Creating subject competence translation tests with GPT-4: A case study in English-to-Turkish translations in the engineering domain

Marina Sánchez-Torrón
marina.sanchez@unbabel.com

Unbabel

Egemen Ipek
Ghent University

Vanessa Enríquez Raído
Macquarie University

Research Article

Keywords: Machine Translation Post-Editing, Human-in-the-Loop, Subject Competence Assessment, ChatGPT, English-Turkish Translation, Artificial Intelligence

Posted Date: April 8th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4187415/v1

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

Additional Declarations: No competing interests reported.
Abstract

As Machine Translation (MT) technologies become more advanced, the translation errors they generate are often increasingly subtle. When MT is integrated in 'Human-in-the-Loop' (HITL) translation workflows for specialized domains, successful Post-Editing (PE) hinges on the humans involved having in-depth subject competence, as knowledge of the specific terminology and conventions are essential to produce accurate translations. One way of assessing an individual's expertise is through manual translation tests, a method traditionally used by Language Service Providers (LSPs) and translator educators alike. While manual evaluation can provide the most comprehensive overview of a translator's abilities, they have the disadvantage of being time-consuming and costly, especially when large numbers of subjects and language pairs are involved. In this work, we report on the experience of creating automated tests with GPT-4 for subject competence assessment in the translation of English-to-Turkish engineering texts in HITL translation workflows. While there may be a level of usefulness in the resulting tests, they are not fit for direct implementation without further refinement.

1 Introduction

Artificial Intelligence (AI) is revolutionizing various sectors, with automatic translation becoming more accurate and efficient. This is particularly important in the context of an increasingly globalized world, where linguistic and cultural barriers often complicate effective communication. The role of AI-enabled translations is thus paramount. AI translation encompasses automatic processes like Machine Translation (MT), speech translation, and, more recently, Large Language Model (LLM) technologies, aiming to achieve significant productivity gains. Essentially, these linguistic technologies are signaling a new epoch wherein intricate interlingual and intercultural communication is increasingly surmounted. AI has been used in translation via neural MT for a considerable period now. In the professional language industry, approximately 74% of the top Language Service Providers (LSPs) now offer neural MT and Post-Editing (PE) services (Hickey, 2023), a figure that has been on an upward trajectory in recent years, due to MTPE showing cost and productivity benefits in a number of translation scenarios.

Despite its widespread adoption in industry and education, MTPE is not without its challenges. Particularly in specialized domains, ensuring that the 'Human-in-the-Loop' (HITL) has the necessary subject competence (meaning domain-specific knowledge and human judgment) to identify and correct MT output can be difficult. The need for subject competence, alternatively known as domain-specific, thematic, or expert knowledge, is critical in professional translation and, by extension, translator education. This notion aligns with various translation competency models (e.g., PACTE, 2005) and those specifically related to PE (e.g., (Wagner, 1987, as cited in O'Brien, 2002), like the DataLitMT Competence Matrix. The latter, situated at the interface of the Professional MT Literacy Framework and the DataLitMT Framework, distinctly emphasizes subject-area competence's role in the theoretical and practical aspects of translation and MT education.
As generative AI technologies become more prevalent, re-evaluating traditional competence assessments is crucial to preserve their relevance and value in professional and educational settings (Raftery, 2023). This article explores the potential of LLMs for subject competence assessment in the context of HITL translation workflows. We report on our experience using GPT-4 (OpenAI, 2023) to generate automated multiple-choice tests for assessing translators’ subject competence in English-to-Turkish engineering texts. While LLMs have found various applications in creating and evaluating quizzes across numerous domains (e.g., medicine, law, education), a gap exists in the literature regarding their use for translation competence assessment. [2] Our research aims to contribute to this gap, focusing on the potential benefits for both LSPs seeking to streamline translator task assignments and translator education programs prioritizing Situated Learning (González-Davies & Enríquez Raído, 2016). Section 2 examines the existing literature on generative AI for learning assessment, with a specific focus on using ChatGPT for creating multiple-choice quizzes. Section 3 describes our methodology for developing English-to-Turkish translation competence tests with GPT-4, evaluating its potential for automating test generation. As mentioned, this initiative serves a dual purpose: facilitating translator task assignments within LSPs and fostering Situated Learning in translator education. Section 4 presents the results of an expert evaluation of these tests, while Section 5 addresses the challenges encountered, including the high incidence of cognates in the generated test items.

