SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 2351-2361
Open Access. Online & Print

DOI: 10.5281/zenodo.122.126183

LOST IN CULTURAL PRAGMATICS? A MULTIDIMENSIONAL BENCHMARKING OF MACHINE TRANSLATION FOR CHINESE-ENGLISH EDUCATIONAL TERMINOLOGY

Wanyi Liu[1*]

[1]University of Warwick, CV4 7AL, Coventry, United Kingdom, Liuwanyiedu@163.com

Received: 10/11/2025
Accepted: 29/12/2025

Corresponding Author: Wanyi Liu (Liuwanyiedu@163.com)

ABSTRACT

The internationalization of Sino-UK higher education has intensified demand for accurate translation of culturally embedded educational terminology. Yet existing machine translation (MT) systems often reduce such terms to literal equivalents, neglecting their cultural, pragmatic, and institutional dimensions. Current evaluation frameworks predominantly focus on semantic fidelity and fluency, leaving institutional accuracy largely unexamined. To address this gap, this study introduces the CPI (Cultural, Pragmatic, Institutional) framework as a multidimensional model for assessing translation adequacy. Using a stratified corpus of 5,000 bilingual sentence pairs drawn from UNESCO policy documents and Chinese university curricula (2020-2024), we benchmarked MT systems (DeepL, GPT-4o) against human translation (HT). Quantitative results based on repeated-measures ANOVA show that while MT systems approach HT in semantic accuracy (DeepL=4.1, GPT-4o=4.3, HT=4.8), they perform markedly worse in pragmatic (2.9-3.2 vs. 4.5) and institutional dimensions (3.0-3.1 vs. 4.6, $p<0.001$). Qualitative coding further revealed recurrent mistranslations of high-risk terms such as 核心素养 and 双减政策. This study makes three contributions: operationalizing institutional accuracy, producing an annotated glossary of 9 high-risk terms, and proposing a hybrid workflow that integrates MT efficiency with expert-informed post-editing. The findings highlight the necessity of culturally and institutionally sensitive translation for equitable cross-cultural educational policy communication.

KEYWORDS: Machine Translation (MT), Chinese–English Educational Terminology, CPI Framework, Cultural Fidelity, Institutional Accuracy, Cross-Cultural Policy.

1. INTRODUCTION

The internationalization of higher education between China and the United Kingdom has accelerated significantly in recent decades. According to UNESCO statistics, more than 150,000 Chinese students were enrolled in British universities in 2023, making China the largest source of international students in the UK[1]. This mobility has been accompanied by institutional collaborations such as the UK–China Education Partnership, which emphasize curriculum reform, joint degree programs, and graduate employability[2]. In such contexts, translation serves not merely as a linguistic tool but as a crucial medium of cross-cultural policy communication and institutional coordination. Accurate rendering of educational terminology is therefore indispensable for sustaining bilateral cooperation and preventing policy misalignment[3].

Despite its importance, fundamental mismatches persist in the way key educational terms are rendered across Chinese and English[4]. Terms such as **核心素养** (héxīn sùyǎng, core competencies) are frequently mistranslated as "core literacy," reducing the concept to basic reading and writing skills while eliminating its moral, civic, and political connotations central to China's curriculum reform. Similarly, 应试教育 (yìngshì jiàoyù) is often translated as "exam-oriented education," a neutral phrase that fails to convey its critical undertones associated with systemic pressure and rote learning practices around the Gaokao. These culturally entrenched mistranslations obscure the ideological foundations of Chinese education policy and risk distorting pedagogical debates in international discourse[5]. The challenge is further intensified by the widespread reliance on machine translation (MT) systems, which, although capable of producing fluent outputs, tend to reproduce and magnify these cultural and institutional misalignments[6].

Educational terminology constitutes a particularly high-risk domain for translation because it is deeply embedded in cultural values, pragmatic norms, and institutional frameworks[7]. Unlike technical terms in natural sciences, which often allow for relatively direct cross-linguistic mapping, educational terms are hybrid constructs encoding ideological assumptions, policy mechanisms, and communicative rituals. For instance, expressions such as 请多多指教 (qǐng duōduō zhǐjiào, a ritual request for guidance) are often rendered literally as "please advise me a lot," a phrase that violates politeness conventions and professional register in English academic contexts. Institutional concepts such as **保研** (bǎoyán, guaranteed postgraduate admission for high-performing undergraduates) have no direct equivalent in English, leading to confusion or oversimplification when inadequately translated. Errors of this kind extend far beyond lexical inaccuracies: they directly affect cross-cultural policy alignment, institutional trust, and the perceived legitimacy of reforms.

The rapid adoption of MT tools such as Google Translate, DeepL, and GPT-4o raises pressing questions about their adequacy in this domain[8]. While neural MT systems have achieved remarkable progress in semantic alignment and surface fluency, their ability to convey cultural fidelity, pragmatic appropriateness, and institutional accuracy remains uncertain. Policymakers, academics, and international organizations are increasingly exposed to translations that are linguistically smooth but socio-culturally misleading, thereby jeopardizing meaningful dialogue and informed decision-making.

