A Reproducible Human-Ai Collaborative Workflow for Quality Assurance in Domain-Specific Translation: A Case Study Of "Luban Workshop/EPIP" Discourse

Zheng Yunrong[1]*

[1]School of Foreign Languages, Tianjin University of Technology and Education, Tianjin, China
**zhengyunrong@126.com*

Keywords: domain-specific translation, reproducible evaluation, terminology alignment, human-in-the-loop (HITL) optimization, quality assurance, Chinese discourse internationalization

Abstract

Large language models (LLMs) face significant challenges in domain-specific translation, particularly for culturally loaded terminology. This study proposes a reproducible human-AI collaborative workflow for quality assurance in domain-specific translation, emphasizing human-in-the-loop optimization over black-box generation. Using the "Luban Workshop/EPIP" discourse as a case study, we constructed a gold-standard dataset comprising 540 parallel sentence pairs and 76 core terms (including standard translations and variants). A multi-dimensional evaluation pipeline was designed, integrating SacreBLEU, BERTScore, and a LaBSE-based Term Alignment Module (TAM) to systematically diagnose translation errors. Based on diagnostic feedback, we implemented a two-stage intervention: terminology forcing followed by context-aware post-editing. Experimental results with three LLMs (GPT-5.1, Kimi 2.5, Qwen 3) demonstrated consistent improvements: average Term Alignment F1 (TAF1) increased by 9.6 percentage points (a relative improvement of 27%), while SacreBLEU and BERTScore remained stable. Human evaluation confirmed that TAF1 correlated moderately yet significantly with expert judgments on terminological consistency ($r = 0.397$, $p < 0.001$), validating TAM's diagnostic validity. This work offers a reproducible, interpretable alternative to both end-to-end NMT systems and autonomous agent paradigms for domain-specific translation, balancing accuracy against auditability in high-stakes discourse internationalization.

1 Introduction

China is committed to channeling its cutting-edge academic achievements into authoritative international platforms, fostering deeper academic exchanges (Wu & Pan, 2024). As a flagship program promoting China's vocational education globally, the Luban Workshop, not only transfers technical expertise but also embodies Chinese educational philosophies, teaching models, and cultural values. Its core teaching model, EPIP (Engineering Practice Innovation Project), transforms China's localized educational concepts and technical terminology into an academic discourse system that is both distinctly Chinese and internationally accessible. This system, characterized by the principles of "engineering-driven, practice-oriented, innovation-targeted, and project-based," integrates Chinese vocational education philosophy with global engineering education paradigms. The accurate translation of such discourse is therefore crucial for enhancing international understanding and fostering educational collaboration.

However, translating culturally rich and specialized terms like those associated with the Luban Workshop/EPIP presents a dual challenge: transcoding an original discourse system and achieving equivalent effect for culture-loaded concepts. While the advancement of large language models (LLMs) and artificial intelligence (AI) has catalyzed a paradigm shift in natural language processing (NLP) (Vaswani et al., 2017), their performance in translation remains insufficiently evaluated through systematic and reproducible frameworks. Existing research tends to cluster around general-purpose texts or relies heavily on isolated automated metrics, which fail to comprehensively reveal the limitations of AI tools in domain-specific translation tasks. Consequently, the development of a systematic workflow for quantitative evaluation and guided improvement of such translation tasks constitutes an urgent priority.

To address this gap, the present study proposes a human-AI collaborative workflow tailored to domain-specific discourse translation, emphasizing human-in-the-loop optimization over autonomous black-box generation. This workflow integrates a multi-dimensional reproducible evaluation framework that identifies systematic biases in AI-generated translations through quantitative metrics, subsequently informing dynamic optimization pathways. Through an empirical investigation of representative Luban Workshop/EPIP corpora, this research seeks to establish a verifiable and transferable quality assurance workflow for domain-specific translation.

1.1 Research question

How can we design a reproducible, human-AI collaborative workflow that

(1) systematically diagnoses AI translation errors in domain-specific discourse via multi-dimensional evaluation, and

(2) automatically optimizes the output based on diagnostic feedback through controlled, interpretable intervention?

1.2 Our work and core contributions

To address these challenges, this study proposes and implements a reproducible, human-AI collaborative workflow

for evaluating and optimizing domain-specific translation, with the "Luban Workshop/EPIP" discourse as a case study.

(1) **Domain-specific data curation:** A gold-standard dataset comprising 540 parallel sentence pairs and 76 core terms with validated **translations** (details in §3.1).

(2) **Reproducible multi-dimensional evaluation pipeline:** Human experts curated the gold-standard terminology set and validated alignment thresholds. The pipeline integrates SacreBLEU, BERTScore, and a LaBSE-based Term Alignment Module (TAM) for comprehensive quality assurance and structured diagnostic reporting.

(3) **Empirical analysis of mainstream LLMs:** By benchmarking against professional human translations, we systematically assessed three representative LLMs (GPT-5.1, Kimi 2.5, Qwen 3). The analysis revealed systematic deficiencies in domain-specific translation, particularly in terminological consistency and cultural concept transfer. Following terminology forcing, Term Alignment F1 (TAF1) improved significantly across all systems (GPT: +13.2 pp; Kimi: +8.1 pp; Qwen: +7.6 pp), while SacreBLEU and BERTScore remained stable.

(4) **Two-stage correction workflow:** We propose and validate a "terminology forcing + context-aware lightweight rewriting" paradigm. This diagnosis-driven intervention targets identified terminology errors through HITL optimization, eschewing end-to-end regeneration in favor of minimal, context-sensitive adjustments.