2 The Impact of AI for Multiple-Choice Assessment

As noted, generative AI technologies, particularly ChatGPT, have seen exponential growth and widespread adoption recently, mirroring the transformative effects of past innovations like the internet, Google, Wikipedia, and calculators in the educational sector (López Martín & Martín Gutiérrez, 2023, p. 4). These technologies promise to revolutionize teaching and learning by potentially reducing the time educators spend on developing and revising instructional materials, thereby allowing more focus on direct instruction and the creation of engaging, multimodal learning experiences (Shah, 2023).

However, the response from higher education institutions and educators to generative AI has been mixed. Some perceive these technologies as tools to alleviate mundane tasks and enhance focus on critical issues, while others raise concerns about threats to academic integrity when used unethically (see, e.g., Ribera & Díaz Montesdeoca, 2024; Shah, 2023). Despite these differing views, the permanence of generative AI in the educational landscape is becoming clear, shifting the debate towards how these tools can be integrated effectively, responsibly, and ethically to augment the learning experience (López Martín & Martín Gutiérrez, 2023, p. 4).

The discussion extends to the application of generative AI tools in creating diverse types of instructional activities, including single-choice, multiple-choice, true/false, fill-in-the-blank, short-answer, and essay-answer quizzes. Multiple-choice questions (MCQs) are a staple in online assessment within higher education, serving both formative and summative functions (Gonsalves, 2023, pp. 1-2; see also Beerepoot, 2023). Recent studies have explored ChatGPT’s effectiveness in answering online quiz questions across various academic fields. Raftery (2023) examined ChatGPT versions 3.5 and 4 in
answering quizzes for first-year quantitative techniques modules in an Irish technological university. The findings showed increasing accuracy across versions: ChatGPT-3.5 scored 35%, ChatGPT-4 achieved 47%, and ChatGPT-4 with the Wolfram plugin reached 78%. After correcting calculation errors, the scores improved significantly, indicating the potential for high success rates in completing online quizzes with ChatGPT’s assistance.

In the medical field, Gilson et al. (2022) found that ChatGPT could correctly answer over 60% of multiple-choice questions from a test bank for the United States Medical Licensing Examination (USMLE), performing comparably to a third-year medical student. Hoch et al. (2023) aimed to determine ChatGPT’s accuracy in otolaryngology board certification quizzes and investigate performance across subspecialties. Using a dataset from the German Society of Oto-Rhino-Laryngology, Head and Neck Surgery, they found that ChatGPT answered 57% of questions correctly, with a higher success rate in single-choice questions and variable performance among subspecialties.

Newton’s (2023a) research in early 2023 highlighted ChatGPT’s moderate performance on MCQs, particularly in areas requiring complex problem-solving, calculations, and interpreting images. However, by the end of March 2023, the introduction of ChatGPT-4 marked a significant improvement (Newton, 2023b). Its performance on MCQs was described as “really good” (as cited in Raftery, 2023), especially in handling the complexities previously found challenging. This rapid shift highlights the rapid advancements in ChatGPT’s capabilities.

Notwithstanding ChatGPT’s advanced capabilities, MCQs remain highly valuable for formative assessments and facilitating self-directed learning. Furthermore, LLMs like ChatGPT can be instrumental in streamlining MCQ creation, a traditionally labor-intensive process involving manual question and answer development. This automation can simplify and expedite test evaluation, showcasing the dual role of MCQs as both educational tools and beneficiaries of AI-driven efficiencies. Research in this area is ongoing, with Tu et al. (2023) exploring the potential of LLMs in learning assessment, particularly within data science education. Their work demonstrates the ability of LLMs to generate questions (e.g., 10 questions on hypothesis testing). However, this study did not include answer generation or an evaluation of the questions’ usability, highlighting the need for further research to ensure the quality and effectiveness of LLM-generated assessment elements.

Dijkstra et al. (2022) and Ionescu and Enescu (2023) showcase the potential of different GPT models for educational assessment. Dijkstra et al. examined GPT-3’s ability to generate educational text completions within reading comprehension quizzes. Their system, EduQuiz, generates a complete MCQ, but faces difficulties creating high-quality distractor options (cf. Section 3.2). On the other hand, Ionescu and Enescu explored ChatGPT-3’s capabilities for creating online MCQs and automating essay-type answer grading. Their implementation successfully generated quiz questions, but automatic essay evaluation proved less consistent, with issues like changing grading styles and inconsistent answer formats. These studies present contrasting findings, therefore also highlighting the need for further development to refine
LLM-based assessment tools, particularly in areas like distractor generation and consistent answer evaluation.

