Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Convention or argument? Rich points, retrieval and the limits of LLM selection in Chinese–English academic translation

Ruixue Tian
Beijing Foreign Studies University
Huijuan Ma
mahuijuan@bfsu.edu.cn

Beijing Foreign Studies University
Xiaoliang Wang
Vimond Media Solutions

Article

Keywords: translation decision, rich points, retrieval-augmented generation (RAG), AI-assisted academic translation, data contamination, AI translation evaluation

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11570431/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Convention or argument? Rich points, retrieval and the limits of LLM selection in Chinese–English academic translation

Ruixue Tian[1], Huijuan Ma[2] & Xiaoliang Wang[3]

1. School of English and International Studies, Beijing Foreign Studies University, Beijing, China

2. School of English and International Studies, Beijing Foreign Studies University, Beijing, China; the corresponding author: mahuijuan@bfsu.edu.cn

3. Vimond Media Solutions, Bergen, Norway

Abstract: Large language models (LLMs) can now produce many viable translations at almost no cost. Their translation performance can be further enhanced by retrieval-augmented generation (RAG) technique which offers relevant translations for LLMs to select. In academic writing, a concept is formed by terms and the line of argument that builds conceptual relations. Yet it has not been tested how LLMs select among viable translations enabled by their pre-training dataset, and among relevant translations offered by RAG technique, compared to expert translators. This study addresses the gap using Fei Xiaotong's *Xiangtu Zhongguo* and its expert English translation *From the Soil*, which have probably been used in AI training. Four expert translation segments were compared with 40 GPT-5.5 translations, half with RAG enhancement and half without. Four supplementary experiments were conducted to test whether RAG setting of retrieval, filtering and ranking influences LLM selection. A concept-based framework was established and codebooks with pre-identified rich points annotated for LLM judges and human expert to evaluate the LLM selection. Results show that GPT selected as experts do, when convention defines the alternatives, and failed to do so when the alternatives must be reasoned from the author's argument. Probable exposure to expert translation passed on the experts' terms but not their decision of how to build the argument, creating a contamination paradox. RAG technique could provide relevant translations for GPT to select, but caused it to select away from expert decisions related to the argument. No RAG setting enabled GPT to select as expert translators do in building the argument. This study specifies selection competence in terms of making settled and situated decision, and proposes the rich-point method as a means of

enhancing the validity of AI translation evaluation, which may otherwise be inflated by data contamination.

Keywords: translation decision; rich points; retrieval-augmented generation (RAG); AI-assisted academic translation; data contamination; AI translation evaluation

1. Introduction

Concepts and concept formation are "the foundation stones of all social scientific inquiry," so translating them is "more important than the translation of any other word" (Ghosh, 2001, p. 61). A concept, however, is not the same thing as a term. It is a "unit of knowledge created by a unique combination of characteristics" (ISO 1087:2019). In academic writing, a concept is formed through two vehicles: the symbolic vehicle consisting of terms and proper names (Cabré, 1999, p. 95; ISO 704:2022); and the logical vehicle of argumentative sequence. It is the line of thought built from claim, grounds, warrant and backing, and signalled by figures of argument and metadiscourse (Williams, 2004; Fahnestock, 2011; Hyland, 2019).

A translator who made the decision to change the sequence of an argument changes its conceptual relations, even if every term stays the same. Translation is regarded as a series of decisions (Levý, 1967/2000). These decisions can be located at rich points: passages of the source text that pose translation problems (PACTE, 2005, 2011). Translators in Chinese-English direction have to make hard decisions to maintain the logical vehicle of argumentative sequence. That is because argument in Chinese texts tends to proceed by induction and leaves readers to infer the conclusions (Yang, 2018, pp. 347–348), and because Chinese concepts tend to be concrete and moral in orientation, while English ones tend to be abstract and logical (Zhang, 2024, p. 8). Therefore, translators need to decide whether, and how far, to make an implicit relation explicit or to reorganise an argument.

The issue of translation decisions may be complicated by seemingly desirable performance of large language models (LLMs). Pym redefines translation competence in the electronic age as two skills: generating more than one viable rendering, and selecting one "quickly and with justified confidence" (2003, p. 489). LLMs have now largely taken over the first skill by generating much more viable renderings. Retrieval-augmented generation (RAG) further

supplies relevant choices from a knowledge base to enhance AI translation performance (Gao et al., 2024), as a human translator consults a translation memory. Yet how will LLMs select among viable and relevant translations compared to expert translators?

This study will address the question by making use of Fei Xiaotong's sociological classic *Xiangtu Zhongguo* (1947) and its expert translation *From the Soil* (Hamilton and Wang, 1992). Fei's book is used for the test for the following four reasons: (1) Fei builds his concepts in both symbolic and logical vehicles. He sets Chinese against Western concepts in four groups (social relations, egocentrism versus individualism, moral systems and family structures), and he develops each concept through definitions, etymologies, contrasts and figures of speech; (2) The book poses a challenge for making translation decisions related to argument. It began as 14 newspaper essays adapted from lectures (Fei, 2019, p. 1). Thus, the arguments are loosely articulated; (3) Fei was trained in the Western tradition, wrote in English himself, and wrote a Chinese shaped by English. His translators therefore had to decide how to maintain "multiple levels of discourse" (Hamilton and Wang, 1992, p. viii); and (4) Both Fei's book and its expert translation have probably been used in AI pre-training[1], since the texts have circulated online for decades, and Fei's coinages, such as "differential mode of association," are established in English-language sociology. The probable data contamination makes the question of how LLMs select among viable and relevant translations more revealing: If exposure to the expert translation were enough, AI selection should fully reflect expert decisions in both vehicles. Any recurring divergence could not then be attributed to lack of exposure.

Inspired by Pym's (2003) redefinition that in the electronic age, generation and selection skills are core components of translation competence, we put forward three research questions, which are: When AI has probably been exposed to an expert translation, which expert decisions at the rich points of concepts do LLMs select? Enhanced by RAG technique, how do LLMs select from viable and relevant translations? Does the varied RAG setting of retrieval, filtering and ranking change how LLMs select? Each question will test one generating source of translation: the AI model's pre-training, RAG retrieval, and the way retrieval is set up in RAG system.