Against this backdrop, the present study introduces the CPI framework, a multidimensional evaluation model integrating cultural fidelity, pragmatic appropriateness, and institutional accuracy. Drawing on a stratified corpus of 5,000 bilingual sentence pairs extracted from UNESCO policy documents and leading Chinese university curricula, the study benchmarks MT performance against human translation (HT). Three research questions guide the inquiry: (1) How do MT systems perform across the CPI dimensions when translating Chinese–English educational terminology? (2) What types of errors dominate in the translation of high-risk terms? (3) How can hybrid translation workflows, combining MT with expert-informed post-editing, mitigate these issues and improve translation adequacy?

The contributions of this research are threefold. Theoretically, it extends MT evaluation beyond surface fluency by operationalizing a multidimensional framework. Methodologically, it applies a mixed-methods design that combines quantitative benchmarking with qualitative error analysis. Practically, it produces an annotated glossary of high-risk educational terms and proposes a hybrid translation workflow designed to enhance the reliability of cross-cultural policy communication. Together, these contributions aim not only to improve translation accuracy but also to foster greater equity and clarity in global educational dialogue.

2. LITERATURE REVIEW

The translation of educational terminology occupies a unique and highly contested space within translation studies. Unlike technical or scientific terminology, which often enjoys relatively stable cross-linguistic mappings, educational terminology carries ideological, pragmatic, and institutional weight[9]. This complexity has generated a diverse body of scholarship, ranging from classical theories of translation to more recent evaluations of machine translation (MT) and human translation (HT) in educational contexts[10]. The following review situates the present study within this evolving literature, emphasizing the limitations of existing approaches and the necessity of a multidimensional framework that incorporates cultural, pragmatic, and institutional dimensions.

2.1. Terminology Translation Theories

Foundational theories of translation have long debated the tension between fidelity to the source text and adaptability to the target audience. Newmark's distinction between semantic translation and communicative translation provides one of the most influential formulations. Semantic translation emphasizes preserving the original meaning and formal structure, while communicative translation privileges naturalness and readability for the target audience[11]. While useful, this binary has proven insufficient for domains such as education, where terms function simultaneously as linguistic units and as carriers of cultural and institutional meaning[12]. For example, rendering **核心素养** (héxīn sùyǎng) merely as "core literacy" satisfies semantic equivalence at a lexical level but fails communicatively, as it misrepresents the ideological intent of China's curriculum reforms.

Vermeer's Skopos theory further advanced translation studies by shifting the focus from linguistic equivalence to translation purpose[13]. According to this framework, the appropriateness of a translation depends on its intended function for the target audience, whether to inform, persuade, or formalize. In educational contexts, Skopos theory enables translators to adapt terms such as 素质教育 (sùzhì jiàoyù) not as a literal "quality education" but as a policy-driven "holistic education reform," aligning with the communicative function of policy discourse[14]. However, Skopos theory also risks excessive adaptation: by privileging purpose over fidelity, translations may drift ideologically, thereby obscuring the political and institutional stakes of the original terminology.

This tension gives rise to a trilemma in educational terminology translation: the need to preserve semantic precision, cultural intelligibility, and institutional specificity simultaneously[15]. Achieving semantic precision alone may reduce cultural nuance; prioritizing cultural intelligibility risks diluting institutional mechanisms; and emphasizing institutional accuracy may compromise readability[16]. The trilemma underscores why existing frameworks often falter when applied to education, where terminology is inseparable from national policy agendas and sociocultural ideologies.

2.2. Machine And Human Translation In Educational Contexts

The advent of neural machine translation has reshaped debates on translation adequacy. MT systems such as Google Translate, DeepL, and GPT-4o achieve remarkable fluency and speed, offering scalable solutions for policy communication and academic exchange[17]. Studies demonstrate that MT models now achieve near-human performance on standard semantic metrics, with BLEU scores approaching those of professional translators in general domains[18]. Yet research also indicates that educational discourse is particularly vulnerable to pragmatic borrowing, whereby source-language speech acts are transferred verbatim into the target language without adaptation[19]. Expressions like 请多多指教 (qǐng duōduō zhǐjiào), when rendered as "please advise me a lot," illustrate how MT fails to recognize culturally specific politeness strategies, resulting in translations that are linguistically correct but pragmatically inappropriate.

By contrast, HT remains more adept at capturing cultural nuance and pragmatic appropriateness. Professional translators often employ glossing strategies or contextual annotations to render institution-specific terms such as **保研** (bǎoyán) in ways that preserve both semantic accuracy and institutional meaning, for instance, as "guaranteed postgraduate admission for high-achieving undergraduates." Empirical studies confirm that HT systematically outperforms MT in conveying institutional intent and cultural fidelity[20]. However, HT suffers from scalability challenges, particularly in large-scale policy communication where multilingual audiences and time-sensitive dissemination are involved.