2 Related Work and Theoretical Foundation

This study aims to develop a reproducible workflow for evaluating and optimizing domain-specific translation. Using the "Luban Workshop/EPIP" discourse as a case study, we seek to enhance translation quality for culturally loaded discourse. We organize related work around three core dimensions: domain-adaptive translation (addressing what to translate), translation quality evaluation (addressing how to evaluate), and human-AI collaborative translation (addressing how to improve). Existing research has yet to achieve a systematic integration of these three dimensions, particularly fine-grained quantitative diagnosis and targeted intervention for culturally loaded terminology and indigenous discourse systems. This section concludes by articulating the theoretical foundation of this study: the paradigmatic distinction between human-in-the-loop optimization and autonomous generation.

2.1 Domain-adaptive translation

The strong instruction-following and contextual understanding capabilities of LLMs have enabled "generative optimization" to supersede traditional corrective post-editing. Researchers have begun exploring how prompt engineering can guide LLMs to automatically refine initial outputs (Raunak et al., 2023). However, existing LLM optimization strategies still face the interpretability bottleneck of "black-box generation" (Palikhe et al., 2025). Most approaches adhere to an end-to-end generation paradigm, lacking the "diagnosis-intervention" closed-loop mechanism analogous to human translators. While this paradigm ensures inference efficiency, its uncontrollability in processing highly specialized or domain-specific discourse—such as "Luban Workshop/EPIP"—introduces potential translation risks, failing to meet stringent quality assurance requirements (Yuksel et al., 2025). This uncontrollability has created an urgent need for diagnosis-driven evaluation.

2.2 Translation quality evaluation

Quality evaluation for AI translation faces the challenge of "metric misalignment"—the disconnect between automated metrics and domain-specific quality requirements. Traditional metrics were originally developed for neural machine translation (NMT): n-gram-based metrics (e.g., BLEU) and character-based metrics (e.g., chrF++) focus on surface-form matching (Callison-Burch et al., 2006; Freitag et al., 2022); BERTScore (Zhang et al., 2019), though capable of capturing semantic similarity through pre-trained language model embeddings, fails to specifically quantify term alignment or cultural concept transfer.

Recent work has explored LLM-as-judge approaches, leveraging LLMs for automated quality assessment (Zerva et al., 2022). However, these approaches often lack reproducibility and structured diagnostic capacity—capabilities essential for human-in-the-loop optimization.

While recent advances in AI agents promise autonomous translation optimization through dynamic tool use and multi-step reasoning (Yuksel et al., 2025), these "agentic" approaches often bypass explicit diagnostic steps, leading to challenges in determinism and accountability. While prior work has explored constrained decoding for terminology enforcement in NMT (Dinu et al., 2019; Post & Vilar, 2018), systematic diagnosis-driven intervention for LLM-based translation remains under-explored, despite growing interest in evaluating LLM translation capabilities (Hendy et al., 2023).

2.3 Human-AI collaborative translation and theoretical foundation

Human-AI collaborative translation centers on workflow optimization that integrates AI efficiency with human expertise. Technologically, collaborative paradigms have evolved from machine translation post-editing (MTPE) to interactive machine translation (IMT) (Knowles & Koehn, 2016), adhering to the core logic of "humans correcting machine output" (Stahlberg, 2020). However, LLMs are reshaping this paradigm: they have shifted from "error generators" awaiting correction to "agents" capable of understanding, reasoning, and generation. This bifurcation has yielded two distinct paradigms: autonomous generation, which leverages dynamic tool use and multi-step reasoning for end-to-end optimization; and human-in-the-loop optimization, which this study advocates—emphasizing human intervention and audit at critical decision points to guide targeted refinement through interpretable diagnostic feedback (Fernandes et al., 2022).

This study is grounded in the paradigmatic distinction between human-in-the-loop and autonomous optimization. Autonomous optimization, exemplified by end-to-end NMT and emerging AI agents, prioritizes efficiency through black-box learning. However, for domain-specific translation requiring accurate conveyance of academic discourse power,

human-in-the-loop optimization—maintaining human oversight at critical decision points—remains essential (O'Hagan, 2020). Our framework emphasizes human irreplaceability at these junctures: experts curate terminology sets, validate alignment thresholds (τ), and audit optimized outputs—distinct from fully autonomous agent paradigms.

These three junctures—terminology curation, threshold validation, and output audit—operationalize the principles of decomposability, intervention specificity, and auditability (Santoni de Sio & Van den Hoven, 2018), to be experimentally validated in §3.4.

3. Research Methodology and Experimental Design

This chapter operationalizes the three design principles of HITL optimization: §§3.1–3.2 detail the data curation and technical configuration underlying **decomposability** (i.e., diagnostic evaluation); §3.3 presents the algorithmic implementation of **intervention specificity** (i.e., targeted correction); and §3.4 outlines the experimental validation of **auditability** (i.e., traceable quality assurance).

3.1 Data resource construction and corpus selection

3.1.1 Gold Standard Data Resources: Based on authoritative publication-grade sources—including Ministry of Education white papers, Luban Workshop monographs, and EPIP literature—we constructed two core resources:

• **Chinese-English parallel corpus:** 540 sentence pairs, automatically aligned and manually validated for strict semantic correspondence.

