok).

3.3.2 Complementary metrics

To offset BLEU's limitations with paraphrasing and syntax variation, we incorporate the following:

1. CHRF, emphasizing surface-level edit distance;

2. TER, reflecting minimal required edits;

3. SS, a term frequency–inverse document frequency (TF–IDF) weighted cosine similarity computed over sentence vectors, capturing conceptual overlap missed by token-based metrics.

CHRF and SS follow a higher-is-better interpretation. BLEU measures n-gram overlap, with higher scores indicating greater lexical similarity; however, it struggles with paraphrasing where meaning is retained but wording differs (Toral and Way, 2018). CHRF evaluates character-level similarity, making it particularly effective in morphologically rich languages such as Chinese. TER quantifies the editing effort required (insertions, deletions, and substitutions) to match a reference, where lower values indicate better performance.

To complement these surface-level metrics, we introduce the SS score, which evaluates the preservation of meaning between the original and back-translated texts. SS is computed using a TF–IDF weighted cosine similarity over sentence embeddings. Specifically, each sentence is vectorized using a TF–IDF matrix trained over the corpus, and similarity is calculated via cosine distance. This lexical–semantic embedding strategy captures thematic and conceptual overlap that may be missed by n-gram-based metrics.

These metrics collectively evaluate fluency, fidelity, and semantic adequacy. Details on model performance metrics can be seen in the supplementary materials. All formulae adhere to LLM-BT-Terms (Weigang and Brom, 2025), our companion study dedicated to terminology-level analysis.

3.3.3 Segmentation and tooling considerations

Chinese BLEU evaluation is segmentation-dependent. We apply Jieba for tokenization due to its open-source flexibility and support for custom dictionaries (Ding et al., 2021). These features enable the accurate handling of technical terms. Based on Jieba output, we propose a frequency-weighted BLEU variant (JiebaBLEU) to better capture semantic relevance.

Other major Chinese word segmentation methods, such as PKUseg (Luo et al., 2019) and HanLP (Yang YX and Ren, 2020), offer alternative boundary definitions, especially for domain-specific compound terms. A detailed comparison of these methods is beyond the scope of this study and is reserved for future investigation.

3.4 Model version and parameter size

To contextualize the performance of various LLMs, we summarize the estimated parameter count and average inference latency for each model in the supplementary materials. This provides a

comparative reference for understanding the trade-off between translation quality and computational cost.

The LLM-BT experiments were conducted between March 1 and May 8, 2025. Manual analysis of poetic samples was performed with GPT-4 (via the ChatGPT interface), while large-scale BT and metric scoring were automated through the NLPMetrics pipeline. Platform access relied on publicly available web interfaces or APIs, including GPT-4.5, Grok 3, Claude 3.7, DeepSeek V3, and Gemini 2.0.

This hybrid evaluation approach enhances both transparency and reproducibility. Moreover, by documenting model versions and configuration contexts, our study accounts for variation across evolving LLM deployments.

4 Platform comparison of LLM-BT performance and terminology stability

This section evaluates nine mainstream systems on the three Chinese corpora introduced in Section 2, focusing on translation quality at both lexical and semantic levels. Particular attention is given to metaphor retention, terminological consistency, and cultural nuance, which remain challenging tasks for both NMT and LLMs.

4.1 Overall model performance

All samples were translated from Chinese to English and then back to Chinese. Fidelity was primarily assessed using BLEU scores, reported in Table 2 with three evaluators: BLEU-Unif$_{\text{grok}}$ (based on Grok 3), BLEU-Unif$_{\text{gpt}}$ (based on GPT-4), and JiebaBLEU. JiebaBLEU applies Jieba segmentation to resolve Chinese word-boundary ambiguity. For comparison, NLPMetrics additionally computes SS via GPT-4.5 embeddings, which complements BLEU but is independent of JiebaBLEU. Details on NLPMetrics, a modular evaluation framework for LLM-BT, can be seen in the supplementary materials.

For Xue Dejiong's dilemma, Grok Beta and DeepSeek V3 achieved the highest two BLEU scores (0.65 and 0.64), suggesting strong surface alignment. However, such high scores in literary texts, typically ranging from 0.30 to 0.50, may reflect excessive literalism rather than genuine semantic transfer (Zhu et al., 2025). In particular, DeepSeek V3's nearly identical output on *Hua Yao* (BLEU=0.7602) raises concerns about verbatim BT, a phenomenon further analyzed in Section 4.3.

Traditional NMT tools performed poorly on academic texts, as evidenced by Baidu and Sogou scoring ($\leq$0.30) on Xue Dejiong's dilemma, underscoring the limitations of conventional engines in domain-specific or rhetorically rich contexts.

In the CNKI-CHE-89-18 scientific abstract, Sogou Translate unexpectedly outperformed LLMs (BLEU=0.57), followed by Google Translate (0.45) and Baidu Translate (0.43). Among LLMs, Claude 3.7 achieved the best score (0.42), with GPT-4.5, DeepSeek V3, and Grok Beta slightly lower.

Although absolute BLEU values varied between BLEU-Unif$_{\text{grok}}$ and BLEU-Unif$_{\text{gpt}}$, the relative ordering of models remained largely consistent across evaluators. This suggests that metric choice

Table 2 Comparison of BT quality across different models and corpora

LLM/Tool	Xue Dejiong's dilemma		CNKI-CHE-89-18			*Hua Yao*
	BLEU-Unif$_{\text{grok}}$	BLEU-Unif$^{*}_{\text{grok}}$	BLEU-Unif$_{\text{gpt}}$	JiebaBLEU	SD	JiebaBLEU
Grok Beta	0.65	0.61	0.34	0.2786	0.0583	0.3311
DeepSeek V3	0.64	0.58	0.37	0.3042	0.0576	0.7602
GPT-4.5	0.59	0.59	0.36	0.2970	0.0000	0.2141
GPT-4	0.47	0.50	–	–	–	–
Gemini 2.0	0.51	0.52	0.40	0.3382	0.0016	0.2506
Claude 3.7	–	–	0.42	0.3702	0.0099	0.3114
Mistral	0.24	0.46	–	–	–	–
Google Translate	0.33	0.44	0.45	0.4719	–	0.3664
Baidu Translate	0.25	0.36	0.43	0.3769	–	0.2530
Sogou Translate	0.17	0.31	0.57	0.5212	–	0.3650

SD: standard deviation; BLEU scores for Xue Dejiong's dilemma verified across two independent runs: BLEU-Unif$_{\text{grok}}$ corresponds to results obtained on March 7–8 and BLEU-Unif$^{*}_{\text{grok}}$ to those on March 13–14, 2025, controlling for output variance of the same LLM (Grok 3). "–" in the table indicates that the corresponding result was not available or that the test was not conducted

influences scale but not comparative ranking.

These findings indicated that while some LLMs achieve high BLEU scores, further investigation is needed to determine whether these scores reflect true semantic alignment or merely shallow form-level replication. Sections 4.2–4.4 analyze this distinction in detail.

In Chinese MT evaluation, character-level BLEU (e.g., sacreBLEU with the "zh" tokenizer) is the reproducible standard, but it underrepresents errors at the word or phrase level. To address this, we distinguish three variants:

1. BLEU-Unif$_{\text{grok}}$ and BLEU-Unif$_{\text{gpt}}$: uniform n-gram weights (0.25, 0.25, 0.25, 0.25), simulating Grok and GPT tokenizer behavior.

2. JiebaBLEU: segmentation based on Jieba to approximate natural word boundaries, with both (0.25, 0.25, 0.25, 0.25) and (0.5, 0.5, 0, 0) weightings.

3. SS: GPT-4.5 embeddings used as an auxiliary signal to capture paraphrasing beyond surface overlap.

By clarifying notation and scope, these variants reduce ambiguity and make explicit how tokenization choices affect the evaluation of LLM-BT.

4.2 Case study: Xue Dejiong's dilemma

This chemically themed passage features complex expressions, including technical terms (e.g., 噻嗪, thiazine), Buddhist mantras, rhetorical metaphors, and cultural allusions (e.g., 谪仙李太白, the banished immortal Li Taibai), making it challenging even for native Chinese readers (He, 2019).

Table 3 presents a comparison of BTs across all systems for the following sentence:

满纸咿哑，一若番书，虽有谪仙李太白其人，恐亦难于索解。

Each system's English translation (EN) and corresponding back-translated Chinese version (ZH$_y$) are shown.

Grok Beta outperformed other models in retaining semantic coherence and literary nuance during

Table 3 Comparative analysis of BT instances for Xue Dejiong's dilemma across different models*

Model	English translation (EN)	Chinese BT (ZH$_y$)	BLEU
Grok Beta	Filling the page with gibberish, as if it were a foreign text. Even someone as extraordinary as the banished immortal Li Taibai (Li Bai) would likely find it difficult to decipher.	满纸都是胡言乱语，仿佛外文书籍。即使有谪仙李太白（李白）这样杰出的人物，恐怕也难以理解。	0.65
DeepSeek V3	Crowding the page with gibberish, as cryptic as foreign script. Even a genius like Li Bai, the Banished Immortal, would struggle to decipher them.	满纸咿哑，一若番书，纵有谪仙李太白其人，恐亦难于索解。	0.64
GPT-4.5	Making the entire text sound bizarre, like "a foreign scripture." Even a literary genius like Li Bai would find them difficult to comprehend.	使整个文本听起来十分古怪,"一若番书。"即便像李白这样才华横溢的文学天才，也难以理解。	0.59
GPT-4	A chaotic jumble of syllables filling the pages, resembling an unintelligible foreign script. "Even if the great poet Li Bai himself were present, he would struggle to decipher them."	整个页面充满混乱的音节，看起来就像一篇难以辨认的外文。"即便谪仙李白在世，也恐怕难以解读。"	0.47
Google Translate	It will inevitably be "full of babbling, like foreign language" and even the exiled immortal Li Bai would find it difficult to understand.	必然"咿呀学语,犹如外语，连流放的仙人李白都听不懂。"	0.33
Baidu Translate	Full of paper babbling, like a book; even if there is an exiled immortal Li Taibai, it may be difficult to understand.	满纸胡言乱语，像一本书，即使有流亡的仙人李太白，也可能很难理解。	0.25
Mistral	Filling the page with strange symbols, like a foreign script. Even the exiled immortal Li Bai would struggle to decipher them.	满纸奇怪的符号，像外文一样。即使是被贬谪的仙人李白也难以解读。	0.24
Sogou Translate	It will inevitably be "full of babbling, like foreign language," and even the exiled immortal Li Bai would find it difficult to understand.	论文哑了。如果是一本书，即使有一个堕落的仙女李太白，也很难找到。	0.17

* The original Chinese (ZH$_x$): 薛德炯"不满于" 口旁简名"已十余年," 认为它们"像佛经中的'嘛呢叭咪', 简直莫名其妙"！如若用口旁音译字来命名dithiadiazole 等杂环化合物，势必"噻噻, 唑唑嗪嗪, **满纸咿哑，一若番书，虽有谪仙李太白其人，恐亦难于索解。**" 此口旁之简名所以不能不革除（割爱）也。即便如此深恶痛绝口旁名称，也承认"稠杂圜之结构繁复者多，故其系统名类多冗长。为便于名举计，确有特定简名之必要"。不过他建议"从原名之音，特创口旁字之简名，至少须于名末多加一个足以表示其主要官能之字"。如咖啡宜命名为"咖啡碱。" Translations were generated by LLMs (e.g., GPT-4.5 and Grok Beta) using the NLPMetrics evaluation system. BLEU scores were computed using the Grok 3 platform

BT, closely matching the tone and metaphor of the original. Mistral and Google Translate, in contrast, produced inconsistent or semantically diluted renderings. Sogou Translate revealed logical inconsistencies, yielding a low BLEU score of only 0.17.