Tlili et al. (2023) report the results of interviews with educators, regarding their perceptions about the use of ChatGPT in learning environments. Concerns raised about quiz generation by ChatGPT are that (a) for some questions it was easy to identify the right answer; (b) in one quiz, the wrong answer was always placed at the end; (c) in another quiz, the correct answer was not given; and (d) there were format inconsistencies. The prompts that generated the quizzes shared in that study, however, are very broad, having no detailed guidelines about expected difficulty, format, or the need to provide an answer, likely leading to unsatisfactory questions and answers.

Fleming et al. (2023) generated 50 questions with GPT-4 in the style of the USMLE and asked physician reviewers to label them as correct or incorrect. Only 32 questions (64%) were deemed correct by all reviewers, while the remaining questions were deemed incorrect by at least one reviewer for having multiple correct answer choices (n=9); the AI-chosen answer being incorrect (n=6), or there being no correct answer choice (n=3).

A significant limitation of all the studies involving commercial LLMs like GPT-4 lies in the proprietary nature of these models. Since companies do not disclose their training data or architecture, and models are being constantly updated, research results are not reproducible. This limitation applies to the findings we present in this study as well. We nevertheless expect that by providing a full description of the experience of creating subject competence tests, we can offer insights that can help similar future studies involving commercial or non-commercial LLMs.

3 Methodology

In this section, we report on the method used to create English-to-Turkish translation subject competence tests focusing on engineering texts using GPT-4 (OpenAI, 2023). Our research goal is to evaluate the feasibility of automating the generation of multiple-choice tests for assessing subject competence in this domain. This serves two purposes: firstly, to aid in modeling translators for task assignments within LSPs, and secondly, to extend our insights to Situated Learning in translator education, thereby delineating the strengths and limitations of this approach. Situated Learning, as defined by González-Davies and Enríquez Raído (2016), is a context-dependent method in translator and interpreter training where learners engage with real-life or highly simulated work environments and tasks, both inside and outside the classroom. The primary objective of Situated Learning is to enhance learners' ability to think and act like professionals.

3.1 Data collection procedure

Data for this study was sourced in July 2023, querying GPT-4 using OpenAI's Playground in chat mode. Prompts were input into the OpenAI Playground, and the resulting model-generated questions were recorded. The process of settling on a prompt to generate the tests was iterative: the first prompt did not
yield desired results, so it was refined until we reached a version that we deemed was aligned with the objectives of this study. Whenever a prompt was input, previous prompts were deleted to disregard the context provided through previous prompts. Overall, and after having defined the guidelines for our tests, the process of coming up with a satisfactory prompt took about 6 hours and went through 18 iterations.

### 3.2 Question type

We generated tests composed of multiple-choice, single answer questions, widely used in educational and professional contexts for learning and skills assessment. In translation, the European Union incorporates this multiple-choice test format for assessing language knowledge and comprehension. Well-designed MCQs can target various subjects and levels of difficulty, enable automatic grading, and ensure consistency across evaluations. As Gonsalves (2023, p.2) explains, MCQs consist of a stem (question context), options (correct and misleading answers), and sometimes additional information like text, images, or audio. Typically, four options are ideal, offering a balance between having enough distractors to challenge understanding and maintaining test efficiency. These distractors aim to pinpoint learners’ gaps in knowledge, based on common errors or misconceptions.

In this study, we generated MCQs to gauge translators’ ability to recognize correct terminological correspondences in different languages. These questions offer insight into the broader spectrum of post-editing skills required in professional translation. Due to the study's exploratory scope, MCQs were considered a suitable and practical method, possibly complementing a broader assessment framework. Specifically, we set the following criteria for constructing the multiple-choice tests, in line with the writing guidelines set out in Haladyna et al. (2002):

- Questions should focus on translating engineering terms from English to Turkish within sentence contexts.
- Each question should clearly identify the term to be translated.
- Four options should be provided per question: one correct answer and three plausible distractors that are semantically similar to the correct option.
- Only one option should be the correct answer, ensuring clarity and precision in assessing the translators' knowledge.

