Generating subject-matter expertise assessment questions with GPT-4: a medical translation use-case

[1,2]**Diana Silveira,** [2]**Marina Sánchez-Torrón,** [1,3]**Helena Moniz**
[1]University of Lisbon, Portugal
[2]Unbabel, Lisbon Portugal
[3]INESC-ID, Lisbon Portugal
{diana.silveira, marina.sanchez, helena}@unbabel.com

Abstract

This paper examines the suitability of a large language model (LLM), GPT-4, for generating multiple choice questions (MCQs) aimed at assessing subject-matter expertise (SME) in the domain of medical translation. The main objective of these questions is to model the skills of potential subject-matter experts in a human-in-the-loop machine translation (MT) flow, to ensure that tasks are matched to the individuals with the right skill profile. The investigation was conducted at Unbabel, an artificial intelligence-powered human translation platform. Two medical translation experts evaluated the GPT-4-generated questions and answers, one focusing on English–European Portuguese, and the other on English–German. We present a methodology for creating prompts to elicit high-quality GPT-4 outputs for this use case, as well as for designing evaluation scorecards for human review of such output. Our findings suggest that GPT-4 has the potential to generate suitable items for subject matter expertise tests, providing a more efficient approach compared to relying solely on humans. Furthermore, we propose recommendations for future research to build on our approach and refine the quality of the outputs generated by LLMs.

1 Introduction

This work presents an approach for developing an assessment tool to evaluate the subject-matter expertise (SME) of professional translators in the field of medical translation, using the large language model GPT-4. As MT becomes more predominant in translation scenarios, including specialized fields, the need for skilled experts who can identify and address quality concerns in MT-generated output is proportionately increasing.

Specialized translators require SME, which involves extensive knowledge and proficiency in a specialized domain in both the source and target languages of the relevant language pair. To measure and evaluate SME effectively, a high-quality test for translators should incorporate translation questions that assess the translator's proficiency in the target language, as well as questions that evaluate their domain-specific expertise in the source language from which they are translating (Montalt, 2007).

Evaluating the SME of potential experts in the loop across different language pairs and domains poses challenges. Implementing a system that uses SME tests to pre-screen and match subject-matter experts with MT texts in the same subject matter could improve quality in a human-in-the-loop flow. However, when done entirely by humans, creating and maintaining a comprehensive and up-to-date question bank for a wide array of language pairs and domains can become expensive and time-consuming. Other challenges are listed in Section 5.

This paper proposes a methodology for automating the creation of SME tests using GPT-4, focusing on the field of medical translation, and the English–European Portuguese and English–German language pairs. Opting for a multiple-choice questionnaire format allowed for the automation of both test generation and test grading, which increases the speed and scalability of the assessment process, while decreasing the costs. Apart from possessing high levels of objectivity, MCQ tests facilitate a fast and effective assessment process, while also being able to cover a broad range of topics (Jovanovska, 2018). Our main objective with the question banks is to distinguish between experts and non-experts, not to evaluate levels of expertise among the expert test-takers, although that might constitute a future point of research (See Section 5 for insights on future work).

We conducted a quality assessment of the generated questions and sets of answers, using an evaluation scorecard structured around five evaluation criteria (detailed in Section 2), which was completed by medical expert translators. Our analysis of the results provides insights into GPT-4's capability in generating high-quality medical expertise test items for English–European Portuguese and English–German. All the data related to this work, including question banks, evaluation and prompts, is available on GitHub[1].

2 Study setup

Our plan for implementing SME tests as part of Unbabel's framework for matching human experts to translation tasks involves generating different assessments for each test-taker, by randomly choosing a set number of questions from a large question bank. As such, our approach centered on compiling a large question bank generated by GPT-4, rather than individual tests, for each language pair. For each language pair, we invited a medical expert translator with more than ten years of professional experience to evaluate the generated questions and answers. The evaluators had no connection to the study and were paid according to their hourly rates with no time restrictions to conduct the experiment.

