Automated Translation of Islamic Literature Using Large Language Models: Al-Shamela Library Application

Mohammad Mohammad Khair[1], Majdi Sawalha[1,2,3]

[1] International Computing Institute for Quran and Islamic Sciences, USA
[2] College of Engineering, Al Ain University, Abu Dhabi, UAE.
[3] King Abdullah II School of Information Technology,
The University of Jordan, Amman, Jordan.
mohammad.khair@gmail.com, sawalha.majdi@ju.edu.jo

Abstract

Large Language Models (LLM) can be useful tools for translating Islamic literature written in Arabic into several languages, making this complex task technologically feasible, providing high-quality translations, at low cost and high-speed production enabled by parallel computing. We applied LLM-driven translation automation on a diverse corpus of Islamic scholarly works including: the Qur'an, Quranic exegesis (Tafseer), Hadith, and Jurisprudence from the Al-Shamela library. More than 250,000 pages have been translated into English, emphasizing the potential of LLMs to cross language barriers and increase global access to Islamic knowledge. OpenAI's gpt-4o-mini model was used for the forward translation from Arabic to English with acceptable translation quality. Translation quality validation was achieved by reproducing Arabic text via back-translation from English using both the OpenAI LLM and an independent Anthropic LLM. Correlating the original source Arabic text and the back-translation Arabic text using a vector embedding cosine similarity metric demonstrated comparable translation quality between the two models.

1 Introduction

Islam is the religion of more than 1.8 billion people on Earth. Yet, only about 20% of that population speak Arabic as their native language, the language of the Quran, and the rest speak their native languages. Islamic literature and the majority of its scholarly writings have been traditionally authored in Arabic with very limited or scarce translations available into other languages, hindered by the manual translation process complexity and effort that requires translators with multilingual proficiency. There exists a large volume of Arabic books in digital libraries with content extending over the last 1450 years of Islamic literature. Mass translation is feasible today using LLM models with professional-grade translation at a fraction of the cost of human translation.

The advent of Large Language Models has enabled the generation of high-quality translations, maintaining the formatting, style, and context of the original source. Parallel computing enables multi-tasking processing of translation for multiple books or to multiple languages simultaneously. LLM models have resulted in significant cost reductions via a cost-efficient API query for translation prompts.

2 Limitations for Translations

The number and type of languages supported by the LLM during its pretraining is a key criterion for selection of LLM to perform the translation task. The ability to understand Arabic language was also required since most of the Islamic books were written in Arabic.

The cost of hardware associated with servicing translation requests was another key criterion. We

Proceedings of the New Horizons in Computational Linguistics for Religious Texts (Coling-Rel), pages 53–58
January 19, 2025.

were able to load multiple small-size models (1B / 3B / 7B parameters) on a single GPU for hardware acceleration for parallel computing of separate model instances, however; the performance was limited due to the maximum GPU speed. Furthermore, the models loaded were open-source models with limited translation quality due to their small parameters size.

Model size, measured in billions of parameters, significantly affects the quality of the translated text. The smaller size models are less capable for translation and use more basic vocabulary as opposed to a more sophisticated expression style. In some instances, the LLM may revert to producing text in its predominant language that it was trained on rather than the language that was requested in the query. For example, English for Llama 3.2, and Chinese for Qwen 2.5.

Another limitation is that many of the sourced Arabic text books are in image-formatted pdf files, and not available in machine readable formatted pdf files. This necessitates the pre-processing step with Optical Character Recognition (OCR) for the recognition of the Arabic text from these files. OCR technology itself is limited in its success rate, with most existing tools are optimized for the English language, and very few OCR tools are available to process Arabic text at high cost, often missing the preservation of diacritic (tashkeel / harakat) marks in the scanned text.

The availability of LLM longer contextual memory is advantageous for continuous information flow, resulting in better translation quality. By comparison, sentence-by-sentence translation using traditional machine translation systems such as Google Translate or Meta's Seamless models do not retain translation context as compared to a LLM with a large prompt token context size.

