Journal of Languages & Translation
P-ISSN: 2716-9359 E-ISSN: 2773-3505
Special Issue December 2025 pp.171-179

AI in the Corner: Evaluating Literary Translation with Artificial Intelligence in Higher Education Setting

Boualem BENGHALEM[1]

University of Ain Temouchent – Algeria

boualem.benghalem@univ-temouchent.edu.dz

0000-0002-6973-7655

Received: *03/06/2025*	***Accepted:*** *25/10/2025*	***Published:*** *01/12/2025*

Abstract

This study investigates the effectiveness of artificial intelligence (AI) in evaluating literary translations within a higher education context. Literary translation, by nature, involves subjective judgment, cultural sensitivity, and stylistic interpretation elements that challenge standardization in assessment. With growing interest in AI-driven educational tools, this study explores whether AI can serve as a reliable evaluator in translation pedagogy. Nine Master's students specialising in literature and civilization were asked to translate Mahmoud Darwish's poem كمقهى صغير هو الحب[Like a Small Café, That's Love] from Arabic to English using various translation techniques studied in class. Each translation was first assessed by a human instructor and then evaluated using GPT-4-mini, a large language model developed by OpenAI. The AI was prompted to assign scores based on four criteria: accuracy, fluency and style, completeness, and grammar and mechanics. Results showed that while the AI consistently assigned lower scores—on average 7.1 points lower than the instructor, it maintained a moderate positive correlation ($r = 0.64$) with instructor rankings, indicating relative reliability in performance differentiation. However, issues such as OCR errors and conservative scoring highlighted limitations in using AI for holistic literary assessment. The findings suggest that AI tools, when properly calibrated and used alongside human oversight, can enhance efficiency and provide formative feedback in translation instruction. Nevertheless, they are not yet suitable as standalone grading solutions for creative, interpretive tasks.

Keywords: artificial intelligence; literary translation; assessment; higher education; GPT-4.

[1] Corresponding author: Boualem BENGHALEM/ boualem.benghalem@univ-temouchent.edu.dz
Journal of Languages & Translation © 2025. Published by University of Chlef, Algeria.
This is an open access article under the CC BY license http://creativecommons.org/licenses/by/4.0/

Introduction

In recent years, the integration of artificial intelligence (AI) into educational environments has gained increasing momentum, particularly in relation to its potential to enhance assessment processes (Popenici & Kerr, 2017; Luckin et al., 2016). Within higher education, one domain that presents distinctive challenges for evaluation is literary translation, which is characterized by its inherent subjectivity, creative expression, and cultural nuance (House, 2015; Venuti, 1995). Traditional assessment of literary translation relies heavily on human judgment, which, while rich in interpretive depth, is often time-consuming and susceptible to inconsistency.

The emergence of AI offers a promising avenue to augment these human-centred methods by introducing more standardised and efficient mechanisms for evaluation. This study investigates the effectiveness of an AI-based system in assessing literary translations produced by Master's students specialising in literature and civilization. Participants were asked to translate كمقهى صغير هو الحب[Like a Small Café, That's Love], a poem by Mahmoud Darwish, from Arabic to English using techniques studied in their coursework. These techniques included literal translation, free translation, adaptive translation, borrowing, calque, transposition, modulation, and equivalence. The core aim of this study was to compare AI-generated assessments with those of a human instructor to determine the degree of alignment between the two.

Research Question

How effectively can AI-based evaluation systems assess literary translations in higher education compared to traditional human grading methods?

Hypothesis

AI-based evaluation systems can reliably assess literary translations, yielding results that are consistent with those obtained through traditional human grading methods.

By exploring this hypothesis, the study seeks to contribute to ongoing discourse on the role of AI in educational assessment particularly in fields that demand subjective interpretation and creative sensitivity.

1. Literature Review

1.1. AI in Educational Assessment

The integration of artificial intelligence (AI) into educational assessment practices has become increasingly prominent, particularly in the humanities and teacher training. Saeed et al. (2025) examined the implementation of AI-based feedback systems in higher education, highlighting their capacity to streamline grading processes, reduce workload, and ensure consistency. Their findings indicate that AI tools can enhance assessment reliability by minimizing subjectivity inherent in human evaluation.