In response to MT's limitations, several evaluation models have been proposed. CSI-Match benchmarks cultural awareness by detecting whether cultural elements are preserved in translation, but it does not

account for the transmission of institutional mechanisms[21]. PTA evaluates pragmatic transfer, focusing on speech act recognition and appropriateness, but it lacks measures for institutional accuracy[22]. Both frameworks make important contributions, yet they remain partial: CSI-Match captures cultural fidelity, PTA addresses pragmatic appropriateness, but neither systematically evaluates how well translations convey the institutional systems embedded in terms such as 双减政策 (shuāng jiǎn zhèngcè, Double Reduction Policy).

The present study responds to this limitation by introducing the CPI framework, which explicitly incorporates institutional accuracy alongside cultural fidelity and pragmatic appropriateness. Unlike prior models, CPI recognizes educational terms as "institutional speech acts" that function within policy, pedagogical, and administrative systems[23]. By operationalizing institutional accuracy as a quantifiable dimension, the framework extends MT evaluation beyond surface fluency, enabling systematic diagnosis of socio-pragmatic failures in the translation of educational terminology.

2.3. Research Gap

Despite advances in both MT evaluation and translation theory, several critical gaps remain. First, existing frameworks inadequately address institutional accuracy, the dimension most essential for policy-relevant translation. Terms such as 985高校 (Project 985 universities) or 双减政策 (Double Reduction Policy) cannot be fully assessed through semantic fidelity or pragmatic appropriateness alone; they require evaluation of how institutional mechanisms are communicated to target audiences. Without this dimension, translation studies risk overlooking precisely those failures that most directly undermine cross-cultural policy communication.

Second, there is a lack of systematic corpus-based studies that examine translation adequacy across large-scale educational datasets. While numerous case studies analyze individual terms or small sets of examples, few have constructed comprehensive bilingual corpora of educational texts with sufficient scope to support robust evaluation[24]. The absence of such corpora hinders replicability, comparability, and scalability in assessing MT performance.

Third, much of the existing research continues to focus on surface-level metrics such as fluency and semantic alignment. While these measures provide useful baselines, they fail to capture the deeper socio-pragmatic and institutional dimensions of translation adequacy. This methodological narrowness limits both the explanatory power of current studies and their practical relevance for educational stakeholders.

By addressing these gaps, the present study contributes theoretically, methodologically, and practically to the field. Theoretically, it advances translation studies by conceptualizing terminology as institutional speech acts embedded within cultural and pragmatic contexts[25]. Methodologically, it employs a mixed-methods design that combines quantitative benchmarking with qualitative thematic coding, supported by a stratified corpus of 5,000 sentence pairs. Practically, it produces an annotated glossary of 9 high-risk educational terms and a hybrid workflow that integrates MT with expert-informed post-editing, thereby offering concrete tools for improving cross-cultural educational communication.

3. METHODOLOGY

This study employed a convergent mixed-methods design to evaluate the adequacy of machine translation (MT) for Chinese–English educational terminology. A mixed-methods approach was necessary because translation adequacy is not reducible to a single dimension: quantitative metrics capture performance trends across systems, while qualitative analyses diagnose cultural, pragmatic, and institutional failures that elude purely statistical approaches[26]. The quantitative strand relied on repeated-measures analysis of variance (RM-ANOVA) to compare MT outputs with human translation (HT) across three evaluation dimensions, cultural fidelity, pragmatic appropriateness, and institutional accuracy, using a five-point Likert scale[27]. For the quantitative strand, cultural fidelity was operationalized through semantic fidelity scores; therefore, "semantic" in statistical tables refers to the measurable proxy of the cultural dimension. The qualitative strand consisted of thematic coding with NVivo, allowing researchers to systematically classify translation errors into cultural, pragmatic, or institutional categories. By integrating these strands, the study combined statistical rigor with qualitative sensitivity, ensuring a comprehensive evaluation design.

3.1. Research Design

The experimental design was structured around three central variables: translation system (DeepL, GPT-4o, human translation), evaluation dimension (cultural, pragmatic, institutional), and sentence source (policy documents vs. curricular texts)[28].

The convergent design allowed for quantitative and qualitative data to be collected in parallel, merged, and interpreted in an integrated manner. Quantitative evaluation generated mean CPI scores and significance tests, while qualitative coding offered insight into why certain translation failures occurred[29]. This dual perspective ensured that translation adequacy was examined both as a measurable outcome and as a socially and institutionally situated phenomenon.