2. Related study

2.1 Translation decisions, the two vehicles to form academic concept and rich points

Classical terminology theory holds that concepts derive meaning from their relations within a concept system: generic–specific, part–whole, cause–effect, opposition and so on (Sager, 1990; ISO 704:2022). This system is assumed to be stable, which is largely true in standardized technical fields. Later approaches locate conceptual relations in how terms are used in texts (Temmerman, 2000; Cabré, 1999; Faber, 2012). In the social sciences, conceptual relations are formed in the argument which is signalled by the markers of metadiscourse and figures of argument. A translation that loses or alters a marker distorts the concept system the author is building.

Levý (1967/2000) regards translation as a choice among a definable set of alternatives. Definitional Instruction governs the range of possible solutions, and Selective Instructions guide the choice within that range. The symbolic and logical vehicles of academic concepts differ in where the defining instruction comes from. In the symbolic vehicle, convention usually defines the range, often narrowing it to one established equivalent. There are exceptions. An author's own coinage, or a word serving its metalinguistic function, may take its meaning only from the argument. In the logical vehicle, the range often comes from the argument. The translator first works out which relation a marker signals. Only then can the alternatives be circumscribed: leave the relation implicit, make it explicit or reorganise the argument.

The necessity and challenge of making such translation choices is well-put by Pym in his two-fold (generation and selection) functional translation competence model. He emphasizes that as "the generative side of business is impeccably technologized," the "reductive reasoning," the problem-solving process of choosing between viable alternatives is "perhaps the hardest part of translation competence" (2003, pp.490-491).

To study how translators choose to solve translation problems, one has to know where they occur. According to Agar (1994), rich points are the places where speakers of one language run into difficulty with another language and its culture. The PACTE Group adapts the term to mean source-text segments that contain translation problems, including problems of coherence and cohesion (PACTE, 2005, 2011), where the logical vehicle of academic concepts examined by this study lies. Performance at rich points separated expert translators from other bilinguals. Because rich points depend on the text, the language pair and the purpose, PACTE selected

them separately for each text, and we do the same. A finished translation does not show translation decisions directly, but they can be reconstructed by comparing source and target texts (Toury, 2012). One feasible method is to pre-identify rich points to present where a decision is needed in source text, and where the decision was taken in target text. As for LLM-generated translations, they can be treated as the trace of selections, without claiming that AI decides as a human translator does.

2.2 LLM translation, its evaluation and data contamination

Scholars can use LLMs to draft in their first language and produce publishable English at lower cost (Bennett, 2023, p. 524). LLMs' weaknesses, however, cluster around concepts. GPT-4 makes fewer meaning errors than novice translators, but it translates unfamiliar terms literally and proper names inconsistently (Yan et al., 2024, pp. 6–7). Fluent output may attach academic concepts to contexts foreign to their disciplinary meaning (Bennett, 2023, p. 524). Quality also drops when the source involves unequal power relations, errors by the author or ambiguous terms (Peeters et al., 2026, pp. 8 – 9). These studies concern mainly with terms whose alternatives are governed by convention. They do not show how LLMs select among viable and relevant translations to build the author's argument.

The usual tools for evaluating machine translation do not answer this question either. Automatic scores that count shared words between a machine output and a reference translation (e.g., BLEU) correlate weakly with human judgment. Newer automatic methods, including using an LLM as a judge, correlate better (Guerreiro et al., 2024, pp. 980 – 981). Multidimensional Quality Metrics (MQM), a widely used human error-classification scheme, supports fine-grained error analysis (Lommel et al., 2013), and GPT-based judging has been shown to agree with MQM ratings, including for Chinese-to-English (Kocmi and Federmann, 2023). All these methods have been validated mainly on general, non-academic texts.

The issue of using a robust method/tool to evaluate LLM translation is further complicated by data contamination when some parts of a test text have been used in AI training, leading to the overestimation of AI translation performance (Sainz et al., 2023; Balloccu et al., 2024), sometimes by a large margin (Kocyigit et al., 2025). This is because the current evaluation methods rely on word overlap, which cannot tell reproducing the wording apart from

reproducing the translation decisions behind it. Only assessment by kind of translation decision can paint a more nuanced picture of how well AI actually translates.

2.3 RAG technique

Retrieval-augmented generation (RAG) is a technique that combines an AI model's pre-training knowledge with examples retrieved from an external knowledge base (Lewis et al., 2020). Compared to fine-tuning, RAG technique is cheaper and easier (Liu et al., 2025, pp. 10–11). In translation, it improved Gemini 2.0's translation of low-resource Taiwanese Hakka while preserving cultural specificity (Chang et al., 2025). Combined with additional training on related tasks, it also improved the translation of sentences that require specialised knowledge (Wang et al., 2025). But these studies retrieve specialised facts or translation equivalents, that is, alternatives governed by convention. None tests how LLMs select at rich points whose alternatives depend on the argument, or whether RAG setting of retrieval, filtering and ranking influences LLM selection among viable and relevant translations as expert translators do.

3. The CB-TQA framework

3.1 Two dimensions

The framework of Concept-Based Translation Quality Assessment (CB-TQA) evaluates translation decisions about academic concepts on two dimensions: Concept Representation (CR) is for the symbolic vehicle, and Concept Logic (CL) for the logical vehicle.

Conceptual relations emphasized by the Terminology discipline are assessed in both dimensions. A relation built into the designations themselves falls under CR-6 (designation relations; ISO 704:2022), for example Fei's paired coinages "differential mode of association" and "organizational mode of association." A relation established in the argument, the typical case in social-science writing, falls under CL. CB-TQA can therefore detect what general-purpose evaluation misses: a translation whose terms are all correct but whose argument has been reorganized, thus distorting the author's intention.

The two dimensions are weighted unequally. When a designation is wrong, the surrounding context can still support understanding. When the line of argument is broken, the relations among concepts change, and context alone cannot facilitate understanding.