1.2 Natural Language Processing (NLP) in Translation Studies

Florou (2025) conducted a comparative analysis of advanced digital tools, including natural language processing (NLP) systems, in teacher training contexts. The study found that while NLP tools are effective in evaluating syntactic accuracy and providing linguistic analysis, they fall short in assessing the subtleties of literary translation, such as tone and metaphor. This underscores a key limitation of AI in evaluating creative and interpretative work. Earlier studies also highlighted these constraints, noting that NLP lacks the capacity for deep cultural and contextual understanding (Somers, 2003; Van den Broeck, 1978).

1.3 AI in Literary Translation

The application of large language models (LLMs) to assess literary translations has been explored by Xu et al. (2025), who analysed AI's capacity to evaluate translated texts using automated metrics like BLEU and METEOR. While these metrics are useful for analysing fidelity and lexical accuracy, the study notes that they are insufficient for capturing the cultural and emotional dimensions of literary texts. This critique is consistent with earlier work by Laviosa (2002), who emphasized the challenges of applying quantitative metrics to literary translation, where interpretative flexibility is key.

1.4 Generative AI in Higher Education

Therón Sánchez et al. (2025) investigated the use of generative AI in multilingual learning environments. Their research focused on how AI tools can produce parallel texts and assess translations in real time. Although the AI demonstrated competency in basic semantic assessment, it struggled with evaluating stylistic decisions and interpretative variations essential to literary expression.

1.5 Ethical and Pedagogical Considerations

Bareh (2025) emphasized the ethical concerns surrounding AI use in the assessment of subjective academic work. The study critiques the over-standardisation risks associated with AI, which may compromise interpretive richness and individual creativity—traits particularly valued in literary translation and the humanities. These concerns were raised in earlier literature regarding the potential dehumanisation of assessment practices when overly reliant on automation (Popenici & Kerr, 2017).

1.6Translation Quality Evaluation Metrics

Both Xu et al. (2025) and Longo et al. (2025) evaluated translation quality using automated scoring systems and LLMs. These studies confirmed that while AI can measure formal properties of translation such as grammar and word choice, it lacks the ability to assess poetic or emotional resonance, a crucial component in literary translations. Papineni et al. (2002), who developed the BLEU metric, acknowledged its limitations for evaluating translations involving complex literary forms, which require more than surface-level accuracy.

1.7 Human Assessment in Literary Translation

Traditional human evaluation of literary translation is typically grounded in a combination of formal criteria (e.g., accuracy, fluency) and qualitative judgment (e.g., literary style, cultural equivalence, tone). According to House (2015), human evaluators often use rubrics or holistic scoring guides that account for aesthetic resonance and faithfulness to the source text's cultural and emotional nuances. Venuti (1995) similarly argued that literary translation should be assessed not merely for linguistic correctness but for how well it "re-creates" the experience of the original. This nuanced human capacity forms a benchmark against which the effectiveness of AI assessment tools must be measured.

1.8 Correlation Between AI and Human Evaluation

Florou (2025) and Ghasemi et al. (2025) reported moderate to high correlation between AI and human grading in technical translation tasks. However, in cases involving literary or philosophical content, human evaluators demonstrated a better grasp of nuance and interpretive intent, resulting in differing evaluations. These findings suggest the need for hybrid evaluation models that combine AI's objectivity with human interpretative depth.

1.9 Machine Learning Models for Grading

Bahçeci et al. (2025) explored how language-based AI systems such as Microsoft Copilot performed in answering academic queries and grading tasks. While the AI produced technically correct and

structured answers, it failed to recognize subtle literary references and metaphorical language, reinforcing the limitations of current AI tools in literary assessment.

Those studies show that there is a growing consensus that AI can support, but not fully replace, human judgment in literary translation assessment. AI excels in standardizing evaluation and providing immediate feedback for syntactic and lexical accuracy. However, its limitations in interpreting cultural, aesthetic, and emotional dimensions highlight the need for complementary human oversight in the grading of literary texts. As such, the integration of AI in literary translation assessment should be approached through hybrid models that balance efficiency with interpretive richness.