3 Detailed Architecture and Design

3.1 LLMs for Translation

LLMs are highly suitable for translation tasks due to several factors: 1) Large context length, typically 4K-128K tokens possible which enables longer scope of text for translation, resulting in less text fragmentation via chunking, and better continuity of information and longer memory due to longer context scope. 2) LLMs support the Transformers architecture with multiple attention heads mechanism enabling it to efficiently map sequence to sequence relationships focusing on key relevant information, making it ideal for translation tasks. 3) Multiple attention heads and deep learning layers are also well suited for parallel computing architectures of GPU hardware enabling computing acceleration. 4) LLMs creates knowledge maps using pre-training on huge volume of text (Trillions of tokens) across multiple languages. 5) Finally, LLMs behavior can be customized using simple prompts, which makes them ideal for ease of use.

3.2 LLM model choice

We subjectively compared translation quality from multiple LLM models that were pretrained on Arabic and other languages, and using different models' parameters sizes. Some of these LLM models are open-source including Llama 3.2 3B, Qwen 2.5 3B, Silma 9B, Jais 13B, and Mistral 7B, and some models are proprietary including OpenAI's "gpt-4o", "gpt-4o-mini", and Anthropic's Claude 3.5 Sonnet and Claude 3.5 Haiku. Finally, the best overall performance LLM "gpt-4o-mini" was selected for low cost, fast response speed, and high quality of translations generated with expressive vocabulary that is contextually relevant and meaningful.

3.3 Prompt Engineering

The LLM prompt query specification is key for driving accuracy and quality in LLM's response to users' requirements. The prompt needs to specify several key elements: 1) Provide detailed proficient translation from Arabic language to English language 2) No transliteration 3) Use Islamic terminology and scientific expressions as possible 4) Keep translation accessible and understandable to the reader 5) Preserve the truthful representation of the source text 6) When translating text from the Quran or Hadith provide both the source Arabic text and its translation 7) Maintain page and paragraph formatting, as well as enumerated or bullet list formatting to ease reading clarity of text structure 8) Use bold headings, and 9) Separate book chapters with a clear heading in bold font and a new page break.

3.4 Parallel Computing Architecture

Accelerating the translation process is key due to the large number of books (more than 50,000 books) available for translation into multiple target languages per book. The main translation process is forked into multithreaded tasks, one task per book per language, whereby each task is responsible for completing translation of all pages of a book in the one of the requested target languages for translation, before picking up another book in sequence. Multiple parallel requests can be made simultaneously as the LLM API requests and responses are uniquely targeted to each subtask process. Using commercially available LLM APIs trades-off the need for significantly high-cost of local GPU hardware with the significantly reduced-cost of remote services, due to scale economics. Figure 1 describes the complete data flow for one translation process task, which is repeated for every page per book per target language. The resulted translations are appended in a Word and Excel output formats, the Word file is finally converted into PDF format as well.

3.5 Database Storage

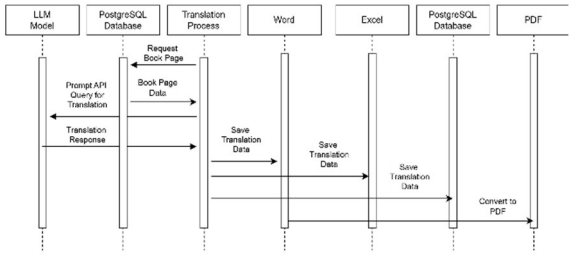

Figure 1: Sequence of Process Steps for Each Translation Task**.**

All input Arabic books text sources as well as translated output text results was stored in PostgreSQL tables. This allows instant API connectivity to the database tables via python-based SQL scripts. Iteration of all available book pages, one page per record, enables automation of an entire book translation, as well as the translation of multiple books simultaneously, via multi-threaded applications.

All text was stored in Unicode utf-8 format. The database schema was kept simple with the book tables containing one page of text per row, which represented the input to the LLM model prompts.

3.6 Output Formats

All translated books were stored in multiple file formats including PostgreSQL tables (one row per page), Microsoft Word format, Microsoft Excel format, and Adobe pdf format. We also added the ability to optionally create mp3 audio books from the translated text as well using the text-to-speech LLM service of OpenAI's "tts-1-hd" model.