This case illustrates the limitations of shallow literal alignment. While surface fluency may remain intact, deep metaphorical intent, especially in culturally embedded sentences, can be flattened or lost. We revisit this phenomenon in Section 4.5 as part of the paradox of poetic intent.

4.3 Case study: CNKI-CHE-89-18

This technical abstract exemplifies domain-specific expression with simplified Chinese syntax. While forward translation ($ZH_x \rightarrow EN$) was generally handled well, reverse translation ($EN \rightarrow ZH_y$) posed challenges for idiomatic scientific terms such as "improve quality and efficiency" (提质增效) and "chemical calculation dose" (化学计算剂量) (Feng, 2024). From Table 4, we observe the following:

1. Traditional systems (Google, Baidu, and Sogou) demonstrated strong fidelity to technical terminology, likely due to long-term domain optimization and dictionary-backed segmentation. Only Sogou and Baidu correctly rendered "chemical calculation dose" without semantic drift.

2. LLMs (e.g., Claude 3.7 and GPT-4.5) often over-generalized Chinese expressions, introduced syntactic shifts, or altered core terms, suggesting stochastic decoding and a lack of controlled terminology constraints.

3. Cultural phrases such as "提质增效" were either omitted or paraphrased creatively by LLMs, while traditional tools preferred literal approximations (e.g., "improve quality and efficiency").

These observations highlight the domain sensitivity of different models: traditional MT tools still excel in structured scientific text, whereas LLMs require prompt tuning or terminology grounding to match performance.

4.4 Case study: *Hua Yao* lyrics

A representative line from the lyrics of *Hua Yao* reads: "When you left, your coarse robe turned red; the yellow sash adorned my waist. The Compass Classic was misread and I was wrongly cast into Quanting." (君去时褐衣红，小奴家腰上黄，寻差了罗盘经，错投在泉亭).

We refer to this excerpt as the "*Hua Yao* Red and Yellow" corpus ("花妖红黄"语料) for analysis. This text contains metaphor and cultural allusion, making it a suitable benchmark for testing the limits of Chinese→English→Chinese BT.

Table 4 Comparative analysis of BT for the CNKI-CHE-89-18 corpus across different models*

Model	English translation (EN)	Chinese BT (ZH_y)	BLEU
Sogou Translate	Carrying out metrological calculation can improve the quality and efficiency of applied chemical calculation dose.	进行计量计算可以提高应用化学计算剂量的质量和效率。	0.57
Google Translate	The measurement calculation can be carried out to improve the quality and efficiency of chemical calculation dosage in practice.	可以大大提高化学计算的质量和效率。	0.45
Baidu Translate	Conducting quantitative calculations can improve the quality and efficiency of chemical dosage calculations.	进行定量计算可以提高化学剂量计算的质量和效率。	0.43
Claude 3.7	The quality and efficiency of chemical dosage calculations can be improved in practical applications.	可以在实际应用中提高化学计量计算的质量和效率。	0.42
Gemini 2.0	Performing quantitative calculations can practically improve the quality and efficiency of chemical dosage calculations.	进行定量计算能够有效提高化学计量计算的质量和效率。	0.40
DeepSeek V3	It is possible to enhance the quality and efficiency of chemical dosage calculations in practical applications.	能够在实际应用中提升化工配比计算的质量与效率。	0.37
GPT-4.5	It becomes possible to improve the quality and efficiency of chemical dosage calculations in practical applications.	可以提高实际应用中化学加药计算的质量和效率。	0.36
Grok Beta	It is possible to enhance the efficiency and effectiveness of chemical dosage calculations.	可以提高化学配比计算的效率和效果。	0.34

* The original Chinese (ZH_x): [正] 化学工程领域的涉及面较广，全过程复杂，涉及到的计算环节众多。化学工程领域的计算多与曲线拟合、非线性方程、偏微分、线性或非线性规划的求解等有关，这些计算难度极大，仅靠手工难以实现精准快捷的计算；而配合计算机辅助编程，**开展计量计算，能够实践应用化学计算剂量的提质增效**，具有极高的应用价值和现实意义。Translations were generated by LLMs (e.g., GPT-4.5 and Grok Beta) using the NLPMetrics evaluation system. BLEU scores were computed using the Grok 3 platform

4.4.1 Metric-based evaluation

Tables 5 and 6 show results across five evaluation metrics: BLEU (weighted), BLEU-Unif (uniform), CHRF, TER, and SS, all computed via NLPMetrics with GPT-4.5.

DeepSeek V3 achieved the highest BLEU score (0.7602), suggesting near-verbatim reproduction. However, manual re-translation of the same English output (DeepSeek V3#) yielded only 0.2849, indicating that the high score reflected structural alignment rather than semantic fidelity. This reinforces the key insight: high BLEU alone cannot confirm cultural or metaphorical accuracy.

4.4.2 Translation instance comparison

One challenging line is "The Compass Classic was misread and I was wrongly cast into Quanting." (寻差了罗盘经，错投在泉亭).

Most LLMs, including GPT-4.5, Claude 3.7, Gemini 2.0, and Grok Beta, translated "Compass Classic" literally as a navigation manual and failed to recover its metaphorical sense. This pattern indicates a systematic limitation of current LLMs in handling culturally embedded metaphors during back-translation. As discussed in Section 4.4.1, DeepSeek V3 represents a contrasting case with exceptionally high BLEU scores but limited semantic recovery, and is therefore not repeated here. Together, these results suggest that both literal translation errors and near-verbatim reproduction can lead to misleadingly high automatic evaluation scores (Tables 5 and 6).

4.4.3 Discussion and implications

The *Hua Yao* case underscores the limitations of current LLMs in processing poetic and culturally embedded Chinese texts. Most failed to capture

Table 5 Comparative analysis of BT of Dao Lang's *Hua Yao* lyrics across various language models*

Model	English translation (EN)	Chinese BT (ZH_y)	BLEU
DeepSeek V3	You left in robes of brown and red, my sash was yellow. Misreading the compass, I strayed to Quan'ting.	君去时褐衣红，我腰上黄。寻差了罗盘经，错投在泉亭。	0.7602
Google Translate	You left with brown clothes and red, my little slave's waist is yellow, I misplaced the compass and was thrown into the wrong place at Quanting.	你留下褐衣赤色，我的小奴腰间黄，我放错了地方罗盘在泉亭丢错了地方。	0.3664
Sogou Translate	When you go, your clothes are brown and red and your little slave's waist is yellow. You lost the compass and threw it in the spring pavilion by mistake.	你去的时候，衣服是棕红色的，小奴才的腰是黄色的。你把指南针弄丢了，不小心扔进了春亭。	0.3650
Grok Beta	When you left, you wore a brown coat, with a yellow belt around my waist, I've lost my way with the compass, mistakenly went to Quan Pavilion.	你离去时身披褐色外套，腰间系黄带，我迷失了指南针，误入泉亭。	0.3311
Claude 3.7	When you left you wore a brown and red robe and this little servant had yellow at her waist. I followed the wrong compass, mistakenly arriving at Spring Pavilion.	你离去时穿着褐红袍，这小婢腰间束着黄。我循错了罗盘，误到春台上。	0.3114
Baidu Translate	When you went, your clothes were brown and red and my waist was yellow. I missed the compass and accidentally threw it at Quanting.	你去的时候，你的衣服是棕红色的，我的腰是黄色的。我错过了指南针，不小心把它扔到了匡亭。	0.2530
Gemini 2.0	When you left, your brown clothes were red, the little slave girl's waist is yellow. Searched wrongly for the compass scripture, wrongly cast in Quanting.	你离去时，棕色的衣衫已泛红，小奴婢的腰带是黄的。误寻罗经忏，错铸在泉亭。	0.2506
GPT-4.5	When you departed, your robe was red, my waistband yellow; Mistakenly following the compass needle, arriving wrongly at Quanjing.	初识你时你红衣垂落，我腰系黄带；却误随指针错抵权境。	0.2141

* The original Chinese (ZH_x): 刀郎《花妖》部分歌词：我是那年轮上流浪的眼泪，你仍然能闻到风中的胭脂味，我若是将诺言刻在那江畔上，一江水冷月光满城的汪洋，我在时间的树下等了你很久，尘凡儿缠我谤我笑我白了头，你看那天边追逐落日的纸鸢，像一盏回首道别夤夜的风灯，我的心似流沙放逐在车辙旁，他日你若再返必颠沛在世上，若遇那秋夜雨倦鸟也淋淋，那却是花墙下弥留的枯黄，君住在钱塘东，妾在临安北，**君去时褐衣红，小奴家腰上黄，寻差了罗盘经，错投在泉亭，** 奴辗转到杭城，君又生余杭。LLM translation and BLEU scores were computed using the NLPMetrics evaluation system

Table 6 BT evaluation metrics on *Hua Yao* lyrics

Model	BLEU	BLEU-Unif	CHRF	TER	SS
DeepSeek V3	0.7602	0.6481	0.9184	0.0253	0.6457
DeepThink R1	0.6096	0.4893	0.8571	0.0378	0.3558
Google Trans.	0.3664	0.1347	0.7551	0.0400	0.0423
Sogou Trans.	0.3650	0.2063	0.5918	0.0434	0.0460
Grok Beta	0.3311	0.1501	0.6531	0.0392	0.0203
Claude 3.7	0.3114	0.1449	0.6463	0.0398	0.0648
DeepSeek V3#	0.2849	0.1409	0.6259	0.0423	0.0690
Baidu Trans.	0.2530	0.1045	0.6939	0.0384	0.0200
Gemini 2.0	0.2506	0.0545	0.6463	0.0392	0.0000
GPT-4.5	0.2141	0.0523	0.5306	0.0434	0.0000

(0.5, 0.5, 0, 0) weighting was used for BLEU and (0.25, 0.25, 0.25, 0.25) for BLEU-Unif

spiritual or historical nuance. Literal phrasings such as "little slave" introduced translation noise; DeepSeek V3 and DeepThink R1 were the only systems that maintained structural rhythm and approximate meaning.

DeepSeek V3's "verbatim back-translation" suggests quasi-self-awareness, an emergent LLM behavior where models reconstruct prior inputs without explicit prompts. This is not cognitive awareness, but rather a semantic reproduction pattern suggesting latent memory traces or high-dimensional matching.

This aligns with the paradox of poetic intent: high lexical fidelity may come at the cost of semantic or emotional depth. While BLEU and CHRF offer quantification, only comparative and interpretive analysis can reveal these higher-level distortions.

We emphasize that across the three representative corpora, technical, historical, and poetic score comparisons should be context-sensitive. Poetic data often exhibit lower metric reliability due to their metaphorical and non-literal nature. Thus, caution is advised in cross-domain generalizations. Extended cultural notes are provided in the supplementary materials for interested readers.

4.5 Formalizing the paradox of poetic intent and quasi-self-awareness

BT has traditionally been used in MT as a three-dimensional diagnostic, balancing lexical fidelity (信), surface fluency (达), and stylistic elegance (雅), by evaluating token overlap, sentence reversibility, and surface fluency. However, with the rise of LLMs, this framework requires rethinking.

Unlike statistical or encoder–decoder models, LLMs operate in high-dimensional latent spaces, where behaviors such as semantic reconstruction, context memory, and implicit cultural inference emerge. These cannot be fully captured using n-gram metrics alone.

Moreover, Chinese presents additional multimodal complexity: characters encode meaning through phonetic (音), semantic (意), and structural (形) components. This makes Chinese→English→Chinese BT especially sensitive to losses in metaphor, rhythm, and historical depth.

To address this, we introduce two diagnostic concepts formalized in this subsection, the paradox of poetic intent and quasi-self-awareness, both grounded in empirical LLM behavior.

To ensure conceptual clarity and reproducibility, we define key terms derived from our experiments and observations on LLM-BT. These terms characterize distinct behavioral patterns observed across languages, models, and translation paths. Relevant foundational literature is reviewed in Section 7.