• **Gold-standard terminology set:** We employed a two-round expert annotation protocol. In Round 1 (open-set nomination), two domain experts independently reviewed source corpora and nominated critical terms with suggested translations. Expert A nominated 66 terms, Expert B 64; merging and deduplication yielded 76 candidates. Given the free-response nature of the task, traditional agreement coefficients (e.g., Kappa, Alpha) were inapplicable. We assessed inter-expert consensus via set similarity: nomination overlap was 71% (54/76), with Dice coefficient of 83% (108/130). For the 52 jointly nominated terms, core translation agreement was 67% (35/52, allowing synonymous variants). In Round 2 (adjudication), experts discussed all 76 terms—including 22 singly nominated and 17 discrepant—to establish gold-standard translations for each entry (see Appendix A).

3.1.2 Core Experimental Corpus Selection: To ensure corpus representativeness, we adopted a two-stage strategy combining purposive sampling with expert verification:

(1) **Purposive Consecutive Passage Sampling**: Vocational education domain experts selected candidate passages from three sub-domains—Chinese vocational education policy documents, Luban Workshop, and EPIP—based on discourse functions (policy statements vs. theoretical expositions). This approach balanced evaluation representativeness with the global coherence of discourse-level translation.

(2) **Data Partitioning**: The 540 parallel sentence pairs selected by experts were randomly divided into an evaluation set (432 sentences, 80%) and a validation set (108 sentences, 20%). The validation set was used exclusively for threshold sensitivity analysis of terminological consistency evaluation, while the test set was strictly held out for final evaluation and never accessed during parameter tuning. This separation eliminates data leakage and ensures the validity of all reported results.

3.2 Engineering implementation of the evaluation pipeline

3.2.1 Technical Configuration of Three-Dimensional Metrics:

Table 1 Technical Configuration of Three-Dimensional Metrics

Dimension	Tool/Model	Key Parameters	Limitations
Lexical Surface	SacreBLEU	Tokenizer=zh/13a, n-gram order=1-4	Oversensitive to exact term matching; fails to recognize semantically equivalent yet differently phrased term translations.
Semantic Depth	BERTScore	Multilingual BERT-base, IDF weighting	Focuses on sentence-level semantic similarity; lacks explicit term alignment measurement.
Term Alignment	TAF1 (Ours)	τ= 0.65 (optimized on validation set); LaBSE layer 6, L2-normalized representations; variant-aware matching	Computational cost higher than traditional metrics; relies on predefined terminology bank.

Note: TAF1 = Term Alignment F1; IDF = Inverse Document Frequency.

TAF1 is defined as the harmonic mean of term precision and recall, calculated on the extracted terminology set based on the alignment decisions from TAM (see §3.2.2 for details).

The three-dimensional metrics—SacreBLEU (lexical surface), BERTScore (semantic depth), and TAF1 (term alignment)—form complementary coverage at the lexical, semantic, and domain-knowledge levels. Integrated evaluation ensures comprehensive assessment of both general translation quality and domain specificity. Notably, LaBSE's 6th layer encodes cross-lingual lexical representations particularly effective for aligning short terms, achieving an AER (Alignment Error Rate) of 18.8%—significantly outperforming mBERT (22.3%) and XLM-R (27.9%) (Wang et al., 2022). TAF1, the primary metric for terminological consistency, operationalizes the alignment decisions of TAM (detailed in §3.2.2) and is computed as the F1 score over gold-standard terms with threshold $\tau = 0.65$. This tripartite integration ensures evaluation coverage of general quality while highlighting domain-specific characteristics.

3.2.2 TAM Implementation Details and Threshold Optimization: TAM employs a rigorous variant-aware alignment strategy to ensure reproducibility and interpretability in term evaluation. Unlike conventional approaches that directly encode source-language terms for

cross-lingual matching, TAM constructs canonical term representations based on predefined English equivalents and their authorized variants.

Term Representation Encoding. For each Chinese term t in the gold-standard terminology set, we retrieve its gold-standard English translation e_{main} along with variant forms $\{e_1,e_2,\ldots,e_n\}$ from the terminology database. All variants are encoded using LaBSE to form a multi-vector representation set:

$$E_t = \{ LaBSE(e_{main}), LaBSE(e_1), \ldots, LaBSE(e_n)\}$$

Similarity Computation. Given an AI-generated translation s, we compute its maximum semantic similarity to term t as:

$$sim(s,t) = \max_{e \in E\_t} \cos(LaBSE(s), LaBSE(e))$$

where all vectors represent L2-normalized embeddings of the LaBSE [CLS] token. This "max-over-variants" design accommodates legitimate terminological diversity (e.g., "Jing Shang" vs. "Canon I") without penalizing valid alternative translations.

Alignment Decision. The alignment indicator function **align**(·,·) is defined as:

$$align(s,t) = 1[sim(s,t) \geq \tau]$$

Threshold optimization was conducted on the validation set (n = 108). Analysis revealed that system-specific optimal thresholds ranged between 0.6 and 0.7. To balance performance across the three systems and ensure evaluation stability, this study adopted a unified threshold of τ = 0.65 for all subsequent evaluations. This threshold effectively controls false positives while constraining F1 loss to below 3% for each system. All final results reported in §4 are based on the test set (n = 432), which was never used during threshold optimization. Figure 1 illustrates the impact of threshold selection on term alignment performance; detailed data are provided in Appendix B.