3.2 Subparameters

CR assesses whether each designation carries the source concept into the translation (Faber, 2012). Drawing on ISO 704:2022, it has six subparameters (CR-1 to CR-6): designation type accuracy, consistency, derivability/compoundability, linguistic correctness, preference for established designations, and designation relations.

CL has layers A and B. Layer A, Argument Schema (CL-A1 to A4), consists of Williams's (2004) claim, grounds, warrant and backing. It is not scored, because social-science claims rest on ethical reasoning, shifts of perspective and cultural critique rather than on measurable data (Hyland, 2005, pp. 187–190). Thus, Layer A serves as a map that helps judges follow the argument and locate decisions within it. Layer B, Rhetorical Topology (CL-B1 to B15), is scored. It covers Layer A's visible markers. These are Fahnestock's (2011) five figures of argument (parallelism, strategic repetition, antithesis, antimetabole and definition) and Hyland's (2019) ten metadiscourse devices (transition markers, frame markers, endophoric markers, evidentials, code glosses, hedges, boosters, attitude markers, self-mention and engagement markers). Hedges include Williams's qualifiers and rebuttals.

3.3 Scoring

To operationalize the CB-TQA framework which evaluates translation decisions of social-science concepts on two dimensions, i.e., Concept Representation (CR) and Concept Logic (CL), we identified rich points in four concept groups in Fei's source text that may pose translation problems, along with acceptable and unacceptable alternatives. The acceptable set includes the expert translation and any alternative that well represents the experts' translation decisions. A rendering outside this set is described as a divergence, forming the unacceptable set.

Divergences are weighted by severity, based on the MQM (Lommel et al., 2013). Because distinguishing "major" from "critical" errors in social-science texts is difficult, CB-TQA only has two levels: CR divergences are minor (×1) and CL-B divergences are major (×5). Two counting rules apply: (1) Each locus, the smallest stretch of source text responsible for a divergence, counts once, under its highest-weighted subparameter; (2) Each divergence counts once per text, however often it recurs. Inconsistency is captured separately under CR-2.

The translation quality score (TQ) therefore reflects how many distinct wrong decisions a

translation contains:

$$TQ = 100 - (P_{CR} + P_{CL}) \times 100$$

$$P_{CR} = \frac{n_{CR} \times 1}{W}$$

$$P_{CL} = \frac{n_{CL} \times 5}{W}$$

Here n is the number of distinct divergences and W is the word count of the translation.

Deciding whether to correct a flaw in the source text is itself a translation decision, and MQM credits translators who improve on the source. CB-TQA therefore allows two-way assessment. Penalties found in the source text, normalised by its length, are subtracted from those in the translation:

$$TQ = 100 - (P_{CR-T} - P_{CR-S}) \times 100 - (P_{CL-T} - P_{CL-S}) \times 100$$

A translation that corrects defects in the source text can therefore score above 100.

4. Research design

4.1 Test materials and codebooks

The source text consists of four passages from *Xiangtu Zhongguo* (c. 7,400 Chinese characters), one for each concept group: social relations (Ch. 4, c. 1,720 characters); egocentrism versus individualism (Ch. 4, c. 1,420); moral systems (Ch. 5, c. 1,950); and family structures (Ch. 6, c. 2,310). The expert translation (ERT) is *From the Soil*.

Each group has a closed codebook. It contains eight coding rules; the rich points, each with a subparameter code and lists of acceptable and unacceptable alternatives; and the defects in the source text together with the ERT's remedies. Judges score only the listed rich points, so TQ measures translation decisions at a fixed set of points, not overall quality. The ERT informed the lists but is not treated as a perfect standard. Scored under the same codebook, the expert translation fails several items.

One author, a professor of Chinese–English translation studies, drafted Version 1.0 of the codebooks. Another author, a translator with 10 years' experience in Chinese–English academic translation and a doctoral student of the first, revised it into Version 2.0. Because any stretch of the source text could in principle become an entry, the usual chance-corrected agreement statistics cannot be calculated. We therefore report positive specific agreement (PSA), which

measures how far two coders agree on the entries they identify (Hripcsak and Rothschild, 2005).

Version 2.0 kept 175 of 225 entries and added none (PSA = 0.88; groups 1–4: 0.84, 0.83, 0.90, 0.97). The 50 deletions were redundant source-defect entries (12), proper names or quotations that do not designate concepts (19) and wrongly segmented CL-B spans (19). The two versions assigned retained entries to the same subparameters in every case (Cohen's κ, a chance-corrected agreement measure, = 1.00), and 19 entries (10.9%) had their acceptability revised. Because Version 2.0 revised Version 1.0 rather than being drawn up independently, these figures are an upper bound.

4.2 Knowledge base and LLM generation

The knowledge base, the collection from which relevant translations were retrieved, contains aligned Chinese–English sentence pairs from two conceptually related works by Fei. It does not include the source text or the ERT.

Peasant Life in China (1939; 4,370 pairs) was written in English and translated into Chinese as *Jiangcun Jingji* (Dai Kejing, 2010); *Xiangtu Chongjian* (1948; 419 pairs) was translated orally by Fei and revised by Margaret Redfield into *China's Gentry* (1953).

In other words, most of the relevant translations enabled by RAG technique represent the author's earlier translation decisions about related concepts. It cannot supply the expert translators' decisions, except where they adopted his wording. In essence, retrieving relevant translations through the RAG technique mirrors a translator's behaviour of consulting relevant resources.

Each sentence pair was converted into a numerical representation of its meaning with the BGE-M3 embedding model and stored with the LlamaIndex software library. Each Chinese concept group was then used as a search query. The search retrieved the sentence pairs closest to it in meaning (dense retrieval), with no second-stage re-ranking. Pairs had to reach a minimum similarity score of 0.4 (on a 0–1 scale), and at most 10 pairs were retrieved (Top-K = 10), roughly the number of references a translator consults when revising.

GPT-5.5 (API version gpt-5.5-2026-04-23), translated each passage under two conditions: (1) RAG-enhanced GPT translation (RGT): the retrieved pairs and their similarity scores were

followed by the prompt “Using the reference translations above for terminology and style, provide the best English translation of the Chinese text”; (2) GPT translation without RAG enhancement (GT) as control group: the AI model only received the prompt “Using the terminology and style befitting the translation of academic text, provide the best English translation of the Chinese text.”