### 3.3 Prompt formulation and iteration

Writing effective prompts for LLMs is crucial for eliciting high-quality and pertinent responses. Clear and specific prompts yield more accurate answers. Providing adequate context ensures the LLM's responses are well-directed. Using precise language thus minimizes ambiguity, while articulating the prompt's purpose explicitly guides the LLM towards the desired outcome. For complex tasks, decomposing them into smaller, manageable segments aids in generating coherent and controlled responses. Finally, indicating the desired response length helps tailor the output's detail. We kept these strategies in mind, along with the guidelines in Section 3.1, to craft our initial prompt:

"We are going to evaluate the domain expertise of translators, in this case, engineering. You are to generate multiple choice questions. There will be four choices as answers per question, only one will be
correct. The question will include a sentence and this sentence will include a term specific to engineering. Then the question will ask the correct translation of this term. The question and the sentence included in the question will be in English and the choices will be in Turkish. Incorrect answers must be terms with similar meaning to the correct answer. Plus, make sure answers do not fall into a pattern, for example the first choice being correct in all cases.

The prompt was submitted with OpenAI's preset playground options: a temperature and a Top P of 1, and both frequency and presence penalties at 0. Maximum token length was set to its upper limit, 4096. Upon analyzing the data generated by the model, several issues were identified. In half of the questions generated (n=4), the question asked for the translation of an English term where the correct answer was a cognate in Turkish. The similarities meant that any English speaker with no knowledge of Turkish could guess the right answer. This is one such example: *The torque of the engine was measured using a dynamometer. What is the Turkish translation of “torque”?*

a) hız  
b) tork  
c) güç  
d) enerji

We then appended to the prompt above *“It is very important that the Turkish translation and the English term aren’t cognates.”*, but with this new information, the percentage of questions with cognates increased to 70% (N = 10). Here’s one such example:

*The sentence: “The engineer used a multimeter to measure the voltage.” What is the Turkish translation of “multimeter”?*

a) voltmetre  
b) ampermetre  
c) multimetre  
d) direnç ölçer

Interestingly, when submitting one of the model-generated questions containing cognates and asking “Is the right answer a cognate?”, the model correctly identified them as such.

One-shot learning involves giving the model a single example or instruction from which it can learn or adjust its responses. When we applied one-shot learning by incorporating a model-generated example question without cognates into the initial prompt, the approach was ineffective. Not only did the percentage of answers with cognates not decrease, the generated questions were biased towards the terms used in the example.
After some trial and error, adjusting the temperature to 0.8 yielded a 10% reduction in answers containing cognates. Further reducing the temperature led to incorrect and incomplete suffixes, likely exacerbated by the agglutinative nature of Turkish.

A second issue we encountered were slow response times or incomplete responses, due to OpenAI’s length limits. Our initial prompt did not include a specific number of questions to generate, so the model would generate text until reaching the limit of 4096 tokens (which includes the text in our prompt). After some additional trial and error, testing different numbers of questions, we found that the model could consistently generate 20 questions in each call.

A third and final issue was that we did not specify a format for how questions and answers should be returned. With the original prompt, questions and answers were returned in a format that was appropriate for presenting them to the test taker, but for ease of processing, they would need to be in a structured format.

We finally settled on a zero-shot prompt to obtain the data presented in this study:

The task is to evaluate the engineering domain expertise of translators who translate from English to Turkish by generating a multiple-choice test with four choices. Use the term in an example sentence. Ask for the term’s translation. It is very important that the Turkish translation and the English term aren’t cognates. False answers must be terms with similar meaning to the correct answer. Present results in CSV format.

We generated five subject-area assessment tests, each composed of 20 questions. The tests were evaluated by an expert translator. The evaluation methodology is presented in Section 3.4, and the evaluation results in Section 4.