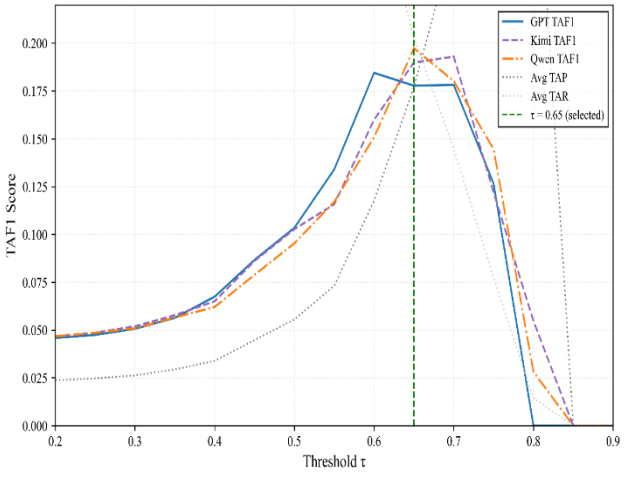

Figure 1 Sensitivity Analysis of Term Alignment Threshold (τ)

3.2.3 Human Validation Design: To validate TAM's automated assessment, we conducted a blind human evaluation on a stratified sample of 90 sentences (20.8%) drawn from the test set, ensuring coverage across the three AI systems, both pre- and post-optimization conditions, and the three sub-domains. The sample was allocated as follows: 3 systems × 2 conditions × 3 sub-domains × 5 sentences = 90 sentences.

Validation Objective: This procedure aims to validate TAM's automated assessment by verifying the concordance between TAF1 scores and expert judgments, thereby establishing criterion-related validity evidence for subsequent automated evaluations.

- **Evaluators:** Two MA students in Translation Studies with domain expertise in vocational education.
- **Dimensions:** Fluency (1–5), fidelity (1–5), and terminological consistency (binary: 0/1).
- **Inter-rater Reliability:** Cohen's κ (terminological consistency); ICC (fluency and fidelity).
- **Validity Assessment:** Pearson correlation between TAF1 and expert terminological judgments.

Sample Size Justification: A sample of 90 sentences achieves statistical power ≥ 0.80 for detecting a medium effect size (Cohen's d ≥ 0.2) using paired-samples t-tests (α = 0.05, two-tailed). Critically, this human validation set is disjoint from the validation set (n = 108) used for TAM threshold optimization (§3.2.2), ensuring validation independence.

3.3 Engineering implementation of intervention optimization

3.3.1 Phase 1: Terminology Forcing (Deterministic Rule-Based System): Based on the diagnostic report, the system first executes terminology forcing: for each core term flagged as "omission" or "mistranslation," fuzzy matching is performed to locate the term in the translation. If an approximate expression is identified, it gets replaced with the gold-standard term; if the term is completely omitted, it is inserted according to grammatical rules. All modified positions are marked for reference in Phase 2. This process intervenes only with the 76 core terms to avoid over-correction.

3.3.2 Phase 2: Context-Aware Rewriting (Controlled Generation):

- **Model:** Large language models were accessed via APIs (this study employed GPT-5.1-2026-01-09, Kimi2.5-thinking-2026-01-09, and qwen3-max-2026-01-09; temperature = 0.3, top_p = 0.9).
- **Prompt Design:** Prompts were constructed to anchor the model to the provided terminology markers and source context. The model was instructed to integrate the fixed terms into natural syntactic structures, modifying only the minimal necessary segments to preserve overall fluency and meaning.
- **Generation Control:** Maximum output length is constrained to ≤120% of source text length, ensuring lightweight rewriting.

Each optimization decision was logged with the diagnostic signal source, intervention type, and pre-/post-intervention metric comparisons, creating a detailed log of intervention types and metric shifts. This transparency ensures the workflow adheres to the principle of auditability. (Implementation code is available in Appendix C.)

3.4 Experimental design and hypothesis

3.4.1 Research Hypotheses: To operationalize the three design principles of HITL optimization (§2.3), we formulate three research hypotheses (Table 2). H1 validates the diagnostic fidelity of our evaluation pipeline, H2 tests the efficacy of the intervention mechanism, and H3 tests whether the auditable intervention maintains general translation quality. Together, these hypotheses address RQ1 (systematic error diagnosis) and RQ2 (HITL optimization).

Table 2 Research hypotheses and validation framework

Hypothesis	Validation Logic (Operationalization)	Theoretical Grounding
H1 (Diagnostic validity): If the proposed Term Alignment Module (TAM) validly decomposes and measures terminological consistency, then its score (TAF1) will show a significant positive correlation with expert human judgment.	Target: Diagnostic accuracy. Method: Pearson correlation between TAF1 and expert ratings on a 90-sentence sample. Success Criterion: r>0.30, p<0.001.	§3.2.2 (Threshold optimization on validation set).
H2 (Intervention efficacy): If the controlled two-stage intervention (terminology forcing + context-aware rewriting) is effective, then it will yield a statistically significant improvement in terminological accuracy (TAF1) compared to its own unoptimized baseline.	Target: Intervention effectiveness. Method: Paired-sample t-test on TAF1 scores before vs. after optimization (n=432). Success Criterion: Δ>7% (average improvement across systems), p<0.001, Cohen's d>0.2.	§2.3 (Intervention Specificity); §3.3 (Two-stage intervention design).
H3 (Stability of General Translation Quality): If the workflow satisfies the auditability principle, then the HITL optimization will improve terminological accuracy (TAF1) significantly while maintaining or improving general translation quality (SacreBLEU, BERTScore).	Target: Stability of general translation quality. Method: Paired t-tests on TAF1, SacreBLEU, and BERTScore. Success Criterion: TAF1 improves significantly (p<0.001); SacreBLEU and BERTScore do not significantly decrease (p>0.05 for one-tailed test of decrease).	§2.3 (Auditability); §3.4 (Decision-point logging).