Both conditions guide register, one by example and one by instruction. This mirrors the choice a translator faces when setting up a workflow. Each condition was run five times per group at the default temperature (1.0), a setting that allows the AI model’s output to vary from run to run.

4.3 Retrieval experiments

Four experiments were conducted to test whether retrieval setting in RAG system influences how LLMs select among viable and relevant translations (Table 1). Each experiment varies one design choice on concept group 4, the group whose argument extends over the longest stretch of text. Sensitivity experiments vary a setting across several values; ablation experiments remove or replace a component to see what difference it makes. Settings not under test were held at a baseline: Chinese query, meaning-based (dense) retrieval, minimum similarity 0.4 and Top-K = 100. This baseline gives a large pool of examples at an acceptable processing cost of tokens.

Table 1. Retrieval experiments (concept group 4)

Design choice	Type	Conditions	n
(1) Retrieval language	Sensitivity	Chinese, English, bilingual	3
(2) Similarity measure	Ablation	Keyword matching (BM25), meaning-based (dense), hybrid (BM25 + dense), no retrieval	4
(3) Retrieval pool	Sensitivity	Minimum similarity {0.4, 0.5, 0.6} × Top-K {10, 20, 40, 80, 100}	15
(4) Ranking	Ablation	Top-K retrieval; + GPT-5.5 reranking	2

Note. BM25 retrieves examples that share keywords with the query like the Google search engine; dense retrieval retrieves examples similar in semantic meaning; hybrid combines the two. In condition (4), GPT-5.5 re-ranks the

retrieved examples by relevance before GPT generation.

The baseline appears once in each of Experiments 1, 2 and 4, and twice in Experiment 3, where minimum similarities of 0.4 and 0.5 retrieve the same 100 items. These five runs estimate how much output varies by chance. Experiment 3 also includes the main-experiment setting (0.4 × 10). Together with the 44 main-experiment texts (4 ERT passages, 20 RGT, 20 GT), 68 translations were evaluated through the CB-TQA framework and its codebooks with pre-identified rich points which reflect translation problems in Fei's text.

4.4 Evaluation procedure

Two LLMs, Claude Opus 4.8 (claude-opus-4-8) and GPT-5.5, acted as judges. Using the same evaluation prompt, each independently scored every translation against the relevant codebook. Each judge read Layer A first to follow the author's argument, and then scored CR and Layer B. Each returned divergence counts, a provisional TQ and a list of items flagged for human expert's review.

The author who produced codebook Version 2.0 decided on the flagged items, and further prompted the LLM judges to revise the initial evaluation reports. The final decision on every divergence item therefore rests with the human expert. The adjudicator was not blind to condition, since any expert would recognize the published ERT, but the closed codebooks constrained her rulings. Because GPT-5.5 judged its own translations, it might have favoured them. Section 5.4 tests for this.

5. Results

5.1 LLM outputs vs expert decisions

All 40 LLM outputs scored below the ERT (Table 2). RGT and GT differed little (Section 5.2), so they are pooled here (n = 10 per group).

Table 2. TQ scores of the ERT and LLM outputs (RGT and GT pooled)

Group	ERT	LLM mean (SD)	LLM range	ERT - LLM mean	ERT - best LLM
1	98.37	92.33 (0.49)	91.59–93.35	6.04	5.02
2	99.26	94.71 (0.78)	93.26–95.48	4.55	3.78

3	98.77	95.51 (0.47)	94.46–96.26	3.26	2.51
4	98.93	97.05 (0.28)	96.54–97.53	1.88	1.40

LLM scores clustered tightly (SD = 0.28−0.78), and even the best LLM output fell short of the ERT in every group. Most of the gap lies in the logical vehicle (Table 3). In every group, the LLM translations contained about three times as many Concept Logic (CL) divergences as the ERT. The CL gap exceeded the Concept Representation (CR) gap in every group even before weighting, and after weighting CL accounted for 87−98% of the gap.

Table 3. Mean divergences by vehicle

Group	CR: ERT / GT / RGT	CL: ERT / GT / RGT	CL ratio (LLM / ERT)
1	0 / 9.2 / 10.6	6 / 18.3 / 19.4	3.1
2	2 / 3.0 / 5.0	5 / 15.0 / 16.5	3.2
3	1.3 / 2.0 / 2.1	4.2 / 13.5 / 13.7	3.2
4	3 / 5.0 / 6.5	3.5 / 10.5 / 10.5	3.0

Note. GT and RGT values are means over five runs.

The LLM's CL divergences fell mainly on metadiscourse rather than on figures of argument. They clustered in transition markers (especially those expressing contrast), frame markers, hedges, boosters, attitude markers, code glosses and engagement markers. The ERT's few CL divergences were concentrated in group 1, where five of its six were self-mentions.

In the symbolic vehicle, whether GPT selected among viable and relevant translations as experts did depends on where the alternatives for a term came from. The author's coinages established in English-language sociology were selected. For example, 差序格局 was rendered "differential mode of association" in all 10 LLM outputs. Established renderings of cultural and classical terms also fell within the acceptable set. Where the GPT departed from the ERT, it sometimes did better. It kept 墨翟 "Mo Di," which the ERT omits, in all 10 outputs for group 2. It rendered 仁 in 亲仁 as "humane(ness)," closer to the Confucian sense than the ERT's "the good." But terms that take their meaning from Fei's argument were not selected. Where the ERT has "organization" for 团体, the GPT gave "group" or "corporate (social)

group" in 30 outputs, all unacceptable. Terms that Fei discusses as words (metalinguistic terms) fared worst: the series of terms derived from 家 (dictionary equivalent "family") in group 1, and 大家庭 (*da jiating*) "large family" and 小家族 (*xiao jiazu*) "small lineage" in group 4. Of 60 instances of these rich points across the LLM outputs, only one was acceptable. It is worth mentioning that the ERT's advantage in the symbolic vehicle was basically only in group 1.

