Do Evaluation Metrics Detect Errors in Classical Chinese to English Translations?

Osvaldo Quinjica[1] **Eric Bennett**[1,2] **Xinchen Yang**[1]
Andrew Schonebaum[2] **Marine Carpuat**[1]

[1]Department of Computer Science
[2]Department of East Asian Languages and Cultures
University of Maryland, College Park
{quinjica,ebenne92,xcyang,schone,marine}@umd.edu

Abstract

Although large language models can translate some historical languages surprisingly well, their usefulness in digital humanities workflows is limited by the lack of reliable evaluation. We investigate whether existing automatic evaluation metrics developed for modern languages are reliable in this setting, using translation from Classical Chinese to English as a test case. We introduce a diagnostic framework based on minimal pairs capturing error types salient in scholarly use, probing both reference-based and reference-free metrics for error sensitivity and tolerance to valid variation. We find that all metrics exhibit blind spots, however MetricX24 performs best overall. Our findings highlight the need for more robust and interpretable metrics for historically and culturally distinct translation settings[1].

1 Introduction

Large language models (LLMs) have demonstrated unexpectedly strong performance on tasks involving historical languages such as Latin (Volk et al., 2024), Hanbun (Son et al., 2022) or Classical Chinese (Jin et al., 2023), likely due to incidental exposure to bilingual texts during pre-training (Briakou et al., 2023). This capability opens new opportunities for digital humanities, enabling large-scale search, translation, and computational analysis of vast historical corpora, and has the potential to democratize access to classical texts (Nehrdich & Keutzer, 2026; Sturgeon, 2019; Song et al., 2025).

Yet, evaluating the quality of LLM-based machine translation (MT) in these settings remains a challenge. Current evaluation practices rely on metrics validated primarily on modern languages, and their robustness to the linguistic, cultural and historical divergences remains unclear. For instance, Classical Chinese poses unique difficulties: its extreme conciseness and reliance on context lead to wide variability in acceptable translations, complicating both translation and evaluation. For scholars who lack proficiency in the source language or need to process corpora at scale, reliable evaluation of LLM translations is essential.

Despite progress in MT evaluation, from string-based metrics (Papineni et al., 2002a; Popovic, 2015) to learned metrics trained on human ratings (Rei et al., 2020; Juraska et al., 2024) and LLM-based judges (Kocmi & Federmann, 2023c; Fernandes et al., 2023), their robustness in historical and culturally distinct settings remains largely untested. Existing studies primarily assess aggregate correlations with human ratings (Papineni et al., 2002b; Rei et al., 2020; Nehrdich et al., 2025), but real-world applications demand finer-grained insights: Which errors do these metrics reliably detect, and how tolerant are they of legitimate variation? Can they help determine whether a translation is sufficiently accurate to present to scholars unfamiliar with the source language, or to construct corpora for semantic search? Or do they risk allowing errors that could distort scholarly interpretation or learning?

[1]We release the code and dataset here: https://github.com/cx-olquinjica/cc-mt-eval

To address this problem, we introduce an error diagnosis framework tailored to Classical Chinese–English MT evaluation. Building on challenge sets based on minimal pairs (Warstadt et al., 2020; Isabelle et al., 2017) and automatic perturbations (Karpinska et al., 2022; Bennett et al., 2025), our framework probes both reference-based and reference-free metrics, assessing their sensitivity to error-inducing perturbations and their tolerance for acceptable paraphrastic variation. Our contributions are threefold: (1) a diagnostic framework for evaluating MT metrics in this domain; (2) benchmarks of widely used metrics on Classical Chinese-English translation; and (3) insights into improving evaluators for linguistically and culturally distinct settings.

By situating MT evaluation in a practical and historically significant scenario, this work aims to support reliable AI-assisted scholarly translation, while advancing our understanding of the generalizability of commonly used evaluation metrics.

2 Background

Classical Chinese (*wenyanwen*) provides an unusually demanding and revealing test case for LLM-based language understanding, translation and the metrics used to assess them. As the written medium of governance, philosophy, religion, science, and literature in East Asia for over two millennia, Classical Chinese underlies an enormous and culturally foundational textual record, comparable in scope and historical importance to Latin in Europe (Sturgeon, 2021). Yet despite its centrality, it remains largely inaccessible to non-specialists, making high-quality translation a key enabling technology for digital humanities research.

Linguistically, Classical Chinese diverges sharply from the modern languages on which most MT systems and evaluation metrics are developed. It lacks inflectional morphology, overt tense or agreement marking, and relies on highly flexible, predominantly monosyllabic lexemes whose syntactic and semantic roles are determined almost entirely by context (Schonebaum et al., 2024). Meaning is frequently implicit: arguments are omitted via pervasive zero anaphora, temporal and causal relations must be inferred, and short clauses often admit multiple plausible readings (Schonebaum et al., 2024). These properties create systematic failure modes for translation models, which may default to modern character meanings, fail to disambiguate polysemous items, or inadequately integrate broader discourse and cultural context when selecting an interpretation. Such errors can produce fluent, well-formed English translations that are nevertheless semantically misleading. For evaluation, the challenge is therefore two-fold: metrics should be both tolerant of legitimate paraphrase and sensitive to subtle, context-dependent interpretation errors. Metrics optimized for lexical overlap or generic semantic similarity may overlook these failures, allowing mistranslations that distort historical meaning while appearing acceptable under conventional automatic evaluation.

Despite these challenges, LLMs are already being used to translate Classical Chinese (Sturgeon, 2021), yet there has been comparatively little research on methods for assessing the quality of these translations. Their translation capabilities are often attributed to incidental exposure to historical and bilingual materials during large-scale pretraining, as well as to models' capacity to generalize from modern Chinese and related high-register texts (Briakou et al., 2023; Chang et al., 2023). Growing research documents LLM performance on Classical Chinese understanding tasks, including sentence interpretation, question answering, and historical knowledge probing (Zhou et al., 2023; Cao et al., 2024b; Li et al., 2024). Knowledge-grounded and retrieval-augmented approaches have proven useful to inject historical context, canonical references, or curated commentaries, as demonstrated in systems such as TongGu and MITRA-zh (Cao et al., 2024a; Nehrdich et al., 2023). Translation-specific studies report that LLMs can generate fluent and often plausible English renderings of Classical Chinese prose and poetry, sometimes rivaling or surpassing earlier neural MT systems (Wang et al., 2023; Chen et al., 2025; You et al., 2025). At the same time, these studies consistently note important translation failures: models may over-modernize meanings, collapse distinct classical senses into a single contemporary gloss, or fail to resolve ambiguity using broader discourse or genre-specific conventions. These errors are difficult to detect automatically, motivating the need for reliable translation quality evaluation.

Evaluation of Classical Chinese translation remains comparatively underexplored. Most existing work adopts standard MT evaluation metrics like BLEU, or relies on expert human judgment along dimensions like adequacy, fluency, and literary quality, particularly for poetry (Chen et al., 2025). A small number of studies have begun to assess modern automatic metrics in this domain. Nehrdich & Keutzer (2026) curated evaluation testbeds for Buddhist Chinese. Bennett et al. (2025) evaluated both string-based and neural translation quality metrics on Pre-Qin Classical Chinese texts (pre-221 BCE), and found that neural metrics are more reliable than BLEU or chrF when faced with artificial perturbations. Howeveer, it remains unclear to what degree those metrics detect naturally occurring errors and acceptable variations that matter for scholarly interpretation. This gap motivates the need for fine-grained, diagnostic evaluation frameworks that probe metric behavior under controlled, linguistically meaningful perturbations, rather than assuming that performance in modern, high-resource settings will generalize to historically and culturally distant languages.