3.2. Corpus Construction

To evaluate translation adequacy in a representative and contextually relevant manner, the study constructed a bilingual corpus of educational texts drawn from authoritative and policy-sensitive sources. Two primary data streams were used. The first consisted of UNESCO educational policy documents published between 2020 and 2024, which provided internationally recognized policy discourse. The second stream comprised official undergraduate curriculum documents collected from Tsinghua University and Peking University websites between 2021 and 2024. Together, these texts represented both policy-level discourse and institutional-level implementation, ensuring coverage across the most critical domains of educational terminology.

The corpus construction process proceeded in several steps. First, original PDF files were converted into UTF-8 encoded text using OCR tools[30]. This process was automated through a custom Python script, which ensured reproducibility:

Code Snippet 1: Converting PDF files to plain text

```
from pdfminer.high_level import extract_text
import os

pdf_dir = "data/pdf/"
txt_dir = "data/txt/"

for fname in os.listdir(pdf_dir):
    if fname.endswith(".pdf"):
        text = extract_text(os.path.join(pdf_dir, fname))
        with open(os.path.join(txt_dir, fname.replace(".pdf", ".txt")),
                  "w", encoding="utf-8") as f:
            f.write(text)
```

This script automated the batch conversion of PDF policy and curriculum files into plain text, ensuring consistency of file encoding and creating a clean dataset ready for alignment. Its primary function was to reduce manual preprocessing effort and guarantee reproducible text extraction across thousands of pages.

Second, sentence alignment was conducted using Trados Studio and LF Aligner, ensuring that parallel sentence pairs were preserved. Third, a stratified sampling procedure was implemented, prioritizing sentences containing high-risk educational keywords such as 核心素养 (core competencies) or 双减 (Double Reduction). The final corpus contained 5,000 bilingual sentence pairs comprising approximately 150,000 Chinese tokens and 180,000 English tokens. To maximize transparency and replicability, all processed files were stored in an open directory structure, including raw PDF data, aligned sentence pairs, MT outputs, and scoring sheets.

3.3. Translation Systems

Two leading MT systems were selected for benchmarking: DeepL v3.1 and GPT-4o. DeepL was chosen for its reputation in producing fluent translations in European and Asian language pairs. It was used with default parameters, producing outputs without domain-specific customization. GPT-4o, an advanced large language model released in May 2024, was tested with parameters optimized for cultural sensitivity: temperature = 0.3 and a system-level prompt instructing the model to "preserve cultural nuance." This configuration was intended to simulate real-world use of generative MT tools in educational policy translation, where stakeholders increasingly rely on GPT-based systems for draft translations. Both MT systems were benchmarked against human translations produced by professional translators with domain expertise.

3.4. Human Evaluation

To ensure rigorous evaluation, five assessors were recruited: three certified translators with more than five years of professional experience in educational translation, and two doctoral researchers in education. All evaluators were trained in the use of the CPI framework, which requires separate ratings of cultural fidelity, pragmatic appropriateness, and institutional accuracy on a five-point Likert scale (1 = very poor, 5 = excellent). Each sentence pair was evaluated independently, and discrepancies were discussed in calibration sessions to maintain scoring consistency. Inter-rater reliability was assessed using Krippendorff's alpha ($\alpha = 0.82$), which exceeded the conventional threshold (0.80). The scoring rubric was designed to capture fine-grained distinctions, such as whether cultural connotations were retained, whether pragmatic functions (e.g., politeness strategies) were adapted to the target language, and whether institutional mechanisms (e.g., admission policies, national reforms) were accurately conveyed.

3.5. Data Analysis

The quantitative strand relied on repeated-

measures ANOVA (RM-ANOVA) to compare machine translation (MT) outputs with human translation (HT). A two-step computational workflow was used: (1) Python was employed to preprocess CPI scores, ensuring data were structured consistently across evaluators, systems, and dimensions; (2) R was then used to conduct RM-ANOVA, post-hoc pairwise comparisons, and visualization.

Code Snippet 2: Preparing CPI scores for statistical analysis

```
import pandas as pd

# Load evaluator ratings
df = pd.read_csv("scores.csv")

# Compute mean CPI scores by system and dimension
summary = df.groupby(["System", "Dimension"]).mean()

print(summary)
```

This preprocessing step standardized evaluator ratings across cultural fidelity, pragmatic appropriateness, and institutional accuracy before importing the dataset into R for RM-ANOVA. In practice, this Python routine served to clean raw evaluator inputs, calculate preliminary means, and ensure the dataset was complete and correctly formatted for subsequent analysis.

The subsequent statistical testing and visualization were conducted in R. The full script is provided below to ensure replicability.