3.4.2 Comparative Systems and Experimental Settings:

Table 3 Performance Comparison of Translation Workflows Across Different LLM Backbones

System	Description	Purpose	Corresponding Hypothesis
GPT-5.1	Direct inference	Baseline for unoptimized LLM performance	—
GPT-5.1	Chain-of-thought prompting	Baseline for autonomous optimization paradigm (§2.3)	H2 (contrast)
Kimi-2.5-32k	Direct inference	Cross-system validation of decomposability	H1
Qwen-3-72B	Direct inference	Cross-system validation of intervention transferability	H2
Ours (HITL optimization)	Application of the two-stage intervention to each of the three base LLMs (GPT-5.1, Kimi-2.5, Qwen-3)	Framework validation: controlled vs. autonomous paradigm	H1, H2, H3

Control Conditions: A unified prompt template was used, with only the source text provided—without access to the terminology bank or reference translations—to simulate real-world usage scenarios.

Human Decision Points: To validate the principle of "meaningful human control" (§2.3), expert supervision was introduced at three critical junctures:

- **Threshold Setting:** The unified threshold $\tau = 0.65$ was determined through sensitivity analysis on the validation set (§3.2.2) and expert-validated (approx. 2 hours).
- **Terminology Curation:** The 76 core terms underwent two rounds of expert review (approx. 8 hours).
- **Output Audit:** A random sample of 10% of optimized outputs (43 sentences) underwent expert blind review (approx. 3 hours).

Total human intervention cost per project: approximately 13 hours, ensuring auditability at critical decision points.

3.4.3 Statistical Protocol : The statistical design follows reporting standards in psychology and computational linguistics (American Psychological Association, 2020; Brysbaert, 2019), with analyses operationalizing the three hypotheses in Table 2: H1 (diagnostic validity) via Pearson correlation between TAF1 and expert ratings; H2 (intervention efficacy) via paired-sample t-tests on pre- and post-optimization TAF1 scores; and H3 (stability) via multivariate paired t-tests with Bonferroni correction.

Sample Size Justification: The test set (n = 432) provides statistical power >0.80 to detect small-to-medium effect sizes ($d \geq 0.2$) using paired-samples t-tests; the human validation set (n = 90) provides power >0.90 to detect moderate correlations ($r \geq 0.30$). The validation set (n = 108, §3.2.2) is independent from main hypothesis testing and used solely for threshold optimization.

Multiple Comparison Correction: Holm-Bonferroni correction was applied to five core comparisons (pre-post optimization for three models, plus autonomous optimization contrast). The significance level for primary comparisons (our method vs. each baseline) was set at $\alpha = 0.05$; for secondary comparisons (between models) at $\alpha = 0.0125$. All p-values reported in Table 3 are corrected.

Quality Control: All statistical tests were preceded by normality checks (Shapiro-Wilk) and outlier detection (Tukey's method, 1.5×IQR). No missing data were present.

Reproducibility: All analyses were implemented in Python 3.10 using SciPy 1.11 and Pingouin 0.5. Full code and data are available as supplementary materials (see Appendix).

4 Experimental Results and Quality Analysis

4.1 Baseline evaluation and systematic error diagnosis

4.1.1 Baseline Performance of Three-Dimensional Metrics:

Table 4 Baseline Performance Comparison of Three Models (n = 432)

System	TAF1	SacreBLEU	BERTScore
GPT-5.1	0.3289	24.95	0.8850
Kimi-2.5	0.3880	25.07	0.8706
Qwen-3	0.3519	17.45	0.8522
Mean	0.3563	22.49	0.8693

Note: TAF1 was computed with $\tau = 0.65$.

Key Finding: All three models achieved BERTScores exceeding 0.85, yet averaged only TAF1 0.3563, revealing systematic unreliability in current LLMs for translating culturally loaded terminology—approximately 74% of core terms exhibited drift or omission. These results support the central argument of this study: standard sentence-level metrics (SacreBLEU and BERTScore) mask critical terminological inaccuracies. This underscores the need for term-specific evaluation tools when assessing domain-specific translations.

4.1.2 Analysis of Terminological Consistency Deficiencies:

Table 5 Taxonomy of Terminology Errors

Error Type	Typical Manifestation	Example
Terminological Drift	Near-synonym substitution for gold-standard term	"三谛" rendered as "Three Truths"
Mistranslation	Literal rendering or conceptual distortion	"衍" rendered as "transformation"

Diagnostic Conclusion: Current LLMs exhibit systematic terminological instability (mean recall = 0.26), establishing the need for intervention.

4.2 Intervention effects of HITL optimization

4.2.1 Main Effects and Cross-Model Generalization:

Table 6 Pre- and Post-Optimization Comparison (n = 432)

Metric	System	Initial	+Ours	Δ	Cohen's *d*	*p*
TAF1	GPT-5.1	0.3289	0.4610	+0.132	0.3292	$2.69E^{-11}$
	Kimi-2.5	0.3880	0.4688	+0.0808	0.2567	$1.54E^{-07}$
	Qwen-3	0.3519	0.4282	+0.0763	0.2272	$3.17E^{-06}$
	Mean	0.3563	0.4527	+0.0964	—	—
SacreBLEU	Mean	22.49	22.74	+0.25	—	—
BERTScore	Mean	0.8693	0.8703	+0.0010	—	—

Note: SacreBLEU and BERTScore showed small standard deviations; hence only means are reported. Model-specific TAF1 differences are discussed in §5.1.