3 Diagnostic Framework

We adopt a controlled meta-evaluation framework inspired by DEMETR (Karpinska et al., 2022), using *minimal pairs* to probe MT evaluation metrics. Each pair contains an original translation and a minimally edited variant that differs by exactly one targeted perturbation. A well-behaved metric should assign a higher score to the original than to the perturbed output, allowing us to attribute score differences directly to a single error type.

For our case study on Classical Chinese-English translation, we extend prior diagnostic setups in two ways: (i) we derive perturbation categories from expert philological inspection of real LLM outputs and instantiate them via semi-automatic rewrites vetted by trained annotators; and (ii) we quantify metric sensitivity through mixed-effects modeling of standardized score deltas, rather than normalizing against catastrophic baselines (e.g., empty outputs), which we find unreliable in this setting.

3.1 Data

Source texts are drawn from the Chinese Text Project (CTP), a curated digital library of premodern Chinese writings spanning diverse genres and historical periods (Sturgeon, 2021). CTP also provides English translations by James Legge, which we treat as high-quality human references to anchor evaluation and support validation. While some of Legge's translations have been criticized for interpretive bias and 19th-century scholarly conventions (MacKenzie, 2024), he remains a foundational scholar of Classical Chinese whose translations are widely used in academic research today.

To construct baseline machine translations, we translate selected CTP segments into English using GPT-4o-mini. These outputs serve as the starting point for all minimal pairs. Because CTP texts and their translations are widely available online, we conduct a memorization analysis following Chen et al. (2024) to assess whether results are driven by training data overlap rather than genuine translation evaluation. As reported in Table 6 in Appendix, we find little evidence of memorization.

3.2 Expert-Grounded Error Categories

A central contribution of this work is the selection of error categories informed by domain expertise. Several co-authors are scholars trained in Classical Chinese philology who have experimented with LLM-based machine translation in their own research. Their experience revealed translation failures that influence scholarly interpretation. To formalize these observations, we conducted a systematic manual analysis of an internal benchmark. Three experts reviewed translations of 150 source passages translated by diverse state-of-the-art LLMs (Yang et al., 2024; OpenAI et al., 2024; Comanici et al., 2025; GLM et al., 2024; Yang et al., 2025; Team), targeting difficult examples, including passages containing numbers, titles, and Classical/Modern Chinese ambiguities, as well as cases with substantial disagreement across systems or evaluation metrics (e.g., high GEMBA and low COMET scores).

Category	Example	Description
ERRORS		
Modern Chinese translation	Su Laoquan, twenty-seven. He began to feel passionate and read books. 苏老泉，二十七岁。他开始感到热情并阅读书籍。	The Classical Chinese source is first translated to English, and then to modern Chinese.
Omitted object	Zhang Gai said, "Please allow me to persuade the state of Lu to remain neutral." Zhang Gai said, "Please allow me to persuade _____ to remain neutral."	The object the state of Lu is omitted, making it unclear what is being persuaded to remain neutral.
Omitted subject	Zi Lu then said to the family, "Not to take office is not righteous." _____ Then said to the family, "Not to take office is not righteous."	The subject Zi Lu is omitted, making it unclear who is speaking.
Incorrect lexical translation	They all say, 'We are wise; But who can distinguish the male and female crow?' They all say, 'We are wise; But who can distinguish the male and female pigeon?'	Crow is mistranslated as pigeon, substituting the referenced entity.
Pronoun substitution	The Master said, "There is Yung! He might occupy the place of a prince." The Master said, "There is Yung! I might occupy the place of a prince."	He is replaced by I, introducing a coreference error.
Incorrect tense	I am not concerned that I am not known. I am not concerned that I was not known.	I am is shifted to I was, creating a tense inconsistency.
Title substitution	The duke of the state of Zhao conferred Wucheng on Lord Mengchang. The king of the state of Zhao conferred Wucheng on Lord Mengchang.	Duke is replaced with king, misrepresenting the rank.
Modern sense substitution	Fu Xie, looking very serious, turned him down: "If I did well and no one noticed, that is simply a matter of luck." Fu Xie, looking very colorful, turned him down: "If I did well and no one noticed, that is simply a matter of luck."	The Classical sense (serious/complexion) is replaced with the modern sense (colorful).
ACCEPTABLE VARIATIONS		
Name formatting	Zhong Yong: The Master said, "Perfect is the virtue which is according to the law!" ZhongYong: The Master said, "Perfect is the virtue which is according to the law!"	Whitespace removed from a romanized name.
Name annotation	Zi Lu then said to the family, "Not to take office is not righteous." Zi Lu (子路) then said to the family, "Not to take office is not righteous."	Added name gloss within parentheses.
Sentence segmentation	He, having grown old, still regrets his tardiness. You, young one, should think early. Having grown old, he still regrets his tardiness; you, young one, ought to think early.	Meaning preserving segmentation and punctuation changes with minor paraphrase.

Table 1: Errors and acceptable variations for Classical Chinese–to–English metric benchmarking. Original and perturbed spans are highlighted.

This process yields a set of 8 error categories that reflect risks in downstream use, whether for distant reading (e.g., semantic search) or close reading (e.g., philosophical or historical analysis): *(i) unintended translation into modern Chinese rather than English*, *(ii) omitted object*, *(iii) omitted subject*, *(iv) incorrect lexical translation*, *(v) pronoun substitution and misattribution*,

(vi) incorrect tense realization, *(vii) title or rank substitution*, and *(viii) modern-sense substitution*. By explicitly targeting these error categories through controlled perturbations, we introduce human expertise at scale without requiring large amounts of manual annotation.

We also examine three categories of acceptable variations of translations, including *(i) name formatting*, *(ii) name annotation*, and *(iii) sentence segmentation variation and minor paraphrasing*, selected to assess whether metrics appropriately tolerate variations which do not affect the underlying meaning of a source. Examples of all variations (whether acceptable or errorful) are provided in Table 1.

3.3 LLM-based perturbations

For each baseline translation, we generate minimally edited variants that instantiate exactly one error type. Perturbations are constructed using a semi-automatic pipeline that combines rule-based edits, LLM-based minimal rewrites, and hybrid procedures that leverage both source- and target-side information.

Automatic Generation LLM-based rewrites are used for phenomena that require semantic or discourse-level manipulation, such as omitting an implicit subject or substituting a modern sense for a classical one while preserving fluency. Because English is the target language, we exploit LLM strengths in controlled English rewriting while holding content constant. Rule-based edits are used where deterministic modifications suffice (e.g., pronoun or title substitution). Hybrid methods first identify candidate locations using source-side cues and then prompt an LLM to realize the minimal change. See Appendix A for prompts.

Manual Vetting To ensure perturbations faithfully instantiate the intended error and introduce no additional changes, we select a subset of at least 100 examples per category using length-based heuristics, and have it manually vetted by two co-authors trained in both modern and Classical Chinese. The results on the validated subset show a moderate agreement between annotators, with a raw agreement of 67% and a Cohen's κ of 0.58. Variation in κ is concentrated in perturbations with extreme class imbalance, where most instances are labeled as accepted (see Appendix D). This validation setup proved more efficient for annotators than traditional translation annotation tasks, as it required only targeted judgments compared to full translation assessments.

Final Benchmark Only pairs that clearly and only realize the intended perturbation are retained in our benchmark, resulting in 1031 minimal pairs for errors, and 334 for acceptable perturbations (see Appendix D for details).

