Liaozhai through the Looking-Glass[1]: On Paratextual Explicitation[2] of Culture-Bound Terms in Machine Translation

Sherrie Shen, Weixuan Wang, Alexandra Birch

Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh

Correspondence: sherrie.shen@ed.ac.uk

Abstract

The faithful transfer of contextually-embedded meaning continues to challenge contemporary machine translation (MT), particularly in the rendering of culture-bound terms—expressions or concepts rooted in specific languages or cultures, resisting direct linguistic transfer. Existing computational approaches to explicitating these terms have focused exclusively on in-text solutions, overlooking paratextual apparatus in the footnotes and endnotes employed by professional translators. In this paper, we formalize Genette's (1987) theory of paratexts from literary and translation studies to introduce the task of paratextual explicitation for MT. We construct a dataset of 560 expert-aligned paratexts from four English translations of the classical Chinese short story collection *Liaozhai* and evaluate LLMs with and without reasoning traces on choice and content of explicitation. Experiments across intrinsic prompting and agentic retrieval methods establish the difficulty of this task, with human evaluation showing that LLM-generated paratexts improve audience comprehension, though remain considerably less effective than translator-authored ones. Beyond model performance, statistical analysis reveals that even professional translators vary widely in their use of paratexts, suggesting that cultural mediation is inherently open-ended rather than prescriptive. Our findings demonstrate the potential of paratextual explicitation in advancing MT beyond linguistic equivalence, with promising extensions to monolingual explanation and personalized adaptation.

1 Introduction

The intricacies in faithfully rendering meaning across languages form a cornerstone of translation, particularly when dealing with literary devices and contextual concepts. These linguistic elements are deeply embedded within the cultural fabric of their origin, carrying connotations that can be difficult to convey into another language. Consequently, the act of translation transcends mere word-for-word substitution, requiring a nuanced understanding of the subtle interplay between language and meaning.

This challenge becomes more pronounced in the field of machine translation (MT). While modern MT systems are capable of producing fluent translations at the surface level (Hassan et al., 2018) and are increasingly proficient at translating idiomatic and metaphorical expressions (Dankers et al., 2022; Karakanta et al., 2025), they still struggle with contextual content (Moghe et al., 2025). This limitation becomes especially apparent in the handling of *culture-bound terms*: expressions or concepts closely associated with a particular language or culture, where the more culturally bound an item is, the greater the difficulty in translating it (Newmark, 1988; Aixelá, 1996).

Translation studies has long recognized that effective cross-cultural communication requires distinct strategies tailored to a wide range of expressions, from calques and loanwords to idioms and culture-bound terms (Nida, 1964; Bassnett, 1980). Consider, for instance, the Chinese idiom '锦上添花' (*jǐn shàng tiān huā*; lit. add flowers to brocade), which means to unnecessarily embellish something already beautiful. The phrase can be translated into English as 'gild the lily' since both cultures share this concept, albeit in different linguistic forms.

By contrast, culture-bound terms come with implicit nuances that often resist translation entirely. The Chinese term 江湖 (*jiāng hú*; lit. rivers and lakes) exemplifies this challenge:

[1] An allusion to footnote 44 of Giles's (1880) translation of *Liaozhai*:

> Which will doubtless remind the reader of *Alice through the Looking-glass, and what she saw there.*

[2] In his seminal work *Seuils* (1987), Genette defines paratexts as 'thresholds' that surround and extend a text, including prefaces, glossaries, and notes like this. In translation studies, paratextual explicitation refers to clarifications and commentary appearing in such peripheral devices, rather than within the translation itself.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34400–34416
November 4-9, 2025 ©2025 Association for Computational Linguistics

> *A semi-mythical, often romanticized social world of martial artists, wanderers, and outlaws who—despite existing outside of state authority—follow their own codes of honor and justice.*

The literal translation conveys little meaningful information to a non-Chinese speaking audience. Unlike 'gild the lily', no English equivalents exist for this concept, and including a full explanation within the translation would excessively disrupt narrative flow. In such cases, explicitation through external commentary offers a more elegant alternative.

Genette (1987) provides a formal framework for such interventions in his theory of *paratexts*: ancillary materials that accompany a primary text, mediating between the author, publisher, and reader. These liminal devices include titles, forewords, epigraphs, footnotes, glossaries, appendices, and other supplementary elements that, while seemingly marginal, exert a substantial influence on reader comprehension. Yet despite their pivotal role in human translation, paratexts remain entirely absent from contemporary MT—a gap that motivates our subsequent computational exploration.

Contribution 1: Task. In this paper, we take an initial step towards paratextual explicitation in literary MT by examining how LLMs can generate paratexts for culture-bound terms. Building upon seminal theories in literary, translation, and cultural studies, we present what is to the best of our knowledge the first formulation of this problem within computational linguistics.

Contribution 2: Dataset. As this task has yet to be formally studied, we introduce a dataset[3] comprising the original text and paratextual material from five published translations of the classical Chinese work *Liaozhai zhiyi*. Our dataset features 560 manually-aligned paratexts across four of the translations, with expert annotations linking each paratext to the classical Chinese term it explicitates.

Contribution 3: Evaluation. We conduct both automatic and human evaluation of LLM-generated paratexts on our dataset, assessing how prompting and agentic strategies informed by translation theory affect performance. While these methods yield improvements of up to 5 percentage points, current LLMs continue to struggle on this task, with even the strongest model identifying fewer than 24% of all translator-marked culture-bound terms.

Contribution 4: Analysis. Finally, we find that even the professional translators of our dataset systematically disagree on both choice and content of paratextual explicitation. Computational analysis reveals the inherently subjective nature of this task and provides a human upper-bound for interpreting system results.

We hope these findings will inspire further research on interdisciplinary approaches to MT and paratextual explicitation.

2 Theoretical Foundations

The translation of meaning across linguistic and cultural boundaries has been a subject of theoretical inquiry since antiquity. Classical scholars distinguished between *verbum e verbo* (lit. word-for-word) and *sensum de sensu* (lit. sense-for-sense) approaches to translation (Cicero, -46; Horace, -19; St Jerome, 395), establishing a tension that would persist through centuries of scholarly discourse.

Contemporary theories have reformulated this dialectic through frameworks for equivalence and equivalent effect, most influentially articulated in Nida (1964); Nida and Taber (1969), and emerging alongside advances in structural linguistics (Chomsky, 1957; Catford, 1965), comparative stylistics (Vinay and Darbelnet, 1958), and translational typology (Newmark, 1981; Koller, 1995, *inter alia*). Beyond lexical and syntactic fidelity, however, theorists increasingly recognized the importance of contextual and pragmatic factors, paving the way for later reconceptualizations of translation as a cultural practice.

2.1 From Linguistic to Cultural Transfer

In the 1990s, translation studies experienced the 'cultural turn', a shift that aligned the field with broader intellectual movements across the humanities (Bassnett and Lefevere, 1990). Drawing upon poststructuralism, postcolonial theory, and cultural studies, scholars began to question traditional assumptions of textual authority, cultural negotiation, and the politics of representation, with translators no longer seen as invisible agents (Venuti, 1995) but active mediators (Lee, 2013).