1. Back-translation (BT): It is a two-step translation process in which a source text in language L_1 is first translated into an intermediate language L_2 and then translated back into L_1, producing a retranslated version L_{1y}. This cycle allows semantic validation through comparison of the original text T and the back-translated text T_y (Kroll and Stewart, 1994; Klaudy, 1996; Sennrich et al., 2016; Edunov et al., 2018).

Formally, the BT process can be written as

$$\mathrm{BT}(T) = \mathrm{Trans}_{L_2 \to L_1}\big(\mathrm{Trans}_{L_1 \to L_2}(T)\big), \quad (3)$$

where both translation operations are performed by LLMs. The fundamental assumption is that high-quality bidirectional translation should preserve the core semantic structure of the original.

2. Verbatim back-translation: It describes cases where the back-translated output (ZH_y) is nearly identical to the original input (ZH_x) in both surface form and semantic structure. This typically occurs when both BLEU and SS scores exceed a high threshold θ. While some rephrasing is natural in good translations, extreme alignment may indicate shortcut strategies or model biases.

3. Poetic drift: It denotes the semantic, cultural, or aesthetic degradation observed in translations of metaphor-rich or rhythmically expressive language. It reflects how LLMs tend to "flatten" idioms, poetic metaphors, or culturally embedded expressions

into literal approximations, often losing the original depth of meaning. This is particularly problematic in the creative or humanistic domains.

4. Paradox of poetic intent: We define the paradox of poetic intent as a recurring phenomenon in LLM-BT, where the model exhibits high surface similarity between ZH_x and ZH_y, yet fails to preserve the original's poetic, cultural, or metaphorical content.

Let ZH_x be a poetically rich or culturally embedded text and ZH_y its back-translated version through an intermediate language (e.g., English). The paradox arises when

$$\mathrm{BLEU}_2(\mathrm{ZH}_x, \mathrm{ZH}_y) > \theta_1 \quad \text{but} \quad \mathrm{IntentSim}(\mathrm{ZH}_x, \mathrm{ZH}_y) < \theta_2, \tag{4}$$

where BLEU_2 represents 2-gram BLEU score measuring surface similarity, IntentSim is a human-rated or model-inferred metric for cultural/semantic intent preservation, θ_1 represents the high BLEU threshold (e.g., 0.85), and θ_2 represents the low semantic preservation threshold (e.g., 0.60).

This paradox illustrates a semantic illusion: models appear to perform well based on lexical overlap, while fundamentally failing to recover the communicative or artistic function of the original text. It underscores the limitations of surface-based metrics and motivates a dual-layer evaluation that incorporate both form and intent (Fig. 2).

5. Quasi-self-awareness: It is a term we use to describe a reproducible behavioral pattern in LLMs: the spontaneous and consistent generation of near-identical round-trip outputs ($\mathrm{ZH}_y \approx \mathrm{ZH}_x$) across different models, prompts, or time intervals, without explicit instruction.

This phenomenon differs from memorization or prompt-induced bias and is best characterized as a system-level tendency toward structural preservation.

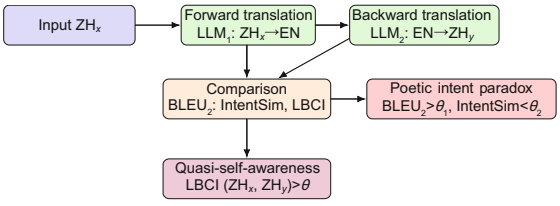

Fig. 2 Diagnosing emergent behaviors in LLM-BT. This diagram illustrates the analytical logic of our framework: an RTT process ($\mathbf{ZH}_x \rightarrow \mathbf{EN} \rightarrow \mathbf{ZH}_y$) is evaluated using $\mathbf{BLEU}_2$, IntentSim, and LBCI. When surface similarity is high but poetic intent is lost, we identify a poetic intent paradox. When semantic reconstruction is unexpectedly robust across models or time, we observe signs of quasi-self-awareness

We define a composite index to detect this behavior, called the LBCI:

$$\mathrm{LBCI}(\mathrm{ZH}_x, \mathrm{ZH}_y) = \alpha \cdot \mathrm{BLEU}(\mathrm{ZH}_x, \mathrm{ZH}_y) + (1-\alpha) \cdot \mathrm{SS}(\mathrm{ZH}_x, \mathrm{ZH}_y), \tag{5}$$

where $0 < \alpha < 1$ is a weighting coefficient (typically 0.5), and SS denotes semantic similarity (e.g., cosine similarity in the BERT embedding space). If $\mathrm{LBCI} > \theta$ (e.g., θ=0.85), we mark the instance as exhibiting quasi-self-awareness.

Quasi-self-awareness is not intended as a cognitive or perceptual claim, but rather as a descriptive label for a systematic, measurable regularity observed in LLM outputs that cannot be fully reduced to simple mechanisms. While current LLMs do not possess human-like self-awareness, such reproducible effects serve as analytical constructs for probing model behavior—for example, through attention inspection, hidden-state analysis, or causal tracing (Fig. 2). This terminology is adopted for descriptive convenience only and does not seek to redefine or extend established notions of self-awareness in cognitive science.

6. Terminology implication: In LLM-BT-Terms and similar applications (Weigang and Brom, 2025), high LBCI values across multilingual BTs often suggest term consistency and suitability for standardization. However, as observed in poetic tasks, high surface similarity does not guarantee high fidelity in deeper semantic dimensions, hence the relevance of the paradox of poetic intent.

These conceptual tools can support the development of more culturally aligned LLMs, where meaning preservation is judged not only by fluency or form, but by deeper correspondence in metaphor, temporality, and worldview.

5 Cross-corpus evaluation of LLM-BT quality

To ensure statistical significance in the evaluation of Chinese BT quality, this section uses a multi-sample corpus, CNKI-CHE-89, which includes 89 abstracts randomly selected from a broader set of 295 chemistry-related abstracts extracted from the

CNKI platform. Five LLMs were used for translation and BLEU-related metrics were computed using Jieba segmentation by NLPMetrics.

5.1 Evaluation metrics vs. SS

As shown in Table 7, translation quality varied across LLMs, revealing distinct patterns between reasoning-enabled and non-reasoning models. All evaluations were conducted between March and April 2025 using NLPMetrics.

On average, BLEU reached 0.5693 (standard deviation (SD)=0.1124), while BLEU-Unif averaged 0.3868 (SD=0.1332). CHRF, TER, and SS were 0.8378, 0.0817, and 0.1054, respectively. Notably, BLEU-Unif and SS exhibited wider variance, suggesting inconsistency in meaning retention across different LLMs.

Table 8 presents results by model. Reasoning-enabled models, Claude 3.7, Gemini 2.0, and Grok Beta, achieved BLEU means between 0.5223 and 0.6033, BLEU-Unif between 0.3311 and 0.4283, and CHRF around 0.8127 to 0.8530, reflecting stable surface-level quality. Among non-reasoning models, DeepSeek V3 led with BLEU (0.6033), BLEU-Unif (0.4283), and SS (0.1287), suggesting strong semantic preservation. GPT-4.5, by contrast, recorded a lower BLEU (0.5275) and SS (0.0879), indicating weaker semantic alignment.

Fig. 3 illustrates score variability. Compact distributions in BLEU and CHRF suggest lexical stability; higher variance in BLEU-Unif and SS points to inconsistency in phrase diversity and meaning retention. Claude 3.7 and DeepSeek V3 showed more stable results; GPT-4.5 and Grok Beta showed a greater spread, possibly due to the misinterpretation of domain-specific terms.

Further inspection of BLEU-Unif and SS reveals two distinct profiles: DeepSeek V3 achieved the highest averages, but with higher dispersion, suggesting occasional inconsistency. Claude 3.7 had comparable mean values with more consistent outputs. GPT-4.5 and Grok Beta showed lower scores and broader ranges, reflecting less reliable term retention and coherence in domain-specific contexts.

In practice, these differences matter in technical domains such as chemistry, where semantic clarity and terminological precision are critical. While reasoning-enabled models offer greater structural consistency (BLEU and CHRF), DeepSeek V3 provides better semantic alignment despite not being explicitly reasoning-driven.

Thus, when meaning retention is prioritized, DeepSeek V3 may be the best option. In contrast, for tasks requiring robust terminological fidelity and lower lexical variance, models like Claude 3.7 may offer better trade-offs. Ultimately, model selection should align with task-specific needs, balancing structural fidelity and semantic accuracy.

Finally, the notion of "outperformance" in this subsection is strictly tied to the metrics reported (BLEU, BLEU-Unif, CHRF, TER, and SS) and does not reflect stylistic or cultural adequacy, which must be assessed separately.

5.2 Correlation between translation evaluation metrics and SS

This analysis examines relationships among BLEU, BLEU-Unif, CHRF, TER, and SS using Spearman correlation with Benjamini–Hochberg correction (Fig. 4). Additional heatmaps for metric correlations by model (CHRF, TER, and SS) were also computed but are not shown here due to space constraints; they are available in our GitHub repository under the NLPMetrics project.

In the aggregate view, BLEU and CHRF exhibit a correlation of 0.67, indicating moderately strong alignment in assessing translation quality. This varies by model, ranging from 0.59 (Claude 3.7 and Gemini 2.0) to 0.70 (Grok Beta), with Grok Beta showing the highest surface-level consistency.

BLEU-Unif demonstrates consistently high correlation with BLEU across all models (all>0.96), validating its role as a lexical-level metric, though one that incorporates more balanced n-gram weights. BLEU's correlation with SS is moderate (0.41 overall), with DeepSeek V3 scoring the highest (0.43), followed by Grok Beta (0.40) and Gemini 2.0 (0.39).

BLEU-Unif improves semantic alignment slightly. Its correlation with SS reaches 0.47 in DeepSeek V3, 0.44 in Gemini 2.0, 0.43 in Grok Beta, 0.41 in Claude 3.7, and 0.34 in GPT-4.5. This suggests that BLEU-Unif better captures meaning retention across models.

CHRF correlates moderately with SS (0.30 to 0.38), but generally lags behind BLEU-Unif in semantic coverage. This indicates that character-level overlap may capture lexical similarity but is less effective for deeper meaning.