3.4 Statistical Analysis of Metric Sensitivity

Our analysis examines whether MT evaluation metrics reliably prefer an unmodified translation over a minimally perturbed one, and how their sensitivity varies across perturbation types. For each segment, metric, and perturbation condition, we compute a *score delta* defined as the difference between the score assigned to the baseline translation and the score assigned to its perturbed counterpart.

Because evaluation metrics operate on different scales, we standardize score deltas (z-scoring) within each metric prior to analysis. This places all metrics on a common scale while preserving relative differences in sensitivity across perturbation types. Next, we model standardized score deltas with a single linear mixed-effects (LME) model implemented with the Python `statsmodels` package, fitting error types and acceptable-variation types jointly with fixed effects for condition, metric, and their interaction, and segment ids as random effects:

$$\delta^* \sim \underbrace{\texttt{condition} \times \texttt{metric}}_{\text{Fixed effects}} + \underbrace{(1 \mid \texttt{segment})}_{\text{Random effect}}, \tag{1}$$

The fixed effects `condition` and `metric`, together with their interaction, model the effects of perturbation type, evaluation metric, and whether metric sensitivity varies across perturbation types, while the random effect $(1 \mid \texttt{segment})$ accounts for repeated observations from the same source segment. From the fitted model we report, for each (metric, condition) combination, an estimated mean standardized delta $\hat{\mu}$ and its significance test obtained by re-levelling the model so that the target condition is the reference category, indicating whether the metric consistently distinguishes baseline translations from their modified variants. For the time-period analysis (Section 5.3), we additionally include the source text's time period and its interaction with metric as fixed effects.

4 Translation Quality Metrics

We evaluate how well off-the-shelf translation quality metrics assess Classical Chinese→English translations. Metrics are drawn from three families: surface overlap, learned neural metrics, and LLM-as-a-judge approaches—and are evaluated in both reference-based (`src+ref+hyp`, `ref+hyp`) and reference-free quality estimation (QE; `src+hyp`) settings. Access to the source and/or a reference directly constrains the evidence available for scoring and shapes the assumptions each metric makes about adequacy. We focus on widely adopted metrics from recent WMT evaluations, where metrics are often applied across source languages and domains beyond their original training conditions(Lavie et al., 2025; Zouhar et al., 2024). This setting allows us to assess whether current translation evaluation paradigms transfer to Classical Chinese.

Surface overlap metrics. BLEU (Papineni et al., 2002a) and ChrF (Popovic, 2015) measure n-gram or character overlap between a hypothesis and a reference. While widely used, their core assumption that lexical overlap tracks adequacy might be misaligned with Classical Chinese translation, where faithful renderings may diverge substantially in phrasing.

Learned neural metrics. We evaluate COMET-22, XCOMET/XCOMET-QE, and MetricX24/MetricX24-QE (Rei et al., 2020; Guerreiro et al., 2024; Juraska et al., 2024). These metrics rely on multilingual encoders (XLM-R or mT5) fine-tuned on human ratings of translation quality as Direct Assessment or MQM error annotation primarily on MT between modern languages in the news domain from WMT evaluations. While prior work shows that such metrics can generalize across languages (Rei et al., 2020), their training data does not include Classical-Chinese sources. We therefore test each metric in its standard configuration, as well as, where applicable, in reference-only and QE variants. For COMET and MetricX24, we additionally evaluate reference-only variants (`ref+hyp`) to isolate the contribution of source conditioning

LLM-as-a-judge metrics. Finally, we evaluate GEMBA, which uses zero- or few-shot prompting of GPT-family models to produce translation quality judgments following DA- or MQM-style instructions (Kocmi & Federmann, 2023b;a). Unlike supervised neural metrics, GEMBA does not learn from WMT human scores via fine-tuning, instead inheriting the LLM's general pretraining and instruction-following priors.

See Appendix B for implementation details.

5 Results

As a preliminary sanity check, we evaluate whether the metrics can distinguish reasonable translations from clearly incorrect ones. For each source segment, we compare the original translation against both an unrelated translation of the source and a word-shuffled version of the original translation. All metrics consistently prefer the original translation (Table 8), demonstrating that they can identify catastrophic translation errors in Classical Chinese-to-English translation.

Having established this basic validity, we turn to the benchmark's central questions. We first analyze the sensitivity of translation quality metrics to translation errors (Section 5.1)

	Error Perturbations														Acceptable Variations					
	Lex. error		Anc./Mod. sense		Mod. Ch. out.		Omit. obj.		Omit. subj.		Pronoun		Tense		Title		Ch. annot.		Name fmt.		Sent. seg.	
Metric	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.
Src + Ref + MT																						
COMET	−0.171 (0.062)	**	−0.286 (0.077)	***	**+2.075** (0.056)	***	**+0.402** (0.067)	***	**+0.145** (0.065)	*	+0.043 (0.054)		−0.448 (0.056)	***	−0.496 (0.054)	***	**−0.488** (0.066)	***	**−0.530** (0.065)	***	**−0.673** (0.064)	***
XCOMET	**+0.279** (0.090)	***	−0.235 (0.113)	*	**+0.973** (0.081)	***	**+0.219** (0.096)	*	**+0.456** (0.094)	***	−0.036 (0.080)		−0.319 (0.080)	***	−0.209 (0.080)	**	**−0.626** (0.096)	***	**−0.299** (0.095)	***	**−0.395** (0.092)	***
MetricX-24	**+0.277** (0.077)	***	+0.143 (0.095)		−1.266 (0.069)	***	**+0.721** (0.083)	***	**+0.462** (0.081)	***	**+1.133** (0.067)	***	−0.149 (0.069)	*	−0.051 (0.067)		**−0.314** (0.082)	***	**−0.249** (0.081)	***	**−0.557** (0.079)	***
GEMBA-DA-ref	**+0.874** (0.081)	***	−0.111 (0.103)		**+1.057** (0.073)	***	+0.149 (0.088)		+0.015 (0.086)		**+0.342** (0.072)	***	−0.447 (0.073)	***	−0.189 (0.072)	**	**−0.728** (0.086)	***	**−0.558** (0.086)	***	**−0.742** (0.083)	***
GEMBA-MQM-ref	**+0.598** (0.085)	***	−0.159 (0.110)		**+0.979** (0.078)	***	+0.098 (0.091)		+0.019 (0.089)		**+0.167** (0.077)	*	−0.278 (0.076)	***	−0.110 (0.077)		**−0.510** (0.091)	***	**−0.415** (0.091)	***	**−0.726** (0.087)	***
Src + MT (QE)																						
XCOMET-QE	**+0.561** (0.096)	***	−0.162 (0.122)		−0.248 (0.087)	**	**+0.217** (0.103)	*	**+0.315** (0.101)	***	+0.055 (0.086)		+0.040 (0.086)		−0.085 (0.085)		−0.470 (0.103)	***	−0.153 (0.102)		−0.108 (0.099)	
MetricX-24-QE	**+0.210** (0.084)	**	+0.125 (0.103)		−0.962 (0.075)	***	**+0.603** (0.090)	***	**+0.495** (0.088)	***	**+1.057** (0.073)	***	−0.121 (0.075)		−0.120 (0.073)		**−0.472** (0.088)	***	**−0.172** (0.087)	*	**−0.564** (0.086)	***
GEMBA-DA	**+0.795** (0.079)	***	−0.097 (0.096)		**+1.230** (0.070)	***	−0.003 (0.085)		−0.154 (0.083)		**+0.526** (0.069)	***	−0.489 (0.070)	***	−0.196 (0.068)	**	**−0.688** (0.084)	***	**−0.653** (0.082)	***	**−0.784** (0.082)	***
GEMBA-MQM	**+0.513** (0.094)	***	−0.190 (0.116)		**+0.783** (0.084)	***	+0.168 (0.101)		+0.050 (0.098)		+0.151 (0.082)		−0.238 (0.084)	**	−0.157 (0.082)		**−0.512** (0.099)	***	**−0.334** (0.098)	***	**−0.461** (0.096)	***
Ref + MT only																						
COMET-noSrc	−0.254 (0.054)	***	−0.326 (0.067)	***	**+2.276** (0.048)	***	**+0.300** (0.058)	***	+0.057 (0.057)		+0.014 (0.047)		−0.438 (0.048)	***	−0.492 (0.047)	***	**−0.447** (0.057)	***	**−0.541** (0.057)	***	**−0.663** (0.055)	***
chrF	−0.356 (0.043)	***	−0.365 (0.053)	***	**+2.553** (0.038)	***	−0.089 (0.047)		−0.180 (0.045)	***	−0.523 (0.038)	***	−0.389 (0.039)	***	−0.330 (0.037)	***	**−0.349** (0.045)	***	**−0.355** (0.045)	***	**−0.415** (0.044)	***
BLEU	−0.273 (0.091)	***	−0.264 (0.119)	*	**+0.881** (0.084)	***	−0.189 (0.097)		−0.207 (0.096)	*	+0.062 (0.083)		−0.289 (0.082)	***	−0.120 (0.083)		−0.109 (0.098)		+0.502 (0.098)		**−0.230** (0.093)	*
MetricX-24-Ref	**+0.205** (0.073)	**	**+0.242** (0.090)	**	−1.523 (0.065)	***	**+0.701** (0.079)	***	**+0.386** (0.077)	***	**+1.149** (0.064)	***	−0.079 (0.065)		+0.009 (0.064)		**−0.241** (0.077)	***	**−0.176** (0.077)	*	**−0.439** (0.075)	***