3.4 Test Evaluation

The tests were evaluated by one English-to-Turkish translator specialized in the engineering domain, one of the specializations that Unbabel, the LSP within which the study was conducted, uses to model external linguists. The evaluator holds a degree in Translation and Interpreting, and has 17 years of experience in translation, of which 1 is in specialized engineering translation. Self-reported yearly volumes translated in the past year are over 400K words, of which over 120K are in the engineering domain. We prepared a test evaluation scorecard with criteria for assessing both the individual questions and the overall tests. Table 1 shows the scorecard for evaluating the test questions individually (Questions 1–9), focusing on aspects like clarity of the question stem and quality of answer choices. The test as a whole was also evaluated using the scorecard (Questions 10–12). The evaluator took 6.5 hours to evaluate all five tests, each composed of 20 questions. They were not informed the tests had been generated by GPT-4.
<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Evaluation scope</th>
<th>Question Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Is the question statement factual?</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q2</td>
<td>Does the question use specific engineering terms in the context of an example sentence?</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q3</td>
<td>The suggested answer choice is correct</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q4</td>
<td>There is only one correct choice among the set of possible answers</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q5</td>
<td>If you’ve answered “No” to Question 4, is GPT-4’s answer choice the more widespread or accepted translation in Turkish?</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q6</td>
<td>If you’ve answered “No” to Question 4, please state which other term or terms are correct, and provide an explanation</td>
<td>Individual</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q7</td>
<td>Is the term used in the question a cognate of any of the answer choices?</td>
<td>Individual</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q8</td>
<td>Please add here any comments or observations you have about this question and answer choices. All your feedback is greatly appreciated</td>
<td>Individual</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q9</td>
<td>On a scale of 1 (Not helpful at all) to 5 (Extremely helpful), please rate how suitable do you consider this test for assessing an English-Turkish translator’s knowledge of engineering-related terminology?</td>
<td>Whole test</td>
<td>Likert scale</td>
</tr>
<tr>
<td>Q10</td>
<td>Is the test comprehensive enough to cover key areas of Engineering translation?</td>
<td>Whole test</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Q11</td>
<td>Please add here any feedback or suggestions for improvement you have about this test. Your expert view is extremely helpful for us</td>
<td>Whole test</td>
<td>Open-ended</td>
</tr>
</tbody>
</table>

### 4 Results

This section presents an expert evaluation of the tests, focusing on the accuracy and relevance of the model-generated questions.

As Table 2 shows, in 100% of the cases, the model-generated questions are factual (Q1) and follow the specified format (Q2). As for Q3, all answer choices identified as correct for Test 1 and Test 3 are indeed correct. However, in one instance in Test 2, the correct option has a spelling error: it is given as kollektör instead of kolektör, as per the Turkish Language Institution (TDK). Test 2, Test 4, and Test 5 each have one answer choice incorrectly identified as correct. Specifically, in Test 2, the choice incorrectly identified as a correct translation of “welding torch” was the less common kaynak feneri, when kaynak hamlacı, part of the generated options, is the most common translation. In Test 4, the question with no correct answer asks for the translation of “level” with the example sentence “Use a level to ensure the shelf is
straight.”, but among the options given, none is for the translation of “level” with that meaning. Finally, in Test 5, none of the answer choices for the translation of “ductility” (özgül ısı, yoğunluk, süzülme and dokulabilirlik) are correct, with süneklik being the expected translation.

Table 2
Percentage of Yes/No Questions answered with “Yes”

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Q2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Q3</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>95</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>Q4</td>
<td>60</td>
<td>100</td>
<td>90</td>
<td>85</td>
<td>55</td>
<td>78</td>
</tr>
<tr>
<td>Q5</td>
<td>87.5 (n = 8)</td>
<td>NA</td>
<td>100 (n = 3)</td>
<td>75 (n = 4)</td>
<td>70 (n = 10)</td>
<td>76</td>
</tr>
<tr>
<td>Q7</td>
<td>50</td>
<td>40</td>
<td>35</td>
<td>30</td>
<td>35</td>
<td>38</td>
</tr>
</tbody>
</table>

Answers to Q4 show more variation. Having one single correct answer choice in a multiple-choice question is desirable because it reduces ambiguity for the test taker, while providing a consistent benchmark, but only Test 2 met this criterion. Across the remaining tests, a qualitative look at the questions with multiple valid answers reveals that in some cases, the context provided is not enough to distinguish between potentially correct choices (e.g., delme vs. sondaj for “drilling”); in other cases, options provided are equivalent, but some are more colloquial (e.g., beton direkler as a translation of concrete “pillars”, vs. the more specialized beton sütunlar). In one case, one of the presented answer choices, transformator, is not even Turkish; in another, it has a grammar mistake (ortografi projeksiyonu).

In most cases, however, there is not a single unified translation to the chosen terms, resulting in various correct options, some of which are more frequent than others. Nevertheless, when there are multiple correct options, the answer choice identified by the model as being correct is the most frequent one in at least 70% of the cases, at the test level (Q5).

