https://doi.org/10.48417/technolang.2026.03.14
Research article

AI Translation between Interpretation and Hallucination – Evidence for the Evaluative Competence of Vietnamese Students

Quang Nhat Nguyen (✉) and Minh Vinh Hien Huynh
Ho Chi Minh University of Banking. 36 Tôn Thất Đạm Street, Saigon Ward, HCMC, Vietnam
nhatnq@hub.edu.vn ; hienhmv@hub.edu.vn

Abstract

Generative AI complicates translation evaluation because fluent outputs may introduce unsupported meaning, while non-literal renderings may represent legitimate interpretation. This study examined Vietnamese EFL students' evaluation of AI-generated translations and its associations with digital competence, translation competence, and critical-evaluation confidence. A cross-sectional study involved 213 final-year English majors who completed a 25-item performance task and self-report scales. Students were considerably more accurate in recognising problematic outputs than in diagnosing their type (68.0% vs. 38.4%). Critical-evaluation confidence showed the strongest association with performance (r=.85), although its conceptual overlap with the performance task warrants caution. In the full regression model, translation competence (β=.31) and digital competence (β=.16) contributed significantly alongside critical-evaluation confidence (β=.64), whereas AI-use frequency did not. When the overlapping critical-evaluation measure was removed, the model still explained 55% of performance variance, with translation competence showing the strongest association (β=.61). The findings indicate that evaluating AI-generated translation depends less on frequency of AI use than on students' translation knowledge and evaluative judgement, particularly their ability to distinguish unsupported generation from defensible interpretation. Rather than simply as a translation tool Generative AI should thus be used as an object of critical inquiry. Alternative AI renderings may expose students to interpretations they had not considered, including potentially productive expansions of source meaning. This approach preserves the interpretive character of translation while making students accountable for the decisions they ultimately accept.

Keywords: AI-assisted translation; AI hallucination detection; Translation evaluation; Translation competence; Critical evaluation; EFL students

Citation: Nguyen, Q. N. & Huynh, M. V. H. (2026). AI Translation between Interpretation and Hallucination – Evidence for the Evaluative Competence of Vietnamese Students. *Technology and Language, 7*(3), 207-232. https://doi.org/10.48417/technolang.2026.03.14

© Nguyen, Q. N., Huynh, M. V. H.. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

УДК 004.8:378.016
https://doi.org/10.48417/technolang.2026.03.14
Научная статья

Перевод с помощью ИИ между интерпретацией и галлюцинацией: Данные об оценочной компетенции вьетнамских студентов

Куанг Нят Нгуен (✉) и Минь Винь Хиен Хуинь
Банковский университет Хошимина, ул. Тон Тхат Дам, дом 36, район Сайгон, г. Хошимин, Вьетнам
nhatnq@hub.edu.vn; hienhmv@hub.edu.vn

Аннотация
Генеративный искусственный интеллект усложняет оценку перевода: безупречно звучащий машинный текст может привносить значение, не подтверждённое исходным текстом, тогда как небуквальный перевод может оказаться законной интерпретацией. В настоящем исследовании рассматривается, как вьетнамские студенты, изучающие английский язык как иностранный, оценивают переводы, сгенерированные ИИ, и как эта оценка связана с цифровой компетенцией, переводческой компетенцией и уверенностью в собственной способности к критической оценке. В поперечном исследовании приняли участие 213 студентов выпускного курса направления «Английский язык», выполнивших практическое задание из 25 пунктов и заполнивших шкалы самоотчёта. Студенты значительно успешнее распознавали проблемные фрагменты, чем определяли их тип (68,0% против 38,4%). Наиболее сильная связь с результативностью обнаружена у уверенности в критической оценке ($r = 0,85$), однако её концептуальное пересечение с самим заданием требует осторожной интерпретации. В полной регрессионной модели значимый вклад наряду с уверенностью в критической оценке ($\beta = 0,64$) внесли переводческая компетенция ($\beta = 0,31$) и цифровая компетенция ($\beta = 0,16$), тогда как частота использования ИИ – нет. При исключении перекрывающегося показателя критической оценки модель по-прежнему объясняла 55% дисперсии результативности, причём наиболее сильная связь приходилась на переводческую компетенцию ($\beta = 0,61$). Полученные результаты свидетельствуют о том, что успешность оценки ИИ-перевода зависит не столько от частоты обращения к ИИ, сколько от переводческих знаний студентов и их оценочного суждения – прежде всего способности отличать неподтверждённую генерацию от обоснованной интерпретации.

Ключевые слова: ИИ-перевод; обнаружение галлюцинаций ИИ; оценка перевода; переводческая компетенция; критическая оценка; студенты, изучающие английский как иностранный

Для цитирования: Nguyễn, Q. N. Huỳnh, M. V. H. AI Translation between Interpretation and Hallucination – Evidence for the Evaluative Competence of Vietnamese Students. Technology and Language. 2026. № 7(3). P. 207-232. https://doi.org/10.48417/technolang.2026.03.14

© Нгуен, К. Н., Хуинь М. В. Х. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Generative artificial intelligence has altered the relationship between translation production and translation evaluation. Neural systems and large language models can now produce fluent target texts rapidly, shifting part of the translator's work from producing language to evaluating, verifying, and taking responsibility for machine-generated output (Ehrensberger-Dow et al., 2023; Jiménez-Crespo, 2025; Shormani & Al-Sohbani, 2025). This shift is particularly consequential in translator education because linguistic fluency does not necessarily indicate fidelity. AI-generated translations may appear grammatical and convincing while introducing meanings that the source does not support, making surface quality an unreliable basis for evaluation (Abdelhalim et al., 2025; Huang & Cheung, 2026; Zhang et al., 2025).

Hallucination presents a specific form of this problem. In machine translation, hallucinated output contains material that is insufficiently grounded in the source and may involve additions, semantic distortions, or other meaning shifts concealed by otherwise well-formed language (Guerreiro et al., 2023; Ji et al., 2023; Xu et al., 2023). Detecting such problems therefore requires more than noticing grammatical errors. Evaluators must compare source and target meanings and determine whether apparently plausible wording has altered or introduced information (Al-Tarawneh et al., 2025; Guerreiro et al., 2023; Xu et al., 2023). This distinction is important for learners, who may otherwise mistake fluency for translation accuracy.

However, not every non-literal rendering constitutes hallucination. Translation is inherently interpretive: meanings may be reformulated, made explicit, or adapted to context rather than reproduced through direct linguistic correspondence (Jakobson, 1959; Eco, 1990; Steiner, 1998). Explicitation, for example, can make contextually recoverable information accessible to target readers without introducing unsupported meaning (Han et al., 2023; Su, 2023; Tian, 2024). The central evaluative question is therefore not simply whether a translation differs from its source, but whether the difference can be justified by textual evidence, context, and communicative purpose (Ehrensberger-Dow et al., 2023; Jiménez-Crespo, 2025; Steiner, 1998). This provides a more defensible boundary between legitimate interpretation and unsupported AI generation.

This problem also raises a methodological question about how students' evaluative competence should be assessed. Research on AI-assisted learning has often relied on perceptions, attitudes, intentions, and self-reported competence, yet these measures do not necessarily establish what learners can demonstrate when evaluating an actual output (Ren, 2025; Zhang et al., 2025; Xu et al., 2025). Performance-based assessment is therefore needed alongside self-reports, particularly when digital competence, translation competence, and critical-evaluation confidence are examined as correlates of evaluation performance. The distinction is relevant in Vietnamese higher education where AI use has expanded rapidly while pedagogical approaches to critical evaluation are still developing (Nguyen & Dang, 2023; Nguyen & Nguyen, 2025; Xu et al., 2025).