Central to the cultural turn was recognition of translation as a form of rewriting, with translators inevitably reshaping texts to abide by literary conventions and ideological norms. Historical examples illustrate the ethical complexities: early Victorian and Edwardian translators systematically soft-

[3] https://github.com/sherrieshen/liaozhai

ened explicit themes of classical Greek comedies like *Lysistrata* (Lefevere, 1992); Edward Fitzgerald's celebrated rendering of Omar Khayyam's *Rubáiyát* reimagined the Persian source according to his own aesthetic ideals (Davis, 2000); and colonial-era translations frequently marginalized indigenous voices through assimilation into Western interpretive frameworks (Spivak, 1978).

These cases underscore translation as a site of cultural negotiation, where adaptation to a target context can easily blur into distortion of the source. The tension between mediation and manipulation has since motivated subsequent theoretical models, particularly those addressing the role of translation as both a cultural bridge and an ideological filter.

2.2 Theories in Translation Studies

Two theoretical frameworks emerging from the cultural turn prove particularly relevant to understanding strategies for adaptation. Polysystem theory, for one, challenges the traditional notion of texts as an isolated entity by conceptualizing translated literature as part of a dynamic system of cultural products constantly interacting within target societies (Even-Zohar, 1978; Toury, 1995).

Central to polysystem theory is the idea that the position of translated literature within a given literary system fundamentally determines translational strategies (Munday, 2016). When occupying a primary role within the polysystem, translators tend to adopt experimental approaches that challenge existing conventions. Conversely, when relegated to a peripheral role, translations typically conform to established target-culture expectations (Tymoczko, 2010).

Complementing this perspective, skopos theory foregrounds the *skopos*—or intended function—of the translated text within its target context (Reiß and Vermeer, 1984). This communicative approach acknowledges that the purpose of the translated text may not always match that of the original and questions the traditional view of a translator's task in producing 'equivalent effect' between two languages (Munday, 2016).

Together, these frameworks illuminate how translation strategies are conditioned by both the literary systems they enter and the functional purposes they fulfill. This theoretical foundation provides a conceptual basis for understanding translator use of paratexts and analyzing translations within their broader cultural contexts, as we explore in the following sections.

3 Dataset

Constructing a dataset of paratextual explicitations requires identifying a source text that can meaningfully challenge computational models. The text should demand deep cultural and contextual understanding, so that success depends on more than surface-level translation. It should exist within a long history of translation, offering diverse human-authored paratexts against which model outputs can be evaluated. Finally, it ideally takes the form of short narratives, thereby reducing confounds such as the long-range dependencies and memory limitations of current LLMs. Guided by these principles, we choose the following dataset for evaluating culture-bound paratextual explicitation.

3.1 Source: *Liaozhai zhiyi*

We select the classical Chinese short story collection《聊斋志异》(*Liáozhāi zhìyì*, or *Liaozhai* for short) as the source text for this task. Composed by Pu Songling during the Qing dynasty (1766), this canonical collection of 494 stories epitomizes the *zhiguai* (supernatural) literary genre.

Liaozhai interweaves fantastical storytelling on encounters with fox spirits and ghosts through social satire and philosophical reflection, frequently using idiomatic expressions, historical allusions, and culture-bound terms (Yi et al., 2025). It holds remarkable status in Chinese literature, and since its introduction to the West, has been the subject of over forty English translations (Jin, 2021). This yields a rich body of paratextual commentary that reflects the diverse backgrounds and interpretive styles of its various translators across the past three centuries.

3.2 Translation Selection and Processing

From the wide corpus of *Liaozhai* translations, we filter for English editions that (1) include at least fifty stories; (2) contain substantial paratextual material; (3) are publicly available; and (4) were published at least five years prior to this study.

Of the five remaining translations, we notably decide to exclude Giles (1880) from evaluation, though still release it to support future research. While Giles established the primary point of entry for *Liaozhai* in the Western literary polysystem, he diverges substantially from the source material and makes culturally-outdated commentary. Due to this and reasons detailed in Appendix A, we deem his translation unsuitable for comparison.

Translator(s)	Year	Profession(s)	Direction	Stories	Paratexts	Tokens	Types
Giles	1880	diplomat, professor	FL→NL	164	657	28,076	footnotes, appendices
Lu et al.	1982	professors	NL→FL	59	102	2,210	footnotes
Mair and Mair	1989	sinologists	FL→NL	43	95	1,695	footnotes
Minford	2006	professor, sinologist	FL→NL	86	238	14,434	endnotes, glossary
Sondergard	2008	professor	FL→NL	76	257	6,364	footnotes

Table 1: For the five English translations of *Liaozhai* in our dataset, we report translator background, direction of translation (native language: NL; foreign language: FL), number of stories, number of paratexts, token count of paratexts, and type of paratexts.

The remaining four translations form the basis of our evaluation. Each edition was OCR-processed and aligned with its corresponding classical Chinese source following Zhang's (1978) index. To comply with fair use regulations, we omit the story translations themselves and retain only paratexts and their associated metadata. A professional editor reviewed the corpus in full and identified thirty-seven typographical and factual issues—including errors in spelling, grammar, punctuation, formatting, and historical detail—all of which were corrected and documented. For experiments and analyses, we combine all terms identified by the four translators into a single, comprehensive reference set, ensuring that no culture-bound explicitation introduced by any individual translator is overlooked. Comprehensive dataset statistics, including counts for the unused Giles (1880), are detailed in Table 1.

3.3 Annotation

A professional translator manually aligned each of the 692 paratexts across the four non-Giles translations with their corresponding classical Chinese source. In instances where multiple translators explicated the same phrase, entries were consolidated, resulting in 560 unique culture-bound terms.

To facilitate downstream analysis, we classified these paratexts according to a five-part framework derived from polysystem theory:

- **literary** explicitations clarify stylistic devices and narrative techniques;
- **historical** explicitations provide information on figures, events, and periods;
- **cultural** explicitations address beliefs, customs, celebrations, and folklore;
- **social** explicitations explain social structures, hierarchies, or relationships; and
- **supplementary** explicitations cover any other form of contextual commentary.

All annotations are provided in our released dataset, with a representative example in Figure 1 and examples of each classification in Appendix D.

Classical Chinese Term: 青山白云人

Literal translation (GPT-4o): 'person of green mountains and white clouds'

Paratextual explicitation: *Fu Yi*: A Tang dynasty scholar (555-639), who wrote his own epitaph (Mair and Mair, 1989).

Subsystem: Historical

Figure 1: Paratext alignment and classification in the *Liaozhai* dataset. While the source term appears to be a descriptive phrase, Mair and Mair (1989) connects it to historical figure Fu Yi.