2.1 Question bank typology

For each language pair, we generated four separate question banks, of 50 unique questions each. Each question bank was generated with a specific prompt. The question banks are categorized by:

a) Topic: each focuses on a different area/type of document within the medical translation field. These are: a) clinical trials and clinical trial protocols; b) general medical information; c) clinical studies and d) terminology translation (which encompasses the previous three topics).

b) Language: they are either a source language only question bank or a translation question bank.

c) Question type: each bank is based on a different format of MCQ: a) multiple choice with four options, one of which is correct, b) alternate-choice questions/true or false, and c) fill-in-the-blanks questions, with four options.

Each medical translation expert, one per language pair, assessed all four question banks. Three of those question banks —fully in English, the source language— were shared across both language pairs. Only one question bank included translation-related questions.

	Question type	Subfield	Language
QB1	Multiple choice A), B), C), D)	Clinical trials and clinical trial protocols	Source language (EN)
QB2	Alternate choice	General medical information	Source language (EN)
QB3	Fill in the blanks	Clinical studies	Source language (EN)
QB4	Multiple choice, true and false, fill-in-the-blanks and alternate choice	Terminology translation	EN–PT
QB5	Multiple choice, true and false, fill-in-the-blanks and alternate choice	Terminology translation	EN–DE

Table 1. Question banks divided by categories.

2.2 Prompts and model parameters

The prompts were fed onto the GPT-4 model on November of 2023 via OpenAI's Playground, with specific parameters to shape the output: temperature, which adjusts the randomness of the model's predictions, was set to 0.4 to enhance the accuracy of the response; Top P was set to 1, ensuring the model's predictions included the whole range of possibilities; and presence penalty, which discourages repetition, was set to 0.5, encouraging the model to introduce new ideas and topics for a varied response.

The process of constructing the ideal prompt was incremental, performed in a trial and error manner. We started with a simple instruction: *"I want to test the subject matter expertise of translators in the domain of [chosen domain]. Create a questionnaire containing [chosen number of items] multiple choice questions."* On subsequent iterations, several instructions were added to the prompts, in order to curtail issues as they arose. To elevate the difficulty of the question banks, we instructed the model to produce items with a level of complexity that would

[1] https://github.com/mstorron/subject-matter-expertise-assessment-questions-with-GPT-4.

make it difficult for non-experts to answer the questionnaire correctly, and to clearly identify the correct answer in the choices given.

To ensure the output matched our language and format expectations, we specified, for QB1, QB2 and QB3, that test items were to be fully in English, with no translation items, and detailed the format of the stem and answer choices, according to each QB's typology (See Table 1).

In order to increase the quality of the distractors, we added the instruction to include plausible distractors, which we further detailed as: *"the answer choices should be similar to each other and the question in category, morphology or syntax"*. Additionally, we instructed the model to provide distractors with the same length and complexity, as well as to ensure that the correct answer and the question stem did not share words with the same word root.

For generating QB4 and QB5, we developed glossaries with domain-specific word pairs using GPT-4 beforehand, to ensure the relevance of the terms used in these question banks. Each glossary contained 50 word pairs related to general medical information, clinical trials and clinical studies, with each word pair used to create one test item.

We used zero-shot prompts for all the question banks, except for QB4, the English–European Portuguese translation question bank, for which we used a few-shot prompt, containing two examples of the ideal type of output. We used few-shot prompting only on QB4, as a means of comparing the quality of the distractors compared to zero-shot prompting; few-shot prompting yielded better distractors (See Section 5.5). As GPT-4 is a commercial model, with charges based on the combined number of input and output tokens, prioritizing zero-shot prompts is generally more cost-effective. To see the complete prompts and respective outputs, refer to the GitHub link provided in the Introduction.

2.3 Evaluation scorecard

We created a question scoring system based on five multiple choice quality criteria, which we curated based on the works of Town (2014) and Jovanovska (2018).