Hypothesis Validation:

- **H2:** All three models showed significant TAF1 improvements post-optimization (GPT-5.1: $\Delta = +0.132$, $p = 2.69 \times 10^{-11}$; Kimi-2.5: $\Delta = +0.0808$, $p = 1.54 \times 10^{-7}$; Qwen-3: $\Delta = +0.0763$, $p = 3.17 \times 10^{-6}$), with Cohen's d effect sizes ranging from 0.23 to 0.33—indicating small-to-medium effects that confirm intervention efficacy.
- **H3:** Significant TAF1 gains ($p < 0.001$) were achieved with no compromise to general translation quality, as evidenced by stable SacreBLEU (mean $\Delta = +0.25$) and BERTScore (mean $\Delta = +0.0010$) scores, thereby satisfying the principle of overall translation quality stability.
- **Generalizability:** The workflow yielded consistent gains across three architecturally different models (GPT-5.1, Kimi K2.5, and Qwen3-Max), indicating that the approach is robust and not tied to any specific LLM architecture.

4.3 Human validation and TAM validity

Following standard practices in meta-evaluation of MT metrics (Graham et al., 2013; Ma et al., 2019), we assessed inter-annotator reliability and criterion-related validity. Reliability was substantial for terminological judgments (Cohen's $\kappa = 0.82$, 95% CI [0.74, 0.90]) and good for fluency and fidelity (ICC = 0.79, 95% CI [0.70, 0.86]). Table 7 presents validity correlations between automated metrics and human ratings.

Table 7 Criterion-Related Validity

Comparison	Pearson *r*	95% CI	*p*
TAF1 vs. human term judgment	0.397	[0.207, 0.558]	<0.001
BERTScore vs. human fidelity	0.408	[0.219, 0.567]	<0.001
SacreBLEU vs. human fluency	0.092	[-0.118, 0.293]	0.390

H1 Validation: TAF1 correlated significantly with human terminological judgments ($r = 0.397$, $p < 0.001$), validating TAM's diagnostic validity. This moderate correlation supports TAM's practical utility. In contrast, SacreBLEU showed no significant correlation with human fluency ($r = 0.092$, $p = 0.390$), underscoring the necessity of terminologically dedicated metrics.

4.4 Failure case analysis and boundary conditions

Table 8 Illustrative Failure Cases

Case	Source Term	Initial Translation	Optimized Translation	Failure Type
1	名	So-called, name	So-called, Name	Insufficient contextual integration
2	三谛	Achieving the "three truths" realm through "action" to generate the harmony realm is the unity of knowledge and action.	Achieving the "Three Realms" realm through "Action" to generate the harmony realm is the Unity of Knowledge and Action.	Absence of discourse-level optimization

Theoretical Reflection: The "conservative" optimization observed in Case 2 does not represent a technical error but rather a deliberate application of the Intervention Specificity principle—the system modifies only identified terminological deficiencies, avoiding the introduction of new errors. This design ensures optimization auditability but also exposes the

boundary of this principle: when the source sentence itself exhibits structural awkwardness, isolated term replacement cannot resolve holistic quality issues. Future work may introduce lightweight global rewriting modules—while preserving auditability—to achieve a balance between terminological accuracy and discourse-level fluency.

4.5 Summary

(1) **RQ1 (Systematic Diagnosis)**: TAM effectively identifies terminological errors in LLM outputs. The LaBSE-based Term Alignment Metric (TAF1) demonstrates a moderately strong, significant correlation with human judgments (r = 0.397, p< 0.001), confirming its diagnostic validity. Baseline evaluation reveals a striking discrepancy: while the three models achieved a mean BERTScore of 0.869, their average TAF1 was merely 0.356—exposing a substantial "masking effect" whereby general-purpose semantic metrics obscure terminological errors (H1 supported).

(2) **RQ2 (HITL Optimization)**: The two-stage intervention significantly improves terminological consistency, with TAF1 increasing by an average of 0.096 (a relative improvement of approximately 27%). Gains are statistically significant across all systems (p < 0.001), with small-to-medium effect sizes (Cohen's d= 0.23–0.33). Crucially, SacreBLEU and BERTScore remain stable despite these terminological enhancements, thereby validating both the principle of Intervention Specificity (H2) and the Stability of General Translation Quality (H3).

(3) **Boundary Awareness**: Failure case analysis (Table 8) exposes the framework's current limitations in compound term recognition and deep cultural transcoding. These boundary conditions furnish empirical targets for refinement, grounding the improvement directions elaborated in Chapter 5.

5 Discussion

5.1 Theoretical implications of findings

The HITL optimization paradigm proposed in this study offers an alternative pathway for human-AI collaboration in domain-specific discourse translation. Unlike end-to-end autonomous generation paradigms, its core value lies in shifting quality control upstream—from post-hoc outcome inspection to in-process intervention. Experimental results demonstrate that the diagnosis-intervention closed loop, grounded in multi-dimensional evaluation, significantly improves terminological consistency (absolute mean improvement: 0.096; relative gain: 27%) while maintaining stable SacreBLEU and BERTScore values. This validates the principle of Intervention Specificity (H2) and confirms the Stability of General Translation Quality (H3).