4 Experiments

We design a two-step pipeline for generating paratextual explicitations in literary MT. Given a story from *Liaozhai*, the system proceeds as follows:

1. Identifies source terms requiring explicitation
2. Generates a paratext from the story context

4.1 Culture-Bound Term Identification

The first step identifies candidate terms from the source text requiring explicitation. For the 150 stories in our dataset containing at least one translator paratext, we prompt the model with the complete classical Chinese story and instruct it to extract expressions likely to demand additional explanation in translation. We compare three prompt variants:

- **Default:** A baseline prompt that directs the model to identify terms requiring explanation when translated, with no additional framing.
- **Theoretical:** The baseline prompt augmented with an explicit reference to and explanation of culture-bound terms. Its design is informed by polysystem theory, as introduced in §2.2.
- **Practical:** The baseline prompt augmented with instructions on the communicative function of translation. Its design follows Skopos theory, as similarly introduced in §2.2.

Complete prompt templates are provided in Appendix E. All experiments use the ChatML framework, with non-default extensions appended as additional user turns. We opt for zero-shot prompting to evaluate a model's intrinsic ability to identify culture-bound terms, deliberately avoiding few-shot exemplars to reduce bias.

Model outputs are evaluated via partial substring matching: a prediction is considered correct if and only if it contains or is contained by any translator-annotated term for the corresponding story. Performance is reported using standard information retrieval (IR) metrics: true positives, false positives, false negatives, precision, recall, and F1.

4.2 Culture-Bound Term Explicitation

For each of the 560 deduplicated terms extracted from the classical Chinese source text, we prompt the LLM to generate paratextual explicitations under conditions corresponding to the *Default*, *Theoretical*, and *Practical* formats described in §4.1. Complete prompts are provided in Appendix F.

Model outputs are evaluated against the pool of all translator paratext(s) for that story using a suite of four lexical, semantic, and LLM-based metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and LLM-as-a-Judge (Zheng et al., 2023). Where multiple references exist, all are included in the evaluation.

4.2.1 Agentic Retrieval

To mimic the way human translators consult external resources, we extend the strongest-performing model with search capabilities to scrape the web for relevant knowledge. Implemented using LangGraph, the agent executes a query given a culture-bound term and its surrounding story context.

First, it generates candidate English translations of the term and queries both Chinese (Baidu) and English (Google) search engines to retrieve information from bilingual sources. This approach enables access to long-tail knowledge in the source language, where culture-specific references are often better documented, while also capturing terminology conventions in the target language.

Returned results are appended as additional context for paratext generation. Importantly, we do not provide the agent with ground-truth translations, as identifying contextually-appropriate renderings is itself a central part of explicitation. Finally, we opt not to perform retrieval over a specific knowledge base to maintain broader applicability.

4.3 End-to-End Evaluation

To assess overall system performance, we conduct an integrated evaluation in which the LLM must both identify and explicitate culture-bound terms without access to gold-standard annotations. In the direct inference setting, the model executes identification and explicitation prompts sequentially. In the explicit chain-of-thought setting, the model is instructed to perform both steps in a single generation step while producing intermediate reasoning traces. This end-to-end evaluation offers a more holistic measure of system capabilities, capturing performance across the full pipeline rather than isolated subtasks.

5 Results

Our experiments are structured into three stages: first, culture-bound term identification across the 150 stories in our *Liaozhai* dataset (§5.1); second, paratextual explicitation of the 560 deduplicated translator-annotated terms (§5.2); and third, combined evaluation of term identification and explicitation in an end-to-end pipeline (§5.3). We report results under two inference modes: `non-thinking`, in which the model generates outputs directly, and `thinking`, in which the model produces intermediate reasoning tokens before final output. This follows the terminology of Qwen3-8B (Yang et al., 2025), the model we conduct all experiments on (unless otherwise specified). Experimental setup details are provided in Appendix G.

	Qwen3-8B (non-thinking)						**Qwen3-8B (thinking)**					
	TP	**FP**	**FN**	**P**	**R**	**F1**	**TP**	**FP**	**FN**	**P**	**R**	**F1**
Default	161	2351	399	6.41%	28.75%	10.48%	201	1699	359	10.58%	35.89%	16.34%
Theoretical	**191**	**1869**	**369**	**9.27%**	**34.11%**	**14.58%**	215	**1037**	345	**17.17%**	38.39%	**23.73%**
Practical	182	2071	378	8.08%	32.50%	12.94%	**216**	1482	**344**	12.72%	**38.57%**	19.13%

Table 2: **Culture-bound term identification results for Qwen3-8B under non-thinking and thinking modes.** Results are reported as true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R), and F1. Best-performing settings within are in bold.

	Qwen3-8B (non-thinking)				**Qwen3-8B (thinking)**			
	BLEU	**ROUGE-L**	**BERTScore**	**LLM**	**BLEU**	**ROUGE-L**	**BERTScore**	**LLM**
Default	1.40	12.91	83.37	64.91	1.31	15.25	84.61	68.72
Theoretical	1.25	12.94	83.51	68.42	1.08	15.14	85.21	71.70
Practical	**1.57**	**13.88**	**84.36**	**69.39**	1.62	16.82	85.56	72.83
Agentic	–	–	–	–	**2.14**	**20.59**	**86.08**	**75.69**

Table 3: **Automatic evaluation of paratextual explicitation quality.** Qwen3-8B is evaluated under non-thinking and thinking conditions using BLEU, ROUGE-L, BERTScore, and LLM-as-a-Judge (LLM). *Practical* prompting yields the strongest results without retrieval, and the agentic setup achieves the highest overall performance (in bold).

	Qwen3-8B (non-thinking)				**Qwen3-8B (thinking)**			
	BLEU	**ROUGE-L**	**BERTScore**	**LLM**	**BLEU**	**ROUGE-L**	**BERTScore**	**LLM**
Default	**0.01**	**0.09**	53.34	21.11	0.01	0.11	53.03	30.73
Theoretical	0.01	0.08	55.45	22.28	0.01	0.10	55.54	32.89
Practical	0.01	0.09	**57.49**	**24.82**	**0.02**	**0.15**	**57.65**	**35.41**

Table 4: **End-to-end evaluation combining term identification and explicitation.** Qwen3-8B performance under non-thinking and thinking conditions, assessed with BLEU, ROUGE-L, BERTScore, and LLM-as-a-Judge (LLM). While absolute scores remain low, *Practical* prompting produces the most consistent gains (in bold).

5.1 Culture-Bound Term Identification

To evaluate LLM performance on term identification, we test the three prompting strategies in §4.1 against our pooled set of 560 culture-bound terms. Use of standard IR metrics allows us to assess the degree to which translation studies-informed instructions improve a model's ability to detect culturally-significant expressions.

As shown in Table 2, the *Theoretical* prompt consistently outperforms both *Default* and *Practical* prompts across inference modes. Under the non-thinking setting, it achieves the highest F1 through balanced gains in precision and recall. This improves further in the thinking setting, with precision nearly seven points higher than that of the *Default* baseline. Although the *Practical* prompt yields the highest recall, its lower precision results in an overall weaker F1.

These results indicate that prompts grounded in translation theory enable the model to recover culture-bound terms while minimizing noise. This aligns with the intuition that theoretical frameworks provide explicit criteria for identifying culturally-marked elements and assessing their translational relevance. Thinking mode further reduces false positives, reflecting the model's two-stage hypothesis-and-filter process.