2. Methodology

2.1 Participants

Nine Master's students (M1 level) specialising in Literature and Civilizations at the department of letters and English language, Faculty of Letters, Languages, and Social Sciences, University of Ain Temouchent, participated in the study. All students were enrolled in the "Literary Translation" module and had previously completed assignments involving the translation and in-class analysis of Arabic poetry using a range of techniques. Participation was voluntary, and all students provided informed consent. As the study was conducted within a regular course activity and involved no intervention beyond standard instructional practice, it was exempt from formal institutional ethics review.

2.2 Materials

The source text used for this study was Mahmoud Darwish's poem (كمقهى صغير هو الحب) "Like a Small Café, That's Love" in its original Arabic form. A published English version of the poem from Simple Arabic Poems: Easy to Read (Arabic Language Online) served as a reference translation. The translation techniques covered in the course, and referenced in this study, included broad strategies such as literal, free, and adaptive translation, as well as specific procedures like borrowing, calque, transposition, modulation, and equivalence.

2.3 Procedure

Each student was given the Arabic poem and instructed to produce an English translation within one hour. They worked individually, without access to digital tools, dictionaries, or external assistance, simulating a controlled, in-class testing environment. Their translations were handwritten to maintain authenticity and uniformity.

For manual assessment, the instructor compared each student's translation with both the original Arabic text and the published English reference. Evaluation was based on four primary criteria: accuracy, fluency and style, completeness, and grammar and mechanics. Each criterion contributed to a score on a 20-point scale. The instructor recorded comments and preliminary scores on handwritten evaluation sheets and annotated the student work directly.

The AI-based assessment was conducted using GPT-4-mini (OpenAI, 2025), a lightweight variant of the GPT-4 large language model. Scanned images of each handwritten student translation were input into the system. The AI was prompted to extract the student's name and translated text, assess the translation using the same four criteria applied by the instructor, assign a score out of 20, and generate a brief justification for each score.

Following this, the instructor reviewed both the AI's output and the original manual annotations to finalise the grades. These final scores, informed by both sources, were recorded as the definitive grades for analysis in the study.

2.4 Data Analysis

To evaluate agreement between human and AI assessments, inter-rater reliability was calculated using Cohen's kappa (κ). This statistical measure allowed the researchers to quantify the consistency between the instructor's preliminary grades and those produced by the AI system.

In instances where the scores assigned by the AI differed from the instructor's preliminary grades by three points or more, a qualitative discrepancy analysis was conducted. This involved reviewing the AI-generated justifications and the instructor's comments to identify specific translation elements such as misuse of modulation or misinterpretation of metaphor that may have contributed to the disagreement.

Finally, the translations were coded according to the types and frequency of translation techniques used. A correlation analysis was conducted to determine whether the application of particular strategies (e.g., use of calques or modulation) was associated with higher or lower AI-generated scores.

3. Results

This section presents the outcomes of the comparison between AI-assigned grades and the final grades given by the instructor for nine Master's students who translated an Arabic literary text. The analysis focuses on overall score differences, the correlation between grading sources, and patterns in score divergence. The aim was to assess whether the AI evaluation system could replicate or approximate the human instructor's grading in a meaningful and pedagogically reliable way.

3.1 Comparison of AI and Instructor Grades

Table 1 provides a detailed comparison of grades assigned by the AI system and those assigned by the instructor. The instructor's final grades were determined after reviewing the AI-generated feedback alongside their initial manual remarks. Across all nine students, the AI assigned consistently lower scores, with an average grade of 9.0 out of 20. In contrast, the instructor's average grade was 16.5. The average difference was +7.1 points in favour of the instructor scores.