The necessity of dimensional decomposition receives compelling empirical support. Despite GPT-5.1's impressive BERTScore (0.885), its TAF1 of merely 0.329 exposes the "masking effect" whereby general semantic metrics obscure terminological inconsistencies. This resonates with Fernandes et al.'s (2023) diagnosis of metric misalignment, yet our study advances this critique: decoupled diagnosis of terminological and semantic layers constitutes a prerequisite for identifying systematic bias. Traditional MT evaluation conflates terminological errors with general lexical errors, yielding ambiguous optimization directives; by contrast, TAM's explicit alignment judgments enable "terminology forcing," thereby achieving a paradigm shift from "detecting errors" to "locating errors."

The moderate correlation between TAM and human terminological judgments (r = 0.397, p< 0.001) carries significant methodological implications. This result confirms that semantic similarity based on LaBSE's 6th-layer embeddings can, to a certain extent, approximate expert judgments of terminological consistency. Given the inherent divergence between semantic similarity and human form-based criteria, this validity level is sufficient to support TAM's practical utility. Notably, BERTScore exhibits a moderate correlation with human fidelity judgments (r = 0.408), whereas SacreBLEU shows no significant correlation with human fluency ratings (r = 0.092, p = 0.390)—further underscoring the necessity of terminology-specific metrics for targeted evaluation dimensions.

Threshold sensitivity analysis (Figure 1) delineates the boundaries of embedding-based alignment strategies: at τ = 0.40, TAF1 achieves a recall of 0.78 for short terms (1-2 words), yet recall plummets to 0.21 for long, compound terms such as "工程实践创新项目" (Engineering Practice Innovation Project). This disparity indicates that while LaBSE's 6th-layer cross-lingual lexical representations perform robustly for short terms, they exhibit structural limitations when processing compound term formations prevalent in Chinese institutional discourse. Future research may explore sub-term decomposition or hierarchical alignment mechanisms to enhance recognition rates for long, multi-word terms.

5.2 Engaging with existing research

The "two-stage intervention" design of this study addresses the efficiency-controllability tension inherent in human-AI collaborative translation. While Fomicheva et al. (2022) emphasize "meaningful human control," they do not operationalize actionable intervention points; conversely, AI agent solutions pursue end-to-end automation at the expense of auditability. The middle-ground approach adopted in this framework—Terminology Forcing (deterministic rule-based system) and Context-Aware Rewriting (controlled generation)—retains human experts at three critical decision points: Threshold Setting, Terminology Curation, and Output Audit. Achieving an average TAF1 improvement of 27% (Cohen's d = 0.23–0.33) within approximately 13 hours of expert labor per project, this design demonstrates both the principle of Intervention Specificity and the principle of Auditability, validating the practical feasibility of these design principles.

This architecture holds particular value for the international dissemination of academic discourse. Translating Luban Workshop/EPIP discourse involves not merely linguistic transfer but the cross-cultural transcoding of educational philosophies and institutional cultures. Fully autonomous AI generation risks conceptual distortion through terminological drift, which can compromise the accurate transmission of

domain-specific concepts. By contrast, the enforced alignment mechanism of this framework ensures systematic consistency of core concepts, satisfying O'Hagan's (2020) accountability requirements for high-stakes translation scenarios.

5.3 Operational boundaries and future directions

The proposed framework is subject to three operational boundaries, which concurrently delineate optimization pathways.

Technical Limitations. LaBSE's n-gram sensitivity constrains recall of long compound terms (e.g., "工程实践创新项目" / Engineering Practice Innovation Project), as illustrated by Case 2 in Table 8. While "三谛" was correctly replaced, the sentence's verbose, repetitive structure remained unaddressed—revealing the deliberate trade-off of the Intervention Specificity principle: targeted intervention averts error introduction yet cannot resolve inherent syntactic deficiencies. Future research may investigate hierarchical TAM architectures that decompose compounds into sub-lexical units (e.g., "工程" + "实践" + "创新" + "项目") for sub-unit alignment, or adaptive thresholding mechanisms that dynamically adjust τ as a function of term length.

Register Adaptation Gap. The framework demonstrates efficacy in contextual reconstruction of vocational education terminology; however, adaptation to formal registers of policy discourse remains suboptimal. Certain optimized outputs exhibit colloquialisms, indicating a need for domain-adaptive fine-tuning of the rewriting model—for instance, incorporating register-preserving constraints (e.g., "maintain formal register") into prompts or deploying domain-specific language models at the rewriting stage.

Cost Constraints. Terminology curation relies on domain experts (~8 hours), engendering deployment latency for rapidly evolving fields. A hybrid workflow integrating automated terminology expansion with expert adjudication may mitigate manual costs: LLMs extract candidate terms and variants from domain corpora, with experts subsequently validating selections. For discourse-level deficiencies exemplified by Case 2 in Table 8, lightweight syntactic reconfiguration modules could be introduced to enhance readability while preserving terminological accuracy.

5.4 Limitations

This study is subject to several limitations. First, although the human-validated samples (90 sentences) meet the requirements for statistical power, effect size estimates may exhibit sampling variation; future research should expand the sample to verify result robustness. Second, while the correlation between TAF1 and human terminological judgments ($r = 0.397$) achieves statistical significance, it falls short of conventional thresholds for a strong correlation (e.g., $r > 0.5$), suggesting the need to further optimize semantic alignment strategies or introduce composite metrics. Third, the optimization effectiveness of this framework has been validated on only three LLMs; its generalizability to other models (e.g., Claude, ERNIE) remains to be examined in future studies. A further limitation of our experimental design is that the baseline models did not have access to the same terminology set during inference. This may have contributed to part of the performance gain. Future work should compare the two-stage intervention against a baseline that is also provided with the term bank but without the controlled rewriting logic, to isolate the effect of the intervention itself.