5.2 Culture-Bound Term Explicitation

Once culture-bound terms have been identified, the next step is then to generate paratexts. We evaluate outputs using lexical (BLEU, ROUGE-L), semantic (BERTScore), and LLM-based metrics (LLM-as-a-Judge with Qwen3-235B-A22B).

In contrast to term identification, results in Table 3 show that the *Practical* prompt performs best for paratextual explicitation, with gains especially pronounced in the thinking mode. We therefore extend the best-performing setting (*Practical* prompt in thinking mode) with agentic retrieval.

Incorporating bilingual web search yields consistent improvements across all metrics (BLEU: +0.52, ROUGE-L: +3.77, BERTScore F1: +0.52, LLM-as-a-Judge: +2.86). This confirms that access to external resources enables more accurate and contextually-enriched explicitations. The trade-off, however, is efficiency: this retrieval pipeline incurs over ten times the runtime of its non-agentic counterpart.

5.3 End-to-End Evaluation

Having examined culture-bound identification and explicitation independently, we next assess model capacity in an end-to-end setting. For terms not identified, automatic metrics pair translator references with an empty string, resulting in a score of zero for that paratextual explicitation.

Table 4 shows that the *Practical* prompt achieves the highest scores, suggesting that while it captures fewer terms overall (shown in Table 2), the paratexts it generates are of higher quality. Thinking improves performance further across all prompts,

Term	泮 (***pàn***)
Reference(s)	**becoming a *xiucai***: a scholar who has passed the county level of the imperial civil service examination (Sondergard, 2008).
LLM generation	**'enter the pond'**: Often associated with scholarly study and the education of young boys. In ancient times, it was common for students to gather by such waters to study and recite classical texts. The term is also used in the context of entering formal education, as students would 'enter the pond' to begin their studies at a local academy.

Term	解语花 (***jiě yǔ huā***)
Reference(s)	**'Intelligent Flower'**: The name given by Emperor Xuan Zong of the Tang Dynasty to his pet concubine Yang Guifei, who was known for her affectations (Lu et al., 1982). **'flower that understands speech'**: First used by the Tang emperor Xuanzong in reference to his beloved concubine Yang Yuhuan (Mair and Mair, 1989).
LLM generation	**Decoding Flower**: A mythical flower in Chinese folklore, said to have the power to understand and respond to human speech, symbolizing deep communication and emotional connection.

Table 5: Illustrative cases of LLM-generated paratextual explicitations compared to translator ones. The first term (泮) shows a contextually accurate but non-canonical explicitation, and the second (解语花) a hallucinated one. But these examples raise a key question: how do non-expert readers determine whether generations are hallucinated?

enhancing both identification coverage and explicitation quality.

Taken together, the three evaluation stages reveal the nature of this task: identification benefits from theoretically-grounded criteria for recognizing culture-specific expressions, whereas explicitation is inherently more audience-oriented, favoring prompts that foreground communicative clarity. Reasoning-enabled inference further amplifies both effects, suggesting that structured intermediate processing helps align system behavior with human translation practices. These findings highlight how no single prompting strategy suffices across both tasks and that effective paratextual explicitation requires adapting instructions to meet the demands of identification and explanation alike.

Table 5 presents examples of LLM-generated paratexts under the *Practical* setting. The first example is accurate, while the second is hallucinated; we provide their corresponding translator explicitations to facilitate interpretation.

5.4 Human Evaluation

To complement these automated results, we conduct two-stage human evaluation to measure the impact and quality of LLM-generated paratexts. This evaluation builds on previous experiments by examining how paratextual explicitations affect perceived translation quality and their alignment to human judgments.

All LLM outputs were generated using Qwen3-235B-A22B, and three native English speakers with no fluency in Chinese (the target audience of the translation of such a text) served as evaluators. Content was anonymized and presented in randomized order to reduce bias; evaluators indicated their preferred translation or explicitation, or selected 'no preference' if both were deemed comparable.

Paratext Impact. We first assess whether human evaluators prefer paratexts in LLM translations. Two subsets of stories from the *Liaozhai* dataset were selected: (1) four stories translated by all four human translators, all of whom included paratexts; and (2) five stories translated by two human translators, neither of whom included paratexts. For each story, evaluators compared a baseline LLM translation against a paratext-enriched version, where explicitations were produced using the *Theoretical* prompt for culture-bound term identification and the *Practical* prompt for explicitation. Results shown below in Figure 2.

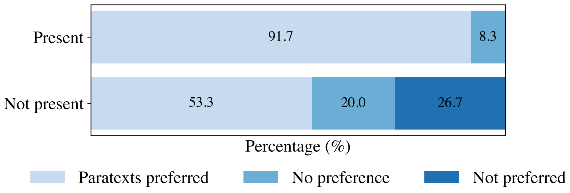

Figure 2: Human preference for LLM translations with versus without paratexts. `Present`: Stories where all human translators used paratexts. `Not present`: Stories where all human translators chose not to use paratexts.

Explicitation Quality. Next, we evaluate the contextual quality of individual explicitations. From the nine *Liaozhai* stories translated by all four translators, we extracted 73 culture-bound terms. For each term, an LLM-generated paratext produced with the *Practical* prompt was compared against one randomly selected translator paratext, both embedded within the full LLM translation. Results shown below in Figure 3.

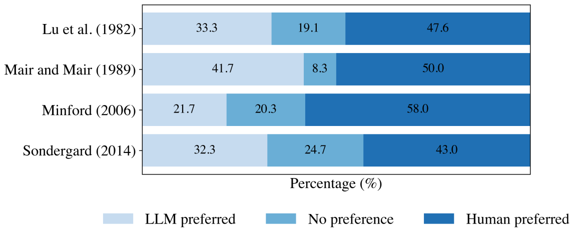

Figure 3: Human preference for LLM versus translator paratexts, reported for each of Lu et al. (1982), Mair and Mair (1989), Minford (2006), and Sondergard (2008).

Discussion. In the paratext impact study, evaluators rated paratext-enriched translations as preferable or equivalent to non paratext-enriched translations in 100% of stories where human translators included paratexts. Notably, even for stories where no human translators employed paratexts, evaluators still generally preferred the augmented version, suggesting that paratextual explicitations substantially improve clarity for target readers.

In the explicitation quality evaluation, translator paratexts were consistently preferred over LLM-generated ones, indicating that while models can produce plausible explicitations, they still fall short of translator-authored ones. Together, these results highlight the utility of paratexts in improving translation quality as well as the limitations of current LLMs relative to humans.

6 Analysis

To better understand the task of paratextual explicitation, we conduct a statistical analysis of patterns in the translator paratexts underlying our dataset. Although translators were not instructed to annotate terms explicitly, each binary choice—whether to provide a paratextual explicitation for a source term—can be treated as a unit of analysis.