5.2 Effects of RAG technique

Each RGT output was paired with the GT output produced at the same time, so that passage, codebook and judge were held constant. A rich point was discordant if GPT's rendering was acceptable in one condition only. It was a gain (b) if only GPT with RAG enhancement selected among viable and relevant translations as experts did, and a loss (c) if only GPT without RAG enhancement did. Counts come from each judge's error tables after expert adjudication. The consensus set contains only the discordances that both LLM judges identified at the same rich point, in the same run and in the same direction. To test whether gains and losses differed, we used McNemar's test, a standard test for paired data, with continuity correction when $b + c \geq 25$, and an exact binomial test otherwise. We applied the Holm correction to adjust for testing four groups at once (Table 4).

Table 4. Gains and losses between paired RGT and GT outputs (b/c; p, Holm-adjusted p in brackets)

Group	Claude Opus 4.8	GPT-5.5	Consensus
1	1/14 (.001 [.004])	3/15 (.008 [.023])	1/7 (.070 [.21])
2	7/20 (.021 [.064])	10/29 (.004 [.016])	4/18 (.004 [.017])
3	5/10 (.302 [.360])	8/5 (.581 [.604])	3/5 (.727 [.727])
4	2/7 (.180 [.360])	5/10 (.302 [.604])	2/7 (.180 [.359])
Total	15/51, $p < .001$	26/59, $p < .001$	10/37, $p < .001$

RAG technique changed GPT's selections in both directions. But GPT did not select among viable and relevant translations more often than it did. Both LLM judges found this

independently. They shared 47 discordances and never scored the same rich point in opposite directions. RGT also scored below GT in 16 of 20 run pairs under each judge (sign test, $p = .012$).

The effect varied by group. Group 2 showed net losses in every analysis. The result remained significant after Holm correction for GPT-5.5 and for the consensus set, and the mean RGT−GT gap was largest here (0.68 points under Claude Opus 4.8; 1.09 under GPT-5.5). Group 1 showed significant net losses under each judge but not in the consensus set. Groups 3 and 4 showed no reliable effect.

Gains and losses reflecting whether GPT selected among viable and relevant translations like experts did can be seen at different rich points. Gains were mostly in the symbolic vehicle, where RAG technique enabled GPT to select established terms such as 团体格局 "organizational mode of association," and in the Dense configuration of the supplementary Experiment 2 to select the designation 家族 "lineage." Losses occurred in both vehicles. In the symbolic vehicle, they involved pinyin used to discuss words and idioms. In the logical vehicle, they involved figures of argument and self-mention. CL losses produced about 80% of the net penalty.

5.3 Influence of RAG setting on LLM selection

Given the negative influence of RAG technique on GPT selection, four experiments were conducted to test whether the result is caused by the language to retrieve the author's own translations, by similarity measure of keyword matching and semantic meaning, by the number of relevant translations, or by the relevancy of the author's translations to the present translation task.

The results show that every setting did change what the relevant translations actually received by GPT 5.5. Retrieval language highlights how the key concept 家 (dictionary equivalent "family") was translated for GPT to choose from. The Wade-Giles romanization "Chia" appeared in 31 of 100 items under Chinese queries, 4 under English and 22 under bilingual queries. Keyword matching (BM25) ranked highest the passages that shared general academic vocabulary with the source text (e.g., 概念 lit. "concept," 性质 "nature/character," 西洋 "the West") but not its kinship concepts. These were mostly long passages on towns, the

pao-chia system and land rent. Minimum similarity and Top-K changed the list from 10 to 100 items and made the prompt 3.8 times longer (2,759 to 10,374 tokens, the units of text processed by a model). Because all retrieved items had similarity scores of 0.574 – 0.660, the 15 configurations yielded only six distinct lists. Re-ranking by GPT-5.5 replaced 59 of the 100 examples and removed all six long off-topic passages. No list contained the expert translation of group 4's key terms, such as 事业组织 "a way of organizing activity."

Despite these changes in how RAG offered different numbers, and language- and topic-related translations for GPT to choose from, what the LLM actually selected at rich points stayed largely the same. In the logical vehicle, the same six CL divergences recurred across the four experiments: (1) CG4-526 但是 (lit. "but") and (2) CG4-527 因为 ("because") were translated as contrast and explanation, although here they only add information. The experts left both untranslated; (3) CG4-529: the experts added "however" at the opening of paragraph 11 to mark a contrast with the preceding argument. The AI model did not; (4) CG4-530 一字的相差 should be made explicit as "small lineage," not translated literally as "the difference of a single word"; (5) CG4-534 我们的家, a self-mention, should take the English reader's position, "Chinese families," not "our families"; (6) CG4-537 我们乡土社会中, also a self-mention, should likewise be "in Chinese rural society," not "in our rural society."

In the symbolic vehicle, GPT's selections changed at only two rich points, 家族 and 家法. Without RAG enhancement, the AI model rendered them "lineage" and "family rules," both acceptable. Enhanced by RAG, however, GPT selected unaccepted translations for both terms: hybrid retrieval gave 家族 as "clan-family," dense retrieval gave 家法 as "family discipline," and BM25 gave "family-clan" and "family law."

In terms of scores, most results of the four experiments are close to the control condition (TQ = 97.80) in which GPT 5.5 is not enhanced by RAG. Retrieval language spread TQ by 0.11 points (bilingual 97.91; English 97.87; Chinese 97.80). Re-ranking changed TQ by 0.03 (Top-K retrieval 98.18; with GPT re-ranking 98.15), a difference that reflects a 25-word difference

in output length (W). The 15 retrieval pool configurations produced 12–14 divergences, with TQ ranging from 97.46 (minimum similarity 0.4, Top-K 10) to 97.80 (minimum similarity 0.6, Top-K 20). Only the latter equalled the control condition without RAG enhancement. BM25 condition shows the relatively widest gap (15 divergences; TQ = 97.32), compared to the control condition.

5.4 Reliability and validity of the evaluation

The above results rest on the two-stage evaluation (Table 5).

