Li WEIGANG, Pedro Carvalho BROM, 2025. Paradox of poetic intent in back-translation: evaluating the quality of large language models in Chinese translation. *Frontiers of Information Technology & Electronic Engineering*, 26(11):2176-2203. https://doi.org/10.1631/FITEE.2500298

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

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

Corresponding author: Li WEIGANG
E-mail: weigang@unb.br
ORCID: https://orcid.org/0000-0003-1826-1850

Motivation

1. Spoken by more than 1.6 billion people worldwide, 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.

2. Chinese⇔English translation, in particular, continues to pose a major challenge for multilingual AI systems. 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, we argue that large language model (LLM) behaviors must now be understood in higher dimensions, where semantic recoverability, cultural alignment, and terminology consistency interact.

Main idea

1. 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. This observation extends beyond traditional fidelity–fluency trade-offs and introduces a high-dimensional framework for evaluating translation quality.

2. Verbatim BTs further reveal model-specific behaviors. GPT-4.5 and DeepSeek V3, often regenerate the original input even without prompting. Building on memory-related research, we describe this as a form of quasi-self-awareness, an emergent, non-memorized, but stable recovery behavior across model boundaries and time delays.

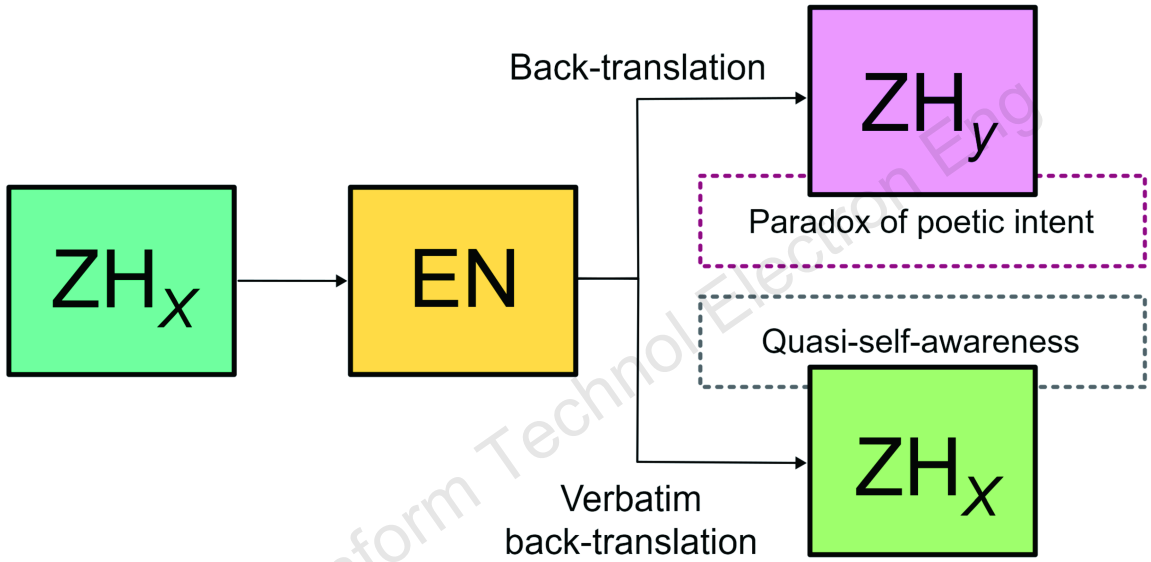
Corpora

1. Xue Dejiong's (薛德炯) dilemma: 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. 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.

2. CNKI abstract corpus of the scientific literature: 开展计量计算，能够实践应用化学计算剂量的提质增效. We selected 295 Chinese-language scientific papers to ensure statistical robustness, and all metadata have been released via GitHub (https://github.com/pcbrom/bt-conference) for reproducibility.

3. Dao Lang's Hua Yao lyrics (刀郎《花妖》歌词): 君去时褐衣红，小奴家腰上黄，寻差了罗盘经，错投在泉亭. 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, inter-weaving poetic sentiment, historical geography, Buddhist cosmology, and nonlinear temporality.