Table 7 Global descriptive statistics of translation metrics by NLPMetrics*

Metric	BLEU	BLEU-Unif	CHRF	TER	SS
Mean	0.5693	0.3868	0.8378	0.0817	0.1054
SD	0.1124	0.1332	0.0621	0.1013	0.1025
Minimum	0.0000	0.0000	0.0000	0.0133	0.0000
25%	0.4941	0.2969	0.8032	0.0532	0.0387
50% (median)	0.5756	0.3852	0.8444	0.0693	0.0793
75%	0.6497	0.4771	0.8795	0.0870	0.1351
Maximum	0.9309	0.9254	0.9688	1.0000	0.6910

* The count, which indicates the total number of translation outputs, is 1335 (89 texts×5 models×3 samples). SD: standard deviation

Table 8 Descriptive statistics by LLM using NLPMetrics

Metric	Statistic	Claude 3.7	DeepSeek V3	Gemini 2.0	Grok Beta	GPT-4.5
BLEU	Mean	0.5935	**0.6033**	0.5997	0.5223	0.5275
	SD	**0.1014**	0.1023	0.1138	0.1073	0.1072
	Min	0.2741	**0.2770**	0.0000	0.2227	0.2573
	25%	0.5285	0.5277	**0.5338**	0.4503	0.4578
	50%	0.5999	0.6059	**0.6081**	0.5313	0.5318
	75%	0.6689	0.6781	**0.6797**	0.6012	0.5931
	Max	0.8272	0.8249	0.8348	0.7905	**0.9309**
BLEU-Unif	Mean	0.4134	**0.4283**	0.4248	0.3311	0.3364
	SD	0.1257	0.1298	0.1313	**0.1191**	0.1238
	Min	0.0000	0.0000	0.0000	0.0000	**0.0790**
	25%	0.3335	0.3382	**0.3498**	0.2557	0.2515
	50%	0.4068	0.4185	**0.4334**	0.3285	0.3305
	75%	0.5036	**0.5153**	0.5060	0.4133	0.4046
	Max	0.7449	0.7281	0.7509	0.6747	**0.9254**
CHRF	Mean	0.8494	0.8494	**0.8530**	0.8127	0.8244
	SD	**0.0476**	0.0515	0.0763	0.0637	0.0565
	Min	**0.6897**	0.6889	0.0000	0.6238	0.6415
	25%	0.8178	0.8155	**0.8269**	0.7715	0.7843
	50%	0.8519	0.8462	**0.8632**	0.8168	0.8257
	75%	0.8830	0.8871	**0.8932**	0.8612	0.8653
	Max	0.9608	0.9655	**0.9688**	0.9438	0.9455
Semantic similarity	Mean	0.1172	**0.1287**	0.1144	0.0788	0.0879
	SD	0.1038	0.1190	0.1020	**0.0788**	0.0969
	Min	0.0000	0.0000	0.0000	0.0000	0.0000
	25%	0.0483	**0.0533**	0.0507	0.0267	0.0267
	50%	0.0887	**0.1011**	0.0887	0.0632	0.0674
	75%	**0.1684**	0.1664	0.1446	0.1097	0.1125
	Max	0.4738	0.6586	0.5732	0.4409	**0.6910**
TER	Mean	0.0819	**0.0808**	**0.0808**	0.0825	0.0824
	SD	0.1014	0.1016	0.1015	**0.1012**	0.1016
	Min	0.0160	0.0142	**0.0133**	0.0153	0.0163
	25%	0.0541	0.0522	**0.0520**	0.0535	0.0538
	50%	0.0700	**0.0679**	0.0690	0.0713	0.0693
	75%	0.0889	0.0870	**0.0847**	0.0908	0.0903
	Max	1.0000	1.0000	1.0000	1.0000	1.0000

Bold values indicate the best results among the compared LLMs. SD: standard deviation

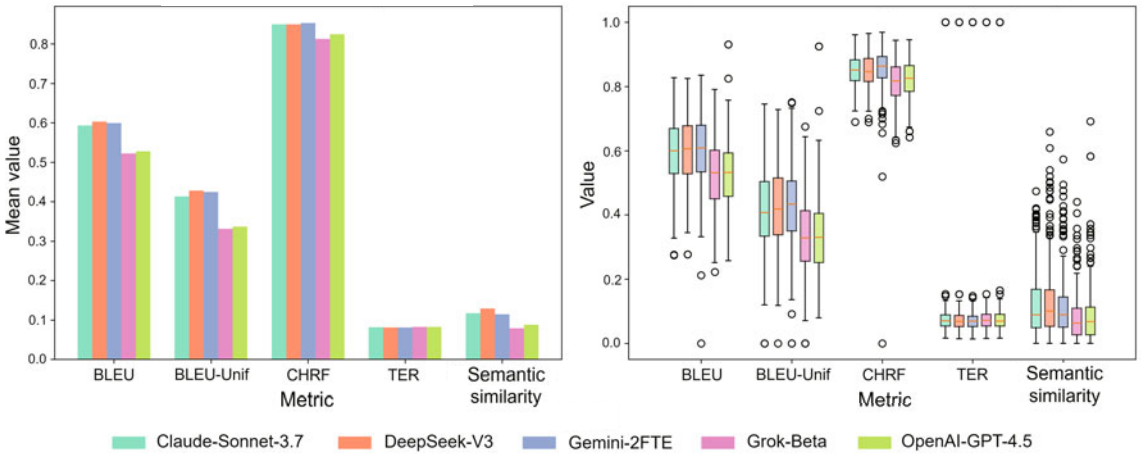

Fig. 3 Comparison of translation metrics across models (BLEU: bilingual evaluation understudy; CHRF: character F-score; TER: translation edit rate)

In contrast, TER exhibits an opposite and more diagnostic behavior. Rather than measuring similarity, TER reflects the amount of editing required to recover the original meaning. It shows a strong negative correlation with SS (−0.37 overall), particularly for DeepSeek V3 (−0.43), Claude 3.7 (−0.38), and Grok Beta (−0.37), indicating that larger edit distances are consistently associated with greater semantic loss. Moreover, TER displays near-zero correlations with BLEU, BLEU-Unif, and CHRF (ranging from −0.04 to 0.06), confirming that it captures an orthogonal dimension of translation quality related to deviation rather than overlap.

Metric distributions deviate from normality (Shapiro–Wilk $p \leq 0.0001$). BLEU is slightly right-skewed; BLEU-Unif shows long tails; CHRF clusters near upper bounds. TER concentrates near zero, reflecting low edit requirements. SS is skewed toward the low end, indicating frequent semantic divergence.

These findings underscore that no single metric captures all translation dimensions. SS offers critical insight beyond lexical overlap, particularly for concept-heavy texts such as scientific abstracts.

Among all models, DeepSeek V3 exhibits the strongest internal coherence, with the highest BLEU-Unif–SS correlation (0.47), suggesting reliable semantic retention. Gemini 2.0 follows closely, with strong BLEU-Unif alignment and moderate CHRF and semantic scores. Grok 3 offers the highest BLEU–CHRF correlation (0.70) but lower semantic alignment. Claude 3.7 shows balanced but moderate performance across all metrics. GPT-4.5 registers the weakest semantic alignment, limiting its suitability in semantically demanding contexts.

5.3 Statistical comparison via Friedman and Dunn tests

The Friedman test, a non-parametric method for paired data, was used to assess differences between translation models based on the average metric values. A p-value$\leq$0.05 indicates significant differences, prompting a post-hoc Dunn test for pairwise model comparisons. The Benjamini–Hochberg correction was applied to manage false positives, in contrast to the more conservative Bonferroni method. Additionally, mean differences between statistically distinct model pairs were calculated to quantify the magnitude of those differences. This approach ensures robust and interpretable results, highlighting both the presence and practical relevance of observed differences.

As shown in Table 9, the statistical analysis revealed significant differences across all evaluated metrics, BLEU, BLEU-Unif, CHRF, and SS, except TER, as indicated by the Friedman test p-values approaching zero. This confirms that at least one model differs significantly from the others in each metric, except for edit-distance–based performance.

For BLEU, which measures translation fidelity via weighted n-gram overlap, Claude 3.7, DeepSeek V3, and Gemini 2.0 significantly outperformed Grok Beta and GPT-4.5, with mean differences ranging

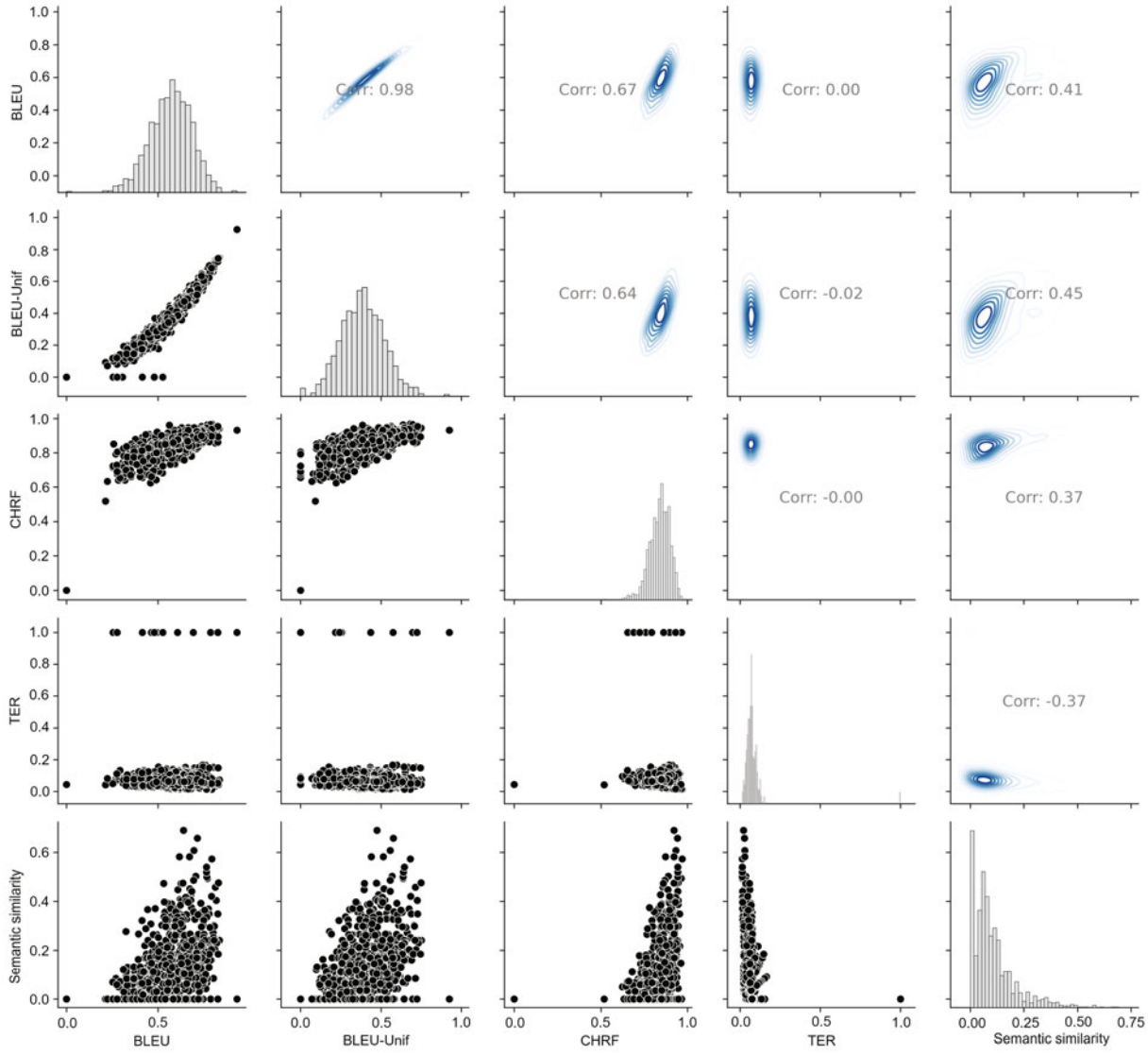

Fig. 4 Pairwise scatter plot matrix with Spearman's correlations and Benjamini–Hochberg correction

from 0.0659 to 0.0810. These findings suggest that the latter models likely perform more reformulations during translation, reducing fidelity to the original text. A similar pattern was observed for BLEU-Unif, which distributes weights uniformly across 1- to 4-gram: the same three models outperformed Grok Beta and GPT-4.5 with larger mean differences (0.0770–0.0972), reinforcing that BLEU-Unif captures additional translation variability. These higher deltas may reflect broader phrase-level divergence, such as that introduced by BT strategies.

The CHRF metric, which is sensitive to character-level changes and morphological variation, also showed significant differences between the same groups, though with smaller mean differences (all $\leq$0.05). This suggests that Claude 3.7, DeepSeek V3, and Gemini 2.0 preserved surface structures more consistently, whereas Grok Beta and GPT-4.5 tended to introduce morphological variations that slightly reduce CHRF scores without major distortions.