1. Question accuracy: is the question worded clearly and unambiguously, so that the correct answer could be clearly identified by an expert?

2. Correct answer factuality: is the correct answer choice, also known as *key*, scientifically true?

3. Non-ambiguous answer choices: is there more than one correct answer?

4. Prevalence of correct answer: is the correct answer the most prevalent and commonly applied option in the context of the question?

5. Plausible distractors in the answer choices: do the incorrect answer choices constitute plausible distractors for non-expert test-takers?

The scoring system is binary, relying on "Yes" or "No" responses to evaluate each question and its answer choices according to the above criteria. The scorecard was created using Google Sheets, containing one test item per row and one question bank per sheet. The subject-matter experts had access to the test items in the following format: question, answer choices, key (selected by the model); followed by the five criteria presented above and ending with a column for comments and a column for the score. They assess and evaluate each question and set of answer choices on each criterion using a drop-down menu with "Yes" and "No" options, which results in an automatic score based on their evaluation. To further clarify, for criterion 3 "*Is there more than one correct answer?*", the ideal answer would be "No", as ambiguity is not desired in these types of tests. In the case of criterion *5, "Do the incorrect answer choices constitute plausible distractors for non-expert test-takers?*", the ideal answer would be "Yes", as this would prevent test takers from achieving high scores simply from "guessing".

3 Results

After generating the question banks, the next step was assessing and scoring their quality. Section 3.1 presents the overall score, results and considerations for English–European Portuguese, conducted by Evaluator A, while Section 3.2 does the same for English–German, conducted by Evaluator B.

3.1 English–European Portuguese question banks scores

On Table 2, it can be observed that "Question accuracy" achieved the highest possible score on every question bank. Conversely,"Plausible distractors in the answer choices", is the overall lowest scoring parameter across all four question banks. When this was the case, it was often because the correct answer would be a term that shared the same root as a word in the question, but the distractors did not, as you can see in the following example:

What is the medical term for inflammation of the pancreas?

A) Pancreatitis
B) Gastritis

EN–PT Medical translation question banks	QB1	QB2	QB3	QB4
Question accuracy	100	100	100	100
Correct answer factuality	100	100	96	80
Non-ambiguous answer choices	96	100	98	74
Prevalence of correct answer	98	100	98	88
Plausible distractors	96	80	80	100
Average	98	96	94.4	88.4

Table 2: Scores of the EN–PT question banks.
QB1 - Clinical trials and clinical trial protocols
QB2 – General medical information
QB3 – Clinical studies
QB4 – EN-PT medical terminology

From this, we infer that GPT-4 frequently fails to follow the instruction "Do not include correct answers that share the same root as words in the stem", when given a zero-shot prompt. However, when the model was given a few-shot prompt, as is the case with QB4, the plausible distractor category achieved the highest score. Despite the plausibility of the distractors, Evaluator A stated that QB4 had a few items with ambiguous answer choices, meaning that more than one answer choice could be considered correct. On QB1, QB2 and QB3, "Correct answer factuality" scored highly, and so did "Prevalence of correct answer": in the majority of test items, the answer indicated by the model as the key (correct answer) was the most prevalent within the context of the question, according to Evaluator A.

3.2 English–German question banks scores

EN–DE Medical translation question banks	QB1	QB2	QB3	QB4
Question accuracy	100	100	100	100
Correct answer factuality	92	100	98	98
Non-ambiguous answer choices	96	98	98	86
Prevalence of correct answer	100	100	98	100
Plausible distractors	92	80	78	92
Average	96	95.6	94.4	95.2

Table 3: Scores of the EN–DE question banks.
QB1 - Clinical trials and clinical trial protocols
QB2 – General medical information
QB3 – Clinical studies
QB4 – EN-DE medical terminology