This study should be interpreted as a pilot investigation for the translation of Luban Workshop/EPIP. The modest sample size reflects the inherent challenges of data collection: the limited availability of authoritative English translations for EPIP texts currently published primarily in Chinese, which precludes the establishment of reliable gold-standard references for terminological validation. Utilizing source materials available only in Chinese would introduce unverifiable ambiguity into human judgments, compromising the validity of our criterion measures. While our sample meets statistical power requirements for detecting medium effect sizes, future work should validate on larger EPIP corpora (e.g., > 2000 pairs) as more reference data become available.

6 Conclusion

Using the Luban Workshop/EPIP discourse as a critical case study, this research has developed and validated a human-AI collaborative workflow for domain-specific translation. Its core contributions constitute a threefold methodological innovation.

Evaluative Dimension Innovation. The proposed TAM addresses a critical gap in term-level evaluation for vocational education translation. Complementing SacreBLEU and BERTScore, TAM enables systematic diagnosis of terminological precision, semantic fidelity, and cultural appropriateness. Criterion-related validity analysis confirms its diagnostic validity: TAF1 demonstrates a moderate yet significant correlation with human terminological judgments ($r = 0.397$, $p < 0.001$).

Intervention Mechanism Innovation. The two-stage design—Terminology Forcing + Context-Aware Rewriting—translates diagnostic signals into auditable optimization instructions, striking an empirically validated balance between efficiency and controllability. Experiments demonstrate an absolute TAF1 improvement of 0.096 (relative gain: 27%), statistically significant across all systems ($p < 0.001$), with stable SacreBLEU and BERTScore values.

Governance Model Innovation. The three-decision-point architecture (threshold setting, terminology curation, output audit) ensures substantive human intervention—distinct from fully autonomous AI agents—offering an interpretable and accountable alternative for domain translation. A total human labor cost of approximately 13 hours per project validates the feasibility of the "meaningful human control" principle.

Empirical evidence confirms that this workflow significantly enhances translation quality within acceptable human effort, with full traceability of the optimization trajectory. Failure cases in Table 8 further indicate refinement opportunities in contextual integration and discourse-level reconstruction—

these constitute specific entry points for subsequent algorithmic optimization.

As AI agent technologies evolve, achieving a dynamic balance between maintaining controllability and reducing upfront labor costs—particularly between automating terminology curation and ensuring precise core-term alignment—will remain central to this research agenda. The open-source implementation of this workflow (code and data provided in the Appendix) furnishes a reproducible platform for advancing this exploration.

References

[1] American Psychological Association (2020). *Publication manual* of *the American Psychological Association* (7th ed.). Washington, DC: American Psychological Association. https://doi.org/10.1037/0000165-000

[2] Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 16. https://doi.org/10.5334/joc.72

[3] Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249–256). Association for Computational Linguistics. https://aclanthology.org/E06-1032/

[4] Dinu, G., Mathur, P., Federico, M., & Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (pp. 3063–3068). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1294

[5] Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F. T., Neubig, G., Garg, A., Clark, J. H., Freitag, M., & Firat, O. (2023). The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)* (pp. 1066–1083). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.100/

[6] Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., & Bojar, O. (2022). Results of the WMT22 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation* (WMT) (pp. 89–109). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.2/

[7] Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. https://aclanthology.org/W13-2305/

[8] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*. https://doi.org/10.48550/arXiv.2302.09210

[9] Knowles, R., & Koehn, P. (2016). Neural interactive translation prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas: MT Researchers' Track* (pp. 107–120). Association for Machine Translation in the Americas. https://aclanthology.org/2016.amta-researchers.9/

[10] Ma, Q., Wei, J., Bojar, O., & Graham, Y. (2019). Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1), 62–90. Association for Computational Linguistics. https://aclanthology.org/W19-5302/

[11] O'Hagan, M. (2020). Translation and technology: Disruptive entanglement of human and machine. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 1–18). Routledge. https://doi.org/10.4324/9781315311258

[12] Palikhe, A., Wang, Z., Yin, Z., Guo, R., Duan, Q., Yang, J., & Zhang, W. (2025). Towards Transparent AI: A Survey on Explainable Language Models. *arXiv preprint arXiv:2509.21631*. https://doi.org/10.48550/arXiv.2509.21631

[13] Post, M., & Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)* (Vol. 1, pp. 1314–1324). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1119

[14] Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12009–12024). https://doi.org/10.18653/v1/2023.findings-emnlp.804

[15] Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

[16] Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69, 343-418. https://doi.org/10.1613/jair.1.12007.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 6000–6010). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1706.03762

[18] Wang, W., Chen, G., Wang, H., Han, Y., & Chen, Y. (2022). Multilingual Sentence Transformer as A Multilingual

Word Aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 2952–2963), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.215

[19] Wu, Y., & Pan, L. (2024). Chinese vocational education internationalization: The Luban Workshop model. *Chinese Education & Society, 56*(2), 89--105.

[20] Yuksel, K. A., Ferreira, T. C., Al-Badrashiny, M., & Sawaf, H. (2025). A Multi-AI Agent System for Autonomous Optimization of Agentic AI Solutions via Iterative Refinement and LLM-Driven Feedback Loops. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)* (pp. 52–62). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.realm-1.4

[21] Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., de Souza, J. G. C., Eger, S., Kanojia, D., Alves, D., Orăsan, C., Fomicheva, M., Martins, A. F. T., & Specia, L. (2022). Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 69–99). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.3

[22] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. OpenReview. https://openreview.net/forum?id=SkeHuCVFDr