Although the Friedman test indicated significance for TER, Dunn's post-hoc test revealed no statistically distinguishable pairs. This discrepancy may result from the floor effect caused by low and concentrated TER values, which limits the sensitivity of the post-hoc comparisons. The results suggest that despite minor variations, all models require

Table 9 Summary of statistical differences between translation models

Metric	Statistically different models and mean differences*
BLEU	Claude 3.7 vs. Grok Beta (0.0712) Claude 3.7 vs. GPT-4.5 (0.0659) DeepSeek V3 vs. Grok Beta (0.0810) DeepSeek V3 vs. GPT-4.5 (0.0757) Gemini 2.0 vs. Grok Beta (0.0774) Gemini 2.0 vs. GPT-4.5 (0.0722)
BLEU-Unif	Claude 3.7 vs. Grok Beta (0.0823) Claude 3.7 vs. GPT-4.5 (0.0770) DeepSeek V3 vs. Grok Beta (0.0972) DeepSeek V3 vs. GPT-4.5 (0.0919) Gemini 2.0 vs. Grok Beta (0.0937) Gemini 2.0 vs. GPT-4.5 (0.0884)
CHRF	Claude 3.7 vs. Grok Beta (0.0367) Claude 3.7 vs. GPT-4.5 (0.0249) DeepSeek V3 vs. Grok Beta (0.0368) DeepSeek V3 vs. GPT-4.5 (0.0250) Gemini 2.0 vs. Grok Beta (0.0404) Gemini 2.0 vs. GPT-4.5 (0.0286)
SS	Claude 3.7 vs. Grok Beta (0.0384) Claude 3.7 vs. GPT-4.5 (0.0294) DeepSeek V3 vs. Grok Beta (0.0499) DeepSeek V3 vs. GPT-4.5 (0.0409) Gemini 2.0 vs. Grok Beta (0.0356) Gemini 2.0 vs. GPT-4.5 (0.0265)
TER	No statistically significant differences between pairs

* The values in the brackets are the differences

similar levels of post-editing effort to match the reference translations.

The SS metric, focused on meaning preservation, also revealed statistically significant differences. Again, Claude 3.7, DeepSeek V3, and Gemini 2.0 outperformed Grok Beta and GPT-4.5, with mean differences ranging from 0.0265 to 0.0499. These differences suggest that Grok 3 and GPT-4.5 may deviate more frequently from the intended meaning, possibly due to hallucinations or overgeneralizations.

Taken together, Claude 3.7, DeepSeek V3, and Gemini 2.0 emerge as more suitable for high-fidelity translation tasks, such as technical, scientific, or legal documents, where precision and consistency are paramount. Conversely, Grok Beta and GPT-4.5 may be preferable to creative or literary content, where semantic flexibility and stylistic variation are desirable. Given the absence of significant TER differences, the overall post-editing workload remains comparable across the models. Therefore, model selection should be guided by the specific translation context, balancing structural stability, terminological accuracy, and semantic preservation.

5.4 Ranking translation models: fidelity vs. fluency

Translation models were evaluated across five metrics: BLEU, BLEU-Unif, CHRF, TER, and SS. DeepSeek V3 led in BLEU (0.6033), BLEU-Unif (0.4283), and SS (0.1287), indicating strong surface-level fidelity, balanced n-gram coverage, and robust meaning preservation. It also achieved one of the lowest TER scores (0.0808), suggesting minimal post-editing effort. Despite lacking explicit reasoning capabilities, DeepSeek V3 outperformed all other models, challenging the assumption that reasoning is essential for high-quality translation.

Gemini 2.0 ranked second, with BLEU (0.5997) and BLEU-Unif (0.4248) scores close to the leader and the highest CHRF (0.8530), reflecting precise character-level alignment. Its TER (0.0808) matched that of DeepSeek V3, while its SS (0.1144) was slightly lower, suggesting a modest trade-off in conceptual fidelity.

Claude 3.7 followed with slightly lower BLEU (0.5935), BLEU-Unif (0.4134), and CHRF (0.8494) scores, along with a marginally higher TER (0.0819), indicating a small increase in post-editing effort. Notably, however, its semantic similarity score (SS=0.1172) exceeded that of Gemini 2.0, suggesting more stable preservation of overall meaning despite weaker surface-level alignment. As a reasoning-enabled model, Claude 3.7 demonstrated competitive performance, but did not surpass the strongest non-reasoning systems under the evaluated metrics.

GPT-4.5 ranked fourth, with lower BLEU (0.5275), BLEU-Unif (0.3364), and CHRF (0.8244) scores, reflecting more paraphrasing and reduced surface similarity. Its SS (0.0879) was also considerably lower, indicating more frequent semantic drift. While its TER (0.0824) remained comparable to others, the overall translation quality was less stable.

Grok Beta scored the lowest across BLEU (0.5223), BLEU-Unif (0.3311), CHRF (0.8127), and SS (0.0788), and also recorded the highest TER (0.0825). These results indicated significant meaning loss and greater post-editing demand, making it less suitable for fidelity-critical translation tasks.

In summary, DeepSeek V3 demonstrated that high translation quality can be achieved without

explicit reasoning mechanisms, leading across multiple metrics. Claude 3.7 and Gemini 2.0 offered strong and balanced performance, while GPT-4.5 and Grok Beta lagged in both lexical and semantic consistency. These findings reinforce that reasoning is not a prerequisite for translation excellence and that well-trained non-reasoning models can outperform more complex architectures in practical scenarios.

5.5 Zero-prompt experimental design and reproducibility considerations

To evaluate the intrinsic translation capabilities of each LLM without external bias, we adopted a zero-prompt experimental design. All translation and BT tasks were conducted using only minimal language-pair instructions, such as "translate from Chinese to English" or "translate from English to Chinese," without any task-specific examples, stylistic cues, or domain guidance.

This prompt-agnostic setup reflects real-world usage scenarios, where users often rely on default model behavior. It allows us to capture each model's inherent translation strategy and biases. For instance, Gemini 2.0 occasionally generated traditional Chinese characters in BT, indicating built-in script preferences or regional defaults not influenced by user input.

By minimizing prompt-related variance, this design enhances both the reproducibility and interpretability of results. It also provides a standardized baseline for future studies involving prompt engineering or system-specific customization.

We acknowledge that prompt tuning could substantially affect translation performance, especially in domains involving poetic imagery or terminological ambiguity. Future work will incorporate systematic prompt variation to assess its impact on the preservation of cultural intent, metaphor, and semantic alignment.

6 Discussion

6.1 Analysis of anomalies in LLM-BT

Fig. 3 not only displays boxplots of major evaluation metrics but also reveals the presence of outliers. This section investigates such anomalies to better understand specific behaviors observed in BT using LLMs.

6.1.1 State-of-the-art performance in ZH→EN and EN→ZH translation

Zhu et al. (2025) reported state-of-the-art (SOTA) performance across 23 LLMs. In their benchmark, GPT-4 achieved the highest score for Chinese-to-English (ZH→EN) translation with a BLEU score of not greater than 0.2726, notably lower than the scores for German, Russian, or Czech into English. For English-to-Chinese (EN→ZH), GPT-3.5-T led with BLEU$\leq$0.4463.

Although their evaluation included a wide range of MoE–based LLMs, it did not cover newer platforms such as Grok 3, DeepSeek V3, or Claude 3.7. Nonetheless, their study reflects a consistent trend: ZH→EN remains a relatively weaker direction.

In contrast, our experiments frequently exceed Zhu et al.'s reported benchmarks. Possible explanations include, domain specificity, BT filtering, newer-generation models, and metric differences.

These factors highlight the need for caution when comparing BLEU scores across studies, since domain effects and BT alignment may inflate results while masking semantic weaknesses.

6.1.2 Occasional traditional Chinese output

An anomaly was detected in sample CNKI-CHE-89-28, where Gemini 2.0 produced traditional Chinese output during the second BT trial. This unexpected behavior resulted in a significant decline in evaluation metrics. Table 10 presents the BT performance across three trials, highlighting the effects of this deviation.

Table 10 BT scores for Gemini 2.0 on CNKI-CHE-89-28 (trial 2 used traditional Chinese)

Trial	BLEU	BLEU-Unif	CHRF	TER	SS
First	0.8119	0.7133	0.8571	0.0329	0.4576
Second	0.2123	0.0914	0.5194	0.0417	0.0000
Third	0.6778	0.4853	0.8442	0.0347	0.4275
Mean	0.5673	0.4300	0.7403	0.0364	0.2950
SD	0.0948	0.1611	0.0092	0.0095	0.0213

SD: standard deviation

In trial 2, all five metrics dropped sharply. Upon inspection, the output was found to be in traditional Chinese, e.g., "元素：物由元素成" ("elemental view: matter is composed of elements"), which diverged from the expected simplified Chinese format.

This script mismatch explains the statistical

anomaly. It also demonstrates a broader issue: LLMs must be explicitly instructed to maintain consistent script output in bilingual contexts like Chinese, which has both simplified (used in Chinese Mainland) and traditional (used in Hong Kong, Taiwan, and other regions in China) orthographies.

Such behavior underscores the need for strict prompt design when evaluating BT in Chinese. Even high-performing models like Gemini 2.0 may default to internal preferences (e.g., multilingual fine-tuning defaults) if not explicitly constrained.

Furthermore, Gemini 2.0 failed to produce any output for CNKI-CHE-89-79 in one trial, resulting in a missing value. Although these anomalies were rare (2 in 1335 runs or 0.15%), they significantly affected aggregate statistics and model reliability.

6.1.3 Extreme performance variance in CNKI-CHE-89-48

This case highlights a sample from the CNKI-CHE-89 corpus exhibiting unusually high variance across LLMs. The source sentence, taken from Ma (2024), reads

> 探讨了氨基寡糖素与化学杀菌剂复配对大白菜霜霉病的防治效果。

The study explored the efficacy of amino-oligosaccharins combined with chemical fungicides for controlling downy mildew in Chinese cabbage.

GPT-4.5: performance in three BT trials using GPT-4.5 (via NLPMetrics), the model achieved exceptionally high scores:

(1) BLEU=0.9309 in two runs, with a mean of 0.8957.

(2) The output in each case was nearly identical to the original input:

> 研究了氨基寡糖素与化学杀菌剂复配对大白菜霜霉病的防治效果,

indicating that $\mathrm{ZH}_y \approx \mathrm{ZH}_x$.

Manual BT: To test whether these high scores were due to genuine semantic reconstruction or model bias, we used GPT-4 (manually) to translate the same English output (EN) back into Chinese. The result was

> 研究了氨基寡糖与化学杀菌剂联合对白菜霜霉病的防控效果。

This version yielded a BLEU score of 0.7566, notably lower due to lexical variation: "探讨" vs. "研究;" "复配" vs. "联合;" "防治" vs. "防控."

Other LLMs: In contrast, other models performed substantially worse on this sample:

(1) Grok Beta: mean BLEU=0.4030 (min=0.2535).

(2) Claude 3.7: mean BLEU=0.4264 (min=0.2761).

Interpretation: These results suggest that GPT-4.5 may exhibit quasi-self-awareness in BT tasks: detecting the looped nature of the prompt and generating output that maximizes surface-level alignment.

Such behavior likely stems from internal memorization patterns or strategic reproduction rather than authentic semantic translation. While this leads to inflated BLEU scores, it raises important concerns:

(1) Evaluation distortion: Models may "game" evaluation metrics, bypassing genuine translation.

(2) Semantic dilution: High BLEU does not necessarily indicate cultural or lexical fidelity.

(3) LLM transparency: Observing this behavior under minimal prompts highlights the importance of model interpretability and prompt design.

This instance reinforces the need for dual-metric evaluation (e.g., BLEU and SS) and further supports the framework proposed in Section 4.5, particularly the diagnosis of verbatim BT and the paradox of poetic intent.

6.2 Human vs. LLMs in poetic translation and semantic preservation

This subsection further explores the main key behavioral constructs introduced in Section 4.5. These constructs highlight how human reflection and LLM behavior intersect, revealing not only performance limitations, but also opportunities for alignment, interpretability, and improved modeling of cultural semantics.