3.7 Optical Character Recognition (OCR)

To enable translation of books in the format of image-based pdf files, an OCR process was performed using a multi-modal LLM, gpt-4o, that accepts as input text and images and produces text as output. Such a multi-modal LLM configuration is ideal to input prompt instructions for processing an image of a pdf page, in order to recognize the text in the image and provide it as an output via the LLM's response to the query.

The pages of each book were first saved in a high-resolution images, and these images were processed with image processing filters to remove background noise, sharpen the text quality of the words, diacritic marks, and punctuations. Then each of these page images were converted into a base64 string that is passed to the prompt query for the LLM along with a text prompt instruction requesting to scan the image; transcribe the Arabic text; maintain diacritic marks if they appear; while preserving the accuracy, and other Quranic symbols such as pause marks.

3.8 Using Retrieval Augmented Generation (RAG) to preserve Quran and Hadith script accuracy.

To preserve the accuracy of transcription of the Quran verses and the Hadith of the Prophet (SAW) is considered critical as any errors could change the meaning significantly. The original Arabic source text of the Ayas from the Quran or the Hadith text is requested to be repeated in the LLM response in addition to the English translation to preserve the reference source in the translated text. The accuracy of the source text for OCR scanned books is also important to ensure its accurate translation steps afterwards. Thus, in order to avoid potential translation errors or OCR processing errors of omission or insertion of a letter or even diacritic marks of the word, we therefore recommend applying Retrieval Augmented Generation (RAG) to the LLM query using trusted validated databases

of Quran verses and Hadith source Arabic text for OCR processing or translations for English text. RAG enables search and match using a cosine similarity metric of embedding vectors equivalent to the desired Quran verse or Hadith from a validated trusted database. Once matched, the Quran verse or Hadith in the OCR result are replaced with the RAG equivalent retrieved result. This process ensured the accuracy of the sacred texts of Quran and Hadith.

Furthermore, building this translation application is part of a larger project that aims to build a database of Islamic-specific terminology dictionary that includes translations of the Islamic terms in several languages that can improve the clarity of the translation description when submitted to the LLM via a RAG process. This represents future enhancement to this project's implementation. Finally, incorporating contextual relevance is important for generating accurate translation of ambiguous terms or phrases that could have multiple interpretations in the target language. The ability of an LLM prompt context to recall historical information within the prompt and from past prompts greatly advances the contextual relevance of the translation output. This has demonstrated advantages in contextual accuracy over sentence-by-sentence chunking and translation.

3.9 Languages Translations and Validation

Target languages for translations of interest that the gpt4o-mini LLM is capable of include: Turkish, Persian, Urdu, Malay, Bengali, Indonesian, Swahili, French, German, Russian, Spanish, and English. More languages can be added depending on the proficiency level of the LLM model. Language bias in LLMs output exists when there exists an imbalance in the training languages used, particularly in open-source models. For example, bias towards English in Llama-2.0 (but less so in Llama-3.2) or bias towards Chinese in Qwen-2.5. This can be detected through the response character set and the query retried with emphasis on the target language in the instructions. However, the used commercial LLMs , the OpenAI-4o-mini and the Anthropic's Clause-Haiku-3.0, did not display detectable language bias between Arabic and English. Validation of translation quality is accomplished using both automated and manual methods. The manual method (human-in-the-loop) includes crowd-sourcing reviews and soliciting feedback from interested readers and reviewers proficient in both the Arabic and the target language (English) who provide either acceptance or correction to the translation. Any correction requires at least two separate reviewers to accept it. Once the allowed review period ends, which varies depending on the book length, then we produce a pdf file, which is also secured from further alterations of page deletions or insertions.