Table 2: **Metric Sensitivity to Error Perturbations and Acceptable Variations.** LME estimates $\hat{\mu}_{m,e}$, fit jointly on Δ_{std}, the z-scored signed change in metric score between the original and modified translations, for each metric m and perturbation type e. Significance levels: * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). **Blue** = correct penalization (error columns) / tolerates variation (variation columns); **Orange** = failure (error columns) / over-penalizes variation (variation columns); no color = not significant. Metrics show variable sensitivity across error types, with blind spots for tense errors, title substitution, and modern-sense substitution. No single metric is reliable across all error categories, whereas acceptable variations are tolerated by nearly all metrics.

and acceptable perturbations (Section 5.2), before turning to temporal variation across texts from different historical periods (Section 5.3).

5.1 Metric Sensitivity to Errors

To study the sensitivity of metrics across error and acceptable-variation types (Table 1), we fit a single linear mixed-effects model predicting the difference in scores between an original translation and a perturbed version jointly over both error and acceptable-variations as described in Section 3.4. There was a significant interaction between error/variation type and metric $\chi^2(120) = 5{,}677.90$, $p < .001$, indicating that the effect of metric on δ^* differed across error and acceptable-variation types, as expected given the diverse nature of metrics considered (Section 4).

To understand the error sensitivity patterns per metric, and how they might impact scholarly use cases, we report per-metric LME estimates of $\hat{\mu}_{m,e}$ across the eight error types in Table 2. These values are the model intercept obtained by re-levelling each error, acceptable variation type and metric as the reference category in the main LME model (see Equation 1), such that $\hat{\mu}$ directly estimates the mean standardized score difference δ^* for metric m on error type e.

First, metric sensitivity to outputs in the wrong language varies widely. The reference-based metrics correctly penalize *unintended translation into modern Chinese*, except for MetricX24, which along with XCOMET-QE also fails in the reference-free version. Lack of sensitivity to language mismatch is a known weakness of metrics based on multilingual encoders (Zouhar et al., 2024; Lavie et al., 2025; Knowles et al., 2024), although interestingly COMET does not suffer from this issue when it has access to a reference in our settings.

Second, COMET, XCOMET, and MetricX24 variants adequately detect omitted content (subject and object). GEMBA variants and surface metrics such as chrF and BLEU are the main failure cases, incorrectly scoring hypotheses that drop an subject or object higher than the original translations. This behavior is particularly problematic when translations are

used to support distant reading practices, such as semantic search in English over historical texts where omitted content will lead to systematic recall errors.

Finally, sensitivity to disambiguation and substitution errors paints a more complex picture. No metric reliably detects *incorrect tense realization, and title or rank substitution* errors. For *modern-sense substitution*, MetricX24-Ref is the only metric that reliably detects the error. For *incorrect lexical translation*, XCOMET and MetricX24 variants join the GEMBA variants in reliably detecting the error, while COMET variants and surface metrics such as chrF and BLEU fail to penalize these mistranslations. For *pronoun substitution and misattribution*, MetricX24 variants detect it most strongly, while GEMBA only shows moderate sensitivity. Together, these results show that all metrics have blindspots when it comes to detecting errors that matter when interpreting Classical Chinese texts, including errors that distort meaning and might mislead readers who do not have the training to verify whether translations are supported by the source text.

Across error types, MetricX24 emerges as the metric that is most sensitive to errors overall, despite blind spots with respect to *unintended translations into modern Chinese*, *incorrect tense, title or rank substitution*. Notably, only the reference-based variant, MetricX24-Ref, reliably detects *modern-sense substitution*, while the base and quality-estimation variants show only weak signal for this error type. This suggests that the extensive use of synthetic data in training to improve its robustness to under- and over-translation between modern languages (Juraska et al., 2024) also improves generalization to an unseen source language compared to other metrics. Interestingly, this behavior is not limited to reference-based configurations where MetricX24 can directly compare outputs and references in English, a target language seen at training time: MetricX24 sensitivity patterns remain similar in the quality estimation mode where assessments are based on the source and translation hypothesis alone.

5.2 Metric Sensitivity to Acceptable Variations

The same joint model, signed δ^*, and re-levelling procedure described in section 5.1 were also used to generate estimates $\hat{\mu}_{m,v}$ for the three acceptable-variation types (Table 2).

Under this formulation, the interpretation of the sign is reversed relative to the error analysis. Because a good metric should not penalize a legitimate variation, a positive $\hat{\mu}$ indicates undesirable over-penalization, whereas a negative $\hat{\mu}$ indicates tolerance, the metric assigns the variation a quality score at least as high as the original translation.

Across all three variation types, $\hat{\mu}$ is negative for essentially every metric, reflecting broad agreement that these variations should not be penalized. All metrics tolerate *Chinese annotation* and *sentence segmentation*, while all except BLEU tolerate *name formatting*, making BLEU the only metric to penalize this variation. *Sentence segmentation* is the most tolerated variation type overall (mean $\hat{\mu} = -0.52$), suggesting that metrics generally rate re-segmented translations at least as favourably as the originals.

It is encouraging that most metrics are tolerant of acceptable variations, despite not having been explicitly trained to recognize them. The strong tolerance for sentence segmentation is particularly important for the evaluation of Classical Chinese translation, where sentence boundaries are often ambiguous and segmentation decisions frequently reflect scholarly interpretation rather than objective correctness. However, it remains to be seen whether this holds across a broader range of perturbations.

