Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 559-575.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

PerQ: A Transformer-based Multi-reference Semantic Translation Framework for Low-Resource Arabic Qur'anic Texts to Persian

Fatemeh Moodi and Hassan Majidi*

Department of Arabic Language and Literature, Hakim Sabzevari University, Sabzevar, Iran.

Article Info

Article History:
Received 16 April 2026
Revised 01 August 2026
Accepted 17 August 2026

DOI:10.22044/jadm.2026.17614.2910

Keywords:
Low-Resource Machine Translation, Multi-Reference Translation, Qur'anic Translation, Arabic–Persian Machine Translation, Neural Machine Translation.

Corresponding author: h.majidi@hsu.ac.ir (H. Majidi).

Abstract

Machine translation of low-resource and domain-specific texts remains a challenging problem, particularly in the absence of appropriate evaluation methodologies. In this study, we propose a multi-reference data augmentation framework for low-resource text translation. Two publicly available, pre-trained Transformer-based machine translation models, mBART-50 and M2M100, are employed and fine-tuned on multi-reference, domain-adapted data. A comprehensive set of automatic evaluation metrics—including BLEU, chrF, BERTScore, METEOR, and COMET—is used in both single-reference and multi-reference settings to assess translation quality. In addition, the translations are evaluated by two human evaluators with expertise in Qur'anic studies and Persian linguistics. The experimental results demonstrate substantial improvements, particularly in terms of BLEU and COMET scores, indicating enhanced translation performance for this low-resource, domain-specific task. The results further suggest that multi-reference training can improve translation quality and that multi-reference evaluation provides a broader assessment of model performance. However, automatic metrics cannot fully capture the theological, cultural, and interpretive fidelity required for Qur'anic translation; therefore, their results are complemented by human expert evaluation.

1. Introduction

With the rapid advancement of deep neural networks and natural language processing (NLP) models, the dependence of these models on large-scale data has become undeniable. Nevertheless, data scarcity persists in certain domains, necessitating effective solutions to address the associated challenges. Low-resource NLP scenarios constitute one of the four major open problems in contemporary NLP research [1]. These scenarios share a common motivation: overcoming the lack of labeled data by leveraging auxiliary resources [2].

Various approaches have been proposed to mitigate data scarcity, among which data augmentation is one of the most widely adopted. Data augmentation techniques include replacing words with their synonyms [3], semantically similar entities [4,5], or words sharing the same morphological structure [6,7]. Such substitutions can also be guided by context-aware language models [8,9]. Despite their effectiveness in general settings, the applicability of these methods is limited in specialized and semantically sensitive texts.

The Holy Qur'an can be considered a low-resource domain for machine translation due to the scarcity of high-quality parallel data and the absence of large standardized corpora (e.g., WMT) [10]. The most effective strategy for data augmentation in low-resource religious texts is to leverage the expert knowledge of domain specialists. The need for models capable of capturing subtle semantics, stylistic nuances, and cultural context in the specialized translation of religious and literary texts such as the Qur'an remains a significant challenge, one that is relevant across many

languages. Over the past two decades, considerable efforts have been made to apply machine translation techniques to Qur'anic translation into various languages [11].

Early machine translation systems were primarily based on rule-based and statistical approaches. Although these methods dominated the field for a period, they were eventually superseded by neural machine translation (NMT) models due to their substantially superior performance [12]. More recently, Transformer-based models have revolutionized machine translation through their ability to process data in parallel and capture rich semantic context, offering clear advantages over sequential models such as recurrent neural networks (RNNs) [13].

Despite significant progress in neural machine translation, applying these models to domain-specific and semantically sensitive Qur'anic translation remains challenging. The most critical challenges in this domain include:

- **Data scarcity:** One of the primary challenges in neural machine translation, particularly for sacred and classical texts, is data scarcity. The Qur'an consists of 6,236 verses, which provides a relatively small amount of parallel data for training neural machine translation models. Moreover, Persian and Arabic constitute a low-resource language pair with limited high-quality parallel corpora. Existing Persian translations of the Qur'an are heterogeneous and often diverge in interpretation, making the training of Transformer-based models difficult due to the lack of consistent, high-quality parallel data.