As a result, this study was conducted to examine how students of English as a foreign language (EFL) evaluate AI-generated English–Vietnamese translations and how their performance relates to self-reported competence after accounting for the frequency

of AI-use. Recognition of problematic output is distinguished from diagnostic classification, following established approaches to translation-quality evaluation rather than being proposed as a new distinction (Freitag et al., 2021; Mariana et al., 2015). Specifically, the study aims to answer the following research questions:

RQ_1: To what extent can EFL students distinguish problematic AI-generated translations from acceptable outputs and accurately classify the identified problems?

RQ_2: How are students' self-reported digital competence, translation competence, and critical evaluation skills associated with evaluation performance after accounting for the frequency of AI-use?

Figure 1 presents the conceptual framework of the study as follows.

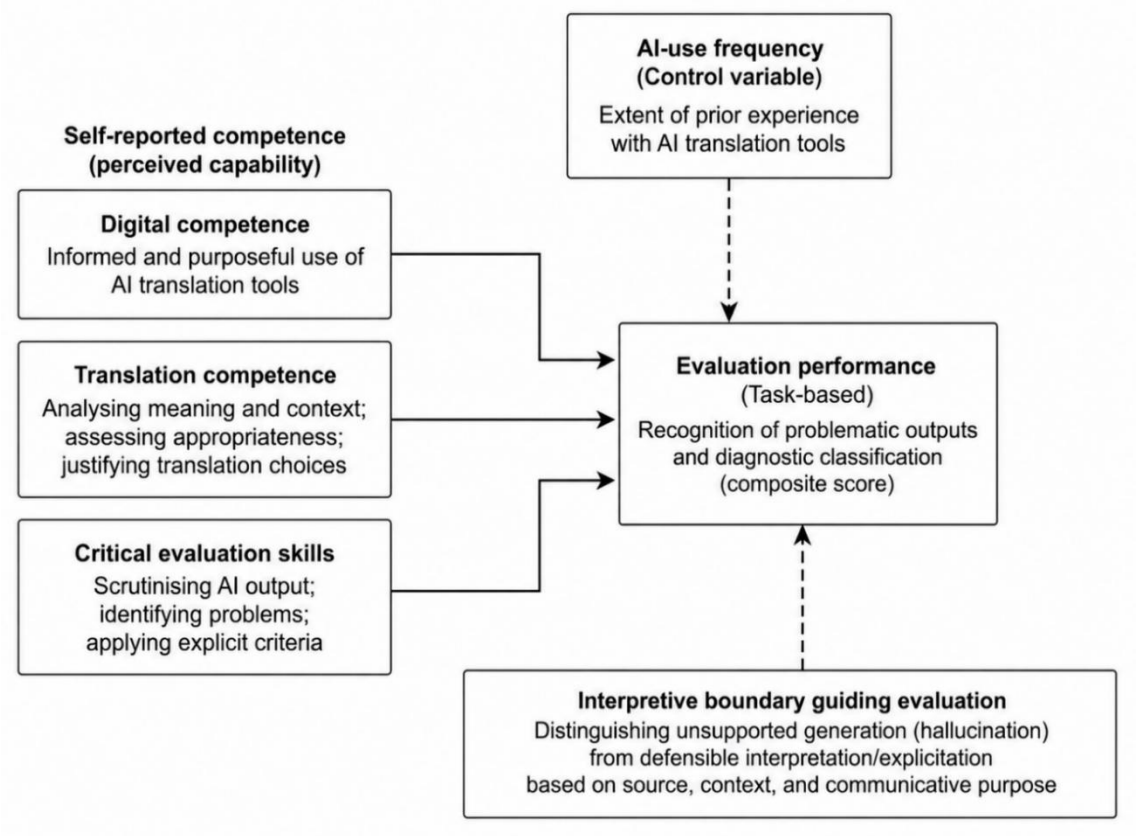

Figure 1. Conceptual framework of the study

The framework distinguishes between students' self-reported competence and their task-based evaluation performance. Three self-reported domains (digital competence, translation competence, and critical evaluation skills) are examined in relation to evaluation performance, while AI-use frequency is included as a control variable to account for differences in prior exposure to AI-assisted translation. Evaluation performance is operationalised through students' recognition of problematic outputs and diagnostic classification of the identified problems, combined into a composite score. Importantly, the framework does not assume that every departure from literal source–target correspondence constitutes a hallucination. Evaluation is situated within an

interpretive boundary that distinguishes unsupported generation from defensible interpretation or explicitation according to the source text, relevant context, and communicative purpose. The framework therefore conceptualises AI-translation evaluation as a task requiring both identification of problematic output and judgement of whether a particular meaning shift is sufficiently warranted.

LITERATURE REVIEW

AI-assisted translation

Generative AI has altered the distribution of human and technological work in translation. Neural machine translation and large language models increasingly support drafting, lexical selection, reformulation, and revision, reducing the amount of text that translators need to produce independently (Ehrensberger-Dow et al. 2023; Huang & Cheung, 2026; Jiménez-Crespo, 2025). Human expertise has consequently become more concentrated in post-editing, verification, contextual interpretation, and responsibility for the final text. Rather than diminishing translation competence, automation changes where that competence is exercised: the translator must judge output whose linguistic fluency may exceed its semantic or pragmatic reliability (Ehrensberger-Dow et al., 2023; Jiménez-Crespo, 2025; Shormani & Al-Sohbani, 2025).

This redistribution of work has important implications for translator education. Students may readily acquire operational familiarity with AI systems, particularly when these technologies offer immediate gains in speed and perceived usefulness, but familiarity does not necessarily entail critical engagement with their output (Abdelhalim et al., 2025; Ren, 2025; Zhang, Li, & Wu, 2025). Translation technology competence therefore extends beyond knowing how to generate or revise text. It includes understanding technological limitations and maintaining human oversight when evaluating meaning, terminology, audience, and communicative purpose (Su, 2023; Tian, 2024; Xu et al., 2025). AI-assisted translation thus places increasing weight on evaluative expertise rather than tool operation alone.

AI hallucinations in translation

Hallucination has become a central concern in research on generative language systems, although its definition varies across tasks and disciplinary traditions. Broadly, hallucination refers to generated content that lacks adequate grounding in the relevant input or evidence (Ji et al., 2023). In machine translation, this may take the form of unsupported additions, semantic distortions, altered propositions, or other source–target divergences that remain linguistically plausible (Guerreiro et al., 2023; Ji et al., 2023; Xu et al., 2023). This combination of fluency and inadequate grounding makes hallucination particularly difficult to detect because target-language well-formedness provides little indication that the underlying source meaning has been changed.

The problem therefore differs from conventional surface-level translation errors. Lexical or grammatical failures often provide visible evidence of poor quality, whereas hallucinated content may preserve coherence while altering semantic relations, stance, certainty, or contextual information (Guerreiro et al., 2023; Ji et al., 2023; Xu et al., 2023).

Evaluation consequently requires source-grounded comparison rather than judgement based primarily on readability. Evidence from AI-assisted translation further suggests that learners can respond positively to fluent output even when systematic verification is limited, reinforcing the danger of treating linguistic naturalness as evidence of fidelity (Abdelhalim et al., 2025; Huang & Cheung, 2026; Zhang, Li, & Wu, 2025).