Table 1: Comparison of AI and Instructor Grades

Student ID	AI Grade (out of 20)	Instructor Grade (out of 20)	Difference (Instructor – AI)
1	8	15	+7
2	6	16	+10
3	7	17.5	+10.5
4	5	16	+11
5	6	16	+10
6	8	13	+5
7	14	18	+4
8	12	18	+6
9	15	19	+4

Source: Mean AI grade = 9.0. Mean instructor grade = 16.5. Mean difference = +7.1

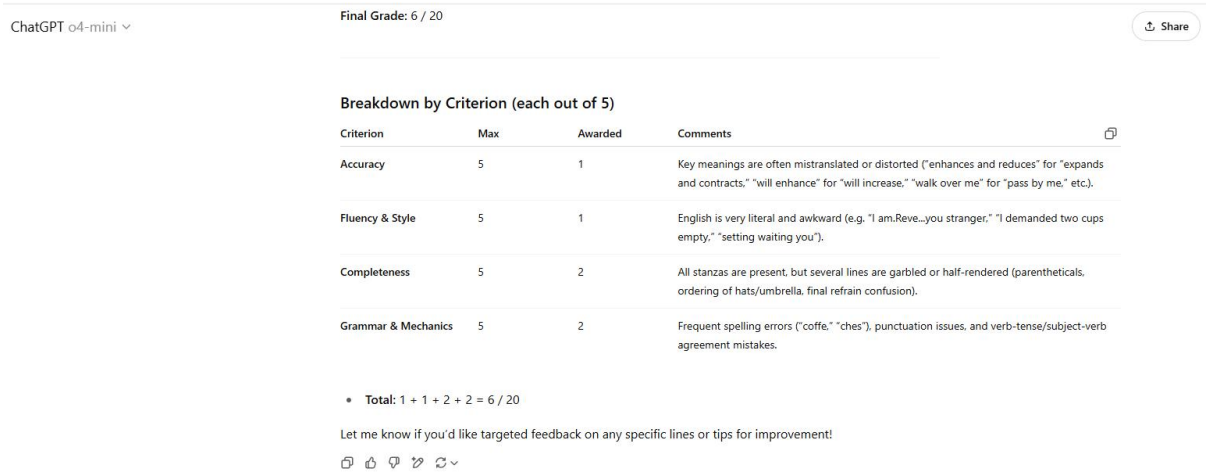
3.2 Correlation Analysis

A Pearson product-moment correlation coefficient was calculated to assess the relationship between AI-assigned and instructor-assigned grades. The result indicated a moderate positive correlation, $r = 0.64$, $p < 0.05$, suggesting that students who received higher grades from the AI also tended to receive higher grades from the instructor. However, the AI consistently assigned lower absolute values, demonstrating a more conservative grading tendency.

3.3 Pattern of Discrepancies

The data revealed two notable trends. First, in every case, the AI assigned lower grades than the instructor, confirming a systematic underestimation. The largest differences, ranging from 10 to 11 points, occurred among mid-range performers (Students 2 through 5).

Figure 1: AI-Based Evaluation Breakdown of a Student's Literary Translation Using a 20-Point Rubric



ChatGPT o4-mini

Final Grade: 6 / 20

Share

Breakdown by Criterion (each out of 5)

Criterion	Max	Awarded	Comments
Accuracy	5	1	Key meanings are often mistranslated or distorted ("enhances and reduces" for "expands and contracts," "will enhance" for "will increase," "walk over me" for "pass by me," etc.).
Fluency & Style	5	1	English is very literal and awkward (e.g. "I am.Reve...you stranger," "I demanded two cups empty," "setting waiting you").
Completeness	5	2	All stanzas are present, but several lines are garbled or half-rendered (parentheticals, ordering of hats/umbrella, final refrain confusion).
Grammar & Mechanics	5	2	Frequent spelling errors ("coffe," "ches"), punctuation issues, and verb-tense/subject-verb agreement mistakes.

- **Total:** 1 + 1 + 2 + 2 = 6 / 20

Let me know if you'd like targeted feedback on any specific lines or tips for improvement!