The automated method for translation validation requires first using a back-translation step to regenerate Arabic text from the English, and then comparing its semantic similarity to the original source in Arabic. Both the original Arabic and back-translated Arabic are converted to vector embeddings. Then, a cosine similarity metric is applied that produces a score 0-1 for their semantic similarity or equivalence. The back-translation step can be performed either by the same primary model "gpt-4o-mini" or by an independent secondary model, such as Claude 3.5 Haiku or "claude-3-haiku-20240307". Any limitations of equivalence between original source Arabic and back-translation Arabic is primarily due to A) lack of available equivalent semantic expressions or equivalent vocabulary between Arabic and English, and then back to Arabic, B) lack of precision of meaning for words or pronouns across languages, such as words indicating singular vs plural or male vs female, and C) if the secondary model used in back-translation is different than the primary model used in forward-translation, then differences in model pre-training data sources and model design can result in performance differences between forward-translation and back-translation steps.

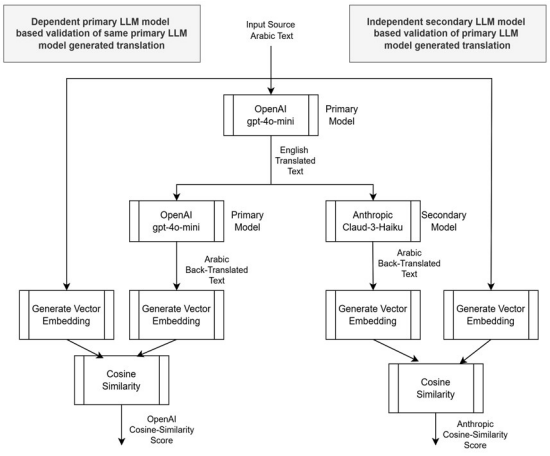

Figure 2: Automated validation of LLM translation

As shown in Figure 2, we applied an analysis using both scenarios of using the same primary model (gpt-4o-mini) for forward-translation and back-translation, and similarly using a different independent model claude-3-haiku-20240307 for back-translation. Both the original source Arabic text and the back-translated Arabic text were converted into vector embeddings and a cosine-similarity is then calculated to evaluate the semantic similarity between them to ensure fidelity of the translation process. A similarity score at or above a threshold level of 0.7-1 indicates the quality of the translation process is acceptable, and below 0.7 is poor similarity, and below 0.2 is a not-accepted outlier due to an error in forward translation. Additionally, the semantic similarity score will depend on the language model design and its pre-training strength and data sources in the Arabic language and the target language (e.g. English). The GPT-4o and the Claude-3.5-Sonnet models are considered the best models for language understanding and expression, while the GPT-4o-mini and the Claude-3.0-Haiku are the cost-effective versions of these models.

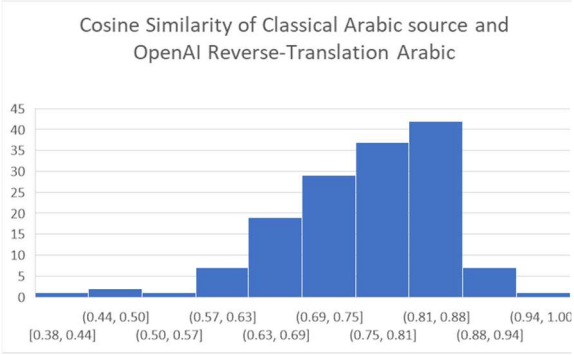

Figure 3: Distribution of OpenAI Back-Translation Cosine Similarity Scores

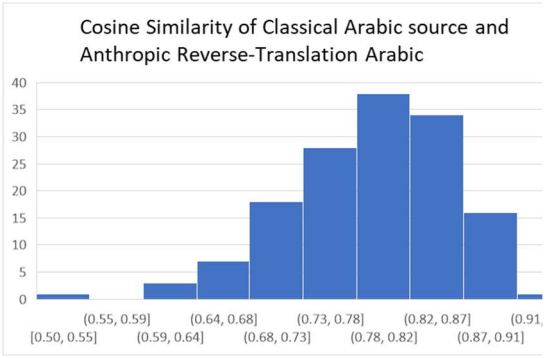

Figure 4: Distribution of Anthropic Back-Translation Cosine Similarity Scores

Furthermore, as shown in Figure 3 and Figure 4, 71% of the OpenAI model results and 86% of the Anthropic model results had cosine similarity scores >= 0.7, acceptable translations quality.