Inter-Annotator Agreement. To assess consistency across translators, we compute inter-annotator agreement using Krippendorff's Alpha (Krippendorff, 2011) and pairwise Cohen's Kappa (Cohen, 1960). Krippendorff's Alpha provides a reliability coefficient for multiple raters, while Cohen's Kappa estimates agreement between rater pairs beyond chance. The resulting Krippendorff's Alpha of **-0.3493** points to systematic divergence rather than random variation, a pattern corroborated by pairwise Cohen's Kappa scores ranging from **-0.45** to **0.029**. Additional details are provided in Appendix I.

F1 Scores. F1 scores for individual human translators, measured against the pooled set of annotations from the other three, reveal that even experienced translators show limited overlap in term selection (Table 6). This highlights the inherent ambiguity in deciding which elements merit paratextual explicitation and provides a rough estimate of the human upper-bound on our dataset.

Translator	F1-score
Lu et al. (1982)	20.74%
Mair and Mair (1989)	24.87%
Minford (2006)	29.70%
Sondergard (2008)	37.11%

Table 6: F1 scores by translator.

Within this context, our best-performing *Theoretical* identification model achieves an F1 score of 23.73%—exceeding the 'worst-performing' translator, Lu et al. (1982), by 2.99 points, and trailing the 'best-performing' translator, Sondergard (2008), by 13.38 points. Notably, scores here diverge from the results of human evaluation in §5.4, where Minford (2006) is rated the most favorably and Mair and Mair (1989) the most poorly.

Consensus and Model Performance. Not all culture-bound terms are equally important to different translators, and this variation also influences model performance. Terms explicated by multiple translators tend to be more salient, and models likewise find them easier to identify. Table 7 reports the distribution of terms by the number of translators providing explicitations, along with corresponding LLM agreement percentages.

Explicitated by	Identified	Percentage
1 translator	172 / 479	35.91%
2 translators	36 / 73	49.32%
3 translators	4 / 5	80.00%
4 translators	3 / 3	100.00%

Table 7: Distribution of terms by the number of translators providing paratexts for that term, along with identification accuracy for the best-performing *Theoretical* prompt under the thinking variation. Higher consensus among translators corresponds to higher LLM accuracy.

Variation in Explicitation Content. Having established how translators differ in term selection, we next examine how translators vary in paratextual explicitation. Pairwise similarity measures (BLEU and BERTScore) computed over the 81 terms with two or more translator paratexts yield:

Bidirectional BLEU:	2.03
BERTScore F1:	87.40

Our best-performing LLM setting produces comparable scores (BLEU: 2.14; BERTScore F1: 86.08), suggesting that model-generated paratexts approximate the variability among human translators.

Summary. Together, these analyses highlight the considerable interpretive variation in paratextual explicitation. The inherent ambiguity of this task contextualizes our LLM results, showing that even when scores appear low, model outputs often fall within the range of human-human variation.

7 Related Work

Cultural and Contextual MT. A central challenge in MT is the handling of culture-specific items (CSIs). Conventional adequacy-based metrics often overlook errors on such items, motivating new benchmarks for capturing contextually-appropriate renderings (Yao et al., 2024). In-domain datasets ground this challenge in concrete contexts, such as food translation where lexical choice reflects cultural expectations (Zhang et al., 2024). Building on these insights, prototype systems extend the focus to end users by detecting cultural references and providing explanations (Pandey et al., 2025).

Adaptation and Localization. Complementary research in cultural MT focuses on strategies for adapting content once CSIs are identified. One line of work localizes named entities to maintain coherence in the target culture (Peskov et al., 2021). Others enrich translations with additional context, such as through curated explicitation corpora (Han et al., 2023) or by aligning background facts with external knowledge (Lou and Niehues, 2023). Beyond text, adaptation similarly extends to other modalities, where image transcreation poses a parallel challenge (Khanuja et al., 2024).

Modeling Human Translation. Recognizing that professional translators rely on reasoning beyond surface patterns, recent research has sought to computationally model these decision-making processes. This includes retrieval-augmented approaches that incorporate external knowledge during inference (Wang et al., 2025), as well as agent-based methods that simulate collaborative workflows between translators, editors, and proofreaders (Wu et al., 2025).

These directions parallel long-standing concerns in translation studies, where applied research has traditionally emphasized translator training, tools, and quality assessment (Holmes, 1972). Translation aids have then been categorized into software tools, reference resources, and collaborative environments (Pym, 2007), categories that now find computational analogues within the systems emerging in MT research.

8 Conclusion

In this paper, we present a first study of paratextual explicitation in literary MT. Drawing on insights from translation studies, we explore the poetics of paratexts as liminal devices that shape the interaction between a source text and target reader from the borderlands of a work. Methodologically, we formalize this task through the construction of an expert-aligned dataset of classical Chinese stories, evaluate contemporary LLMs on choice and content of explicitation, and analyze variation across professional translators to situate model performance within human practice.

While we focus on literary MT, the relevance of paratextual explicitation extends well beyond such a setting. In monolingual contexts, paratexts can serve as a form of explanatory glossing, clarifying technical or domain-specific terms for non-expert audiences. Similarly, in personalized applications, paratexts can be tuned to a reader's prior knowledge, interests, or expertise. These fine-grained levels of tailoring are impractical for human translators yet may be achievable through computational methods.

Paratextual explicitations thus offer more than just peripheral embellishment. Translations rarely preserve the full cultural, historical, and stylistic fabric of their originals, and paratexts provide a pragmatic means of bridging the contextual gap left by literal renderings. In this spirit, paratextual explicitation functions as a finely calibrated *looking-glass* into adaptation—refracting the interpretive choices, necessary compromises, and subtle negotiations through which texts are continually reshaped for new readers and new worlds.

Limitations

Paratext Types. Genette's (1987) framework for paratexts draws a distinction between between peritexts—elements included in the same volume as the main text, such as notes or glossaries—and epitexts, which exist independently but relate to the main text, such as interviews or promotional materials (Munday, 2016). Subsequent scholarship has extended this framework to include additional elements encompassing material features of the text (e.g., typeface, binding, page layout) as well as digital artifacts (e.g., metadata, hyperlinks); see Batchelor (2018). Translation studies has long recognized that paratexts shape how translated works are received and interpreted, and some scholars consider translations themselves as paratexts relative to the original.

This paper focuses specifically on paratextual explicitation through notes and commentary and does not address other forms within or beyond Genette's framework, such as translator prefaces or author bibliographies. These additional materials are included in the released dataset, and we invite future work to explore a broader range of paratextual forms across both textual and visual modalities.

Language and Domain. Our dataset exclusively focuses on classical Chinese to English translation within the literary domain. This allows us to study paratextual explicitation in a setting that is linguistically complex and culturally rich, but also constrains the generalizability of our findings. Literary texts present a unique challenge in terms of stylistic variation, cultural references, and interpretive nuance, which may not fully reflect patterns in other genres. Nevertheless, the underlying task of paratextual explicitation is not inherently limited to this language pair or domain and can be applied to a wide range of contexts.

Evaluation. Paratextual explicitation involves free-form generation rather than constrained translation, meaning that multiple formulations of the same information can be valid. To assess model output in this setting, we adopted a complementary suite of automated metrics: BLEU (Papineni et al., 2002) to provide a precision-based measurement, ROUGE-L (Lin, 2004) to capture longer subsequence matches, and BERTScore (Zhang et al., 2020) to evaluate semantic similarity across different surface realizations. While these metrics offer a useful baseline for measuring overlap and meaning, they remain limited in capturing the cultural nuance required by paratextual explicitation.