Source: OpenAI ChatGPT (GPT-4-mini), evaluation of student translation, May 2025

Second, the smallest discrepancies (4 to 6 points) were observed among the top three students (Students 7 through 9), indicating greater convergence between AI and instructor evaluations at higher performance levels. This pattern implies that while the AI may not align with human evaluation in absolute terms, it still captures relative ranking effectively.

4. Discussion

This study highlights both the potential and the current limitations of AI-assisted grading in higher education settings, specifically within the domain of literary translation. The following subsections examine key themes and issues that emerged during the analysis, including technical constraints, scoring consistency, and implications for teaching practice.

4.1 Name Recognition and Input Quality

The instructor observed that the AI occasionally produced incorrect or fabricated student names when analysing the scanned handwritten translations. These errors likely stemmed from optical character recognition (OCR) limitations, particularly when faced with poor handwriting or uneven scan quality. Such mistakes pose serious risks for fairness and data integrity. Future systems should implement a pre-processing validation step that checks extracted names against a verified class list or, alternatively, adopt digital text submissions to reduce dependence on OCR altogether.

4.2 Systematic Underestimation and Calibration Needs

The AI assigned lower scores across all cases, with an average underestimation of 7.1 points compared to the instructor's grades. While this suggests a bias toward stricter grading particularly in areas such as fluency, idiomatic expression, and mechanical accuracy the moderate correlation (r =0 .64) indicates that the AI captured the general ranking of student performance reasonably well. This supports the idea that AI systems may be useful as diagnostic tools or as secondary graders. However, calibration is

essential. Adjusting the weight of specific criteria or fine-tuning the grading rubric to reflect literary nuances could improve alignment with instructor expectations.

The hypothesis proposed that AI-based evaluation systems can reliably assess literary translations, yielding results that are consistent with those obtained through traditional human grading methods. The findings offer partial support for this claim. On one hand, the moderate positive correlation ($r = 0.64$) between AI and instructor-assigned grades suggests that the AI system was effective in capturing the relative performance of students. On the other hand, the consistent underestimation of scores by the AI averaging 7.1 points lower than the instructor's grades indicates a significant divergence in absolute scoring. This suggests that while the AI system may be a useful tool for comparative ranking or formative feedback, it does not yet reliably match human evaluation in summative grading for literary translation tasks. Therefore, the hypothesis is not confirmed, highlighting the need for calibration and human oversight in high-stakes assessment contexts.

4.3 Implications for Pedagogy and Fairness

Automating portions of the grading process could reduce faculty workload and provide more immediate feedback to students. Nevertheless, fairness remains a concern. If AI graders penalize non-standard idiomatic usage or fail to appreciate adaptive cultural strategies, certain students especially those with creative or unconventional approaches may be unfairly evaluated. Integrating AI into literary assessment must therefore include human oversight, especially for subjective tasks where interpretation, tone, and cultural sensitivity are central to the evaluation.

4.4 Limitations and Future Directions

Several limitations of this study should be acknowledged. First, variability in handwriting likely contributed to OCR errors, including the misidentification of student names and misinterpretation of content. Requiring typed submissions would improve input consistency. Second, the rubric used in this study emphasized accuracy, fluency, completeness, and mechanics but did not fully address stylistic innovation or creative transposition key elements in literary translation. Expanding the evaluation criteria to include poetic or cultural sensitivity could enhance the depth of AI assessment. Finally, with a sample size of only nine participants, the findings are preliminary. Future research should involve larger, more diverse cohorts across different literary genres and levels of translation expertise to validate and refine the conclusions drawn here.

Conclusion

This study evaluated the effectiveness of an AI-based grading system in assessing literary translations produced by Master's students in a higher-education context. The AI tool demonstrated moderate reliability in ranking student performance, as evidenced by a positive correlation with instructor-assigned grades. However, it systematically underestimated student achievement and struggled with input-related errors, particularly when processing handwritten submissions.

These findings suggest that while AI can assist in providing consistent and rapid feedback, it is not yet suitable as a standalone evaluator for literary translation, where nuanced interpretation, creativity, and contextual sensitivity are critical. The hypothesis that AI could reliably assess literary translations in a manner consistent with traditional human grading was only partially supported and not entirely confirmed.