The concept nevertheless becomes problematic when hallucination is equated with any departure from literal correspondence. Translation routinely involves restructuring, explicitation, contextual adaptation, and shifts in linguistic form. A meaningful distinction must therefore be made between generated content that lacks source support and non-literal translation that remains contextually justified (Han et al., 2023; Su, 2023; Tian, 2024). Hallucination is more usefully understood as a failure of grounding or warrant than as textual difference itself.

The boundary between hallucination and productive meaning in translation

The boundary between hallucination and legitimate variation becomes clearer when considered against hermeneutic and semiotic accounts of translation. Such traditions challenge the assumption that translation consists of transferring a fixed meaning through formally corresponding linguistic units. Roman Jakobson (1959) conceptualises translation through interpretation across signs, while Steiner (1998) places understanding and interpretation at the centre of translational activity. Umberto Eco (1990) similarly foregrounds the interpretive possibilities and constraints involved in meaning-making. Across these perspectives, translation inevitably involves choices through which meaning is reconstructed in relation to linguistic form, context, prior knowledge, and communicative purpose.

Non-literalness therefore cannot provide a sufficient criterion for hallucination. Explicitation offers a clear example: translators may verbalise information that remains implicit but contextually recoverable in the source, particularly where target readers do not share the same background knowledge (Han, Boyd-Graber, & Carpuat, 2023; Su, 2023; Tian, 2024). Reformulation may likewise alter lexical or syntactic form while preserving the relevant communicative meaning. The critical distinction lies between an interpretive expansion for which textual or contextual justification can be provided and an expansion that introduces information, stance, certainty, or causal relations unsupported by available evidence (Eco, 1990; Jiménez-Crespo, 2025; Steiner, 1998).

Generative AI complicates this boundary because generativity can produce both legitimate alternatives and unwarranted elaborations. An unexpected rendering may draw attention to a possible interpretation overlooked by the human reader, giving AI potential heuristic value in exploratory or creative reading. Yet interpretive novelty does not establish translational validity. Generated alternatives acquire value when they prompt comparison, reconsideration, and justification against relevant evidence; they become problematic when plausibility substitutes for such verification (Guerreiro et al., 2023; Ji et al., 2023; Xu et al., 2025). The literature therefore points towards interpretive warrant as a useful principle for distinguishing productive meaning-making from unsupported generation.

Recognition and diagnosis in translation quality assessment

Translation quality assessment provides an established framework for distinguishing the recognition of a problem from its diagnosis. Analytic approaches have long moved beyond holistic judgements by identifying individual errors and classifying them according to dimensions such as accuracy, terminology, fluency, and style. The Multidimensional Quality Metrics framework systematises this approach through a hierarchical error taxonomy, allowing evaluators to specify not merely that a translation is problematic but how and where its quality is compromised (Mariana et al., 2015). Professional evaluation of advanced machine translation has similarly demonstrated the value of explicit error annotation and classification (Freitag et al., 2021).

Recognition and diagnosis are therefore related but analytically distinguishable. Recognition requires an evaluator to identify a source–target relationship as problematic; diagnosis additionally requires the problem to be characterised using an appropriate conceptual category. The latter places greater demands on explicit translation knowledge because an evaluator who senses that a rendering is inadequate may not necessarily be able to determine whether the inadequacy concerns terminology, semantic accuracy, pragmatic meaning, or another dimension (Freitag et al., 2021; Mariana et al., 2015). Diagnostic classification consequently provides information that global acceptability judgements cannot capture.

This distinction acquires renewed relevance with generative AI because fluent output can obscure qualitatively different source–target problems. Error recognition indicates sensitivity to problematic output, whereas classification reveals whether that sensitivity is supported by sufficiently differentiated knowledge to characterise the problem. The distinction is therefore not a new consequence of generative AI, but an established principle of translation assessment whose importance increases as surface fluency becomes less informative about underlying translation quality (El-Farahaty, 2025; Freitag et al., 2021; Xu et al., 2023).

Competence, self-assessment, and evaluative performance

Several competence domains may contribute to the evaluation of AI-generated translation. Digital competence concerns purposeful engagement with technological systems, including awareness of their capabilities and limitations and the strategic use of digital resources (Nguyen & Dang, 2023; Nguyen & Nguyen, 2025; Tian, 2024). Translation competence more directly supports source–target analysis through knowledge of meaning, terminology, audience, context, and communicative purpose (Alshaikhi, 2026; Su, 2023; Tian, 2024). Critical evaluation, meanwhile, concerns the capacity to scrutinise plausible information, apply explicit criteria, and resist treating linguistic fluency as sufficient evidence of accuracy (Al-Tarawneh et al., 2025; Jiménez-Crespo, 2025; Xu et al., 2025). These domains are complementary, but none is conceptually equivalent to successful performance on a translation-evaluation task.

A further distinction concerns competence and its self-assessment. Self-report measures capture individuals' perceptions of what they can do rather than direct demonstrations of the relevant capability. Research on self-assessment has repeatedly shown that perceived and demonstrated competence do not necessarily coincide,

particularly where successful self-evaluation itself requires domain knowledge (Kruger & Dunning, 1999; Ren, 2025; Xu et al., 2025). Self-reported digital or translation competence should therefore be interpreted as perceived capability rather than an objective measure of performance.

The issue becomes more pronounced when self-report items resemble the behaviour assessed by a performance task. A scale asking respondents whether they can detect subtle distortions or recognise fluent but inaccurate translations approaches the conceptual territory of a task requiring those same judgements. Strong associations between such measures may reflect genuine competence and accurate self-perception, but they may also partly arise from shared construct content (Kruger & Dunning, 1999; Mariana et al., 2015). Distinguishing perceived competence from demonstrated performance is consequently essential when interpreting relationships between self-report constructs and task-based translation evaluation.

AI exposure, experience, and evaluative performance

Frequency of AI use represents another possible source of variation in evaluative performance. Repeated interaction with generative systems may expose users to recurrent problems and thereby develop expectations about when apparently fluent output requires closer inspection. Greater experience could therefore support sensitivity to typical AI failures. At the same time, frequent use may simply increase familiarity, convenience, or confidence without improving the quality of evaluation. Research on technology adoption indicates that AI use is often shaped by perceived usefulness, efficiency, and behavioural intention, none of which necessarily entails systematic verification (Ren, 2025; Zhang et al., 2025; Nguyen & Nguyen, 2025).

Exposure and competence should therefore remain conceptually separate. Frequency measures the extent of contact with AI, whereas competence concerns the resources available for using and evaluating it. The distinction is pedagogically significant because experience can take different forms: repeated acceptance of generated output and deliberate comparison of AI translations with source texts may involve similar frequencies of use but very different evaluative practices (Ren, 2025; Zhang et al., 2025; Xu et al., 2025). What remains unclear is whether exposure itself is meaningfully associated with stronger evaluation once relevant competence is considered..

Research gaps

The literature establishes the importance of human judgement in AI-assisted translation, but three issues remain insufficiently resolved. First, much research on student engagement with generative AI has emphasised attitudes, perceptions, adoption, or self-reported capability, while comparatively less evidence links these perceptions to performance on source-grounded translation-evaluation tasks (Ren, 2025; Zhang et al., 2025; Xu et al., 2025). This leaves the relationship between perceived competence and demonstrated evaluative ability insufficiently understood.