5.3 Metric Sensitivity Across Time Periods

To test whether metric sensitivity to errors differs across time periods, we fit a separate LME model where the time period in which the source text was written was added as a fixed effect, separating two levels: Pre-Qin and Han (1600BC - 220AD), and Post-Han (220AD-1912), since the language of Pre-Qin and Han texts is thought to have been not very different from cultured speech of the time, while the gap between the written and spoken language began to develop in the Han dynasty and increased with time thereafter (Peyraube, 2004).

	Error Perturbations														Acceptable Variations					
	Lex. error		Anc./Mod. sense		Mod. Ch. out.		Omit. obj.		Omit. subj.		Pronoun		Tense		Title		Ch. annot.		Name fmt.		Sent. seg.	
Metric	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.
Src + Ref + MT																						
COMET	−0.158 (0.064)	*	−0.288 (0.075)	***	**+2.005** (0.056)	***	**+0.398** (0.066)	***	+0.092 (0.067)		+0.037 (0.053)		−0.477 (0.056)	***	−0.499 (0.053)	***	**−0.530** (0.066)	***	**−0.527** (0.065)	***	**−0.715** (0.067)	***
XCOMET	**+0.291** (0.095)	**	−0.222 (0.114)		**+0.974** (0.084)	***	**+0.208** (0.098)	*	**+0.410** (0.099)	***	−0.032 (0.080)		−0.352 (0.084)	***	−0.203 (0.080)	**	**−0.606** (0.098)	***	**−0.310** (0.097)	**	**−0.444** (0.099)	***
MetricX-24	**+0.335** (0.082)	***	+0.153 (0.096)		−1.286 (0.072)	***	**+0.723** (0.086)	***	**+0.463** (0.087)	***	**+1.128** (0.068)	***	−0.153 (0.073)	*	−0.058 (0.068)		**−0.318** (0.085)	***	**−0.243** (0.084)	**	**−0.585** (0.086)	***
GEMBA-DA-ref	**+0.871** (0.086)	***	−0.114 (0.103)		**+1.014** (0.076)	***	+0.159 (0.089)		+0.002 (0.091)		**+0.319** (0.073)	***	−0.465 (0.076)	***	−0.184 (0.072)	**	**−0.730** (0.089)	***	**−0.563** (0.089)	***	**−0.719** (0.090)	***
GEMBA-MQM-ref	**+0.623** (0.090)	***	−0.154 (0.111)		**+0.942** (0.081)	***	+0.058 (0.092)		+0.011 (0.095)		**+0.157** (0.078)	*	−0.285 (0.080)	***	−0.107 (0.077)		**−0.546** (0.094)	***	**−0.415** (0.094)	***	**−0.700** (0.094)	***
Src + MT (QE)																						
XCOMET-QE	**+0.562** (0.101)	***	−0.158 (0.122)		−0.197 (0.090)	*	**+0.218** (0.105)	*	+0.277 (0.106)	**	+0.077 (0.086)		+0.073 (0.090)		−0.078 (0.085)		**−0.420** (0.106)	***	−0.157 (0.104)		−0.108 (0.105)	
MetricX-24-QE	**+0.265** (0.090)	**	+0.127 (0.105)		−0.976 (0.078)	***	**+0.632** (0.093)	***	**+0.494** (0.094)	***	**+1.063** (0.074)	***	−0.129 (0.079)		−0.128 (0.074)		**−0.487** (0.092)	***	−0.167 (0.091)		**−0.593** (0.093)	***
GEMBA-DA	**+0.791** (0.083)	***	−0.094 (0.097)		**+1.249** (0.073)	***	−0.010 (0.087)		−0.189 (0.087)	*	**+0.527** (0.069)	***	−0.495 (0.074)	***	−0.205 (0.069)	**	**−0.683** (0.086)	***	**−0.655** (0.084)	***	**−0.804** (0.087)	***
GEMBA-MQM	**+0.556** (0.099)	***	−0.051 (0.104)		**+0.727** (0.087)	***	+0.185 (0.103)		−0.254 (0.088)	**	+0.118 (0.083)		−0.150 (0.082)		−0.150 (0.082)		**−0.445** (0.102)	***	**−0.354** (0.100)	***	**−0.485** (0.103)	***
Ref + MT only																						
COMET-noSrc	−0.240 (0.056)	***	−0.326 (0.066)	***	**+2.218** (0.049)	***	**+0.307** (0.059)	***	+0.023 (0.059)		+0.013 (0.047)		−0.459 (0.050)	***	−0.496 (0.047)	***	**−0.491** (0.058)	***	**−0.535** (0.057)	***	**−0.695** (0.059)	***
chrF	−0.372 (0.037)	***	−0.365 (0.043)	***	**+2.372** (0.032)	***	−0.104 (0.039)	**	−0.185 (0.038)	***	−0.322 (0.031)	***	−0.400 (0.033)	***	−0.331 (0.031)	***	**−0.358** (0.037)	***	**−0.355** (0.038)	***	**−0.446** (0.038)	***
BLEU	−0.287 (0.088)	**	−0.256 (0.113)	*	**+0.685** (0.082)	***	−0.221 (0.089)	*	−0.214 (0.092)	*	+0.073 (0.079)		−0.278 (0.079)	***	−0.118 (0.079)		−0.179 (0.094)		**+0.532** (0.094)	***	**−0.296** (0.092)	**
MetricX-24-Ref	**+0.247** (0.078)	**	**+0.252** (0.092)	**	−1.544 (0.068)	***	**+0.690** (0.081)	***	**+0.373** (0.082)	***	**+1.147** (0.065)	***	−0.083 (0.069)		+0.007 (0.065)		**−0.220** (0.081)	**	**−0.172** (0.079)	*	**−0.436** (0.081)	***

Table 3: **Metric Sensitivity in Pre-Qin and Han Texts.** LME estimates $\hat{\mu}_{m,e}$, fit on Δ_{std} within Pre-Qin and Han texts only (1600BC–220AD), for each metric m and error/variation type e. Significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. **Blue** = correct penalization (error columns) / tolerates variation (variation columns); **Orange** = failure (error columns) / over-penalizes variation (variation columns); no color = not significant.