More recent metrics such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) aim to evaluate semantic quality and general translation adequacy, yet still fall short of assessing whether paratexts are factually accurate, contextually relevant, useful to the reader, or expressed in an appropriate manner. Human evaluation frameworks such as MQM (Lommel et al., 2013) provide hierarchical error annotation, but do not fully reflect the qualitative aspects that make explicitation meaningful. These limitations underscore the need for task-specific evaluation methods capable of assessing the context-sensitive and interpretive nature of translation commentary.

Ethical Considerations

This project obtained approval from the University of Edinburgh's Informatics Research Ethics committee, application number 2024/160527.

Fair Use. The original classical Chinese source text for *Liaozhai zhiyi* is openly accessible online. In accordance with fair use provisions, we release only the paratextual materials and metadata associated with the four human translations, explicitly excluding the stories themselves. These materials are provided in a transformative manner and intended solely for research purposes, ensuring that the dataset does not infringe upon the rights of the original publishers or translators.

Institutional and Ideological Concerns. Beyond questions of fair use, ethical considerations also involve the role of publishing agents in shaping how translations are produced, framed, and received. Translators, editors, and publishers alike make deliberate choices about which paratexts to include and how to present them, shaping the ways in which certain perspectives are amplified and others minimized. While our work explores paratextual explicitation as a computational task, it remains situated within broader institutional and ideological contexts which influence how knowledge and norms are transmitted across cultures. More broadly, discussions of gender, feminism, and queer representation in translation (Godard, 1990; Simon, 1996; Harvey, 2012, *inter alia*)—though beyond the scope of this work—illustrate how paratextual practices can influence the voices marginalized or silenced in translation.

Acknowledgments

This work received funding from the EU's Horizon Europe (HE) Research and Innovation programme [grant numbers 101070631, 101070350] and from UK Research and Innovation under the UK government's HE funding guarantee [grant numbers 10039436, 10052546].

References

Javier Franco Aixelá. 1996. Culture-Specific Items in Translation. In Rosa Alvarez and M. Carmen-Africa Vidal, editors, *Translation, Power, Subversion*, pages 52–78. Multilingual Matters, Clevedon.

Susan Bassnett. 1980. *Translation Studies*. Routledge, London and New York.

Susan Bassnett and André Lefevere. 1990. *Translation, History and Culture*. Pinter Publishers, London.

Kathryn Batchelor. 2018. *Translation and Paratext*. Routledge, London.

John C. Catford. 1965. *A Linguistic Theory of Translation*. Oxford University Press, Oxford.

Martha Pui Yiu Cheung and Wusun Lin, editors. 2006. *An Anthology of Chinese Discourse on Translation: From Earliest Times to the Buddhist Project (Volume 1)*. Routledge, London and New York.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton & Co., The Hague.

Marcus Tullius Cicero. -46. *De Optimo Genere Oratorum*. Translated as *On the Best Kind of Orators*, by Harry Mortimer Hubbell. 1960. Harvard University Press, London.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Dick Davis. 2000. 'Omar Khayyam'. *Encyclopedia of Literary Translation into English*, 2:1019–1020.

Itamar Even-Zohar. 1978. The Position of Translated Literature within the Literary Polysystem. *The Translation Studies Reader*.

Gérard Genette. 1987. *Seuils*. Éditions du Seuil, Paris. Translated as *Paratexts: Thresholds of Interpretation*, by Jane E. Lewin. 1997. *Literature, Culture, Theory*, vol. 20. Cambridge University Press, Cambridge.

Herbert A. Giles. 1880. *Strange Stories from a Chinese Studio*. T. De La Rue & Co., London.

Barbara Godard. 1990. Theorizing Feminist Discourse / Translation. *S. Bassnett and A. Lefevere (eds)*, pages 87–96.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging Background Knowledge Gaps in Translation with Automatic Explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.

Keith Harvey. 2012. Translating Camp Talk: Gay Identities and Cultural Transfer. *The Translation Studies Reader*, pages 344–364.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, and 5 others. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *Preprint*, arXiv:1803.05567.

James S. Holmes. 1972. The Name and Nature of Translation Studies. In J. Qvistgaard et al., editor, *Third International Congress of Applied Linguistics (Copenhagen, 21–26 August 1972): Congress Abstract*. Ehrverskøkonomisk Forlag, Copenhagen.

Horace. -19. *Ars Poetica*. Also known as *The Art of Poetry*, *Epistula ad Pisones*, or *Epistle to the Pisos*.

Yanan Jin. 2021. An Overview of the Translation and Introduction of "Liaozhai Zhiyi" in the English World. In *7th International Conference on Humanities and Social Science Research (ICHSSR 2021)*.

Alina Karakanta, Mayra Nas, and Aletta G. Dorst. 2025. Metaphors in Literary Machine Translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 276–286, Geneva, Switzerland. European Association for Machine Translation.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? On image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Werner Koller. 1995. The Concept of Equivalence and the Object of Translation Studies. *Target: International Journal of Translation Studies*, 7(2).

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Tong-King Lee. 2013. *Translating the Multilingual City: Cross-lingual Practices and Language Ideology*. Peter Lang.

André Lefevere. 1992. *Translation, Rewriting and the Manipulation of Literary Fame*. Routledge, London and New York.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Renhan Lou and Jan Niehues. 2023. Audience-specific Explanations for Machine Translation. *Preprint*, arXiv:2309.12998.

Yunzhong Lu, Tifang Chen, Liyi Yang, and Zhihong Yang. 1982. *Strange Tales of Liaozhai*. Commercial Press, Hong Kong.

Denis C. Mair and Victor H. Mair. 1989. *Strange Tales from Make-do Studio*. Foreign Languages Press, Beijing.

John Minford. 2006. *Strange Tales from a Chinese Studio*. Penguin, London.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, 51(1):73–137.

Jeremy Munday. 2016. *Introducing Translation Studies: Theories and Applications*, fourth edition. Routledge, London.

Peter Newmark. 1981. *Approaches to Translation*. Pergamon Press, Oxford.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall International, New York.

Eugene A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill, Leiden.

Eugene A. Nida and Charles R. Taber. 1969. *The Theory and Practice of Translation*. Brill, Leiden.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A Reading Assistant for Cross-Cultural Content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yongming Pei. 2018. *Re-presenting China through Retranslation: A Corpus-based Study of Liaozhai Zhiyi in English*. Doctoral dissertation, Kent State University. OhioLINK Electronic Theses and Dissertations Center.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting Entities across Languages and Cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Songling Pu. 1766. 聊斋志异 *(lit. Liaozhai zhiyi)*.