Second, translation quality research has established sophisticated traditions of error identification and classification, yet the recognition–diagnosis relationship warrants further empirical examination under conditions in which AI-generated fluency can

conceal semantic, terminological, pragmatic, and contextual problems (Freitag et al., 2021; Mariana et al., 2015). The issue is not whether recognition and classification are theoretically new, but how effectively learners can perform these established evaluative operations on increasingly plausible AI output.

Third, the concept of hallucination remains difficult to reconcile with the interpretive character of translation if source deviation is treated as intrinsically erroneous. Hermeneutic and semiotic traditions permit reformulation, explicitation, and contextual enrichment, whereas computational accounts emphasise grounding and source support (Eco, 1990; Guerreiro et al., 2023; Ji et al., 2023; Steiner, 1998). A more adequate account of AI-translation evaluation therefore requires attention to the boundary between unsupported generation and warranted interpretation. Taken together, these gaps point to the need for performance-based evidence that distinguishes recognition from diagnosis, separates self-perceived competence from demonstrated evaluation, accounts for differences in AI exposure, and treats source–target evaluation as a problem of interpretive warrant rather than literal correspondence alone.

METHODOLOGY

Research design

This study employed a cross-sectional quantitative design to examine EFL students' task-based evaluation of AI-generated English–Vietnamese translations and its associations with self-reported competence. Digital competence, translation competence, and critical-evaluation confidence were measured through self-report scales, whereas evaluation ability was assessed through a performance-based task. The former were therefore treated as perceived capabilities and the latter as demonstrated performance. Given the non-experimental design, relationships among these variables were interpreted as associations rather than causal effects.

Participants were 213 final-year English majors at Ho Chi Minh University of Banking, recruited through convenience sampling from a cohort of approximately 250 students (response rate ≈ 85%). All had completed three core translation modules and had prior exposure to technology-supported translation. Selecting students at the same stage of the programme provided a relatively comparable disciplinary background while retaining variation in AI use and self-reported competence.

The data were collected through a structured online instrument consisting of three components: participant background information, self-report competence scales, and a performance-based evaluation task.

Self-report questionnaire

Data were collected through an online instrument comprising background information, three self-report competence scales, and a performance-based translation-

evaluation task. Background variables included gender, AI-use frequency for translation, AI tools used, and previous training in post-editing or AI-assisted translation. AI-use frequency was measured on a five-point scale from *Never to Very often* and was retained as a control variable in the regression analysis.

Digital competence, translation competence, and critical-evaluation confidence were each measured using eight items rated from 1 (*Strongly disagree*) to 5 (*Strongly agree*). Digital competence addressed purposeful AI use, output comparison, terminology checking, and awareness of technological limitations. Translation competence covered source–target analysis, terminology, contextual appropriateness, revision, and justification of translation decisions. Critical-evaluation confidence addressed students' perceived ability to scrutinise plausible AI output, detect meaning distortions, and apply explicit evaluative criteria. Mean scores were calculated for each scale.

The performance task consisted of 25 English source sentences paired with Vietnamese AI-generated translations. For each item, students first determined whether the translation contained a problematic deviation and, where applicable, classified the problem as semantic, pragmatic, terminological, cultural/contextual, or added information. Four translations were acceptable and 21 contained predetermined deviations. A deviation was considered problematic when the target meaning could not be adequately warranted by the source and immediate context; non-literal but defensible renderings were not treated as errors.

Scoring protocol

Evaluation performance comprised recognition and diagnostic accuracy. Recognition measured whether students correctly identified an output as problematic or acceptable across all 25 items. Diagnostic accuracy measured whether they correctly classified the principal problem in the 21 items containing predetermined deviations. These measures were retained separately to distinguish initial identification from the more specific classification of translation problems.

Each correct recognition decision contributed one point and each correct diagnostic classification contributed one additional point, yielding 46 scorable responses (25 recognition decisions + 21 diagnostic classifications). The composite score was calculated as the total number of correct responses divided by 46 and transformed to a 0–10 scale. Recognition and diagnostic accuracy were additionally reported as percentages to permit direct comparison despite their different denominators.

Data collection and analysis

The instrument was administered online under supervised classroom conditions. Participants completed it independently without consulting external AI or translation tools, and all followed the same sequence: background information, self-report scales, and the performance task. This procedure reduced the possibility that task scores reflected real-time assistance from the technologies being evaluated.

The dataset was screened before analysis. No duplicate identifiers or out-of-range Likert responses were identified, and performance-task responses were complete.

Missing self-report responses accounted for 1.36% of scale data and were replaced using participants' within-construct means.

Analysis proceeded in four stages. First, descriptive statistics and Cronbach's alpha were calculated, and recognition and diagnostic accuracy were examined separately. Second, regression assumptions were assessed through residual normality, independence, homoscedasticity, and multicollinearity diagnostics. Third, Pearson correlations with 95% confidence intervals were used to examine bivariate relationships between competence measures and performance. Finally, hierarchical multiple regression was conducted with AI-use frequency entered first, followed by digital competence, translation competence, and critical-evaluation confidence. Unstandardised and standardised coefficients, 95% confidence intervals, and explained variance were reported.

Because several critical-evaluation items closely reflected operations required by the performance task, a sensitivity analysis excluding this predictor was conducted to assess whether the associations of digital and translation competence with performance remained when the potentially overlapping construct was removed.

RESULTS

Preliminary analyses and assumption testing

The final sample comprised 213 students, including 137 females (64.3%) and 76 males (35.7%). AI use for translation was common: 11 students (5.2%) reported never using AI, 19 (8.9%) rarely, 75 (35.2%) sometimes, 72 (33.8%) often, and 36 (16.9%) very often.

Table 1 reports the internal consistency of the three self-report scales.

Table 1. Reliability of the questionnaire

Scale	Cronbach's α	No. of items
Digital competence	.92	8
Translation competence	.93	8
Critical-evaluation confidence	.93	8

Note. N = 213. Scores ranged from 1 to 5.

All three self-report competence scales showed high internal consistency, with Cronbach's alpha coefficients ranging from .92 to .93. These values supported the use of composite mean scores for digital competence, translation competence, and critical-evaluation confidence in the subsequent analyses.

Table 2 presents descriptive statistics for the self-report measures and task-based evaluation performance.

Table 2. Descriptive Statistics for self-report measures and task performance

	M	SD	Min	Max
Digital competence	3.20	1.05	1.00	5.00
Translation competence	3.01	1.07	1.00	5.00
Critical-evaluation confidence	2.92	1.06	1.00	5.00

AI-use frequency (1–5)	3.48	1.04	1.00	5.00
Recognition accuracy (%)	68.0	21.5	8.0	100.0
Diagnostic accuracy (%)	38.4	24.7	0.0	90.5
Performance score (0–10)	5.45	2.23	0.43	9.35

Note. N = 213. Performance score = total correct responses (out of 46) / 46 x 10.

The three competence ratings were centred around the midpoint of the five-point scale, with digital competence receiving the highest mean (M=3.20), followed by translation competence (M=3.01) and critical-evaluation confidence (M=2.92). Students reported more frequent use of AI (M=3.48). Task performance showed a different pattern: recognition accuracy averaged 68.0%, whereas diagnostic accuracy averaged only 38.4%. The overall performance score was 5.45 out of 10 (SD=2.23). These descriptive results point to a marked difference between recognising a problematic translation and correctly identifying the type of problem.