Code Snippet 3. RM-ANOVA and Visualization (R)

```
# scripts/04_score_analysis.R

# Load required libraries
library(tidyverse)
library(lmerTest)
library(emmeans)
library(effectsize)


# Read CPI scores
scores <- read_csv("data/processed/cpi_scores.csv")

# Data structure:
# id, sentence, system (DeepL / GPT-4o / HT),
# dimension (Cultural / Pragmatic / Institutional), score

# RM-ANOVA model
anova_res <- aov(score ~ system * dimension +
Error(id/dimension), data = scores)

# Output results
summary(anova_res)

# Effect size
eta_squared(anova_res, partial = TRUE)

# Post-hoc comparisons
posthoc <- emmeans(anova_res, pairwise ~ system | dimension)
summary(posthoc)
```

```
# Visualization
ggplot(scores, aes(x = system, y = score, fill = system)) +
geom_boxplot() +
facet_wrap(~ dimension) +
theme_minimal() +
labs(title = "CPI Scores by System and Dimension",
y = "Mean Score", x = "Translation System")
```

This script implements the RM-ANOVA model, computes effect sizes, conducts pairwise comparisons, and generates boxplots stratified by dimension. Its purpose was to provide statistical testing of group differences while also producing visual diagnostics that confirm robustness across CPI dimensions.

The qualitative strand consisted of thematic coding in NVivo, where errors were systematically classified into cultural, pragmatic, or institutional categories. This ensured that quantitative benchmarking was complemented by interpretive depth.

4. RESULTS

The results of this study are presented in three sections. First, quantitative findings reveal significant performance differences between machine translation (MT) systems and human translation (HT) across the three CPI dimensions. Second, qualitative analyses highlight the main categories of errors. Finally, representative case examples illustrate how high-risk terms were mistranslated and how these can be improved through context-sensitive solutions.

4.1. Quantitative Findings

The comparative evaluation of MT and HT performance is summarized in Table 1. Results demonstrate clear variation across semantic, pragmatic, and institutional dimensions.

Table 1: MT vs. HT Performance (Mean CPI Scores).

Dimension	DeepL	GPT-4o	HT	p-value
Semantic	4.1	4.3	4.8	<0.01
Pragmatic	2.9	3.2	4.5	<0.001
Institutional	3.0	3.1	4.6	<0.001

Semantic performance showed the smallest gap: both MT systems produced largely intelligible translations, though HT remained slightly superior ($F(2, 58) = 6.42$, $p < 0.01$, $\eta^2 = 0.18$). Pragmatic performance diverged sharply, with MT systems scoring between 2.9–3.2 compared to 4.5 for HT ($F(2, 58) = 15.73$, $p < 0.001$, $\eta^2 = 0.41$). Institutional accuracy revealed the largest disparity, with MT averaging around 3.0 and HT achieving 4.6 ($F(2, 58) = 22.11$, $p < 0.001$, $\eta^2 = 0.46$). These effect sizes indicate that the differences in pragmatic and institutional

dimensions were large and practically meaningful ($\eta^2 > 0.25$).

To provide a multidimensional visualization of these differences, Figure 1 presents a radar chart of the three CPI dimensions by translation system. The chart highlights how MT systems approximate human translation on the semantic dimension but fall markedly short on pragmatic and institutional accuracy.

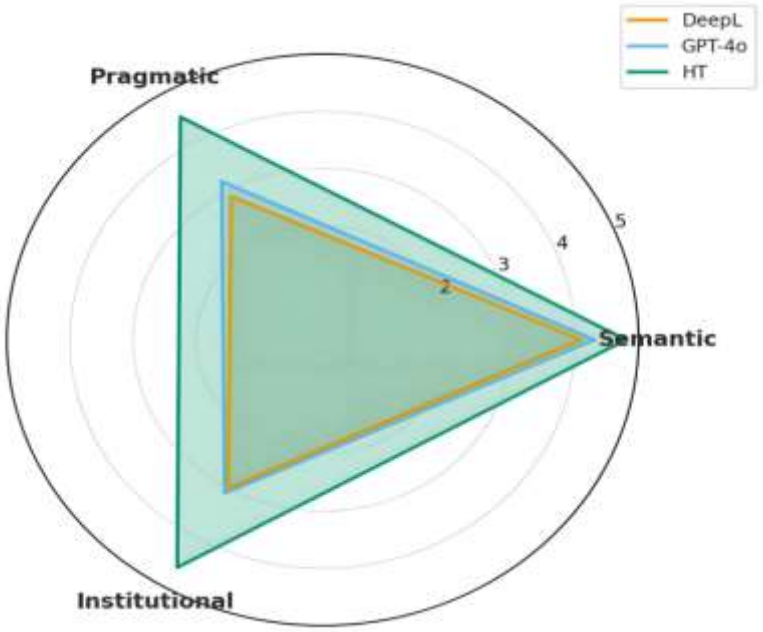

Figure 1: Radar Chart Of CPI Dimensions By Translation System (Deepl, GPT-4o, HT).

4.2. Qualitative Findings

Qualitative coding revealed systematic failures across cultural, pragmatic, and institutional dimensions.