	Error Perturbations														Acceptable Variations					
	Lex. error		Anc./Mod. sense		Mod. Ch. out.		Omit. obj.		Omit. subj.		Pronoun		Tense		Title		Ch. annot.		Name fmt.		Sent. seg.	
Metric	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.	$\hat{\mu}$ (SE)	Sig.
Src + Ref + MT																						
COMET	−0.237 (0.218)		−0.167 (0.819)		**+2.924** (0.273)	***	+0.580 (0.383)		**+0.628** (0.237)	**	+0.266 (0.473)		−0.234 (0.223)		−0.250 (0.579)		−0.035 (0.300)		**−0.689** (0.315)	*	−0.355 (0.202)	
XCOMET	+0.137 (0.282)		−0.970 (0.986)		**+0.939** (0.329)	**	+0.263 (0.484)		**+0.785** (0.310)	*	−0.145 (0.569)		+0.112 (0.281)		−0.574 (0.697)		−0.624 (0.396)		−0.056 (0.439)		−0.110 (0.273)	
MetricX-24	−0.171 (0.187)		−0.391 (0.809)		−1.019 (0.270)	***	**+1.214** (0.457)	**	**+0.454** (0.200)	*	**+1.354** (0.467)	**	−0.309 (0.208)		+0.363 (0.572)		−0.015 (0.247)		−0.273 (0.251)		**−0.479** (0.181)	**
GEMBA-DA-ref	**+0.899** (0.260)	***	−0.014 (0.872)		**+1.591** (0.291)	***	−0.071 (0.444)		+0.116 (0.271)		**+1.308** (0.504)	**	−0.251 (0.261)		−0.548 (0.617)		−0.631 (0.352)		−0.484 (0.401)		**−0.866** (0.239)	***
GEMBA-MQM-ref	+0.401 (0.258)		−0.244 (0.892)		**+1.435** (0.297)	***	**+0.891** (0.440)	*	+0.118 (0.270)		+0.488 (0.515)		−0.188 (0.261)		−0.244 (0.631)		−0.182 (0.363)		−0.380 (0.374)		**−0.908** (0.235)	***
Src + MT (QE)																						
XCOMET-QE	+0.447 (0.278)		−0.401 (1.045)		−0.888 (0.349)	*	−0.059 (0.460)		+0.600 (0.330)		−0.868 (0.603)		−0.184 (0.298)		−0.496 (0.739)		**−1.082** (0.426)	*	−0.181 (0.465)		−0.109 (0.287)	
MetricX-24-QE	−0.159 (0.209)		−0.026 (0.749)		−0.794 (0.250)	**	+0.117 (0.361)		**+0.543** (0.227)	*	+0.814 (0.433)		−0.145 (0.226)		+0.394 (0.530)		−0.216 (0.279)		−0.239 (0.298)		**−0.438** (0.195)	*
GEMBA-DA	**+0.823** (0.255)	**	−0.319 (0.845)		**+0.990** (0.282)	***	+0.119 (0.422)		+0.119 (0.263)		+0.510 (0.488)		−0.429 (0.246)		+0.349 (0.597)		**−0.758** (0.325)	*	−0.632 (0.366)		**−0.667** (0.221)	**
GEMBA-MQM	+0.202 (0.256)		−0.510 (0.907)		**+1.467** (0.302)	***	−0.084 (0.459)		**+0.832** (0.271)	**	**+1.525** (0.524)	**	−0.126 (0.271)		−0.615 (0.641)		**−1.264** (0.354)	***	+0.015 (0.391)		−0.398 (0.247)	
Ref + MT only																						
COMET-noSrc	−0.358 (0.177)	*	−0.316 (0.635)		**+2.987** (0.212)	***	+0.211 (0.308)		+0.362 (0.195)		+0.062 (0.366)		−0.273 (0.183)		−0.214 (0.449)		+0.123 (0.242)		**−0.686** (0.266)	**	**−0.433** (0.166)	**
chrF	−0.237 (0.184)		−0.394 (0.611)		**+4.798** (0.204)	***	+0.201 (0.303)		−0.146 (0.182)		−0.362 (0.353)		−0.279 (0.181)		−0.270 (0.432)		−0.232 (0.236)		−0.361 (0.268)		−0.229 (0.167)	
BLEU	−0.168 (0.338)		−0.368 (1.122)		**+3.352** (0.374)	***	+0.467 (0.561)		−0.145 (0.353)		−0.106 (0.646)		−0.189 (0.316)		−0.316 (0.791)		+0.374 (0.444)		−0.198 (0.500)		+0.115 (0.310)	
MetricX-24-Ref	−0.099 (0.195)		−0.276 (0.697)		−1.272 (0.232)	***	**+1.092** (0.521)	*	**+0.446** (0.216)	*	**+1.230** (0.402)	**	−0.138 (0.245)		+0.146 (0.493)		−0.348 (0.278)		−0.185 (0.282)		**−0.502** (0.188)	**

Table 4: **Metric Sensitivity in Post-Han Texts.** LME estimates $\hat{\mu}_{m,e}$, fit on Δ_{std} within Post-Han texts only (220AD–1912), for each metric m and error/variation type e. Significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. **Blue** = correct penalization (error columns) / tolerates variation (variation columns); **Orange** = failure (error columns) / over-penalizes variation (variation columns); no color = not significant.

We found a significant metric × period interaction, $\chi^2(12) = 57.94$, $p = .001$, indicating that when the source text was written modulates how metrics respond to modifications of the translations. However, this interaction should be interpreted cautiously because Post-Han texts constitute only 6.4% of the dataset. Consequently, many errors are substantially larger in the Post-Han (Table 4) subset(omitted object, omitted subject, pronoun substitution, etc), and many effects that are significant in the larger Pre-Qin/Han (Table 3) sample lose significance despite similar point estimates. Nevertheless, sensitivity to *unintended translation into modern Chinese* remains significant for every metric in both periods.

6 Conclusion

This work examined whether current machine translation evaluation metrics can detect the kinds of errors that matter for Classical Chinese–English translation. We find that neural metrics substantially outperform surface-overlap baselines, but their performance remains uneven across error categories. While some perturbations are detected reliably, sensitivity to omissions and meaning-altering substitutions is inconsistent. MetricX24 performs best overall, but it too exhibits important blind spots. Overall, no single metric reliably captures the full range of distinctions relevant to scholarly evaluation.

These findings suggest that progress in Classical Chinese translation evaluation depends not only on better metrics, but also on better evaluation resources. Metrics are least reliable on the subtle semantic distinctions that matter most for scholarly interpretation, yet such distinctions are often underrepresented in conventional benchmarks. Developing scholar-informed evaluation data that systematically targets these phenomena is therefore a necessary step toward more reliable automatic assessment.

Beyond Classical Chinese, we view the framework itself as a central contribution. Synthetic minimal pairs grounded in a scholar-informed error taxonomy provide a scalable way to convert expert judgment into evaluation signals while minimizing large-scale annotation requirements. Such perturbations can support both evaluation and the training of metrics that better capture domain-specific notions of translation quality, offering a practical interface between scholarly expertise and NLP methodology in low-resource and historically complex language settings.

Acknowledgments

This project was funded in part by the Artificial Intelligence Interdisciplinary Institute at Maryland (AIM). We thank the members of the CLIP Lab at the University of Maryland for their valuable feedback and support throughout this project. We are especially grateful to Calvin Bao, Kartik Ravisankar, HyoJung Han, and Dayeon Ki for their early feedback on the project and an earlier draft of this paper. We also thank the Chinese Text Project, whose openly available corpus of pre-modern Chinese texts we drew on for this work.

References

Eric R. Bennett, HyoJung Han, Xinchen Yang, Andrew Schonebaum, and Marine Carpuat. Evaluating evaluation metrics for Ancient Chinese to English machine translation. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, Marco C. Passarotti, and Rachele Sprugnoli (eds.), *Proceedings of the Second Workshop on Ancient Language Processing*, pp. 71–76, The Albuquerque Convention Center, Laguna, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-235-0. doi: 10.18653/v1/2025.alp-1.9. URL https://aclanthology.org/2025.alp-1.9/.

Eleftheria Briakou, Colin Cherry, and George Foster. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability, May 2023.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4196–4210, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.243. URL https://aclanthology.org/2024.findings-emnlp.243/.

Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. C^3bench: A comprehensive classical chinese understanding benchmark for large language models, 2024b. URL https://arxiv.org/abs/2405.17732.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33019–33036, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1678. URL https://aclanthology.org/2025.emnlp-main.1678/.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. *arXiv preprint arXiv:2407.07087*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.),

Proceedings of the Eighth Conference on Machine Translation, pp. 1066–1083, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100.

Team GLM, Aohan Zeng, Bin Xu et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach to evaluating machine translation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2486–2496, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1263. URL https://aclanthology.org/D17-1263/.