As shown in Figure 5, there exists a correlation

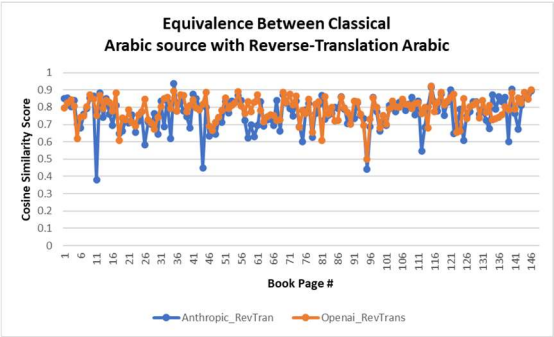

Figure 5: Equivalence Between Classical Arabic source with Back-Translation Arabic (without error outliers with scores < 0.2)

in the semantic similarity scores between the OpenAI gpt-4o-mini cosine similarity results and the Anthropic claude-3.0-haiku model cosine similarity results.

To control cost, quality checks for translation validation can be randomized checks and do not need to be systematic across the entire document being translated.

3.10 Conclusions

LLMs deep learning architectures offer a strong tool for linguistic understanding across wide variety of languages enabling them as ideal tools for translation tasks. Their ability to retain longer historical contextual information, and their attention design to focus on key information enables them to be more capable of producing high-quality translations that are contextually relevant and provide better information continuity. The longer context window also allows us to reduce fragmentation by translating one page at a time instead of one sentence at a time. The availability of a low-cost API interface for querying the LLMs enables us to parallelize the computational loads for translation queries resulting in a faster translation process execution, commercial LLMs economies of scale removes

requirements for localized LLM requirements for GPU hardware costs to perform the computations. Translation quality was validated by examining the source Arabic text and the back-translated Arabic text (produced using either OpenAI or Anthropic models) by applying a cosine similarity metric to their vector embeddings, and the results demonstrate high acceptability of translation quality for both models. We therefore highly recommend use of LLMs as reliable translation tools, specifically the OpenAI's GPT-4o-mini and Anthropic's Claude Haiku LLM models.

Acknowledgments

We would like to extend our gratitude to the Al-Shamela Library team of developers who put great care in assembling and preserving the Islamic literature. All literature works used for translation are royalty-free of copyrights, and we thank the authors of these books who granted their knowledge and work to educate the world on Islam.

References

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, Lei Li. 2024. *Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis*. Computing Research Repository, arXiv:2304.04675v4.

https://arxiv.org/abs/2304.04675

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. *Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pages 7297-7306.

https://aclanthology.org/2021.acl-long.566/

Philipp Koehn and Rebecca Knowles. 2024. *How Good Are LLMs for Literary Translation, Really?* Computing Research Repository, arXiv:2410.18697v1.

https://arxiv.org/html/2410.18697v1

Ali Albashir Mohammed Alhaj. 2020. Translating Islamic texts into English: *A Practical and Theoretical Guide for Students of Translation*. LAP LAMBERT Academic Publishing, Saarbrücken.

https://www.amazon.com/Translating-Islamic-texts-into-English/dp/6202554258

Ronit Ricci. 2011. *Islam Translated: Literature, Conversion, and the Arabic Cosmopolis of South and Southeast Asia*. University of Chicago Press, Chicago, IL.

https://press.uchicago.edu/ucp/books/book/chicago/I/bo11274031.html

Bradley J. Cook. 2010. *Classical Foundations of Islamic Educational Thought: A Compendium of Parallel English-Arabic Texts*. Brigham Young University Press, Provo, UT.

https://press.uchicago.edu/ucp/books/book/distributed/C/bo11698936.html

Abu Bakr Soliman, Kareem Eisa, and Samhaa R. El-Beltagy. 2017. *AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP*. In Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, UAE.

https://www.sciencedirect.com/science/article/pii/S1877050917321749

Haaya Naushan. 2021. *Arabic Sentence Embeddings with Multi-Task Learning*. Towards Data Science.

https://towardsdatascience.com/arabic-sentence-embeddings-with-multi-task-learning-815801024375

A Supplementary Material

All translations will made accessible at https://QuranComputing.org

Sample translation material available at Github: https://github.com/mohammadkhair7/Translations