- **Classical Arabic and linguistic complexity:** The Qur'anic text is written in Classical Arabic, which is highly complex and stylistically unique. Most contemporary Transformer-based models are trained on modern language data that exhibit substantial syntactic, morphological, and lexical differences from Qur'anic Arabic, resulting in limited capability to model classical and low-resource linguistic structures.

- **Polysemy and contextual dependence**: Another defining characteristic of the Qur'an is its high degree of polysemy and strong reliance on context [15]. These properties pose significant challenges for traditional machine translation methods and even many modern neural models in accurately conveying meaning [1]. Many Qur'anic verses and lexical items are semantically interdependent, increasing the risk that translation models may select incorrect or overly superficial meanings.

- **Stylistic variation across translations:** Substantial stylistic variation among existing translations can cause models to generate outputs biased toward a particular translation style. This motivates the use of multiple translations to expose the model to diverse valid lexical and syntactic realizations and enhance generalization.

- **Variation in verse length:** The lengths of verses and chapters in the Qur'an are highly uneven, introducing challenges for Transformer-based models. Selecting an appropriate input length is nontrivial due to this variability, and longer input sequences can significantly increase computational cost and GPU memory consumption.

- **Evaluation sensitivity:** Unlike general-domain texts, even minor translation errors in Qur'anic translation may lead to significant theological or interpretive differences. Consequently, reliance solely on conventional automatic metrics such as BLEU is insufficient. Although semantic metrics such as BERTScore and COMET provide more context-sensitive assessments, they cannot fully capture the theological, cultural, and interpretive fidelity required for Qur'anic translation. Therefore, human expert evaluation is also important for assessing translation quality in this domain.

Proposed solutions to address challenges in machine translation include the use of attention-based and Transformer-based neural models, advanced preprocessing techniques, tokenization and subword segmentation, data augmentation strategies, multilingual learning, and approaches tailored to low-resource languages, such as multi-reference translation methods [14].

Despite the aforementioned limitations and the availability of multiple authoritative, human-produced Persian translations of the Holy Qur'an, the use of multiple Persian reference translations for systematically training and evaluating neural Arabic–Persian Qur'anic translation models has received limited attention. There is a clear need for a comprehensive examination of different machine translation approaches and their comparison against established human translations. Moreover, most existing machine translation studies rely on a single reference translation for training and evaluation. To date, a thorough investigation of the semantic and stylistic diversity inherent in Qur'anic translations—using Transformer-based models for translating Arabic Qur'anic text into other languages, and Persian in particular—has not been conducted.

Although pre-trained multilingual models such as mBART have demonstrated success in low-resource settings, their evaluation in domain-specific and semantically sensitive scenarios, such as sacred texts, especially under multi-reference fine-tuning regimes, has been largely overlooked.

In addition, there is a pressing need to construct a parallel corpus for Qur'an-to-Persian translation that mitigates the low-resource nature of this sacred text and can support a wide range of machine learning tasks, including sentiment analysis, semantic similarity, verse clustering, alignment, and information retrieval.

One effective approach to data augmentation in machine translation is the use of multiple translation sources. In this paradigm, the model is exposed to a single source text in the source language alongside multiple target-language translations, enabling it to learn a broader range of semantic and syntactic patterns rather than relying primarily on a single reference style. This diversity may help the model handle polysemy and interpretive variation and reduce dependence on a particular translation style. The use of multilingual pre-trained models is of particular importance for Qur'anic Arabic–Persian translation due to the scarcity of high-quality parallel data. These Transformer-based models, such as mBART-50 and M2M100, enable cross-lingual knowledge transfer by sharing linguistic representations learned across multiple languages.

In Qur'anic translation, semantic fidelity is more critical than surface-level lexical correspondence. By training on multiple reference translations of the same verse, the model is exposed to different lexical and syntactic realizations of similar semantic content. Furthermore, using multiple translations instead of a single reference reduces dependence on a specific translation style, promotes stylistic diversity, and may enhance model generalization and the learning of syntactic–semantic patterns. Although some verses occur multiple times in the Qur'anic text with different translations and contextual settings, this repetition reflects the original structure of the source text rather than artificial duplication. Retaining these verses therefore allows the model to learn translation patterns in their original contexts rather than removing potentially meaningful contextual information.