Table 5. LLM Judge agreement and the effects of expert adjudication

Group	Paired decisions	κ (95% CI)	Missed divergences per text (GPT / Claude)	TQ inflation (GPT / Claude)	Flagged	Overturned	Toward stricter
1	416	.78 (.72–.84)	4.7 / 8.5	+1.65 / +2.80	103	64.10%	92.40%
2	418	.76 (.69–.82)	2.5 / 5.3	+0.81 / +1.94	132	55.30%	98.60%
3	390	.85 (.79–.90)	0.8 / 1.6	+0.17 / +0.45	73	41.10%	86.70%
4	519	.83 (.79–.88)	0.3 / 3.6	+0.24 / +0.80	79	48.10%	71.10%
Total	1,743	.81 (.78–.84)	2.1 / 4.8	+0.72 / +1.50	387	53.50%	89.90%

Note. κ (Cohen's kappa) measures agreement between the two LLM judges beyond chance (0 = chance, 1 = perfect) and is computed on decisions before adjudication. "Missed divergences" and "TQ inflation" compare values before and after adjudication. "Toward stricter" is the share of overturned items moved from acceptable to divergence.

The LLM judges agreed on 91% of the 1,743 paired decisions (κ = .81). Agreement was lowest in groups 1 and 2, where CL divergences were most frequent (Table 3). The expert overturned 207 of the 387 flagged items (53.5%), and 186 of these (89.9%) moved toward a stricter rating. Missed divergences peaked in group 1, where they inflated TQ by 1.65 (GPT-5.5) and 2.80 points (Claude). The AI judges thus tended to accept the same kinds of wrong translations by AI, mostly in the logical vehicle.

Human expert's adjudication changed the size of the scores, not their order. Both LLM judges ranked the ERT above all 40 LLM outputs, ranked GT above RGT in 16 of 20 run pairs,

and reproduced the CL pattern reported in Section 5.1. Because missed divergences fell mainly on the logical vehicle of the LLM outputs, AI judges working alone would have understated the concept logic gap. Nor was GPT-5.5 lenient toward its own translations. It missed fewer divergences than Claude and showed smaller TQ inflation in every group.

6. Discussion

6.1 Selection by convention or by argument

The LLM outputs fell short of the ERT in every group. Where they fell short matters more than by how much. As the overall gap narrowed from 6.04 to 1.88 points, the CL ratio stayed at 3.0–3.2. The shortfall is a matter of a kind of translation decision, not of a general level of translation quality.

Levý (1967/2000) divides instructions governing the translator's choice into Definitional Instruction and Selective Instructions. The former "gives form to the paradigm" (ibid., p.150) which is limited "either by the lexical possibilities of the language or by the verbal ingenuity of the translator" (ibid., p.152). The latter governs "the translator's choice from the available alternatives" (ibid., p.150). Applied to social-science academic translation in the age of AI, Levý's two distinctions can be updated to the difference between convention and argument.

This study reveals that in the symbolic vehicle, convention had already narrowed the LLMs' selection range to one solution. The convention might be a coinage that has entered English-language, an established translation of a classical or cultural term, or a well-accepted term in an academic discipline. In these cases, the LLM selected the term as the experts did, and where the ERT departed from convention, the LLM sometimes did better (Section 5.1). But it diverged where a term takes its meaning only from the author's argument. Fei defines 团体 in contrast to 差序 ("differential mode"). The ERT's "organization" brings out that contrast, but GPT's "group" does not. Metalinguistic terms did worst, because Fei builds their meaning by discussing the words themselves.

In the logical vehicle, no convention narrows the range of alternatives for LLM to select. The LLM translated the words that mark relations as linguistic symbols rather than conceptual relations that unfold in the argument. Two examples: (1) CG3-536, 神，必须是无私的

(experts: “God is necessarily without a private side”). Here 必须 marks a conclusion Fei draws from his premises. Given what God represents, God has no private interests. Rendered as “must,” the sentence reads as a requirement imposed on God. An inference becomes an obligation; (2) CG3-532, 于是发生了权利的观念 (experts: “Thus, a concept of individual rights is inherently a part of [the morality of the Western mode of organization]”). Fei derives the concept of rights from the Western mode of organization that he has just described. Both LLMs with or without RAG enhancement rendered 于是发生了 as a sequence in time. For example, the GT output at timestamp 20260505_123753 has “Thus, ... the concept of rights arises,” and the RGT output at timestamp 20260505_141932 has “Thus, ... there arises the concept of rights.” As the concept of rights “arises,” a logical relation is distorted as a chronological one.

In both cases, GPT treated a connective as a word with a dictionary equivalent. Chinese argument often leaves relations to be inferred (Section 1), so the translator has to decide whether, and how far, to make them explicit. The experts intervened. The AI model did not.

Correcting defects in the source is another decision that requires expert intervention, and CB-TQA’s two-way scoring credits it. It requires reading the source critically against what it refers to. The codebooks list 10 source-text defects. The ERT corrected eight, and GPT outputs corrected none. In CG2-301, for example, Fei wrongly glosses the *Analects*’ 北辰 as 北斗星 “the Big Dipper,” which was corrected by the experts as “North Star.” Fei is an early example of a type of author that is now common: trained in the Western tradition and writing a Chinese shaped by English. A translator who reproduces such a source faithfully also reproduces its defects. In AI-assisted academic translation, human expertise is therefore needed less for terminology than for rebuilding the argument and checking the source.

6.2 The contamination paradox

If an expert translation were used in GPT-5.5’s pre-training process, its translation should have matched *From the Soil*. Our results show that GPT matched expert translation where terms are established. In the logical vehicle, both GPT with and without RAG enhancement fell short of expert translation in all 40 outputs. Thus, it is clear that exposure passed on the experts’ terms but not their decision of how to build the argument. We call this the contamination paradox.