Framework: LLM-BT



Conceptual diagram: the paradox of poetic intent in back-translation (BT) vs. emergent quasi-self-awareness in LLM verbatim BT, where ZH_x is the original Chinese text, EN is the translated English text, and ZH_y is the back-translated Chinese text.

Method: LLM-BT

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}:

$$\mathrm{BT}(T)=\mathrm{Trans}_{L2\rightarrow L1}(\mathrm{Trans}_{L1\rightarrow L2}(T)).$$

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

Method: LLMs test metrics

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

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

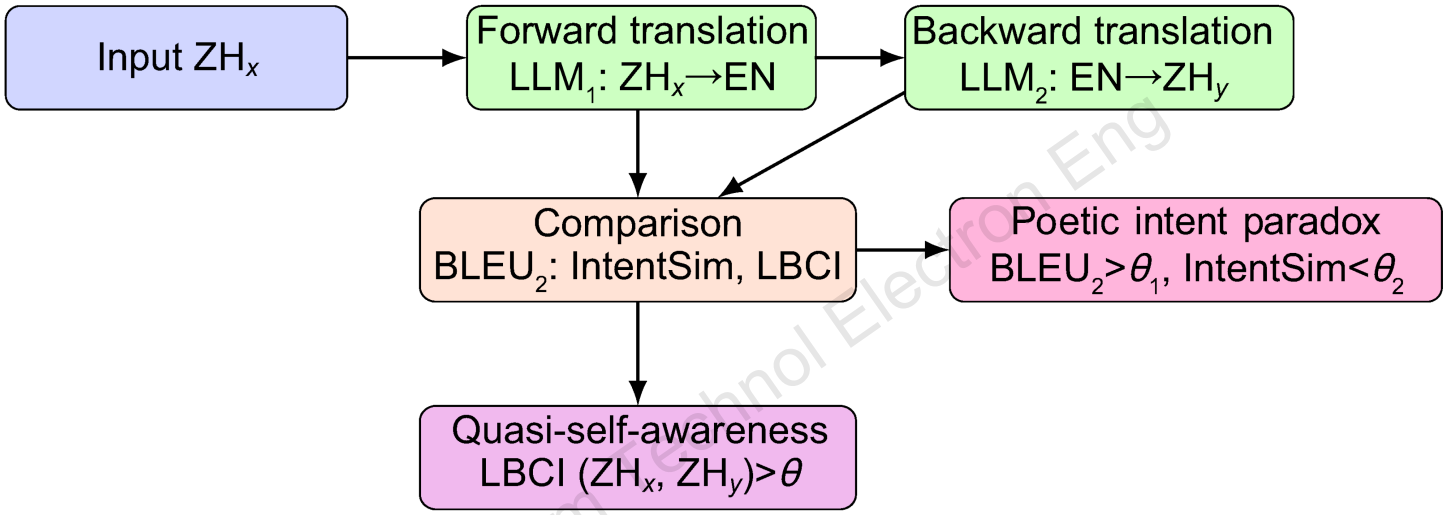
where w_n is weight assigned to the n-gram, $p_{n\text{-gram}}$ is precision of n-gram matches after Jieba-based segmentation, and $\mathrm{BLEU_J}$ thus reflects word frequency-aware n-gram alignment, not just token overlap.

Two BLEU variants are compared:

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

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.

Paradox & Quasi-self-awareness



Diagnosing emergent behaviors in LLM-BT. This diagram illustrates the analytical logic of our framework: an RTT process ($ZH_x \rightarrow EN \rightarrow ZH_y$) is evaluated using BLEU2, IntentSim, and index to detect this behavior (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.