To improve the validity and fairness of AI-assisted assessment, future research should include larger samples, broaden evaluation rubrics to account for adaptive and poetic strategies, and reduce dependence on OCR by using digital submissions. When carefully calibrated and used in tandem with human judgment, AI holds significant potential to enhance assessment practices in translation pedagogy.

References

Bahçeci, T., Elmaağaç, B., & Ceyhan, E. (2025). Comparative analysis of the effectiveness of Microsoft Copilot artificial intelligence chatbot and Google Search in answering patient inquiries about infertility. *International Journal of Impotence Research*. https://www.nature.com/articles/s41443-025-01056-z.pdf

Bareh, C. K. (2025). A qualitative assessment of the accuracy of AI-LLM in academic research. *AI and Ethics, Springer*. https://link.springer.com/article/10.1007/s43681-025-00730-8

Florou, K. (2025). Optimizing language analysis: A comparative study of advanced digital tools in teacher training contexts. *INTED2025 Proceedings*. https://www.researchgate.net/publication/390208905

Ghasemi, N., Rokhshad, R., Zare, Q., Shobeiri, P., & Schwendicke, F. (2025). Artificial Intelligence for Osteoporosis Detection on Panoramic Radiography: A Systematic Review and Meta Analysis. *Journal of Dentistry*, 105650. https://doi.org/10.1016/j.jdent.2025.105650

House, J. (2015). *Translation quality assessment: Past and present* (2nd ed.). Routledge.

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Rodopi.

Longo, U. G., Marino, M., Nicodemi, G., et al. (2025). Artificial intelligence applications in the management of musculoskeletal disorders of the shoulder: A systematic review. *Journal of Evidence-Based Orthopaedics*. https://esskajournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/jeo2.70248

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318.

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 1–13.

Saeed, S., Jhanjhi, P., Khan, M. A., & Yadav, D. K. (2025). Digital transformation and cybersecurity challenge. *Frontiers in Computer Science*. https://www.frontiersin.org/articles/10.3389/fcomp.2025.1631362/full

Somers, H. (2003). *Computers and translation: A translator's guide*. John Benjamins Publishing.

Therón Sánchez, R., Vázquez Ingelmo, A., & García-Holgado, A. (2025). A technology-mediated approach to addressing reading diversity in German classrooms. *GREDOS Repository*. https://gredos.usal.es/bitstream/10366/164969/1/genaichi2025_11.pdf