Kai Jin, Dan Zhao, and Wuying Liu. Morphological and semantic evaluation of Ancient Chinese machine translation. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti (eds.), *Proceedings of the Ancient Language Processing Workshop*, pp. 96–102, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.alp-1.11/.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.35. URL https://aclanthology.org/2024.wmt-1.35/.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9540–9561, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.649. URL https://aclanthology.org/2022.emnlp-main.649/.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 475–491, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.34. URL https://aclanthology.org/2024.wmt-1.34/.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64/.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen et al. (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023b. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Tom Kocmi and Christian Federmann. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Mary Nurminen, Judith Brenner, Maarit Koponen et al. (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023c. European Association for Machine Translation.

Alon Lavie, Greg Hanneman, Sweta Agrawal et al. Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 436–483, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.24.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11260–11285, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.671.

John M. MacKenzie. Alexander chow (ed.). 2021. scottish missions to china: Commemorating the legacy of james legge (1815–1897). *Studies in World Christianity*, 30(3):399–400, 2024. doi: 10.3366/swc.2024.0486. URL https://doi.org/10.3366/swc.2024.0486.

Sebastian Nehrdich and Kurt Keutzer. Mitra: A large-scale parallel corpus and multilingual pretrained language model for machine translation and semantic retrieval for pāli, sanskrit, buddhist chinese, and tibetan, 2026. URL https://arxiv.org/abs/2601.06400.

Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, and Kurt Keutzer. MITRA-zh: An efficient, open machine translation solution for buddhist Chinese. In Mika Hämäläinen, Emily Öhman, Flammie Pirinen, Khalid Alnajjar, So Miyagawa, Yuri Bizzoni, Niko Partanen, and Jack Rueter (eds.), *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pp. 266–277, Tokyo, Japan, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.nlp4dh-1.29/.

Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, and Kurt Keutzer. MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics. In Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar (eds.), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 129–137, Albuquerque, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-234-3. doi: 10.18653/v1/2025.nlp4dh-1.12.

OpenAI, Josh Achiam, Steven Adler et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002a.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002b. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Alain Peyraube. Ancient Chinese. In Roger D. Woodard (ed.), *The Cambridge Encyclopedia of the World's Ancient Languages*, pp. 988–1014. Cambridge University Press, Cambridge, 2004.

Maja Popovic. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pp. 392–395, 2015.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213.

Andrew Schonebaum, Anthony George, David Lattimore et al. *Introduction to Classical Chinese*. February 2024.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. Translating hanja historical documents to contemporary Korean and English. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1260–1272, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.91. URL https://aclanthology.org/2022.findings-emnlp.91/.

Seyoung Song, Haneul Yoo, Jiho Jin, Kyunghyun Cho, and Alice Oh. Heritage: An end-to-end web platform for processing korean historical documents in hanja, 2025. URL https://arxiv.org/abs/2501.11951.

Donald Sturgeon. Chinese text project: a dynamic digital library of premodern chinese. *Digital Scholarship in the Humanities*, 2019. doi: 10.1093/llc/fqy032. URL https://ctext.org.

Donald Sturgeon. Chinese Text Project: A dynamic digital library of premodern Chinese. *Digital Scholarship in the Humanities*, 36(Supplement_1):i101–i112, June 2021. ISSN 2055-7671. doi: 10.1093/llc/fqz046.

Anthropic Team. The claude 3 model family: Opus, sonnet, haiku. URL https://api.semanticscholar.org/CorpusID:268232499.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. LLM-based machine translation and summarization for Latin. In Rachele Sprugnoli and Marco Passarotti (eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pp. 122–128, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lt4hala-1.15/.

Hao Wang, Hirofumi Shimizu, and Daisuke Kawahara. Kanbun-LM: Reading and Translating Classical Chinese in Japanese Methods by Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8589–8601, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.545.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

Aiyuan Yang, Bin Xiao, Bingning Wang et al. Baichuan 2: Open large-scale language models, 2025. URL https://arxiv.org/abs/2309.10305.

An Yang, Baosong Yang, Binyuan Hui et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Mu You, Derek Wong, Jing Zhang, and Kaixin Lan. How well can state-of-the-art machine translation systems render a 16th-century chinese novel? *Cadernos de Tradução*, 45:1–24, 09 2025. doi: 10.5007/2175-7968.2025.e108394.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. WYWEB: A NLP Evaluation Benchmark For Classical Chinese. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3294–3319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.204.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. Pitfalls and outlooks in using COMET. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1272–1288, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.121. URL https://aclanthology.org/2024.wmt-1.121/.

A Prompt Used for Example Generation

All LLM-generated perturbations were produced using GPT-4o-mini. We selected this model on the grounds of cost efficiency and reproducibility: GPT-4o-mini offers sufficient instruction-following capability for the constrained perturbation tasks defined in Section 3.2, injecting a single targeted error into a translation.

Prompt used for OMITTED SUBJECT

Prompt:

Given the following Classical Chinese text: "{source_segment}"
And its English translation: "{machine_translation_text}"
Create a perturbed version of the English translation that contains an OMITTED SUBJECT error.
In Classical Chinese, subjects are often omitted when context makes them clear. However, when translating to English, these subjects must be explicitly stated. An omitted subject error occurs when the translator fails to properly identify or include the subject, making the English translation ambiguous or unclear.

The perturbed version should:

1. Remove or make ambiguous the subject of at least one clause or sentence
2. Create confusion about who or what is performing the action
3. Maintain the overall structure and most of the vocabulary of the original reference including capitalization, spacing, punctuation, and chinese characters if present in the translation
4. Be a realistic error that could occur in translation

Example:

- Original: "The Master said, 'The gentleman studies virtue.'"
- Perturbed: "Said, 'The gentleman studies.'" (subject omitted)

IMPORTANT: You must respond with ONLY valid JSON in this exact format: "{perturbed_translation}": "the perturbed English translation with omitted subject",
"{error_description}": "brief description of how the subject was omitted or made unclear"
Do not include any text before or after the JSON. Only return the JSON object.

Prompt used for OMITTED OBJECT

Prompt:

Given the following Classical Chinese text: "{source_segment}"
And its English translation: "{machine_translation_text}"
Create a perturbed version of the English translation that contains an OMITTED OBJECT error.
In Classical Chinese, objects can be omitted when context makes them clear. When translating to English, these objects must often be explicitly stated for clarity. An omitted object error occurs when the translator fails to properly identify or include the direct or indirect object, making the English translation incomplete or ambiguous.

The perturbed version should:

1. Remove or make ambiguous the object of at least one verb
2. Create confusion about what is being acted upon
3. Maintain the overall structure and most of the vocabulary of the original translation including capitalization, spacing, punctuation, and chinese characters if present in the translation etc.
4. Be a realistic error that could occur in translation

Example:

- Original: "The Master taught the students virtue."
- Perturbed: "The Master taught virtue." (object 'students' omitted)

IMPORTANT: You must respond with ONLY valid JSON in this exact format: "{perturbed_translation}": "the perturbed English translation with omitted object",
"{error_description}": "brief description of how the object was omitted or made unclear"
Do not include any text before or after the JSON. Only return the JSON object.