Anthony Pym. 2007. Natural and Directional Equivalence in Theories of Translation. *International Journal of Translation Studies*, 19(2):271–294.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Katharina Reiß and Hans J. Vermeer. 1984. *Grundlegung einer allgemeinen Translationstheorie*. Max Niemeyer Verlag, Tübingen. Translated as *Towards a General Theory of Translational Action* by Christiane Nord. 2013. Routledge, London.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sherry Simon. 1996. *Gender in Translation: Cultural Identity and the Politics of Transmission*. Routledge, London and New York.

Sidney L. Sondergard. 2008. *Strange Tales from Liaozhai*. Jain Pub., Fremont, California.

Gayatri Spivak. 1978. The Politics of Translation. *The Translation Studies Reader*, pages 312–330.

St Jerome. 395. *De Optime Genere Interpretandi (Letter 101, to Pammachius)*. Translated as *The Best Kind of Translator Letter to Pammachius, #57*, by Paul Carroll. 1997. Routledge.

Gideon Toury. 1995. *Descriptive Translation Studies – And Beyond*. John Benjamins Publishing Company, Amsterdam and Philadelphia.

Maria Tymoczko. 2010. *Enlarging Translation, Empowering Translators*, second edition. Routledge, London and New York.

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*, first edition. Routledge, London and New York.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*. Les éditions Didier, Paris. Translated as *Comparative Stylistics of French and English: A Methodology for Translation*, by Juan C. Sager and Marie-Jo Hamel. 1995. John Benjamins Publishing Company, Amsterdam and Philadelphia.

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2025. Retrieval-Augmented Machine Translation with Unstructured Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. Technical report, Qwen Team, Alibaba Group.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking Machine Translation with Cultural Awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Shuihan Yi, Chwee Fang Ng, and Hazlina Abdul Halim. 2025. A Study on the Translation of Culture-Specific Items in Character Depiction in the English Version of Pu Songling's *Liaozhai Zhiyi*. *World Journal of English Language*, 15(1).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Youhe Zhang. 1978. 聊斋志异会校会注会评本 *(lit. Liaozhai zhiyi huijiao huizhu huiping ben)*.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural Adaptation of Menus: A Fine-Grained Approach. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Exclusion of Giles (1880)

Herbert A. Giles (1845–1935) was a British diplomat and professor of Chinese at the University of Cambridge, renowned for his extensive translations of classical Chinese literature and scholarly work shaping Western understanding of Chinese language and culture. His 1880 translation of *Liaozhai*, *Strange Stories from a Chinese Studio (Volumes 1 and 2)*, is credited with introducing the literary work to Western audiences.

Giles's translation reflects the linguistic conventions of his era, employing his eponymous Wade-Giles romanization system to render the names of people and places. This style has since been largely supplanted by *pinyin* romanization (e.g., 'Peking' in Wade-Giles is 'Beijing' in *pinyin*), resulting in a naming convention that differs substantially from modern usage.

In addition, Giles's translation exhibits certain interpretive liberties, particularly with material he considered inappropriate or sensitive. For instance, depictions of fox spirits entering a bedchamber at night are altered to more innocuous events such as drinking tea. While consistent with the conventions of his time, such editorial interventions reflect broader patterns of ideological rewriting in translation.

Today, Giles's translation is best regarded as a historical artifact. Translations reflect the period they were produced in, and modern Western understanding of Chinese culture and literature has evolved considerably since the nineteenth century. Examples provided below.

On *'bamboo shoots'*, Giles writes:

> Which, well cooked, are a very good substitute for asparagus.

On *the 'Silver River'*, Giles writes:

> The Milky Way is known to the Chinese under this name—unquestionably a more poetical one than our own.

On *Mr. Chang*, Giles writes:

> The surnames Chang, Wang, and Li, correspond in China to our Brown, Jones, and Robinson.

These examples illustrate the dated and interpretive nature of Giles's commentary. For these reasons, we exclude his work from the dataset used in the experiments reported in this paper.

B Characteristics of Other Translations

Lu et al. (1982) adopt a conservative approach that adheres closely to classical Chinese narrative structures, using paratexts sparingly and only when deemed strictly necessary. By contrast, Mair and Mair (1989) weave explicitations directly into the main text, replacing potentially obscure terms with more accessible equivalents instead of heavily relying on external commentary.

Minford (2006) aims to preserve narrative flow while providing paratextual support, including detailed explanations through particularly extensive back matter. Sondergard (2008) pursues the most annotation-intensive strategy, employing frequent footnotes to explicate terms likely to be unfamiliar to the contemporary reader.

We present representative paratexts for each translator below.

> **Flower-Morning:** The twelfth day of the second month, traditionally held to be the birthday of flowers.

Note: Mair and Mair (1989) for story `v11s2`, *Yellow-Bloom*.

> This, one of the best known and most often anthologized and translated of all the *Tales*, is a greatly expanded variation on a brief item in the much earlier collection *In Search of Spirits*, attributed to Gan Bao (*fl.* 320). In the earlier story, the magician is called Xu Guang:
>
>> Once he was performing his magic arts in the marketplace and begged for a gourd from a vendor, who refused to give him one. So he asked for a flower and planted it in the ground, where it immediately started growing, spreading its tendrils over the ground. First it bore flowers, and then fruits. Xu Guang picked one, ate it, and then began handing the fruits out to the spectators. When the vendor turned to look at his own gourds, they had all disappeared.
>
> (My translation of the extract quoted by Zhu Yixuan, *Liaozhai zhiyi ziliao huibian*, revised edition (Tianjin, 2002), p. 17. For the complete tale, see Li Qi and Liang Guofu (eds.), *Soushenji Soushen houji yizhu* (Jilin, 1997), p. 27.)
>
> For obvious reasons this tale has always been popular with Marxist commentators, and is placed first in the popular selection made by Yan Weiqing and Zhu Qikai in 1984. It has been published many times in cartoon-strip form.
>
> The Chronicler of the Strange appends one of his most trenchant comments to this tale, sharply reproaching the nouveaux riches for their meanness, for the way they turn a deaf ear to needy friends or relations coming to them with simple requests for loans of food or money. In other words, his target is far broader than the country bumpkin who is made to look such a fool in the tale.

Note: Minford (2006) for story `v1s14`, *Growing Pears*.

Subsystem	Paratext
literary	**We 'know each other's sound'**: One who 'knows the sound' of another is, as William Acker puts it: a friend whose knowledge of music is such, and whose mind is so attuned to that of the player that he can catch the finest nuances of the performer's thought and feeling, as he listens, and by his speech or by his silence after the playing of a piece shows that he has understood the other's thoughts as though they have been spoken rather than played... (*Some T'ang and Pre-T'ang Texts on Chinese Painting* (Leiden, 1954), p. 10) The expression comes from the story of Bo Ya and Zhong Ziqi, in the Taoist *Book of Liezi* (Minford, 2006).
cultural	**the Cut Sleeve persuasion**: Emperor Ai, last ruler of the Former Han dynasty (206 BC-AD 9), had a number of boy-lovers, the best-known of whom was a certain Dong Xian. Once when the Emperor was sharing his couch with Dong Xian, the latter fell asleep lying across the Emperor's sleeve. When the Emperor was called away to grant an audience, he took his sword and cut off his sleeve rather than disturb the sleep of his favourite. Hence the term 'Cut Sleeve' (*duanxiu*) has become a literary expression for homosexuality among men (Minford, 2006).
social	**The white clothes of the xiucai**: Worn by a scholar who's passed the imperial civil service examination at the county level (Sondergard, 2008).
supplemental	**A notorious place**: The Bu River in Shandong province passed into the vernacular as a "place notorious for profligacy" since it became a popular site for romantic trysts (see Zhu 51n6) (Sondergard, 2008).