Similarly to English–European Portuguese, "Question accuracy" achieves a perfect score on all four question banks for English–German. "Correct answer factuality" scored highly across the question banks, with a few exceptions on QB1 and QB3. These two question banks also presented the highest amount of ambivalent answer choices. When there were more than two possible correct answer choices, the answer identified as correct was still the most prevalent option in QB1, QB2 and QB4, which demonstrates high precision by GPT-4 when identifying factual answers, but a lower capacity for providing answer choices that are both plausible distractors *and* unambiguously incorrect. This can seen in the following example:

The German translation for "Vertigo" is:
A) Vertigo
B) Schwindel
C) Schwindelgefühl
D) Vertigokrankheit

Option C) was selected as the key by GPT-4, but options A) and B) were also correct translations.

Once more, “Plausible distractors” was the lowest scoring parameter across all question banks, with lower scores on QB2 and QB3. Still, it is worth noting that its score, across all question banks, never reaches below 78 points out of 100.

3.3 GPT-4 generated glossaries

As mentioned in Section 2.2, the prompts for the translation question banks included glossaries of 50 word pairs, one for each test item. The English–German glossary was generated by GPT-4 in the following way: we requested a set of 50 medical terms, in English, related to the topics mentioned in Section 2.1. From that output, we asked the model to replace repetitive or irrelevant items, which we singled out from the original list, using the instruction *“Replace items [number of each item in the list]”*

Finally, we asked the model to translate the medical terms into German, resulting in a curated glossary of 50 word pairs. The evaluator considered all the translations in the glossary correct, but pointed out two repeated items that had not been previously detected.

The English–European Portuguese glossary was only partly generated by GPT-4: 26 word pairs were obtained from a publicly available glossary on the medical subfield of clinical trials, L10N Studio[2]. The remaining 24 word pairs were generated in the same way described above. This division showed clear results: the word pairs generated by GPT-4 showed some terminology and mistranslation issues. For example, “clinical pharmacology study” was translated as “estudo clínico de medicamento”, when it should be “estudo clínico de farmacologia”, and “particle therapy” was translated as “terapêutica de partículas” when it should be “terapia de partículas”. The word pairs extracted from the verified source, on the other hand, were deemed much more accurate by our SME evaluator, with only one out of the 26 term pairs not considered the ideal translation.

3.4 Overall results

Table 3 shows the overall quality of each of the five question banks generated by GPT-4 for evaluating medical translators’ subject matter expertise. (Note: for the score of QB1, QB2 and QB3, we calculated the average of the scores of both evaluators, when they differed). English–European Portuguese (four question banks combined) has an overall score of 94.2%, while English–German has an overall score of 95.3%. The results reflect the high level of quality and practical applicability of the generated question banks. In terms of perceived level of difficulty, the evaluators gave the question banks an average of 3.5 out of 5 (1 being very easy and 5 very difficult). For more on the difficulty dimension, see Section 5.4.

Question bank overall quality scores	
QB1 – Clinical trials and clinical trial protocols	97
QB2 – General medical information	95.8
QB3 – Clinical studies	94.4
QB4 – EN-PT medical terminology	88.4
QB5 – EN-DE medical terminology	95.2

Table 3: Overall scores of the five GPT-4 generated question banks. The maximum possible score for each question bank is 100.

4 Limitations

Despite its preliminary positive findings, this study presents several limitations. Firstly, the topics chosen for the question banks only represent a very small portion of medical knowledge, and only address a few of the most commonly translated medical documents. As mentioned in Section 4, reproducing the study with different and perhaps less common language-pair combinations is likely to produce different results. We hypothesize that the less common the language-pair combination, the lower the quality achieved by GPT-4 . Dac Lai et al (2023) state that there is a decrease in performance for languages other than English in natural language processing (NLP) tasks, which might be verified in the use-case of MCQ automation.The same can be said for less common language varieties.

Additionally, the evaluation of the GPT-4 output was done by only one expert per language pair. The sample size evaluated by each expert was substantial (200 question stems and 200 sets of answer choices), but extending the evaluation process to more experts can strengthen the validity of the results.