Prompt used for INCORRECT LEXICAL TRANSLATION

Prompt:

Given the following Classical Chinese source: "{source_segment}"
And its English translation: "{machine_translation_text}"
Create a perturbed version of the English translation that contains an ACCURACY ERROR.
Accuracy errors involve factual or semantic inaccuracies in translation. These can include:

- Incorrect translation of specific terms of concepts
- Misrepresentation of quantities, measurements, or relationships
- Wrong interpretation of causal or logical connections
- Factual errors about people, places, or events
- Semantic shifts that change the core meaning

The perturbed version should:

1. Introduce a factual or semantic inaccuracy while maintaining plausible structure
2. Change the meaning subtly enough to seem like an honest mistake
3. Preserve most of the original wording and structure (including capitalization, spacing, punctuation, and chinese characters if present in the translation etc.)
4. Create a realistic translation error that could mislead readers

Example:

- Original: "He ruled for thirty years and established peace."
- Perturbed: "He ruled for three years and established peace." (quantity error)

IMPORTANT: You must respond with ONLY valid JSON in this exact format: "{perturbed_translation}": "the perturbed English translation with accuracy error",
"{error_description}": "brief description of the specific accuracy error introduced"
Do not include any text before or after the JSON. Only return the JSON object.

Prompt used for INCORRECT TENSE REALIZATION

Prompt:

Given the following Classical Chinese text: "{source_segment}"
And its English translation: "{machine_translation_text}"
Create a perturbed version of the English translation that contains a TENSE ERROR.
Classical Chinese has different temporal markers than English, and tense must often be inferred from context. Tense errors occur when the translator uses incorrect verb tense (past, present, future).

The perturbed version should:

1. Change the tense of at least two verbs to an incorrect tense
2. Maintain the overall structure and vocabulary

Example:

- Original: "The king ruled wisely and his people prospered."
- Perturbed: "The king rules wisely and his people prospered." (tense inconsistency)

IMPORTANT: You must respond with ONLY valid JSON in this exact format: "{perturbed_translation}": "the perturbed English translation with tense error",
"{error_description}": "brief description of the specific tense error introduced"
Do not include any text before or after the JSON. Only return the JSON object.

Prompt used for UNINTENDED TRANSLATION INTO MODERN CHINESE

Prompt:

Translate the following "{english_machine_translation}"[a] to Chinese. Put your translation between $$ signs (e.g., $$your translation here$$)

[a] English machine translation of the Classical Chinese original

Prompt used for MODERN CHINESE SENSE SUBSTITUTION

Prompt:

Given the following Classical Chinese source: "{source_segment}"
And its English translation: "{machine_translation_text}"
The Classical Chinese text contains the character "{character}" which has different meanings:

- Ancient meaning: "{ancient_meaning}"
- Modern meaning: "{modern_meaning}"

Analyze the translation and identify where the ancient meaning of "{character}" appears. Then create a perturbed version of the translation by replacing the ancient meaning with the modern meaning.

The perturbed version should:

1. Replace the ancient meaning with the modern meaning naturally
2. Maintain the overall structure and flow of the original translation
3. Preserve the meaning of the rest of the sentence

Example:

- If character "君" has ancient meaning "gentleman" and modern meaning "monarch"
- And translation contains "The gentleman has wine"
- Then perturbed version should be "The monarch has wine"

IMPORTANT: You must respond with ONLY valid JSON in this exact format: "{perturbed_translation}": "the translation with ancient meaning replaced by modern meaning",
"{character_found}": true/false,
"{ancient_usage}": "the specific ancient meaning usage found in translation",
"{modern_replacement}": "the modern meaning replacement made",
"{explanation}": "brief explanation of the change made"
Do not include any text before or after the JSON. Only return the JSON object.

B Metrics Implementation Details

Metric	# Params	Languages	Source
string-based metrics			
BLEU	–	any	SacreBLEU
ChrF	–	any	SacreBLEU
learned neural metrics			
COMET	580M	94	Unbabel/wmt22-comet-da
COMET-REF-ONLY	580M	94	Unbabel/wmt22-comet-da
XCOMET	3.5B	94	Unbabel/XCOMET-XL
XCOMET-QE	3.5B	94	Unbabel/XCOMET-XL
MetricX-24	3.7B	49	google/metricx-24-hybrid-xl-v2p6
MetricX-24-REF-ONLY	3.7B	49	google/metricx-24-hybrid-xl-v2p6
MetricX-24-QE	3.7B	49	google/metricx-24-hybrid-xl-v2p6
LLM-as-a-judge			
GEMBA-DA	–	–	gpt-5-mini
GEMBA-DA-ref	–	–	gpt-5-mini
GEMBA-MQM	–	–	gpt-5-mini
GEMBA-MQM-ref	–	–	gpt-5-mini

Table 5: Metrics used in the evaluation. The **Source** column links to the metric implementation/model card.

C Memorization Test

Text	R-1 (R)	R-2 (R)	R-L (R)
Classical Chinese (source)	33.94	9.40	27.62
English (reference)	19.61	3.04	13.31

Table 6: **ROUGE Recall scores for memorization testing using the CTP dataset**. Low overlap of text completion with reference suggests that there is no evidence of memorization.

D Inter-Annotator Agreement

Section	Category	N	A1	A2	Agr.	Inv.	Tot. Agr.	Agr. Rate	Fail Rate	κ
Perturbations	Incorrect lexical translation	112	100	102	96	6	102	86%	5%	0.496
Perturbations	Unintended translation into modern Chinese	137	128	127	121	3	124	88%	2%	0.265
Perturbations	Omitted objects	168	100	106	82	44	126	49%	26%	0.474
Perturbations	Omitted subjects	180	100	102	87	65	152	48%	36%	0.684
Perturbations	Incorrect tense realization	137	127	127	120	3	123	88%	2%	0.245
Perturbations	Pronoun substitution and misattribution	136	130	131	126	1	127	93%	1%	0.148
Perturbations	Substitution of classical senses for modern meaning	192	97	76	64	83	147	33%	43%	0.532
Perturbations	Title substitution	136	133	127	127	3	130	93%	2%	0.483
Acceptable Variations	Sentence segmentation	132	99	115	91	9	100	69%	7%	0.229
Acceptable Variations	Chinese annotation	134	112	97	86	11	97	64%	8%	0.210
Acceptable Variations	Name formatting	156	99	96	88	49	137	56%	31%	0.740
Perturbations	**Perturbations total**	**1198**	**915**	**898**	**823**	**208**	**1031**	**69%**	**17%**	**0.622**
Acceptable Variations	**Acceptable Variations total**	**422**	**310**	**308**	**265**	**69**	**334**	**63%**	**16%**	**0.468**
Overall	**Overall**	**1620**	**1225**	**1206**	**1088**	**277**	**1365**	**67%**	**17%**	**0.580**

Table 7: **Inter-annotator agreement statistics** by category, including raw annotator counts, agreement rates, failed perturbation rates, and Cohen's κ. A1/A2 = Annotator 1/2, Agr. = Agreed, Inv. = Both invalid.

E Baseline Accuracy

Metric	Unrelated	Shuffled
Src + Ref + MT		
COMET	92.2	99.7
XCOMET	96.7	97.5
MetricX-24	96.1	100.0
GEMBA-DA-ref	99.9	99.9
GEMBA-MQM-ref	88.5	88.5
Src + MT (QE)		
XCOMET-QE	95.0	97.1
MetricX-24-QE	94.2	99.8
GEMBA-DA	99.9	99.8
GEMBA-MQM	87.6	86.2
Ref + MT only		
COMET-noSrc	90.8	98.8
chrF	77.6	80.7
BLEU	79.8	57.8
MetricX-24-Ref	95.0	100.0

Table 8: **Baseline accuracy (%).** Proportion of pairs where the metric correctly ranks the MT output above the Unrelated and Shuffled baselines.