Though proposed 23 years ago when machine translation was far from desirable, Pym's (2003) warning about the prime impact of technology is still highly relevant in the age of AI. While amplifying the productive side of business, technology often obscures "the necessities of reduction," which means that human translators' "reductive reasoning is perhaps the hardest part of translation competence" (ibid., p.49). Applied Pym's (2003) perspective to social-science academic translation assisted by AI, we would like to highlight the difference of translation decisions between settled decision and situated decision. The settle decision occurs when a discipline has already defined and narrowed the range of translation alternatives, while situated decision is needed when such alternatives must be reasoned from the author's argument.

In generative AI systems, a settled decision is recorded as an equivalent pair, such as 差序格局 = "differential mode of association." The pair can be detached from its context and recurs across the translation and across the discipline. An LLM can learn it from repeated exposure, and select as expert translators do. A situated decision is recorded as a reorganised argument, with no trace of the alternatives rejected, and the translation decision behind it. To select alternatives reflecting such decision, an AI model would have to infer the translation strategy behind the choice. Our study shows that GPT 5.5 cannot select argument-related decisions as experts do.

The contamination paradox matters for AI translation evaluation. Word-overlap evaluation metrics (Section 2.2) reward AI models that have absorbed settled decisions and say little about situated ones. Assessment by kind of translation decision may increase the validity of AI translation evaluation. The AI judges' blind spot (Section 5.4) shows that the LLM judges accepted the same kinds of wrong translations generated by GPT. Therefore, the evaluation of AI translation by situated decision cannot be automated without human judgement.

6.3 Why RAG influences some of LLM selections but not others

The RAG technique works on the generation side. It did provide GPT with relevant translations, but the results are not as positive as supposed.

In the symbolic vehicle, relevant translations offered by RAG competed with the conventions that GPT had learned. RAG's knowledge base records Fei's own earlier decisions, mostly from his 1939 work *Peasant Life in China*. When re-ranking the example translations,

GPT-5.5 gave “Chia, an expanded family” its top relevance score of 10, although in the to-be-translated source text Fei sets “expanded family” aside. In other words, RAG placed the author’s earlier decisions beside the experts’ later ones, and sometimes in their place. When GPT adopted them, it fell short of expert translation, e.g., 家族 (“lineage”) and 家法 (“family rules”) (Section 5.3). Losses recurred across runs at the same rich points (Section 5.2), which suggests a systematic pull from RAG, leading to unacceptable translation.

In the logical vehicle, relevant translations offered by RAG could not supply what GPT lacked. RAG cannot “tell” the AI model whether 于是发生了 in CG3-532 marks inference or sequence (Section 6.1). Re-ranking made the example translations more relevant to the topic. But relevance to the topic is not relevance to a translation decision. RAG could even disturb decisions in the logical vehicle, as the losses in figures of argument and self-mention in groups 1 and 2 show.

In short, the factors limiting RAG’s influence on LLM selection can be attributed to what the technique recorded and what the kind of translation decision at stake. Such factors explain why changing RAG setting of retrieving, filtering and ranking made little difference.

7. Conclusion

Inspired by Pym’s (2003) redefinition that translation competence is more about a process of choice between viable alternatives when the generative side of translation business is technologized, we tested how LLMs select among viable translations enabled by their pre-training dataset, and among relevant translations offered by RAG technique, compared to expert translators. By establishing a concept-based framework and annotating codebooks with pre-identified rich points for LLM judges and human expert to evaluate GPT performance by translation decision, we find that GPT selected as experts do, when convention defines the alternatives, and failed to do so when the alternatives must be reasoned from the author’s argument. Probable exposure to expert translation passed on the experts’ terms but not their decision of how to build the argument, creating a contamination paradox. RAG technique could provide relevant translations for GPT to select, but caused it to select away from the expert decisions in the logical vehicle. No RAG setting of retrieving, filtering and ranking enabled GPT to select as expert translators do in the logical vehicle. That is to say, LLMs can select

among settled decisions, and sometimes improve on them. They do not reliably select from situated decisions.

This study has implications in two aspects. In terms of translation theory and practice, we specify Pym's (2003) skill of eliminating possible translations as selection competence of making settled decision and situated decision in the age of AI-assisted academic translation. Rich-point codebooks may be used in translator training to enhance reflective decision report which records how the trainee encountered translation problems, rejected the alternatives and made final choice, a method advocated by Gile (2004) in his Integrated Problem and Decision Reporting model. In terms of research methodology, this study uses CB-TQA and codebooks with rich points to evaluate AI translation by argument-related translation decision. As data contamination is increasingly becoming inevitable, evaluating by kind of translation decision can be more robust than by word overlap with reference translation.

Future studies can test the above conclusion through more texts, language pairs and AI models. To ensure RAG technique exerts greater influence on LLM selection, the knowledge base may take the form of concept-oriented termbases, knowledge graphs to present structured networks of concepts and their relations, along with prompts that ask AI to flag conceptual relations and argumentative sequence.

Data availability

The datasets generated during and/or analysed during the current study are available in the Rich-points-retrieval-and-the-limits-of-LLM-selection repository: https://github.com/Taya-tian/Rich-points-retrieval-and-the-limits-of-LLM-selection

References:

Agar M (1994) The intercultural frame. Int J Intercult Relat 18(2):221−237

Balloccu S, Schmidtová P, Lango M, Dušek O (2024) Leak, cheat, repeat: data contamination and evaluation malpractices in closed-source LLMs. In: Proceedings of the 18th

conference of the European chapter of the Association for Computational Linguistics (volume 1: long papers), St. Julian's, Malta, March 2024. Association for Computational Linguistics, pp 67–93

Bennett K (2023) Translating knowledge in the multilingual paradigm: beyond epistemicide. Soc Sci Inf 62(4):514–532

Cabré MT (1999) Terminology: theory, methods and applications (ed: Sager JC; trans: DeCesaris JA). Terminology and lexicography research and practice, vol 1. John Benjamins, Amsterdam

Chang CC, Li CF, Lee CH, Lee HS (2025) Enhancing low-resource minority language translation with LLMs and retrieval-augmented generation for cultural nuances. In: Intelligent systems conference (IntelliSys 2025). Springer Nature Switzerland, Cham, pp 190–204

Faber P (ed) (2012) A cognitive linguistics view of terminology and specialized language. Applications of cognitive linguistics, vol 20. De Gruyter Mouton, Berlin