Lastly, for Q7, we see that all tests have a significant percentage of questions—ranging from 30–50%—where the term for which a translation is being asked is a cognate of at least one of the answer choices. In all these cases, cognates were deemed to be correct translations by the evaluator, even if in some cases the model preferred other, also correct options. For example, to the question “What is the Turkish term for “insulation”? The insulation of the house was completed by professionals.” In Test 4, the model presented izolasyon, yalıtım, koruma and engelleme as options, and identified yalıtım as the correct translation. The evaluator agreed it was correct, but pointed out that izolasyon can also be correct. In fact, yalıtım and izolasyon are synonyms: yalıtım is the Turkish origin word while izolasyon is a loanword from French.
In the assessment at the test level, for Q9, the evaluator rated all tests as “Very helpful”, or a 4 in the 5-point Likert scale. For Q10, all tests were deemed comprehensive enough to cover key areas of engineering translation. Answers to Q11 offer helpful qualitative feedback that add nuance to those assessments. Specifically, the evaluator found that in Test 1 “it was not that difficult to find the correct choice”.

Test 2 was also deemed “basic”, with distractor options being inefficiently used. One such example is a question asking for the translation of “gauge” in context, where the presented options were "ölçer ("gauge") dönüştürücü ("converter"), jeneratör ("generator") and transformator ("transformer"). In questions like these, test takers can easily narrow down the choices based on cognate recognition or by dismissing entirely unrelated terms, rather than relying on their actual subject expertise.

As the evaluation progressed, the evaluator noted that questions asked for translation of terms already present in previous tests. Table 3 shows the degree of overlap of English terms across tests, showing there are no repetitions within tests. However, across tests, the highest percentage of repetitions is 20%.

Test 5 was considered by the evaluator as being the most comprehensive of all, and the evaluator pointed out that the high presence of cognates is expected in specialized engineering texts, to differentiate them from words used in more general contexts.

<table>
<thead>
<tr>
<th>Table 3 Percentage of term repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
</tr>
<tr>
<td>Test 1</td>
</tr>
<tr>
<td>Test 2</td>
</tr>
<tr>
<td>Test 3</td>
</tr>
<tr>
<td>Test 4</td>
</tr>
<tr>
<td>Test 5</td>
</tr>
</tbody>
</table>

Regarding the repetition of terms, we see that, across all tests, the 100 questions ask for the translation of 84 unique terms, with the same terms appearing in similar contexts across tests, with different answer choices. For example, “conveyor belt” appears in two tests. In Test 1, the translation identified as correct by the model is indeed correct, but in Test 2, none of the presented translations are correct.

5 Conclusions and Future Work

We identified several key challenges and learnings in our exploration of creating subject competence translation tests using GPT-4 for assessing English-to-Turkish translation competence in the engineering domain. One significant issue was the high presence of cognates. Despite including an instruction to avoid them in our prompt, 38% of questions contained cognates. This reduces the effectiveness of the
tests due to the potential of test-takers to eliminate or guess answers without genuine subject knowledge. Research by Allen (2019), Batista and Horst (2016) and Elgort (2013) shows that including cognates in tests of lexical knowledge in the context of L2 learning significantly increases response accuracy, and it is expected that cognates will also provide an advantage in our tests. The evaluator remarked that frequent use of cognates is common in English-to-Turkish engineering translations. As such, it could be expected that the generated tests should reflect this characteristic, but not when an explicit sentence instructing otherwise is included. As far as we know, there are no studies looking at the frequency of cognates or loanwords in English-to-Turkish terms in the engineering domain. In a related study, Yu (2021) reported that 16% of frequent English words from the general vocabulary having Turkish cognates were present in the list of English engineering terms [Word, 2009, as cited in Yu, 2021]. It is likely the percentage of cognates in the specialized vocabulary is higher. To mitigate the impact of cognates, we suggest excluding questions containing them from the subject assessment tests. Cognate-heavy questions might serve as instructional tools to introduce generalist post-editors to specialized vocabulary, rather than for the purpose of evaluating their subject competence.

On a positive note, GPT-4 effectively generated factual questions, in the format requested, and accurately generated and identified a correct answer to each question. However, except for one test, all tests had questions with multiple correct answers (ranging from 2 questions in Test 3 to 9 in Test 5). In most cases, these multiple correct answers were more or less frequent translations of the English term. This reflects the very nature of translation, where context and usage can lead to multiple valid terms for a single concept.