Furthermore, the ontology of subject matter expertise is vastly complex and multifaceted (Collins and Evans, 2007; Shavelson, 2010) and this paper does not intend to claim that the measurement of subject matter expertise can be fully judged by the results of MCQ assessments. The resulting MCQs of this study are tailored for specific use-cases, not to measure competency in general. They are also designed to be part of a larger assessment process, in which other specific tasks (such as reviewing a

[2] “30 common clinical trial terms (English and Portuguese (PT and BR)”, *L10N*. 05-21-2021. https://www.l10nglobal.com/en/news/30-common-clinical-trial-terms-english-and-portuguese-pt-and-br-. Accessed in November of 2023.

specialized machine translated text, for instance) contribute to a more accurate representation of the expertise level of the human-in-the-loop.

5 Conclusions and future work

MCQ automation using large language models has been a prevalent topic of research in a wide range of fields, such as reading comprehension (Sayin et al. 2024), vocabulary testing (Wang et al. 2024), programming (Doughty et al. 2024) and medical education (Kiyak, 2023), among others. In this study, we observed promising results regarding GPT-4's capability to generate SME tests for specialized translators, in the medical domain, with the English–European Portuguese and English–German language pairs. In order to verify the applicability of the findings in this study, it is recommended to replicate the study with other language pairs and subject-matter domains. While we selected GPT-4 to perform the study, the same methodology might yield high-quality results with other LLMs.

The objective of this study was to determine the viability of automating the generation of MCQs for assessing and labeling the skills of expert translators at Unbabel, to match them to specialized tasks requiring those skills. The overall quality of the four question banks combined was 94,2% for English–European Portuguese and 95.3% for English–German. This indicates that including GPT-4-generated MCQs in our expertise assessment process is a viable option. Our initial aim for the generated questions is to differentiate non-experts from experts. In the future, it may be interesting to assign different levels of expertise based on the percentage of correct answers. To achieve this, we might need to introduce more challenging questions and distractors. This will be considered once we analyze the difficulty of the current question banks.

What follows are recommendations for improving the relevance of the output from an LLM for the use case presented in this paper: generating SME test items to be ultimately used as an optimization method of task assignment in expert-in-the-loop translation flows. They can also be adapted for other contexts, MT related or otherwise. These SME test items might be relevant, for instance, in traditional translation workflows, research surveys to gather data, or in businesses, in the context of assessments and job interviews. Another relevant use case is the integration of the generated test items in self-directed learning methodologies (Loeng, 2020), either in corporate contexts or within freelance translator and reviewer training.

5.1 Language-pair glossaries

When creating the prompt for QB4 and QB5 (the translation question banks) , instead of utilizing glossaries generated by GPT-4, we recommend curating an up-to-date glossary of terms taken from one or more reliable and accredited sources; the number of word pairs to include should be the same as the number of questions requested in the prompt. Requesting a mixed format question bank (including different typologies of multiple-choice questions) resulted in varied and diversified question banks, as was the case for the translation question banks. Including high-quality specialized glossaries does constitute an extra step before crafting the prompts, but it guarantees a superior result and less intervention when it comes to the human step of reviewing, validating and (potentially) correcting the question banks.

5.2 Inclusion of relevant word pairs

For entities and organizations that have similar types of specialized documents and materials with which they work regularly, we recommend extracting the most common and relevant domain-specific terms found in the translated content, and adding them to the word-pair glossaries. That way, the tests generated by GPT-4 or other LLMs will become more tailored to the organization's workflow.

5.3 Elimination of alternate-choice question format

It is less likely to have plausible distractors with only two answer choices (Towns, 2014) and we verified that the distractors provided by GPT-4 often decreased the level of difficulty, making it easier for non-experts to guess the correct answer. Instead, we consider it is more beneficial to replace the alternate-choice question bank, QB2, with a classic four-option MCQ question bank, maintaining, however, the same topic (general medical information). On QB4 and QB5 (the translation section), we would likewise remove the instruction for including alternate-choice questions.