Recognition versus Diagnosis of AI translations

To examine RQ_1 in greater detail, recognition and diagnostic accuracy were compared across the individual translation items. Table 3 reports the percentage of correct responses for each item.

Table 3. Item-level recognition and diagnostic accuracy

Item	Key type	Rec. %	Diag. %	Item	Key type	Rec. %	Diag. %
Q37	Terminology	62.0	32.9	**Q50**	Terminology	62.0	33.8
Q38	Semantic	70.9	43.7	**Q51**	Semantic	70.9	43.2
Q39	Acceptable	76.5	-	**Q52**	Terminology	61.5	35.2
Q40	Pragmatic	66.7	39.9	**Q53**	Added information	73.7	43.7
Q41	Pragmatic	65.7	37.1	**Q54**	Added information	67.6	43.7
Q42	Semantic	78.4	47.9	**Q55**	Acceptable	70.9	-
Q43	Terminology	57.3	28.2	**Q56**	Semantic	69.0	35.2
Q44	Added information	67.1	43.2	**Q57**	Semantic	69.5	41.3
Q45	Acceptable	73.7	-	**Q58**	Terminology	59.2	29.1
Q46	Semantic	72.3	40.4	**Q59**	Terminology	69.0	44.1
Q47	Acceptable	76.1	-	**Q60**	Terminology	63.4	34.7
Q48	Semantic	70.4	43.7	**Q61**	Semantic	64.8	34.7
Q49	Terminology	61.5	30.5				

Note. "-" indicates an acceptable item for which no diagnostic classification was required.

Recognition accuracy was consistently higher than diagnostic accuracy. Across the task, students correctly recognised 68.0% of the outputs but correctly diagnosed only 38.4% of the problematic cases, a difference of 29.6 percentage points. The difficulty was particularly evident for terminology items. For example, Q43 was correctly recognised by 57.3% of students but correctly classified by only 28.2%; similar gaps occurred for Q49 and Q58. Semantic and added-information items generally produced somewhat

higher diagnostic accuracy, although diagnosis remained below recognition throughout. The item-level results therefore show that the principal difficulty was not simply noticing that an output was problematic but identifying the nature of the problem once it had been noticed.

Regression assumption testing

Before the hierarchical regression was conducted, residual distribution, independence, homoscedasticity, and multicollinearity were examined. Table 4 reports the collinearity diagnostics for the full model.

Table 4. Collinearity statistics

	Tolerance	VIF
Digital competence	.728	1.374
Translation competence	.614	1.630
Critical evaluation skills	.716	1.396
AI-use frequency	.922	1.085

Residuals showed no significant departure from normality, Shapiro–Wilk W=.99, p=.133, and the Breusch–Pagan test provided no evidence of heteroscedasticity, LM=2.65,p=.618. The Durbin–Watson statistic was 2.13. As shown in Table 4, tolerance values ranged from .614 to .922 and VIF values from 1.085 to 1.63. These diagnostics indicated no substantial violation of the assumptions underlying the regression analysis.

(1) Correlations between Competence and Performance

To examine the bivariate relationships relevant to RQ_2, Pearson correlations were calculated among performance, the three self-reported competence measures, and AI-use frequency. Table 5 presents the correlation matrix with 95% confidence intervals.

Table 5. Correlations among variables with 95% confidence intervals

	1	2	3	4	5
1. Performance score	-				
2. Digital competence	.52 [.41, .61]	-			
3. Translation competence	.71 [.64, .77]	.49 [.38, .59]	-		
4. Critical-evaluation confidence	.85 [.81, .89]	.34 [.21, .45]	.52 [.42, .61]	-	
5. AI-use frequency	.18 [.05, .31]	.26 [.13, .38]	.21 [.07, .33]	.17 [.04, .30]	-

Note. Competence–performance correlations: p<.001; performance–AI-use correlation: p=.008.

Performance was positively related to all three competence measures. The strongest correlation was observed for critical-evaluation confidence (r=.85), followed by translation competence (r=.71) and digital competence (r=.52). AI-use frequency showed a much weaker relationship with performance (r=.18). The competence measures were themselves moderately correlated, but none of their intercorrelations approached the magnitude of the relationship between critical-evaluation confidence and performance.

Given the content overlap identified between the critical-evaluation scale and the performance task in the Methods, the unusually high r=.85 is examined further in the sensitivity analysis rather than treated as evidence of an independent effect.

Hierarchical regression analysis

Hierarchical multiple regression was used to determine whether the three self-reported competence measures accounted for performance beyond AI-use frequency. Table 6 presents the two regression models.

Table 6. Model summary

Model	R	R^2	Adjusted R^2	SE	F (df)	p
1: AI-use frequency	.18	.03	.03	-	7.2 (1, 211)	.008
2: Competence measures	.92	.84	.84	.89	281.31 (4, 208)	.000

Note. For the full model, $SS_{regression}$=885.86, $SS_{residual}$ =163.75; Durbin–Watson = 2.13.

Table 6 shows that AI-use frequency alone explained 3% of the variance in performance, F(1,211)=7.2,p=.008. Adding the three competence measures increased explained variance by 81 percentage points (ΔR^2=.81), and the full model accounted for 84% of performance variance, F(4,208)=281.31, p<.001. The large increase between the two models indicates that the association with performance was primarily attributable to the competence measures rather than to frequency of AI use.

Table 7 reports the coefficients for each predictor in the full regression model.

Table 7. Coefficients

	B	SE	β	t	p	95% CI
Constant	−1.253	.275	-	−4.56	.000	[−1.79, −.71]
AI-use frequency	−.068	.061	−.03	−1.11	.266	[−.19, .05]
Digital competence	.333	.068	.16	4.87	.000	[.20, .47]
Translation competence	.639	.073	.31	8.79	.000	[.50, .78]
Critical-evaluation confidence	1.356	.068	.64	19.94	.000	[1.22, 1.49]

Note. Dependent variable = performance score (0–10). N = 213.

The result show that Digital competence (β=.16), translation competence (β=.31), and critical-evaluation confidence (β=.64) were all positively associated with performance after AI-use frequency and the other competence measures were controlled. AI-use frequency, however, was no longer significant (β=−.03,p=.266). Critical-evaluation confidence had the largest coefficient in the full model, but this estimate cannot be interpreted in the same way as the other coefficients because several items in that scale closely mirror the operations assessed by the performance task. The reduced model reported below therefore provides an important test of whether the associations with digital and translation competence remain when this overlapping measure is removed.

Sensitivity analysis

Given the conceptual overlap between the critical-evaluation confidence scale and the performance task, a sensitivity analysis was conducted by re-estimating the regression model without critical-evaluation confidence. Table 8 presents the results of the reduced model.

Table 8. Sensitivity analysis excluding critical evaluation skills

	B	SE	β	t	p
Constant	.171	.451	-	.38	.706
AI-use frequency	−.001	.104	−.001	−.01	.994
Digital competence	.465	.116	.219	4.02	.000
Translation competence	1.262	.112	.606	11.30	.000

Note. N = 213. Dependent variable = performance score (0–10). R^2=.545, adjusted R^2=.539; F(3,209)=83.73, p<.001.