Fahnestock J (2011) Rhetorical style: the uses of language in persuasion. Oxford University Press, New York

Fei X (2019) Xiangtu Zhongguo (乡土中国; From the soil). Qingdao Chubanshe, Qingdao (original work published 1947). English edition: Fei X (1992) From the soil: the foundations of Chinese society (trans: Hamilton GG, Wang Z). University of California Press, Berkeley

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H (2024) Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

Ghosh P (2001) Translation as a conceptual act. Max Weber Stud 2(1):59–63

Gile D (2004) Integrated problem and decision reporting as a translator training tool. J Spec Transl (2):2–20

Guerreiro NM, Rei R, van Stigt D, Coheur L, Colombo P, Martins AFT (2024) xCOMET:

transparent machine translation evaluation through fine-grained error detection. Trans Assoc Comput Linguist 12:979–995

Hripcsak G, Rothschild AS (2005) Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 12(3):296–298

Hyland K (2005) Stance and engagement: a model of interaction in academic discourse. Discourse Stud 7(2):173–192

Hyland K (2019) Metadiscourse: exploring interaction in writing, 2nd edn. Bloomsbury, London

International Organization for Standardization (2019) Terminology work and terminology science—vocabulary (ISO 1087:2019). ISO, Geneva

International Organization for Standardization (2022) Terminology work—principles and methods (ISO 704:2022). ISO, Geneva

Kocmi T, Federmann C (2023) GEMBA-MQM: detecting translation quality error spans with GPT-4. In: Proceedings of the eighth conference on machine translation, Singapore, December 2023. Association for Computational Linguistics, pp 768–775

Kocyigit MY, Briakou E, Deutsch D, Luo J, Cherry C, Freitag M (2025) Overestimation in LLM evaluation: a controlled large-scale study on data contamination's impact on machine translation. arXiv preprint arXiv:2501.18771

Levý J (2000) Translation as a decision process. In: Venuti L (ed) The translation studies reader. Routledge, London, pp 148–159 (original work published 1967)

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 33:9459–9474

Liu X, Yun J, Li B, Shi X, Zhang Y (2025) Jiyu daxing yuyan moxing de jiansuo zengqiang shengcheng zongshu (基于大型语言模型的检索增强生成综述; A survey of retrieval-augmented generation based on large language models). Jisuanji Gongcheng yu Yingyong

(计算机工程与应用; Computer Engineering and Applications) 61(13):4−19

Lommel AR, Burchardt A, Uszkoreit H (2013) Multidimensional quality metrics: a flexible system for assessing translation quality. In: Proceedings of Translating and the Computer 35, London, November 2013. Aslib

PACTE (2005) Investigating translation competence: conceptual and methodological issues. Meta 50(2):609−619

PACTE (2011) Results of the validation of the PACTE translation competence model: translation problems and translation competence. In: Alvstad C, Hild A, Tiselius E (eds) Methods and strategies of process research: integrative approaches in translation studies. John Benjamins, Amsterdam, pp 317−343

Peeters K, Daems J, Plieseis C, Rivas Ginel MI, Şahin M (2026) AI for translation and interpreting: a roadmap for users and policymakers. Interpret Soc 6(1):3−33

Pym A (2003) Redefining translation competence in an electronic age: in defence of a minimalist approach. Meta 48(4):481−497

Sager JC (1990) A practical course in terminology processing. John Benjamins, Amsterdam

Sainz O, Campos J, García-Ferrero I, Etxaniz J, Lopez de Lacalle O, Agirre E (2023) NLP evaluation in trouble: on the need to measure LLM data contamination for each benchmark. In: Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 2023. Association for Computational Linguistics, pp 10776−10787

Temmerman R (2000) Towards new ways of terminology description: the sociocognitive approach. Terminology and lexicography research and practice, vol 3. John Benjamins, Amsterdam

Toury G (2012) Descriptive translation studies—and beyond, rev edn. Benjamins translation library, vol 100. John Benjamins, Amsterdam

Wang J, Meng F, Zhang Y, Zhou J (2025) Retrieval-augmented machine translation with unstructured knowledge. In: Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 2025. Association for Computational

Linguistics, pp 5858−5871

Williams M (2004) Translation quality assessment: an argumentation-centred approach. University of Ottawa Press, Ottawa

Yan J, Yan P, Chen Y, Li J, Zhu X, Zhang Y (2024) GPT-4 vs. human translators: a comprehensive evaluation of translation quality across languages, domains, and expertise levels. arXiv preprint arXiv:2407.03658

Yang CN (2018) Shuguang ji (曙光集; Dawn collection) (ed and trans: Weng F). Shenghuo Dushu Xinzhi Sanlian Shudian, Beijing

Zhang X (2024) Zai wenming hujian zhong jianli Zhongguo zizhu de fanyixue (在文明互鉴中建立中国自主的翻译学; Building China's own translation studies through mutual learning among civilizations). Wen Shi Zhe (文史哲; Literature, History and Philosophy) (5):5−8

Author contributions

In the author name order, R. Tian designed the CB-TQA framework, revised the codebooks, adjudicated the evaluation, conducted data analysis and wrote the draft paper; H. Ma, the corresponding author (correspondence to: mahuijuan@bfsu.edu.cn), drafted the first version of the evaluation codebooks, supervised the project, and finalized the manuscript; and X. Wang designed the Retrieval-Augmented Generation (RAG) system, and implemented the main experiment, and the four supplementary experiments to further test the retrieval effects.

Funding declaration

This study received no funding.

Competing interests

The authors declare no competing interests.

Ethical statements

Ethical approval was not required as no human participants were involved in this research.

[1] We input the first paragraph of the source text in the concept group 1 and its expert English translation with the prompt “I want to know whether the following source text and its English translation are included in your pre-training data” to test GPT-5.5. Its answer is “I cannot know whether these exact Chinese and English passages were in my pre-training data; they are publicly available and recognizable as Fei Xiaotong / *From the Soil*, but corpus inclusion cannot be confirmed.” Last tested Oct 30,2026.