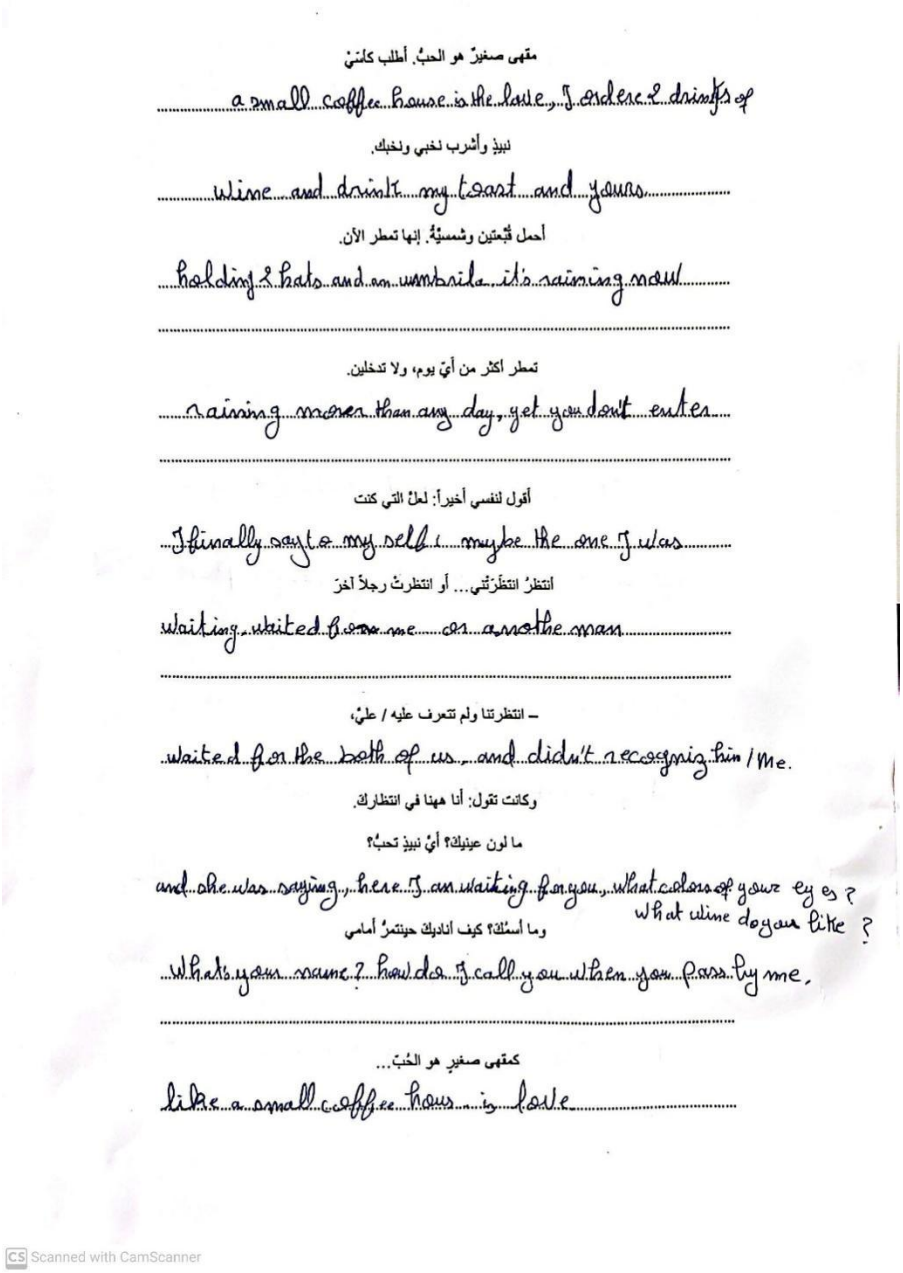
Van den Broeck, R. (1978). The concept of equivalence in translation theory: Some critical reflections. In J. S. Holmes, J. Lambert, & R. van den Broeck (Eds.), *Literature and Translation: New Perspectives in Literary Studies* (pp. 29–47). Leuven University Press.

Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

Xu, Q., Liu, Y., Wang, D., & Huang, S. (2025). Automatic recognition of cross-language classic entities based on large language models. *npj Heritage Science*. https://www.nature.com/articles/s40494-025-01624-y.pdf

Appendices:

Sample of a student's translation before manual translation



مقهى صغيرٌ هو الحبُ. أطلب كأسَيْ

a small coffee house is the love, I ordere 2 drinks of

نبيذٍ وأشرب لخبي ولخبك.

wine and drink my toast and yours

أحمل قبّعتين وشمسيّةً. إنها تمطر الآن.

holding 2 hats and an umbrella it's raining now!

تمطر أكثر من أيّ يوم، ولا تدخلين.

raining more than any day, yet you don't enter

أقول لنفسي أخيراً: لعلّ التي كنت

I finally say to my self: maybe the one I was

أنتظرُ انتظرَتني... أو انتظرتُ رجلاً آخرَ

waiting waited for me or another man

– انتظرتنا ولم تتعرف عليه / عليّ،

waited for the both of us, and didn't recognize him / me.

وكانت تقول: أنا ههنا في انتظارك.

ما لون عينيك؟ أيّ نبيذٍ تحبُ؟

and she was saying, here I am waiting for you, what colors of your eyes? What wine do you like?

وما اسمُك؟ كيف أناديك حينتمرُ أمامي

What's your name? how do I call you when you pass by me.

كمقهى صغيرٍ هو الحُبّ...

like a small coffee house is love

CS Scanned with CamScanner