1. Cultural dimension. Reform-oriented terms such as 素质教育 (sùzhì jiàoyù) were frequently rendered as "quality education," which obscured their policy-driven connotations. Human translators instead used "holistic education reform" or "quality-oriented education policy in China," providing more accurate representations of ideological intent.
2. Pragmatic dimension. MT often produced literal renderings of politeness formulas. For instance, 请多多指教 (qǐng duōduō zhǐjiào) appeared as "please advise me a lot," which was awkward in English. HT versions such as "I look forward to your feedback" captured the intended deference. Similarly, 您辛苦了 (nín xīnkǔ le) became "you've worked hard," while HT produced "thank you for your effort," aligning better with English norms.
3. Institutional dimension. Terms lacking direct equivalents posed the greatest challenges. 保研 (bǎoyán) was rendered as "postgraduate recommendation," omitting the mechanism of guaranteed admission without entrance exams. HT supplied fuller versions such as "guaranteed postgraduate admission for top-performing undergraduates." Likewise, 双减政策 (shuāng jiǎn zhèngcè) appeared as "double reduction policy," a phrase opaque to outsiders; HT provided contextualized translations clarifying its scope in reducing workload and tutoring.
4. These examples show that while MT performs acceptably at the lexical level, it consistently misses cultural nuance, pragmatic intent, and institutional specificity.

4.3. Case Examples

To provide a consolidated view of recurrent translation challenges, Table 2 presents representative high-risk educational terms. Each entry includes the original Chinese term, its literal machine translation, common misunderstandings, and a recommended translation with annotations.

Table 2: Case Examples Of Educational Terminology Translation.

Chinese Term	Literal MT Rendering	Common Misunderstanding	Recommended Translation (with Annotation)
应试教育 (yìngshì jiàoyù)	exam-oriented education	Interpreted neutrally as "exam-focused teaching," ignoring its critique of rote learning	test-driven education system or exam-centric pedagogy under systemic pressure (contextual footnote)
素质教育 (sùzhì jiàoyù)	quality education	Mistaken as "high-quality teaching," overlooking its reform agenda	holistic education reform or quality-oriented education policy in China (annotated for policy scope)
核心素养 (héxīn sùyǎng)	core literacy	Misunderstood as basic reading/writing skills	core competencies or key competencies and values in the Chinese national curriculum
学术不端 (xuéshù bùduān)	academic indiscipline	Sounds vague, downplays seriousness	academic misconduct (standardized English equivalent)
保研 (bǎoyán)	postgraduate recommendation	Confusing to international readers; omits mechanism	guaranteed postgraduate admission for high-achieving undergraduates (with gloss explaining exam exemption)
双减政策 (shuāng jiǎn zhèngcè)	double reduction policy	Opaque policy name, lacks explanatory context	Double Reduction Policy (aimed at easing student workload and regulating after-school tutoring)
985高校 (Project 985 universities)	Project 985	Unintelligible without context	Project 985 key universities in China (retain name with explanatory note)
请假条 (qǐngjià tiáo)	leave note	Misunderstood as informal personal note	formal leave application form (context: required written request to supervisors or teachers)
师生关系 (shīshēng guānxì)	teacher-student relationship	May evoke inappropriate connotations in English	pedagogical rapport or academic mentoring relationship (clarify institutional context)

In sum, Table 2 functions as a reference tool, condensing recurrent mistranslation patterns into a structured format. It complements the narrative findings in Section 4.2 by offering a concise overview of typical pitfalls and context-sensitive solutions.

5. DISCUSSION

The findings of this study reveal that while machine translation (MT) systems perform adequately in semantic equivalence, they fall considerably short in pragmatic appropriateness and institutional accuracy. These shortcomings carry profound implications for cross-cultural policy communication and international educational exchange. The discussion is organized into five subsections: cultural implications, pragmatic implications, institutional implications, theoretical contributions, and practical contributions.

5.1. Cultural Implications

Quantitative results indicated that semantic adequacy was relatively strong, yet qualitative evidence showed that cultural fidelity was frequently compromised. This pattern can be interpreted through translation theory: Newmark's distinction between semantic and communicative translation demonstrates why MT defaults to literal renderings that sacrifice ideological nuance. Similarly, Skopos theory underscores the risks of failing to adapt educational terms to their communicative purpose. The "trilemma" of educational terminology, semantic precision, cultural intelligibility, and institutional specificity, helps explain why cultural connotations are often lost. Without systematic strategies to reconcile these competing demands, MT risks reducing policy-laden expressions to sterile descriptors, distorting the cultural logic of Chinese educational reforms.

5.2. Pragmatic Implications

The pragmatic failures of MT exemplify what He (2022) describes as pragmatic borrowing, where the surface form of a speech act is transferred but its function is lost. This problem reveals a limitation of existing evaluation models such as PTA, which recognize pragmatic misalignment but lack institutional anchoring. When politeness conventions or ritualized formulas are mistranslated, relational trust and academic etiquette are undermined. This highlights the necessity of a multidimensional framework that goes beyond fluency to assess whether translated expressions remain pragmatically appropriate for the target audience.