The reduced model remained significant and explained 54.5% of the variance in performance. Translation competence showed the larger association with performance (β=.606, p<.001), while digital competence also remained significant (β=.219, p<.001). AI-use frequency was not significant (β=−.001, p=.994). Thus, digital and translation competence remained associated with performance after the conceptually overlapping critical-evaluation measure was excluded. The lower explained variance relative to the full model (R^2=.844) also indicates that the contribution of critical-evaluation confidence should be interpreted cautiously because of its measurement overlap with the performance task.

DISCUSSION

Recognition and diagnosis of problematic AI translations

The first finding was the clear gap between recognising problematic AI-generated translations and diagnosing the specific problem. Students correctly recognised 68.0% of the outputs but correctly classified only 38.4% of the problematic cases. This pattern is consistent with evidence that fluent AI output can conceal semantic, terminological, and pragmatic problems, making detailed evaluation more demanding than an initial judgement of acceptability (Guerreiro et al., 2023; Ji et al., 2023; Xu et al., 2023). It also aligns with translation-quality assessment, where error identification and error classification represent related but distinct evaluative operations (Mariana et al., 2015; Freitag et al., 2021). The particularly low diagnostic accuracy for several terminology items suggests that students could often sense that a translation was problematic without having sufficient analytical precision to explain what had gone wrong.

This distinction should not imply that every departure from the source constitutes a hallucination. Translation is interpretive rather than a process of literal substitution, and shifts in wording, explicitation, or contextual adaptation may be legitimate when they are supported by the source and communicative context (Steiner, 1998; Han et al., 2023). Generative AI complicates this distinction because the same generative capacity that

produces useful reformulations can introduce plausible but unsupported meanings (Ji et al., 2023; Xu et al., 2023; Guerreiro et al., 2023). The relevant boundary is therefore not simply literal versus non-literal translation, but whether a target-text interpretation has sufficient textual and contextual warrant. AI-generated alternatives may even stimulate productive or creative readings, but such possibilities remain interpretive hypotheses until they can be justified against the source.

Competence and evaluation performance

All three competence measures were positively associated with performance, but the relationships differed considerably in magnitude. Critical-evaluation confidence showed the strongest correlation and the largest coefficient in the full model, followed by translation competence and digital competence. This ordering is theoretically plausible because evaluating AI translation requires more than operational familiarity with a tool: it requires judgement about meaning, terminology, context, and communicative appropriateness (Ehrensberger-Dow, Delorme Benites, & Lehr, 2023; Jiménez-Crespo, 2025; Xu et al., 2025). Translation competence therefore remains central even in AI-mediated workflows, while digital competence appears to provide a supporting rather than sufficient basis for evaluation (Su, 2023; Tian, 2024; Jiménez-Crespo, 2025).

The very strong relationship involving critical-evaluation confidence nevertheless requires caution. Several items in this scale ask students whether they can recognise fluent but inaccurate translations or detect subtle distortions, which closely resembles what the performance task requires. Its large coefficient may therefore reflect genuine evaluative competence, shared construct content, or both. This concern is supported by the substantial decrease in explained variance when critical-evaluation confidence was removed from the model. Importantly, however, translation and digital competence remained significant in the sensitivity analysis, with the reduced model still explaining 54.5% of performance variance. The substantive relationship between competence and performance therefore cannot be attributed entirely to the overlapping critical-evaluation measure (Kruger & Dunning, 1999; Xu et al., 2025; Zhang et al., 2025).

These findings also call for restraint in invoking the Dunning–Kruger effect. Although discrepancies between self-assessment and demonstrated performance are well established, the present study was not designed to test whether lower-performing students systematically overestimated their competence (Kruger & Dunning, 1999). The results instead show that self-reported competence is meaningfully associated with performance, particularly for translation and critical evaluation, while remaining methodologically distinct from demonstrated ability. The appropriate conclusion is therefore not that a Dunning–Kruger effect was confirmed, but that self-assessment and performance should not be treated as interchangeable measures.

AI use and the role of experience

AI-use frequency was weakly correlated with performance but became non-significant once the competence measures were included. Frequent exposure to AI therefore appears less important than the capabilities students bring to the evaluation of its output. Repeated use may familiarise learners with AI-generated translations, but

frequency alone does not ensure systematic verification or informed judgement (Ren, 2025; Zhang et al., 2025; Jiménez-Crespo, 2025). This finding is particularly relevant in educational contexts where increased access to generative AI may be mistaken for increased AI literacy.

The result supports a distinction between using AI frequently and using AI evaluatively. Effective engagement requires learners to question fluent output, compare source and target meanings, identify terminology and contextual problems, and justify acceptance or revision (Ehrensberger-Dow et al., 2023; Xu et al., 2025; Huang & Cheung, 2026). From this perspective, exposure creates opportunities for learning, but translation and evaluative competence determine how productively those opportunities are used.

Implications for translation training

The findings have several implications for translation education. Most importantly, frequent AI use should not be equated with effective AI-assisted translation practice. Although repeated exposure may familiarise students with recurrent problems in generated output, such as terminological inconsistency, semantic shifts, contextual mismatch, and unsupported additions, experience alone does not ensure careful evaluation. Translation training should therefore move beyond increasing students' access to AI and instead structure that exposure around deliberate comparison, verification, and post-editing (Huang & Cheung, 2026; Jiménez-Crespo, 2025; Xu et al., 2025). Students need opportunities to examine apparently fluent outputs critically and to determine when human intervention is required.

A second implication concerns the distinction between error detection and interpretive judgement. Translation cannot be reduced to literal source–target correspondence, since reformulation, explicitation, and contextual adaptation may represent legitimate translation choices. Training should therefore require students not only to identify potentially problematic output but also to justify whether a source–target difference is warranted by textual evidence, context, genre, audience, and communicative purpose (Steiner, 1998; Mariana et al., 2015). Classroom tasks could ask students to compare alternative AI-generated translations, classify relevant problems, identify supporting source evidence, and explain why a rendering should be accepted, revised, or rejected. This approach develops diagnostic precision without treating every non-literal rendering as a hallucination.

Generative AI can consequently be used as an object of critical inquiry rather than simply as a translation tool. Alternative AI renderings may expose students to interpretations they had not considered, including potentially productive expansions of source meaning. Their value, however, depends on whether students can distinguish a defensible interpretation from unsupported generation. AI outputs can therefore be treated as provisional interpretive hypotheses that require verification rather than as authoritative translations (Ji et al., 2023; Guerreiro et al., 2023; Jiménez-Crespo, 2025). This approach preserves the interpretive character of translation while making students accountable for the decisions they ultimately accept.

In general, translation education should aim to develop calibrated judgement rather than either trust or distrust of AI. Performance-based evaluation should complement self-

reported competence, while instructional tasks should require students to articulate the reasoning behind their decisions. The goal is not merely to teach students to detect hallucinations, but to develop their ability to distinguish defensible reformulation, acceptable explicitation, translation error, and unsupported generation. Such training places interpretation, verification, and professional responsibility at the centre of AI-assisted translation practice (Ehrensberger-Dow et al., 2023; Jiménez-Crespo, 2025; Xu et al., 2025).

CONCLUSION

This study examined Vietnamese EFL students' ability to evaluate AI-generated translations and the roles of digital competence, translation competence, and critical-evaluation confidence. Students were considerably more successful in recognising problematic translations than in diagnosing the specific nature of the problem. All three competence measures were positively associated with task performance. Although critical-evaluation confidence showed the largest association in the full model, this result requires caution because of its conceptual overlap with the performance task. Importantly, the sensitivity analysis showed that translation and digital competence remained significant after this measure was excluded. By contrast, AI-use frequency did not independently predict performance once competence was considered.