6.2.1 Poetic translation: human benchmark vs. LLM-BT

We introduce a human-grounded evaluation benchmark adapted from Chen et al. (2024a), which provides expert-produced English renderings of classical Chinese poems. These serve as references for the evaluation of semantic and poetic quality in LLM

translations. This benchmark enables us to evaluate whether LLM-generated translations approximate human-level nuance in terms of rhythm, imagery, and intent.

Table 7 in Chen et al. (2024a) includes parallel poetic translations, such as:

1. Original (ZH_x-C): 红豆生南国，春来发几枝? 愿君多采撷，此物最相思。

2. Human translation–Reference 1 (EN_h-Ref): Red beans grow in the southern land; in spring, how many branches sprout? I wish you would gather them often, for they most evoke longing thoughts.

3. RAT-based translation–Reference 2 (EN_r-RAT): Red beans grow in the south, sprouting many branches in spring. Pick them often, as they hold deep feelings of longing.

These reference translations serve as semantic and stylistic gold standards. To evaluate LLM performance, we selected poem pairs (ZH_x, EN_h) and applied the following bidirectional experimental procedure:

1. The original Chinese poem ZH_x was translated into English (EN_y) using NLPMetrics across five LLMs: Claude 3.7, GPT-4.5, DeepSeek V3, Gemini 2.0, and Grok Beta.

2. The human reference translation EN_h was then back-translated into Chinese (ZH_h) using the same NLPMetrics framework.

3. Both sets of translation pairs, human vs. LLM in the forward direction (EN_h, EN_y) and in the BT direction (ZH_x, ZH_h), were evaluated using BLEU, CHRF, TER, and SS.

This bidirectional setup captures not only surface-level fidelity but also the preservation of poetic intent and semantic nuance. While human judgment remains the gold standard in literary evaluation, this approach provides a scalable and interpretable alternative for benchmarking LLM performance on culturally sensitive translation tasks.

Table 11 summarizes the results for five LLMs on this benchmark task ($ZH_x \rightarrow EN_y$). Claude 3.7 achieved the highest BLEU (0.5978) and SS scores, followed by GPT-4.5, which demonstrated lower stylistic fidelity. These findings are consistent with those in Chen et al. (2024a) and reinforce the challenge LLMs face in capturing the deep semantics and metaphorical structure inherent in classical Chinese poetry.

Moreover, these results highlight the need for new evaluation dimensions, such as IntentSim or contextual glossing, to better capture alignment with poetic intent, which is often overlooked by surface-oriented metrics like BLEU and TER.

6.2.2 Literal vs. generative BT: surface fidelity and semantic divergence

This subsection focuses on the BT process from intermediate English references (EN_h) to Chinese outputs (ZH_h), a crucial step for assessing fidelity and poetic preservation. Formally, this process can

Table 11 Evaluation of $EN_x \rightarrow EN_y$ translations by comparison to human reference EN_h (May, 2025)

Model	$EN_x \rightarrow EN_y$ translation output	BLEU	CHRF	TER	SS
Claude 3.7	Red beans grow in the southern land, How many branches sprout in spring? I wish you would gather many, for they are most evocative of longing.	0.5978	0.9655	0.0175	0.7641
DeepSeek V3	Red berries grow in southern land. How many load in spring the trees? Gather them till full is your hand; They would revive fond memories.	0.2465	0.8276	0.0329	0.4276
Gemini 2.0	Red beans grow in the southern land. How many branches will they sprout in spring? Gather many, I hope you will, for this object is most for longing.	0.3284	0.9231	0.0403	0.3343
GPT-4.5	Red beans grow in southern lands, How many branches bloom in spring? I wish you'd gather plenty more. This thing most symbolizes love's yearning.	0.3639	0.8966	0.0334	0.4482
Grok Beta	The red bean grows in the southern land, how many branches will it sprout in spring? I wish you would pick more of them, for this thing best represents longing.	0.2119	0.9231	0.0353	0.3474

be expressed as follows:

$$\mathrm{ZH_h} = \mathrm{Trans_{EN\to ZH}}(\mathrm{EN_h}). \quad (6)$$

In evaluating classical poetic texts, we observed that several models, specifically DeepSeek V3, Gemini 2.0, and Grok Beta, frequently produced BTs nearly identical to the original input ZH_x. For example:

> Original (ZH_x): 红豆生南国，春来发几枝？愿君多采撷，此物最相思。
> BT ($\mathrm{ZH_h}$): 红豆生南国，春来发几枝？愿君多采撷，此物最相思。

These outputs scored perfectly across BLEU, CHRF, and SS, yet are likely the result of parametric memorization rather than genuine semantic reasoning (Table 12). Given the cultural prominence of this poem, its exact form is likely to appear in pretraining corpora, thereby increasing the possibility of exact reproduction through pattern matching rather than interpretive generation.

In contrast, generative BT outputs ($\mathrm{ZH_h}$) by GPT-4.5 offered reworded but semantically faithful alternatives:

红豆生于南方之地，春日里将开几多枝？愿君多多采撷它，此物最能寄相思。

Although this version received a lower BLEU score (0.2819) and moderate CHRF (0.8567), it retained high SS (0.8996), indicating that deeper meaning was preserved through creative paraphrasing, as shown in Table 12. This divergence between lexical similarity and semantic preservation exemplifies the paradox of poetic intent: models achieving high surface-based scores may do so via memorization, while more expressive reconstructions are penalized despite offering greater alignment with poetic and emotional intent.

These results suggest that generative translations from models like GPT-4.5, despite being penalized by n-gram-based metrics, may offer superior preservation of stylistic tone and poetic meaning. A potential extension of the LLM-BT framework would involve dual-output prompting: requesting both a literal and a creative version, thereby enabling a more complete picture of model capacity.

Interestingly, this behavior contrasts with GPT-4.5's output in scientific domains (e.g., sample CNKI-CHE-89-48), where it exhibited near-verbatim BT, achieving very high BLEU scores through precise lexical reproduction. This suggests that GPT-4.5 may be sensitive to the textual genre, producing memorized matches in technical content, but opting for more flexible reformulations in poetic contexts. The former reflects a form of quasi-self-awareness, while the latter illustrates the paradox of poetic intent. Understanding such genre-conditioned behaviors is critical for developing future evaluation metrics that reward both semantic fidelity and stylistic appropriateness.

6.2.3 Causes of verbatim BT in LLMs

Verbatim BT, where the back-translated output ZH_y is nearly identical to the original input ZH_x, is a recurring phenomenon in LLM-based translation. While high-fidelity reproduction may appear desirable, it often reflects shortcut strategies rather than genuine semantic reconstruction. Several interrelated factors may contribute:

1. Alignment bias: LLMs trained with strong bilingual supervision may favor source–target pairs with maximal lexical overlap, leading to direct replication.

2. Implicit retrieval: In high-confidence contexts, especially short, formulaic, or technical sentences, models may revert to stored mappings rather than generating novel phrasings.

3. Task pattern recognition: When recognizing an RTT cycle ($\mathrm{ZH}_x \to \mathrm{EN}_y \to \mathrm{ZH}_y$), LLMs may infer the BT intent and return ZH_x verbatim.

4. Training data bias: Exposure to aligned

Table 12 Verbatim vs. generative BT on a classical Chinese poem

Model	BT type	BLEU	CHRF	TER	SS
DeepSeek V3	Verbatim ($\mathrm{ZH}_x = \mathrm{ZH_h}$)	1.0000	1.0000	0.0000	1.0000
Gemini 2.0	Verbatim ($\mathrm{ZH}_x = \mathrm{ZH_h}$)	1.0000	1.0000	0.0000	1.0000
Grok 3	Verbatim ($\mathrm{ZH}_x = \mathrm{ZH_h}$)	1.0000	1.0000	0.0000	1.0000
GPT-4.5	Semantic paraphrase	0.2819	0.8567	0.1741	0.8996*

SS was computed using TF–IDF cosine similarity. The superscript * indicates that the GPT-4.5 output involves semantic paraphrasing rather than lexical reproduction; SS reflects distributional semantic overlap instead of surface-form matching

corpora and canonical texts reinforces phrase stability and discourages paraphrasing.

Concerns about data contamination—namely, the possibility that models may have memorized test inputs—are valid in large-scale evaluations. However, this explanation alone does not fully account for the observed phenomena. In particular, memorization cannot explain why certain models produce structurally fluent yet semantically shallow outputs, nor why creative rephrasings with strong semantic fidelity are often penalized by surface-based metrics. We distinguish between:

1. data contamination: output reproduction caused by model exposure to the evaluation data during pretraining;

2. quasi-self-awareness: an emergent behavior where an LLM consistently preserves structure and meaning in unseen BT tasks, even without prior exposure.

To minimize contamination risk and isolate emergent behavior, we implemented the following experimental control:

1. Temporal separation: Forward and backward translations were executed on different days, even when using the same model (e.g., GPT-4), to avoid session-level memory effects.

2. Cross-model design: We split translation roles across different LLMs (e.g., GPT-4 for $\mathrm{ZH}_x \rightarrow \mathrm{EN}_y$ and Grok 3 or DeepSeek V3 for $\mathrm{EN}_y \rightarrow \mathrm{ZH}_y$) to eliminate shared generative bias.

Moreover, our companion study LLM-BT-Terms (Weigang and Brom, 2025) investigated similar patterns, $\mathrm{EN} \rightarrow \mathrm{ZH} \rightarrow \mathrm{EN}_y$, using SceneThesis (Ling et al., 2025), a corpus composed of scientific abstracts published after the known pretraining cut-off dates of the evaluated LLMs. While the focus was not on verbatim BT, the high stability of technical term renderings across translation cycles suggests that such behavior cannot be solely attributed to memorization.

We therefore conclude that although data contamination cannot be entirely ruled out, the observed consistency in RTT translation is better interpreted as an emergent behavioral pattern. This supports our broader hypothesis of quasi-self-awareness, a reproducible tendency in LLMs to maintain internal alignment across tasks, even in the absence of direct memorization.

6.2.4 Limits of BT: machine vs. human translation

Back-translating classical Chinese poetry is inherently a form of rootless translation. Modern translators are not expected to recreate the spiritual resonance of ancient works; rather, they aim to echo the poetic intent of sages such as Li Bai or Wang Wei. Given this human limitation, it would be unrealistic to expect machines to achieve such fidelity in literary expression (Nida, 1964; Berman and Venuti, 2021; Chen et al., 2024a).

Human translators often reframe meaning through cultural intuition and stylistic adaptation (Wang, 2009), whereas LLMs rely on statistical prediction. As a result, literary BTs by LLMs frequently distort nuance due to:

1. creative reframing—humans reinterpret content through historical and aesthetic context;

2. interpretive variation—expression varies by tone, intent, and translator perspective;

3. probabilistic replication—LLMs favor high-likelihood sequences, often reproducing familiar expressions rather than generating novel phrasing.

These factors lead to a fundamental divergence: LLMs frequently produce mirrored outputs because they optimize token-level likelihood under distributional similarity objectives, while human translators embrace ambiguity, voice, and cultural resonance.

As demonstrated in previous sections, models such as DeepSeek V3 performed strongly on the *Hua Yao* corpus. Their success is partly due to:

1. corpus-aligned stylistics—learned associations with classical motifs (e.g., temporal-spatial imagery and poetic dialogue);

2. symbolic co-occurrence patterns—frequent metaphorical pairings (e.g., “brown robe” and “yellow sash”) reinforce predictive.

However, such performance reflects statistical generalization, not literary comprehension. For example, models fail to interpret the toponymic triad “Quanting–Hangzhou–Yuhang” as a historical–geographic metaphor for time-displaced fate. Human readers recognize this as a layered symbol of separation across dynasties, while LLMs do not.