5.3. Institutional Implications

Institutional inaccuracy emerged as the most damaging type of error. Unlike cultural or pragmatic misalignments, institutional mistranslations risk obscuring the very mechanisms of educational reform. Existing frameworks such as CSI-Match benchmark semantic fidelity but fail to measure institutional transfer. The CPI framework directly addresses this gap by treating educational terms as

institutional speech acts. This theoretical stance allows translation studies to move beyond linguistic or pragmatic adequacy toward evaluating how institutional realities are communicated. The persistence of institutional errors in MT highlights the importance of hybrid workflows, where human translators provide contextual glossing that automated systems cannot generate.

5.4. Theoretical Contributions

The study advances translation theory by introducing institutional accuracy as a quantifiable dimension. While CSI-Match captures cultural elements and PTA evaluates pragmatic transfer, neither accommodates the institutional mechanisms embedded in educational terminology. CPI fills this gap and reframes translation adequacy as inherently multidimensional. This represents a shift in translation studies: from evaluating surface equivalence to examining how translations sustain the socio-pragmatic and institutional functions of discourse.

5.5. Practical Contributions

Although this study focuses on educational terminology, its methodological insights are transferable to other specialized domains such as law, medicine, and policy translation. Practically, the findings highlight the value of combining machine translation with expert-informed post-editing. The proposed hybrid workflow balances efficiency with contextual sensitivity, ensuring that semantic, pragmatic, and institutional dimensions are addressed simultaneously. Furthermore, the annotated glossary of 9 high-risk terms provides a concrete tool for practitioners, enabling translators, educators, and policymakers to avoid recurrent pitfalls. Beyond the educational context, this workflow illustrates a generalizable model that can enhance translation adequacy in other domains where institutional precision is equally critical.

6. CONCLUSION

This study examined the adequacy of machine translation (MT) in rendering Chinese-English educational terminology across three dimensions: cultural fidelity, pragmatic appropriateness, and institutional accuracy. With respect to RQ1 and RQ2, results showed that while MT systems approximate human performance at the semantic level, they fall markedly short in the pragmatic and institutional dimensions, and recurrent error patterns were identified through qualitative analysis.

The main contributions of this research are threefold. Theoretically, it expands evaluation models by incorporating institutional accuracy into the CPI framework. Methodologically, it integrates quantitative benchmarking with qualitative thematic coding, producing robust and replicable evidence. Practically, it delivers both an annotated glossary of high-risk educational terms and a hybrid workflow that balances MT efficiency with human expertise.

Two innovations distinguish this study: the operationalization of institutional accuracy as a measurable construct, and the design of a hybrid translation model that addresses the scalability-accuracy dilemma. With respect to RQ3, the hybrid workflow demonstrated its effectiveness in balancing efficiency and accuracy, validating its role as a practical solution.

At the same time, limitations remain. The corpus was restricted to Sino-English educational texts, raising questions of generalizability to other domains. Evaluators were drawn from education-related fields, which may bias assessments. Moreover, the CPI framework has not yet been embedded into MT systems for real-time application.

Future research should extend the CPI framework to other language pairs, test its applicability across specialized domains, and explore integration into MT systems as a quality-control layer. Investigating how translation adequacy shapes the reception of education policy in multilingual contexts will further enhance understanding.

In conclusion, translation in education cannot be reduced to linguistic equivalence alone. Cultural, pragmatic, and institutional dimensions are equally vital for sustaining accurate and meaningful cross-cultural communication. By validating the CPI framework and proposing hybrid workflows, this study provides both conceptual clarity and practical tools for improving the quality of educational translation.

REFERENCES

Alasmary, A. (2025). Discourse-organising lexical bundles in academic law textbooks: a corpus-based analysis. Humanities and Social Sciences Communications, 12(1), 1-15.

Al-Tarawneh, A., Al-Badawi, M., & Hatab, W. A. (2024). TRANSLATING GOVERNANCE AND LEGAL COMPLIANCE: EXPLORING THE ROLE OF TRANSLATION IN FACILITATING CORPORATE REPORTING AND POLICY IMPLEMENTATION. Corporate Law & Governance Review, 6(3).

Andersen, G. (2014). Pragmatic borrowing. Journal of pragmatics, 67, 17-33.

Bakiyev, F., & Kayimova, M. (2025). COMPLEXITY OF CULTURE-SPECIFIC UNITS AND SEMANTIC AMBIGUITIES IN TRANSLATION. Ижтимоий-гуманитар фанларнинг долзарб муаммолари Актуальные проблемы социально-гуманитарных наук Actual Problems of Humanities and Social Sciences., 5(6), 344-350.

Baratbayevna, N. M. (2025). Comparative Analysis of Educational Terminology in English And Uzbek Languages Structural, Semantic and Functional Aspects. European International Journal of Philological Sciences, 5(04), 11-15.