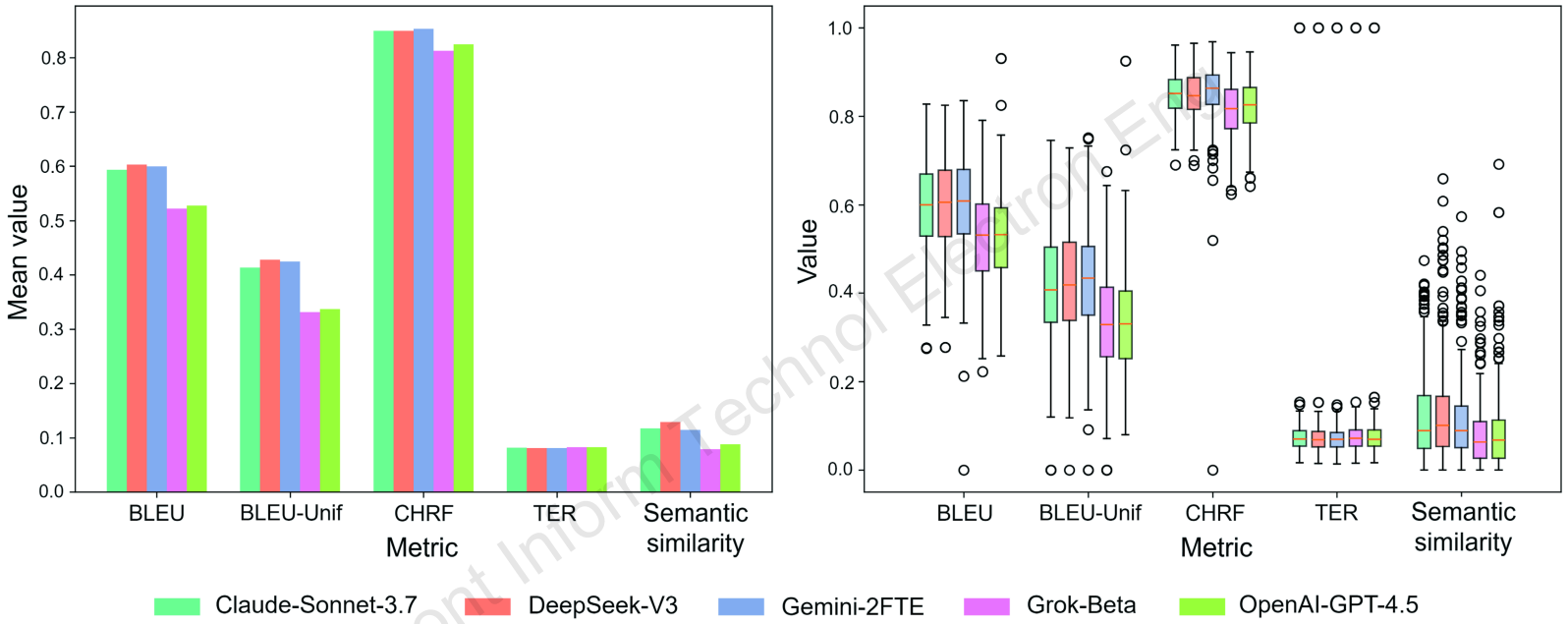
Results: Dao Lang's *Hua Yao*

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

Cross-corpus evaluation of LLM-BT quality



Comparison of translation metrics across models (BLEU: bilingual evaluation understudy; CHRF: character F-score; TER: translation edit rate). To ensure statistical significance in the evaluation of Chinese BT quality, we use a multi-sample corpus, CNKI-CHE-89, which includes 89 abstracts randomly selected from a broader set of 295 chemistry-related abstracts extracted from CNKI.

Conclusions

1. 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.

2. 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.

Li Weingang earned his Doctor of Science degree from the Aeronautics Institute of Technology (ITA), Brazil, in 1994, and pursued postdoctoral research at the University of Calgary, Canada, from 2001 to 2002. He is currently a Full Professor and vice head at the Department of Computer Science (CIC) of the University of Brasilia (UnB). He also coordinates the Laboratory of Computational Modeling and Intelligence for Transport (TransLab.unb.br). His research interests span AI, machine learning, computational modeling for air traffic management, and Chinese natural language processing.

Pedro Carvalho BROM graduated in Mathematics from the State University of Minas Gerais (Feb/01-Dec/03), Bachelor in Statistics from the University of Brasilia (Feb/14-Dec/20), Specialist in Mathematics and Statistics from the Federal University of Lavras (Jul/04-Dec/05), Data Science Specialist from Johns Hopkins University (Nov/16-Feb/17), and Master in Statistics (Jul/21-Jan/23). He is a professor and researcher at the Federal Institute of Brasilia, working in applied mathematics, statistics, modeling, and programming in R language. He is currently working on the Laboratory of Intelligent Processing and Recognition of Analytical Systems and Methods (IFB).