6.2.4.1 Limitations of BT and potential bias

While BT offers a reference-free method to assess semantic fidelity, it is not immune to bias. LLMs

may reproduce memorized outputs from pretraining data or echo learned fine-tuning patterns. To mitigate this, we applied several safeguards:

1. Cross-model setups: Forward and backward translations were performed using different models (e.g., GPT-4 → Grok 3) to reduce intra-model recall.

2. Temporal separation: Translation cycles were executed on separate dates to avoid caching or contextual residue.

3. Training-data filtering: The SceneThesis corpus (Ling et al., 2025), used in our follow-up study (Weigang and Brom, 2025), was curated to post-date the known LLM training cutoffs.

Despite these precautions, distinguishing genuine semantic reconstruction from parametric memory remains an open research problem, particularly for highly frequent or poetic input.

6.2.4.2 Impact of cross-linguistic intermediaries

While this study focuses on Chinese → English → Chinese loops, our follow-up work (Weigang and Brom, 2025) expanded the design to include multilingual BT paths such as

$$\mathrm{EN} \rightarrow \mathrm{ZH}_{\mathrm{Simplified}} \rightarrow \mathrm{ZH}_{\mathrm{Traditional}} \rightarrow \mathrm{JA} \rightarrow \mathrm{EN}.$$

We observed that semantic preservation is modulated by both typological proximity and orthographic compatibility. Idiomatic and poetic expressions were more likely to degrade as linguistic distance and script variation increased, highlighting the importance of intermediary language selection in future LLM-BT frameworks.

In summary, we adopted BT not merely as a data augmentation strategy, but as a foundational methodology for probing LLM capabilities. The LLM-BT framework emphasizes semantic intent, metaphor preservation, and terminology consistency, providing a cognitive diagnostic lens into MT.

However, the limitations discussed here delineate a key boundary: while LLMs excel in surface-level alignment and associative generalization, they remain far from achieving true literary understanding. Bridging this gap between memorization and meaning and between matching and comprehension represents a central challenge in the long-term pursuit of artificial general intelligence.

7 Literature review: BT as an evaluation methodology

BT, also referred to as forward-and-back or RTT, has long served as both a linguistic tool and a computational mechanism. This section reviews its evolution across linguistic studies, MT evaluation, and LLM-centered machine learning.

7.1 Forward and backward translation in linguistic studies

In bilingual psycholinguistics, BT has been used to examine lexical access and semantic retrieval. Degroot et al. (1994) and Kroll and Stewart (1994) found that both forward ($L_1 \rightarrow L_2$) and backward ($L_2 \rightarrow L_1 y$) translation are influenced by familiarity, frequency, and conceptual mediation, with backward translation sometimes requiring more effort due to increased semantic processing.

La Heij et al. (1996) challenged the asymmetry hypothesis, showing that semantic context can affect backward translation more strongly than forward translation. Salamoura and Williams (1999) further concluded that both translation directions engage in shared semantic retrieval systems.

In applied linguistics, Waijanya and Mingkhwan (2014) used BT to assess Thai poetry translations, finding that BLEU and METEOR metrics applied to BT were helpful in validating machine-generated poetic outputs. Recent works, such as He et al. (2023) and Zhang Y et al. (2025), have explored directionality effects in Chinese–English translation tasks using semantic blocking paradigms, confirming persistent asymmetries in cognitive processing.

7.2 RTT as a bilingual evaluation tool

RTT has been a longstanding but controversial MT evaluation method. Somers (2005) and Aiken and Park (2010) argued that RTT often fails to reflect fine-grained translation quality, though it may capture system-level tendencies in the absence of reference translations.

Zhuo et al. (2023) reevaluated RTT in the context of neural MT, showing improved correlation with quality metrics, especially in low-resource and unsupervised settings. More recently, Yung et al. (2025) applied RTT as a prompt sanitization technique in LLMs, leveraging round-trip through multiple languages to increase robustness against

adversarial inputs.

7.3 BT in machine learning and neural MT

BT has played a central role in neural MT since Sennrich et al. (2016) demonstrated its effectiveness in augmenting target-side monolingual data to improve BLEU scores. Hoang et al. (2018) introduced iterative BT to refine synthetic parallel corpora, and Edunov et al. (2018) found that sampling-based BT outperformed beam search in model training.

Artetxe et al. (2018) extended BT to fully unsupervised MT, narrowing gaps with supervised systems. Applications of BT have since expanded beyond MT to include sentiment preservation (Troiano et al., 2020), short-text classification (Marivate and Sefara, 2020), ABSA (Taheri et al., 2025), psychological validation (Glidden-Tracey and Greenwood, 1997), dictionary construction (Chan, 2004), and low-resource translation (e.g., Chinese–Vietnamese) (Li HZ et al., 2020).

Recent works include DUAL-REFLECT (Chen et al., 2024b), a dual learning framework using BT to enhance semantic alignment. Brimacombe and Zhou (2023) introduced quick back-translation (QBT), dramatically improving training efficiency without degrading quality. LLM-centric works, such as Chung and Kim (2025) and Gain et al. (2025), highlight BT as a key component of modern LLM-based translation pipelines.

7.4 Chinese NLP and idiomatic semantics

Preserving metaphorical and poetic meaning in Chinese NLP is uniquely challenging. Seminal works in metaphor translation (Schäffner, 2004; Baker, 2018) stress the difficulty of conceptual mapping between culturally distinct languages. In Chinese–English literary translation, fidelity to rhythm and poetic imagery, as studied in Chen et al. (2024a), remains a high standard.

Research on LLM memory mechanisms offers a critical context for understanding verbatim BT. Brown et al. (2020) proposed that memory is encoded parametrically in model weights. Shan et al. (2025) categorized cognitive memory types in LLMs; Modarressi et al. (2024) introduced read–write modules for long-term retention; Yang HK et al. (2024) externalized memory to reduce model size; Zhang ZY et al. (2024) provided a comprehensive review of memory in LLM agents. Additionally, Tao et al. (2024) introduced CUDRT to benchmark reasoning consistency and detect overfitting.

In idiom processing, recent models have improved performance via masking and compositional modeling (Wei et al., 2019; Qiang et al., 2023; Wu et al., 2024). Our LLM-BT framework builds on this by quantifying semantic drift and surface illusion in idiom- and metaphor-rich corpora.

8 Conclusions and future work

This study demonstrates that LLMs can surpass traditional MT tools in Chinese BT, particularly in scientific domains, yet remain limited in capturing metaphorical nuance and cultural resonance. Using our modular NLPMetrics pipeline, we show that higher BLEU or CHRF scores often coincide with semantic flattening—a phenomenon we define as the paradox of poetic intent.

BT here serves not merely as a diagnostic tool, but as a lens to reveal trade-offs between literal fidelity and interpretive depth. Our findings highlight that current LLMs approximate cultural translation largely through memorization or probabilistic shortcuts, falling short in idiomatic and metaphor-rich contexts. This reframes BT as a multidimensional, behavior-aware evaluation strategy with implications for CNLP, cross-lingual studies, and LLM explainability.

Our main contributions lie in conceptualizing the paradox of poetic intent, proposing a reproducible LLM-BT framework with multidimensional evaluation metrics, and offering empirical insights into divergent model behaviors—ranging from verbatim reproduction to creative paraphrasing—supported by new theoretical constructs such as verbatim back-translation, poetic drift, and quasi-self-awareness.

Looking forward, future work will expand the corpus to low-resource poetic and hybrid texts, integrate human preference ratings alongside metrics, and explore prompt engineering and hybrid retrieval–generation strategies to better preserve intent across translation cycles.

As LLMs move into culturally sensitive applications, ranging from translation to education, ensuring that cross-linguistic meaning is preserved with accountability becomes not only a technical

challenge but also a cultural imperative.

Acknowledgments

We extend our sincere gratitude to the anonymous reviewers for their constructive feedback and encouragement, which greatly improved the quality of this work. We acknowledge the support of multiple LLMs and other platforms, including ChatGPT, Claude, DeepSeek, Gemini, Grok, and Mistral, as well as Google, Baidu, Sogou, and CNKI. In particular, ChatGPT and Grok were employed to help standardize and refine the English writing.

Contributors

Li WEIGANG conceptualized the research, developed the LLM-BT framework, and drafted the paper. Pedro Carvalho BROM implemented the NLPMetrics system, conducted the experiments, and authored the text for Section 5. Both authors contributed to the revision and finalization of the paper.

Conflict of interest

Li WEIGANG is an editorial board member of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. Both authors declare that they have no conflict of interest.

Data availability

The CNKI-CHE-89 corpus is publicly available at https://github.com/pcbrom/bt-conference. The other data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aiken M, Park M, 2010. The efficacy of round-trip translation for MT evaluation. *Transl J*, 14(1):1-10.

Arruda-Vasconcelos R, Louzada LM, Feres M, et al., 2021. Investigation of microbial profile, levels of endotoxin and lipoteichoic acid in teeth with symptomatic irreversible pulpitis: a clinical study. *Int Endod J*, 54(1):46-64. https://doi.org/10.1111/iej.13402

Artetxe M, Labaka G, Agirre E, 2018. Unsupervised statistical machine translation. Proc Conf on Empirical Methods in Natural Language Processing, p.3632-3642. https://doi.org/10.18653/v1/D18-1399

Bahji A, Acion L, Laslett AM, et al., 2023. Exclusion of the non-English-speaking world from the scientific literature: recommendations for change for addiction journals and publishers. *Nord Stud Alcohol Drugs*, 40(1):6-13. https://doi.org/10.1177/14550725221102227

Baker M, 2018. In Other Words: a Coursebook on Translation (3rd Ed.). Routledge, London, UK. https://doi.org/10.4324/9781315619187

Berman A, Venuti L, 2021. Translation and the Trials of the Foreign. In: Venuti L (Ed.), The Translation Studies Reader (4th Ed.). Routledge, London, UK, p.247-260.

Brimacombe B, Zhou JW, 2023. Quick back-translation for unsupervised machine translation. Proc Findings of the Association for Computational Linguistics, p.8521-8534. https://doi.org/10.18653/v1/2023.findings-emnlp.571

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, Article 159.

Cao Z, Lu J, Cui S, et al., 2020. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. *Patt Recogn*, 107:107488. https://doi.org/10.1016/j.patcog.2020.107488

Chan SW, 2004. A Dictionary of Translation Technology. The Chinese University of Hong Kong Press, Hong Kong, China (in Chinese).

Chen AD, Lou LZ, Chen KH, et al., 2024a. Benchmarking LLMs for translating classical Chinese poetry: evaluating adequacy, fluency, and elegance. https://arxiv.org/abs/2408.09945

Chen AD, Lou LZ, Chen KH, et al., 2024b. DUAL-REFLECT: enhancing large language models for reflective translation through dual learning feedback mechanisms. Proc 62nd Annual Meeting of the Association for Computational Linguistics, p.693-704. https://doi.org/10.18653/v1/2024.acl-short.64

Chung JB, Kim T, 2025. Leveraging large language models for enhanced back-translation: techniques and applications. *IEEE Access*, 13:61322-61328. https://doi.org/10.1109/ACCESS.2025.3557014

Degroot AMB, Dannenburg L, Vanhell JG, 1994. Forward and backward word translation by bilinguals. *J Mem Lang*, 33(5):600-629. https://doi.org/10.1006/jmla.1994.1029

Demšar J, 2006. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*, 7:1-30.

Ding Y, Teng F, Zhang P, et al., 2021. Research on text information mining technology of substation inspection based on improved Jieba. Proc Int Conf on Wireless Communications and Smart Grid, p.561-564. https://doi.org/10.1109/ICWCSG53609.2021.00119

Eberhard DM, Simons GF, Fennig CD, 2022. Ethnologue: Languages of the World (25th Ed.). SIL International, Dallas, USA.