Chen, A. (2025). Exploring the effects of international experiences on graduates' employability development: a comparative study on Sino-UK international joint universities. Studies in Higher Education, 50(5), 988-1004.

Ejjebli, S. (2024). Unlocking the Power of Language: Navigating Linguistic Diversity in Cross-Cultural Research. Studies in Pragmatics and Discourse Analysis, 5(1), 46-62.

Elizbarashvili, A., Tsintsadze, M., & Khachidze, M. (2024). Enhancing Georgian Text Processing: Transliteration Techniques. Baltic Journal of Modern Computing, 12(4).

Halverson, S. L., & Kotze, H. (2021). Sociocognitive constructs in Translation and Interpreting Studies (TIS): Do we really need concepts like norms and risk when we have a comprehensive usage-based theory of language?. In Contesting epistemologies in cognitive translation and interpreting studies (pp. 51-79). Routledge.

Hansen, D., & Esperança-Rodier, E. (2022, July). Human-adapted MT for literary texts: Reality or fantasy?. In NeTTT 2022 (pp. 178-190).

Haseeb, M., Akbar, M., & Abbasi, W. S. (2025). Machine translation vs. human translation: A comparative study of translation quality. Social Science Review Archives, 3(1), 885-894.

He, B., Wang, J., & Wang, Y. (2025). Research on quantitative assessment of translation quality from the perspective of phraseology. PloS one, 20(2), e0318804.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. European Journal of Psychological Assessment, 15(3), 277.

Liu, H., Cheng, X., & Zhang, J. (2025). Fidelity in Translating Government White Papers—A Case Study of Youth of China in the New Era.

Liu, P. (2025). From a raciolinguistic perspective: analyzing curricula and assessments in a two-way Chinese-English Dual Language Bilingual Education (DLBE) program. International Journal of Bilingual Education and Bilingualism, 1-28.

Monzó-Nebot, E. (2024). The cultural roots of translation automation: revealing ideologies in machine translation. In The Social Impact of Automating Translation (pp. 12-37). Routledge.

Nykyporets, S. S., Herasymenko, N. V., & Chopliak, V. V. (2024). Cognitive strategies impacting the structural composition of translated technical and scientific texts: An analysis of translation methodologies.

Peters, C., Picchi, E., & Biagini, L. (2000). Parallel and comparable bilingual corpora in language teaching and learning. In Multilingual corpora in teaching and research (pp. 73-85). Brill.

Prykhodko, V., Kulakevych, L., Litkovych, Y., Kanonik, N., & Horodniuk, N. (2024). Cultural transfer in translation: innovative approaches to preserving intercultural aspects of a text. Synesis (ISSN 1984-6754), 16(1), 47-60.

Reiß, K., & Vermeer, H. J. (2010). Grundlegung einer allgemeinen Translationstheorie (Vol. 147). Walter de Gruyter.

Rouhani, H., & Modarresi, G. (2023). The role of translation-based, meaning-based, and hint-based instructions in vocabulary acquisition: A mixed-methods study. Iranian Journal of Applied Language Studies, 15(1), 83-100.

Schuff, H., Vanderlyn, L., Adel, H., & Vu, N. T. (2023). How to do human evaluation: A brief introduction to user studies in NLP. Natural Language Engineering, 29(5), 1199-1222.

Sokol, Y., & Sokol, M. (2025). THE LANGUAGE OF DIPLOMACY: HOW STRATEGIC COMMUNICATION SHAPES INTERNATIONAL RELATIONS. UNIVERSUM, (21), 104-112.

Sultan, K. M. (2007). The semantics, pragmatics and translation of speech acts. Journal of the College of Basic Education, (50), 23-42.

Sun, R. (2024). Evaluating the Translation Accuracy of ChatGPT and DeepL Through the Lens of Implied Subjects.

Taguchi, N., & Li, S. (2020). Contrastive pragmatics and second language (L2) pragmatics: Approaches to

assessing L2 speech act production. Contrastive Pragmatics, 2(1), 1-23.
Wang, J. (2024). Exploring the potential of chatgpt-4o in translation quality assessment. Journal of Theory and Practice in Humanities and Social Sciences, 1(3), 18-30.
Yang, P., Lu, N., & Yin, J. (2025). Student mobility to China: an overview. Handbook on Migration to China, 98-113.
ZHAO, Y., & LI, Y. (2019). CE Translation Strategies of Diplomatic Speech from the Perspective of Skopos Theory. Cross-Cultural Communication, 15(2), 15-18.
Zou, L., Li, K., Lamerton, J., & Mirzapour, M. (2025, June). GenAIese-A Comprehensive Comparison of GPT-4o and DeepSeek-V3 for English-to-Chinese Academic Translation. In Proceedings of the Eleventh Workshop on Patent and Scientific Literature Translation (PSLT 2025) (pp. 1-12).