One way to incorporate these multiple frequent translations into the test-taking process is to instruct test-takers to select not just the correct translation, but the best and most frequent translation, considering usage frequency. This approach aligns with real-world translation decision-making.

In addition, it can be worth exploring generating question types other than multiple-choice, for example, fill-in-the-blanks or short-answer formats. These can help gauge both terminological competence and context understanding for a given subject.

The context generated for the questions was always in the form of simple declarative sentences. However, real-life and pedagogical translations in specialized domains can feature complex terms with little or no context, inconsistent terms, or terms whose meaning is clarified in a different sentence. Creating tests with diverse context types would enhance the assessment's realism and educational value.

The repetition of terms across tests highlights a need for a more varied test design. This observation not only highlights a gap in our current approach but also serves as a learning point for enhancing test creation methodologies. By broadening the spectrum of terms and contextual scenarios used, we can increase test efficacy and ensure a more comprehensive assessment of subject competence. Integrating a wider range of engineering terminology and varied contexts will not only minimize repetition but also deepen the evaluative quality of the tests, providing a richer, more challenging learning and assessment environment for test-takers. This strategy aligns with the overarching goal of refining prompts to better
mirror the complexity and nuance of real-world engineering translation, thereby fostering a more robust evaluation of translators' abilities.

Regarding the broader implications of our study, we note the limited resource status of the Turkish language and the potential variability in outcomes across different domains and language pairs. The opaqueness of current LLMs like GPT-4 makes it difficult to estimate future improvements in terms of assessment-focused capabilities. The variability of results underlines the need for an interdisciplinary approach to developing reliable assessment tools, involving collaboration among linguists, subject-matter experts, translator trainers, and computer scientists. Our findings suggest the need to revisit the design of online quizzes to incorporate generative AI with human supervision. The limitations of current language models can shed light for designing MCQs that promote active engagement and critical thinking. Strategies like incorporating multimedia, using conditional logic, grounding questions in current events, and meticulously crafting distractors can counteract the model's pattern recognition, making it more difficult for the model to simply identify answer choices based on statistical patterns (Gonsalves, 2023).

Finally, despite the challenges to traditional quiz formats, generative AI holds promise for enhancing students’ (self-directed) learning experiences. Developing prompt engineering skills is increasingly crucial, highlighting the dynamic interplay between technological advancements and educational methodologies. As we anticipate further evolution in AI capabilities, with potential enhancements in multimodal interactions and improved content generation, it is essential to continue exploring how these tools can support and transform educational assessments.

Declarations

Author Contribution

“M.S contributed to the study conception and design. Material preparation, data collection and analysis were performed by M.S and E.I. The first draft of the manuscript was written by M.S, V.E and E.I. All authors read and approved the final manuscript.”

Data Availability

https://osf.io/u9sfm/?view_only=13b845ab721a4b8eaceb78230b23153c

References


**Footnotes**

1. The DataLitMT project, conducted at TH Köln – University of Applied Sciences, focuses on teaching data literacy within the context of MT. Specifically, it aims to equip students of translation and specialized communication programs at both BA and MA levels with the necessary skills to navigate data in the field of professional MT.

2. In contrast, the primary applications of LLMs in translation focus on automatic translation, as explored by Briva-Iglesias, Cavalheiro Camargo, and Dogru (2024) and Robinson et al. (2023). Other significant areas of research include the comparison of LLM-based MT with traditional neural MT systems, as studied by Castilho et al. (2023) and Zhang, Haddow, and Birch (2023), the assessment of MT quality (for instance, Kocmi & Federmann, 2023; Wang et al., 2023), the automation of MTPE (refer to Slator, 2023), and the use of LLMs as a resource or tool for translators (see Siu, 2023).

4. “Top P” controls the randomness in the prediction process. Lower temperatures make the model's responses more deterministic and conservative, while higher temperatures increase creativity and diversity. “Frequency Penalty” reduces the model's likelihood of repeating the same words or phrases within a single piece of text. A frequency penalty of 0 means there is no penalty applied for repetition, allowing the model to repeat words or phrases as often as they naturally occur based on the trained patterns. “Presence Penalty” discourages the model from mentioning the same topic or concept multiple times, aiming to increase the diversity of topics in the output. A presence penalty of 0 indicates that the model is not penalized for revisiting the same topics or concepts, allowing for more focused or in-depth exploration of a smaller set of ideas.


6. Unless a different n value is provided, n = 20.