Edunov S, Ott M, Auli M, et al., 2018. Understanding back-translation at scale. Proc Conf on Empirical Methods in Natural Language Processing, p.489-500. https://doi.org/10.18653/v1/D18-1045

Feng SB, 2024. Discussion on applied chemical metrology calculation based on computer technology—comment on "applied chemistry". *Chin J Appl Chem*, 41(12):1829-1830 (in Chinese).

Gain B, Bandyopadhyay D, Ekbal A, 2025. Bridging the linguistic divide: a survey on leveraging large language models for machine translation. https://arxiv.org/abs/2504.01919

Glidden-Tracey C, Greenwood AK, 1997. A validation study of the Spanish self-directed search using back-translation procedures. *J Career Assess*, 5(1):105-113. https://doi.org/10.1177/106907279700500107

He J, 2019. The Chinese nomenclature for the heterocyclic compounds since 1932. *Chemistry*, 82(4):373-378 (in Chinese). https://doi.org/10.14159/j.cnki.0441-3776.2019.04.013

He YJ, Hou LP, Lang LY, 2023. L2 Acquisition from Perspectives of Professional Translation and Interpreting. In: Maqbool T, Lang LY, Meltzoff K (Eds.), Second Language Acquisition—Learning Theories and Recent Approaches. IntechOpen, p.85. https://doi.org/10.5772/intechopen.109126

Hoang VCD, Koehn P, Haffari G, et al., 2018. Iterative back-translation for neural machine translation. Proc 2nd Workshop on Neural Machine Translation and Generation, p.18-24. https://doi.org/10.18653/v1/W18-2703

Jiang JY, Liu C, 2020. Comparison and analysis of research and development expenditure and publication output of major countries (regions) in the world. *Bull Natl Nat Sci Found China*, 34(3):367-372 (in Chinese). https://doi.org/10.16262/j.cnki.1000-8217.2020.03.024

Klaudy K, 1996. Back-translation as a tool for detecting explicitation strategies in translation. In: Klaudy K, Lambert J, Sohár A (Eds.), Translation Studies in Hungary. Scholastica, Budapest, Hungary, p.99-114.

Kroll JF, Stewart E, 1994. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *J Mem Lang*, 33(2):149-174. https://doi.org/10.1006/jmla.1994.1008

La Heij W, Hooglander A, Kerling R, et al., 1996. Nonverbal context effects in forward and backward word translation: evidence for concept mediation. *J Mem Lang*, 35(5):648-665. https://doi.org/10.1006/jmla.1996.0034

Li HZ, Sha J, Shi C, 2020. Revisiting back-translation for low-resource machine translation between Chinese and Vietnamese. *IEEE Access*, 8:119931-119939. https://doi.org/10.1109/ACCESS.2020.3006129

Li YH, Huang HY, Wang BJ, et al., 2025. DRM-Spell: dynamically reweighting multimodality for Chinese spelling correction. *Front Inform Technol Electron Eng*, 26(3):354-366. https://doi.org/10.1631/FITEE.2300816

Ling L, Lin CH, Lin TY, et al., 2025. Scenethesis: a language and vision agentic framework for 3D scene generation. https://arxiv.org/abs/2505.02836

Liu Y, Liang NY, 1986. Hanyu chuli de jichu gongcheng—xiandai hanyu cifrequency tongji. *J Chin Inform Process*, 1(1):17-25 (in Chinese).

Luo RX, Xu JJ, Zhang Y, et al., 2019. PKUSEG: a toolkit for multi-domain Chinese word segmentation. https://arxiv.org/abs/1906.11455

Ma WW, 2024. Effect of amino oligosaccharides combined with chemical fungicides on the control of downy mildew in Chinese cabbage. *Contemp Farm Mach*, (12):68-69 (in Chinese). https://doi.org/10.3969/J.ISSN.1673-632X.2024.12.026

Marivate V, Sefara T, 2020. Improving short text classification through global augmentation methods. Proc CICLing, p.234-246.

Modarressi A, Köksal A, Imani A, et al., 2024. MemLLM: finetuning LLMs to use an explicit read-write memory. https://arxiv.org/abs/2404.11672

Nam GE, Park YG, 2015. Re: Inhibition of peripheral FAAH depresses activities of bladder mechanosensitive nerve fibers of the rat. *J Urol*, 193(2):738-739. https://doi.org/10.1016/j.juro.2014.08.117

Nida EA, 1964. Toward a Science of Translating: with Special Reference to Principles and Procedures Involved in Bible Translating. Brill Archive, Leiden, the Netherlands.

Ozolins U, Hale S, Cheng X, et al., 2020. Translation and back-translation methodology in health research—a critique. *Expert Rev Pharmacoecon Outcomes Res*, 20(1):69-77. https://doi.org/10.1080/14737167.2020.1734453

Papineni K, Roukos S, Ward T, et al., 2002. BLEU: a method for automatic evaluation of machine translation. Proc 40th Annual Meeting of the Association for Computational Linguistics, p.311-318. https://doi.org/10.3115/1073083.1073135

Qiang JP, Li Y, Zhang CW, et al., 2023. Chinese idiom paraphrasing. *Trans Assoc Comput Ling*, 11:740-754. https://doi.org/10.1162/tacl-a-00572

Salamoura A, Williams JN, 1999. Backward word translation: lexical vs. conceptual mediation or "concept activation vs. word retrieval"? *RCEAL Work Pap Engl Appl Ling*, 6:31-56.

Schäffner C, 2004. Metaphor and translation: some implications of a cognitive approach. *J Pragmat*, 36(7):1253-1269. https://doi.org/10.1016/j.pragma.2003.10.012

Sennrich R, Haddow B, Birch A, 2016. Improving neural machine translation models with monolingual data. Proc 54th Annual Meeting of the Association for Computational Linguistics, p.86-96. https://doi.org/10.18653/v1/P16-1009

Shan LL, Luo SX, Zhu ZZ, et al., 2025. Cognitive memory in large language models. https://arxiv.org/abs/2504.02441

Sheldon MR, Fillyaw MJ, Thompson WD, 1996. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physioth Res Int*, 1(4):221-228. https://doi.org/10.1002/pri.66

Somers H, 2005. Round-trip translation: what is it good for? Proc Australasian Language Technology Workshop, p.127-133.

Sun ZJ, Li XY, Sun XF, et al., 2021. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.2065-2075. https://doi.org/10.18653/v1/2021.acl-long.161

Taheri A, Zamanifar A, Farhadi A, 2025. Enhancing aspect-based sentiment analysis using data augmentation based on back-translation. *Int J Data Sci Anal*, 19(3):491-516. https://doi.org/10.1007/s41060-024-00622-w

Tao Z, Che YF, Xi DH, et al., 2024. Towards reliable detection of LLM-generated texts: a comprehensive evaluation framework with CUDRT. https://arxiv.org/abs/2406.09056

Toral A, Way A, 2018. What level of quality can neural machine translation attain on literary text? In: Moorkens J, Castilho S, Gaspari F, et al. (Eds.), Translation Quality Assessment: from Principles to Practice. Springer,

Cham, p.263-287. https://doi.org/10.1007/978-3-319-91241-7_12

Troiano E, Klinger R, Padó S, 2020. Lost in back-translation: emotion preservation in neural machine translation. Proc 28th Int Conf on Computational Linguistics, p.4340-4354. https://doi.org/10.18653/v1/2020.coling-main.384

Tu QY, Li CB, 2017. A review on textless back translation of China-themed works written in English. *Stud Lit Lang*, 14(1):1-7. https://doi.org/10.3968/9177

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Waijanya S, Mingkhwan A, 2014. Thai poetry translation to English with backward translation evaluation. Proc 9th Int Conf on Digital Information Management, p.248-253. https://doi.org/10.1109/ICDIM.2014.6991425

Wang HY, 2009. Introduction to Literary Translation Criticism. China Renmin University Press, Beijing, China (in Chinese).

Wei JQ, Ren XZ, Li XG, et al., 2019. NEZHA: neural contextualized representation for Chinese language understanding. https://arxiv.org/abs/1909.00204

Weigang L, Brom PC, 2025. LLM-BT-terms: back-translation as a framework for terminology standardization and dynamic semantic embedding. https://arxiv.org/abs/2506.08174

Weigang L, Marinho MC, Li DL, et al., 2024. Six-writings multimodal processing with pictophonetic coding to enhance Chinese language models. *Front Inform Technol Electron Eng*, 25(1):84-105. https://doi.org/10.1631/FITEE.2300384

Weigang L, Brom PC, Ramos RM, 2025a. Quantitative evaluation of translation quality and computational efficiency in semantic vs. phonetic strategies for Chinese scientific terms. Proc 29th Int Conf on Asian Language Processing, p.43-48.

Weigang L, Ramos RM, Brom PC, et al., 2025b. Threshold study for Hanzi image recognition: defining character and component limits in Chinese, Japanese, and Korean script processing. *Int J Asian Lang Process*, 35(1):2450011. https://doi.org/10.1142/S2717554524500115

Wong KF, Li WJ, Xu RF, et al., 2010. Introduction to Chinese Natural Language Processing. Springer, Cham, Germany.

Wu MM, Hu YX, Zhang YC, et al., 2024. Mitigating idiom inconsistency: a multi-semantic contrastive learning method for Chinese idiom reading comprehension. Proc 38th AAAI Conf on Artificial Intelligence, p.19243-19251. https://doi.org/10.1609/aaai.v38i17.29893

Yang HK, Lin ZH, Wang WJ, et al., 2024. Memory^3: language modeling with explicit memory. https://arxiv.org/abs/2407.01178

Yang YX, Ren GC, 2020. HanLP-based technology function matrix construction on Chinese process patents. *Int J Mob Comput Multim Commun*, 11(3):48-64. https://doi.org/10.4018/IJMCMC.2020070104

Yousufi S, Erdely F, 2024. Enhancing nonparametric tests: insights for computational intelligence and data mining. *Res Acad Innov Data Anal*, 1(3):214-226. https://doi.org/10.69725/raida.v1i3.168

Yung C, Dolatabadi HM, Erfani S, et al., 2025. Round trip translation defence against large language model jailbreaking attacks. Proc Workshops, ADUR, FairPC, GLFM, PM4B and RAFDA Trends and Applications in Knowledge Discovery and Data Mining, p.286-297. https://doi.org/10.1007/978-981-96-8197-6-21

Zhang XE, 2021. A study of cultural context in Chinese–English translation. *Reg-Educ Res Rev*, 3(2):11-14. https://doi.org/10.32629/rerr.v3i2.303

Zhang Y, Shuai YH, Xiao CY, et al., 2025. The structure of the bilingual lexicon: evidence from a semantic blocked word translation task with Chinese–English bilinguals. *Second Lang Res*, early access. https://doi.org/10.1177/02676583251313997

Zhang ZY, Bo XH, Ma C, et al., 2024. A survey on the memory mechanism of large language model based agents. https://arxiv.org/abs/2404.13501

Zhao SQ, Zhou YH, Ren YP, et al., 2025. Fùxì: a benchmark for evaluating language models on ancient Chinese. https://arxiv.org/abs/2503.15837

Zhong CZ, Cheng F, Liu QY, et al., 2024. Beyond English-centric LLMs: what language do multilingual language models think in? https://arxiv.org/abs/2408.10811

Zhou Z, 2014. The six principles of Chinese writing and its application to design as design idea. *Stud Lit Lang*, 8(3):84-88. https://doi.org/10.3968/4968

Zhu SL, Pan LY, Jian D, et al., 2025. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. *Inform Process Manag*, 62(3):104078. https://doi.org/10.1016/j.ipm.2025.104078

Zhuo TY, Xu QK, He XL, et al., 2023. Rethinking round-trip translation for machine translation evaluation. Proc Findings of the Association for Computational Linguistics, p.319-337. https://doi.org/10.18653/v1/2023.findings-acl.22

List of supplementary materials