5.4 Evaluation of item difficulty and discrimination

A viable next step for this research would be to evaluate the level of difficulty of the generated test items. For this, we would employ a difficulty index, which would indicate the percentage of test-takers who answered each question correctly, as well as a discrimination index, which calculates the relationship between each individual test taker's test item score with the overall scores of all test takers, allowing each test item to discriminate between high and low scorers (Hingorjo and Jaleel, 2012). With this knowledge, we would be able to establish a pass/fail threshold with a percentage (to be defined) of correctly answered test items that distinguishes an expert test-taker from a non-expert, when it comes to the experts-in-the-loop who would be assigned to Unbabel's domain specific translation tasks. With the difficulty index, we could also determine which test items prove to be extremely easy or extremely difficult (therefore not good indicators of SME) and potentially eliminate them from our pool of questions. This is a necessary step before implementing the tests at Unbabel and measuring their impact on the company's workflows, to ensure that we are using only the most accurate and appropriate test materials.

5.5 Few-shot prompting

We strongly recommend using few-shot prompts containing at least two examples of test items with plausible distractors. For this study, we ascertained that distractors considered plausible had to be either semantically plausible (usually in the same category as the key) or morphologically plausible, which means they would also contain terms sharing the same word-root as the key or the question stem. This was most successfully achieved with the use of few-shot prompting on QB4, which led to the conclusion that this prompting technique is the most adequate to generate high quality distractors for the present use case of question bank automated generation.

5.6 LLM and language-pair diversity

Finally, this study should be replicated in the future with different LLMs and language pairs, as well as different areas of specialized translation, to extend its findings and further assess the validity of this type of methodology.

Acknowledgements

This work was developed within the scope of the project nº 62 - "Center for Responsible AI", financed by European Funds, namely "Recovery and Resilience Plan - Component 5: Agendas Mobilizadoras para a Inovação Empresarial", included in the NextGenerationEU funding program and was partially founded by FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020)

References

Doughty, Jacob, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, Christopher Bogart, Eric Keylor, Can Kultur, Jaromir Savelka, Majd Sakr. 2024. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in Programming Education. *Proceedings of the Australian Computing Education Conference 2024.*

Hingorjo, Mozaffer Rahim, and Farhan Jaleel. 2012. Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index, and Distractor Efficiency. *Journal of the Pakistan Medical Association*. 62(2): 142–7.

Jovanovska, Jasmina. 2018. Designing effective multiple-choice questions for assessing learning outcomes. *Infotheca - Journal for Digital Humanities* 18(1): 25–42.

L10N Global. 2018. 30 Common clinical trial terms in English and Portuguese (PT and BR). Accessed on 9/10/2023. https://www.l10nglobal.com/en/news/30-common-clinical-trial-terms-english-and-portuguese-pt-and-br-.

Loeng, Svein. 2020. Self-directed learning: A core concept in adult education. *Education Research International*, 2020: 1-12.

Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Paper presented at the Conference on Empirical Methods in Natural Language Processing 2023.*

Montalt, Vicent, and Maria González-Davies. 2006. *Medical Translation Step by Step: Learning by Drafting.* Oxfordshire, England: Routledge.

OpenAI. Accessed in November of 2023. "Playground." https://playground.openai.com

Sayin, Ayfer, et al. 2024. Using OpenAI GPT to Generate Reading Comprehension Items. *Educational Measurement Issues and Practise*. 43(1): 5–18.

Shavelson, Richard J. 2010. On the measurement of competency. *Empirical Research in Vocational Education and Training*. 2(1): 41–63.

Wang, Qiao, Ralph Rose, Naho Orita, and Ayaka Sugawara. 2024. Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5. *Proceedings of the Joint 3rd International Conference on NLP4DH and 8th IWCLU.*