These findings suggest that effective evaluation of AI-generated translation depends less on frequent AI use than on the knowledge and judgement students bring to the task. For translation education, training should therefore extend beyond general AI exposure towards systematic practice in recognising, diagnosing, and justifying source–target differences. Students should learn to distinguish unsupported generation from legitimate reformulation or interpretation rather than treating every non-literal rendering as a hallucination.

The study is limited by its cross-sectional design, reliance on self-report measures for the competence constructs, conceptual overlap involving critical-evaluation confidence, and the use of sentence-level evaluation tasks. Future research should employ independent performance-based competence measures and more contextualised translation tasks and examine how evaluative judgement develops over time. Such work would clarify how learners can use generative AI while retaining the interpretive judgement and professional responsibility required for translation.

REFERENCES

Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: A comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts & Humanities, 12*(1). https://doi.org/10.1080/23311983.2025.2508031

Alshaikhi, T. (2026). AI-integrated translation training and translators' competence system using neural machine translation (NMT). *SN Computer Science, 7*, 60. https://doi.org/10.1007/s42979-025-04661-3

Al-Tarawneh, A., Abu Hatab, W., & Al-Badawi, M. (2025). Artificial intelligence and the transformation of human translation: A study of the translation industry in Jordan. *Cogent Arts & Humanities, 12*(1), 2598659. https://doi.org/10.1080/23311983.2025.2598659

Eco, U. (1990). *The limits of interpretation*. Indiana University Press.

Ehrensberger-Dow, M., Delorme Benites, A., & Lehr, C. (2023). A new role for translators and trainers: MT literacy consultants. *The Interpreter and Translator Trainer, 17*(3), 393–411. https://doi.org/10.1080/1750399X.2023.2237328

El-Farahaty, H. (2025). Integrating corpora and AI tools in the teaching, learning, and assessing of Arabic/English legal translation. *International Journal for the Semiotics of Law, 38*, 2031–2060. https://doi.org/10.1007/s11196-025-10281-0

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. T. (2023). Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics, 11*, 1176–1194. https://doi.org/10.1162/tacl_a_00615 \

Han, H., Boyd-Graber, J., & Carpuat, M. (2023). Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9718–9735). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.603

Huang, Y., & Cheung, A. K. F. (2026). Exploring AI's performance in literary autobiography translation: How closely do AI models match human translation. *Humanities and Social Sciences Communications*. https://doi.org/10.1057/s41599-026-06630-4

Jakobson, R. (1959). On linguistic aspects of translation. In R. A. Brower (Ed.), *On translation* (pp. 232–239). Harvard University Press. https://doi.org/10.4159/harvard.9780674731615.c18

Ji, Z., Lee, N., Frieske, R. M., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 248. https://doi.org/10.1145/3571730

Jiménez-Crespo, M. A. (2025). Human-centered AI and the future of translation technologies: What professionals think about control and autonomy in the AI era. *Information, 16*(5), 387. https://doi.org/10.3390/info16050387

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Mariana, V., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metrics (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation, 23*, 137–161. https://doi.org/10.26034/cm.jostrans.2015.343

Nguyen, Q. N., & Nguyen, N. P. D. (2025). Measuring digital competence for EFL education in Vietnam: A DigiCompEdu framework analysis. *Technology and Language, 6*(3), 152–180. https://doi.org/10.48417/technolang.2025.03.12

Nguyen, T. H. N., & Dang, T. K. A. (2023). Digital transformation in Vietnamese higher education: Opportunities, challenges, and policy implications. *Education and Information Technologies, 28*(9), 11547–11567. https://doi.org/10.1007/s10639-023-11641-w

Ren, X. (2025). We want but we can't: Measuring EFL translation majors' intention to use ChatGPT in their translation practice. *Humanities and Social Sciences Communications, 12*, 256. https://doi.org/10.1057/s41599-025-04604-6

Shormani, M. Q., & Al-Sohbani, Y. A. (2025). Artificial intelligence contribution to translation industry: Looking back and forward. *Discovery Artificial Intelligence, 5*, 389. https://doi.org/10.1007/s44163-025-00487-3

Steiner, G. (1998). *After Babel: Aspects of language and translation* (3rd ed.). Oxford University Press.

Su, W. (2023). The effectiveness of translation technology training: A mixed methods study. *Humanities and Social Sciences Communications, 10*, 1. https://doi.org/10.1057/s41599-023-02066-2

Tian, X. (2024). Personalized translator training in the era of digital intelligence: Opportunities, challenges, and prospects. *Heliyon, 10*(20), e39354. https://doi.org/10.1016/j.heliyon.2024.e39354

Xu, S., Su, Y., & Liu, K. (2025). Investigating student engagement with AI-driven feedback in translation revision: A mixed-methods study. *Education and Information Technologies, 30*, 16969–16995. https://doi.org/10.1007/s10639-025-13457-0

Xu, W., Agrawal, S., Briakou, E., Martindale, M. J., & Carpuat, M. (2023). Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics, 11*, 546–564. https://doi.org/10.1162/tacl_a_00563

Zhang, W., Li, A. W., & Wu, C. (2025). University students' perceptions of using generative AI in translation practices. *Instructional Science, 53*, 633–655. https://doi.org/10.1007/s11251-025-09705-y

APPENDIX

Survey: AI-supported translation practices and AI hallucination detection among Vietnamese EFL students

Section A. Participant Information

1. Gender: ☐ Male ☐ Female

2. How often do you use AI tools for translation-related tasks?

( ) Never ( ) Rarely ( ) Sometimes ( ) Often ( ) Very often

3. Which AI tools do you use most frequently?
( ) Google Translate ( ) ChatGPT ( ) DeepL ( ) Gemini ( ) Other: _____
4. Have you received any training in post-editing or AI-assisted translation?
( ) Yes ( ) No
Please indicate your level of agreement with the following statements.
1 Strongly disagree 2 Disagree 3 Neutral 4 Agree 5 Strongly agree

Section B. Digital competence

5. I can use AI tools effectively to support translation tasks.	☐1 ☐2 ☐3 ☐4 ☐5
6. I know how to compare outputs from different AI tools.	☐1 ☐2 ☐3 ☐4 ☐5
7. I can adjust prompts to improve AI translations.	☐1 ☐2 ☐3 ☐4 ☐5
8. I can use tools to check terminology and consistency.	☐1 ☐2 ☐3 ☐4 ☐5
9. I can identify when AI output needs revision.	☐1 ☐2 ☐3 ☐4 ☐5
10. I understand strengths and limitations of AI translation.	☐1 ☐2 ☐3 ☐4 ☐5
11. I can use AI tools responsibly.	☐1 ☐2 ☐3 ☐4 ☐5
12. I am confident in integrating AI tools into translation workflow.	☐1 ☐2 ☐3 ☐4 ☐5