Table 8: Representative examples of paratexts illustrating the literary, cultural, social, and supplementary subsystems. The two paratexts from Minford (2006) are quite extensive and have been condensed here for ease of reading.

> ***xiaolian***: An old term for *juren* (举人), a successful candidate in the imperial examination at the provincial level in the Ming and Qing Dynasties.

Note: Lu et al. (1982) for story `v1s6`, *A Wall-painting*.

> **This poem**: This twelve-line poem, consisting of seven characters per line, is structurally reminiscent of the "*jiang shang yin*" ("River Poem") of Li Bo (699-762 C.E.), China's most famous poet. However, its subject and tone are almost precisely the opposite of those in the Li poem, treating Buddhism and nostalgic sadness rather than Daoism and exuberant joy.

Note: Sondergard (2008) for story `v2s40`, *Fourth Lady Lin*.

C Dataset Structure

Our dataset is organized into three subfolders:

- `annotations/` containing all expert-aligned paratexts and their corresponding annotations (`annotations.csv`), as well as the log of typographical corrections (`corrections.md`);
- `source/` including the Chinese source texts in JSON format, organized by classical (`classical/main.json`) and contemporary (`contemporary/main.json`) styles; and
- `translations/` including the five English translations (`1880_giles/`, `1982_lu_etal/`, `1989_mair_and_mair/`, `2006_minford`, and `2008_sondergard`) in `main.json` files.

Each `main.json` contains entries corresponding to texts in the collection, presented in order of appearance. Possible metadata fields for each entry include:

- `id`: the global identifier with regards to the Chinese source text;
- `title`: the story title;
- `content`: the story body;
- `commentary`: curator notes on literary significance; and
- `notes`: translator paratexts, in the form of footnotes or endnotes.

Not all fields appear in every file, and some translator folders include additional paratextual materials, such as glossaries or appendices.

D Subsystem Examples

Representative examples of paratexts classified according to the five-part framework adapted from polysystem theory are provided in Table 8, covering the literary, cultural, social, and supplementary subsystems. Historical examples are given in the main text (Figure 1) and are therefore not repeated here.

Paratexts from Minford (2006) are particularly extensive; the excerpts included in the table have been condensed to highlight the core explanatory content while maintaining readability. Full annotations for all paratexts are available in the released dataset.

E Prompts for Term Identification

Default:
You are a helpful translation assistant. When provided with a story in classical Chinese, identify key terms that require additional explanation when translated into English. Return these terms as a comma-separated list.

Theoretical:
The terms you identify should be culture-bound terms as defined in translation studies: expressions deeply rooted in the literary, historical, or social context of the source culture. Such terms are often unfamiliar to readers from other cultures and may necessitate explicitation to bridge the gap in understanding.

Practical:
Your target audience is composed of native English speakers with limited knowledge of Chinese culture. The terms you identify for additional explanation should therefore help them understand the story or its setting in a more meaningful manner.

F Prompts for Term Explicitation

Default:
You are a helpful translation assistant. Given a classical Chinese story and term from the story, provide (1) an English translation of the term and (2) a clear description of the term's meaning or significance. Format your answer as: {translated_term}: {description}.

Theoretical:
Select an appropriate translation strategy (e.g., domestication, foreignization) for the term and let that choice guide your rendering and explanation. Interpret the culture-bound term with respect to its role within the literary, cultural, historical, or social dynamics of the source culture and present your description as a peritext in the Genettean sense—a translator's footnote intended to support the reader's understanding. Do not explain your reasoning; simply provide the term and description.

Practical:
Translate the term for a target audience of native English speakers unfamiliar with Chinese culture. Your description should preserve the term's cultural grounding while remaining clear and accessible. Keep the description concise but informative, offering just enough context to aid reader understanding without being overwhelming.

Agentic:
You are an expert at identifying relevant information. From the provided search results, extract passages that seem the most relevant to defining the classical Chinese term in the given context. Focus on dictionary definitions, explanations, and contextual usage information.

G Experimental Setup

We follow QWEN3's recommended hyperparameter settings of `temperature` = 0.7, `top_p` = 0.8, `top_k` = 20, `min_p` = 0 for non-thinking mode and `temperature` = 0.6, `top_p` = 0.95, `top_k` = 20, `min_p` = 0 for thinking mode.

For LLM-as-a-Judge, we use QWEN3-235B-A22B under the non-thinking configuration with same hyperparameter settings. In each evaluation, we present the judge with the classical Chinese source term, the LLM-generated explicitation, and all human reference(s), asking it to evaluate accuracy and clarity on a scale of 0 to 100.

H Human Evaluation Details

Instruction for paratext impact evaluation:

For each evaluation, you will be given two English translations of a classical Chinese story from *Liaozhai*. Select the translation you prefer, or no preference if both are comparable. You may optionally provide a brief justification of your choice.

Instruction for explicitation quality evaluation:

Read the given English translation of a classical Chinese story from *Liaozhai*. Throughout the translation you will see terms highlighted, and then two explanations of the term. Select the explanation you prefer, or no preference if both are comparable. You may optionally provide a brief justification of your choice.

Each of the 54 story evaluations was compensated at $3.00 USD, resulting in a total cost of $162 USD for human evaluation.

I Inter-Annotator Agreement

For Krippendorff's Alpha, we adopt a three-way encoding scheme for each source term: `NaN` for stories not translated by a given translator, `0` for terms left unexplicitated, and 1 for terms that received explicitation. For Cohen's Kappa, we compute pairwise, chance-corrected agreement between translators on their set of overlapping stories (results in Figure 4).

This combined approach measures overall agreement among all translators, with pairwise comparisons revealing whether any individual translator disproportionately affects the aggregate score.

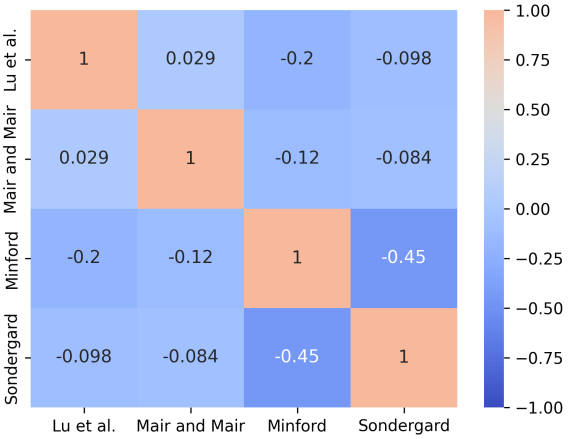

Figure 4: Pairwise Cohen's Kappa heatmap results.