Section C. Translation competence

13. I can identify meaning differences accurately.	**☐1 ☐2 ☐3 ☐4 ☐5**
14. I can choose contextually appropriate vocabulary.	☐1 ☐2 ☐3 ☐4 ☐5
15. I can detect communicative weakness in translations.	☐1 ☐2 ☐3 ☐4 ☐5
16. I can adapt translations to audience and purpose.	☐1 ☐2 ☐3 ☐4 ☐5
17. I can detect terminology problems.	☐1 ☐2 ☐3 ☐4 ☐5
18. I can revise translations effectively.	☐1 ☐2 ☐3 ☐4 ☐5
19. I understand translation as interpretation, not substitution.	☐1 ☐2 ☐3 ☐4 ☐5
20. I can justify translation choices.	☐1 ☐2 ☐3 ☐4 ☐5

Section D. Critical evaluation skills

21. I can detect subtle meaning distortions.	☐1 ☐2 ☐3 ☐4 ☐5
22. I can recognise fluent but inaccurate translations.	☐1 ☐2 ☐3 ☐4 ☐5
23. I can evaluate communicative appropriateness.	☐1 ☐2 ☐3 ☐4 ☐5

24. I can detect tone/register mismatch.	☐1 ☐2 ☐3 ☐4 ☐5
25. I question AI outputs critically.	☐1 ☐2 ☐3 ☐4 ☐5
26. I can detect cultural/contextual errors.	☐1 ☐2 ☐3 ☐4 ☐5
27. I can distinguish acceptable vs optimal translation.	☐1 ☐2 ☐3 ☐4 ☐5
28. I evaluate translation using clear criteria.	☐1 ☐2 ☐3 ☐4 ☐5

Section E. AI use

29. I use AI to generate first drafts.	☐1 ☐2 ☐3 ☐4 ☐5
30. I use AI to check vocabulary.	☐1 ☐2 ☐3 ☐4 ☐5
31. I use AI to revise translations.	☐1 ☐2 ☐3 ☐4 ☐5
32. I usually trust AI output without revision.	☐1 ☐2 ☐3 ☐4 ☐5
33. I compare AI output with my judgement.	☐1 ☐2 ☐3 ☐4 ☐5
34. I have seen AI outputs that look correct but are wrong.	☐1 ☐2 ☐3 ☐4 ☐5
35. AI improves my efficiency.	☐1 ☐2 ☐3 ☐4 ☐5
36. AI may reduce quality if used uncritically.	☐1 ☐2 ☐3 ☐4 ☐5

Section F. AI hallucination detection test

For each item:

- Decide if the translation contains a hallucination (Yes/No).
- If YES, select the main error type

( ) Semantic error ( ) Pragmatic error ( ) Terminology error
( ) Cultural/contextual error ( ) Added information

37. Source: The company is facing liquidity pressure in the short term. AI: Công ty đang đối mặt với áp lực thanh lý trong ngắn hạn. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
38. Source: Your account is temporarily restricted for security reasons. AI: Tài khoản của bạn đã bị chấm dứt vĩnh viễn vì lý do bảo mật. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
39. Source: The borrower must disclose all relevant liabilities. AI: Bên vay phải công bố mọi nghĩa vụ tài chính liên quan. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*

40. Source: This feature is currently unavailable in your region. AI: Tính năng này hiện đã bị cấm tại khu vực của bạn. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
41. Source: The findings are somewhat surprising. AI: Các kết quả này hoàn toàn gây sốc và không thể chấp nhận được. ☐ Yes☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
42. Source: The patient reported mild discomfort after the procedure. AI: Bệnh nhân báo cáo cơn đau dữ dội sau ca phẫu thuật. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
43. Source: The central bank adopted a hawkish stance on inflation. AI: Ngân hàng trung ương áp dụng lập trường giống chim diều hâu. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
44. Source: The proposal was not approved at this stage. AI: Đề xuất đã bị bác bỏ hoàn toàn và sẽ không bao giờ được xem xét lại. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
45. Source: Please consult your physician if symptoms persist. AI: Vui lòng gặp bác sĩ nếu các triệu chứng tiếp tục kéo dài. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
46. Source: The app stores your preferences locally. AI: Ứng dụng bán các sở thích của bạn tại địa phương. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
47. Source: The witness declined to comment. AI: Nhân chứng từ chối bình luận. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
48. Source: The software may experience delays during peak hours. AI: Phần mềm có thể bị sập hoàn toàn trong giờ cao điểm. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*

49. Source: Non-performing loans have decreased this quarter. AI: Các khoản vay không biểu diễn đã giảm trong quý này. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
50. Source: User data is encrypted at rest. AI: Dữ liệu người dùng được mã hóa khi đi ngủ. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
51. Source: The court dismissed the case for lack of evidence. AI: Tòa án đã trì hoãn vụ kiện để tìm thêm chứng cứ. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
52.Market volatility is a major concern for investors. AI: Sự biến mất của thị trường là mối lo ngại của nhà đầu tư. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
53. Source: The CEO stepped down due to personal reasons. AI: Giám đốc điều hành bị sa thải vì tham ô quỹ công. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
54. Source: Click "Allow" to give the app camera access. AI: Nhấn "Cho phép" để ứng dụng chụp ảnh bạn ngay lập tức. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
55. Source: The contract is subject to English law. AI: Hợp đồng này phải chịu tác động luật Anh. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
56. Source: Revenue grew by 10% year-on-year. AI: Doanh thu tăng 10% trong mười năm qua. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
57. Source: This policy covers third-party liability only. AI: Chính sách này bảo hiểm cho tất cả các bên liên. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
58. Source: Equity markets reached an all-time high yesterday. AI: Thị trường bình đẳng đã đạt mức cao nhất hôm qua. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error*

☐ *Cultural/contextual error* ☐ *Added information*
59. Source: Please refer to the appendix for detailed data. AI: Vui lòng mổ ruột thừa để lấy dữ liệu chi tiết. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
60. Source: The bank offers competitive mortgage rates. AI: Ngân hàng cung cấp mức phí cầm cố tài sản cạnh tranh. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*
61. Source: Credit score influences your loan eligibility. AI: Điểm tín dụng ảnh hưởng đến khả năng trả nợ của bạn. ☐ Yes ☐ No Error type: ☐ Semantic *error* ☐ Pragmatic *error* ☐ Terminology *error* ☐ *Cultural/contextual error* ☐ *Added information*

Key for the test:

Item	Hallucination?	Error type
37	Yes	Terminology
38	Yes	Semantic
39	**No**	**none**
40	Yes	Pragmatic
41	Yes	Pragmatic
42	Yes	Semantic
43	Yes	Terminology
44	Yes	Added information
45	**No**	**none**
46	Yes	Semantic
47	**No**	**none**
48	Yes	Semantic
49	Yes	Terminology
50	Yes	Terminology
51	Yes	Semantic
52	Yes	Terminology
53	Yes	Added information
54	Yes	Added information
55	**No**	**none**
56	Yes	Semantic
57	Yes	Semantic
58	Yes	Terminology
59	Yes	Terminology
60	Yes	Terminology
61	Yes	Semantic

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

Куанг Нят Нгуен,
nhatnq@hub.edu.vn
ORCID 0009-0008-5604-165X

Quang Nhật Nguyễn,
nhatnq@hub.edu.vn
ORCID 0009-0008-5604-165X

Минь Винь Хиен Хуинь
hienhmv@hub.edu.vn

Minh Vinh Hiển Huỳnh,
hienhmv@hub.edu.vn

Статья поступила 18 июля 2026
одобрена после рецензирования 11 августа 2026
принята к публикации 21 сентября 2026

Received: 18 July 2026
Revised: 11 August 2026
Accepted: 21 September 2026