In this study, we fine-tune multilingual Transformer-based machine translation models for Classical Arabic–Persian Qur'anic translation using a multi-reference parallel corpus constructed from multiple authoritative Persian translations of each verse. Using multiple independent reference translations exposes the model to a wider range of valid lexical and syntactic realizations of the same semantic content. The availability of multiple authoritative Persian translations for each verse enables the model to learn semantic equivalence across stylistically different renderings while reducing dependence on any single translator's style.

The primary objective of this research is to investigate the effectiveness of multi-reference training, which leverages multiple translations to learn a shared semantic representation across different references and ultimately produce a more semantically robust translation. To this end, authoritative Persian translations of the Qur'an are collected to construct a multi-reference parallel corpus, and Transformer-based models are trained on verses with multiple translations.

The model is trained on multiple references but generates a single translation during inference, which is subsequently evaluated using both single-reference and multi-reference metrics. By focusing on multi-reference Qur'anic translation, this research contributes to the development of intelligent tools for Islamic studies, digital humanities, and broader access to Qur'anic concepts for researchers and the general public.

Machine translation systems are generally evaluated using two main approaches: human evaluation and automatic evaluation [5]. Human evaluation is considered the most reliable approach for assessing translation quality; however, it is costly, time-consuming, and requires qualified bilingual evaluators. In contrast, automatic evaluation is more cost-effective and can be easily applied to compare multiple translation systems [6].

Traditional automatic metrics mainly rely on lexical overlap, whereas recent approaches have employed contextual embeddings derived from Transformer-based language models to provide a deeper semantic assessment of translations [16]. Nevertheless, the application of such evaluation methods to low-resource languages remains challenging due to the limited availability of linguistic resources and evaluation tools. Moreover, for religious texts such as the Qur'an, automatic metrics may not fully capture theological, cultural, and interpretive nuances. Therefore, human expert evaluation is particularly important for assessing the quality and faithfulness of Qur'anic translations.

Existing Arabic–Persian Qur'anic translation studies primarily rely on single-reference corpora and automatic evaluation, which limit the ability of translation models to learn stylistic variation and semantic equivalence across different human translations.

Furthermore, the absence of publicly documented multi-reference corpora has restricted systematic investigation of multi-reference training for Qur'anic machine translation.

The main contributions of this paper are as follows:

- We construct the first publicly documented multi-reference Arabic–Persian Qur'anic parallel corpus containing approximately 37,000 sentence pairs collected from six authoritative Persian translations. The corpus includes structured metadata such as surah number, verse number, Arabic text, and Persian translations, making it highly suitable for fine-tuning, semantic similarity analysis, verse clustering, and evaluation of translation models.
- We investigate the effectiveness of fine-tuning multilingual Transformer models (mBART-50 and M2M100) on the proposed corpus for Arabic–Persian Qur'anic translation.
- We demonstrate that multi-reference training consistently improves translation quality under both automatic and human evaluation.
- We provide a comprehensive evaluation using lexical, semantic, ablation, and human evaluation protocols.

To the best of our knowledge, this is the first study to investigate multi-reference Arabic–Persian Qur'anic translation using multilingual Transformer models trained on a dedicated multi-reference corpus. The experimental results suggest that the proposed multi-reference training approach can improve translation quality and reduce dependence on individual translation styles.

2. Literature Review

In this study, the literature is reviewed across three main categories: (i) research on machine translation, (ii) studies directly related to machine translation for low-resource texts, and (iii) research in natural language processing applied to Qur'anic text.

2.1. Machine Translation for Arabic and Qur'anic texts

Li et al. introduced a multi-style machine translation system, termed MISS, which integrates high-accuracy translation, simultaneous (real-time) translation, multi-style output generation, back-translation for quality evaluation, and grammatical error correction [15]. The primary objective of this work was to enhance user experience in machine translation through dynamic interaction and continuous learning from crowdsourced data. A tag-based model with g-transformations was employed for grammatical error correction. The multi-style translation component was built upon the Transformer architecture and leveraged attention mechanisms to capture inter-word relationships. Additionally, crowdsourcing was used to collect new data for grammar correction through user feedback. The MISS system comprises four main components: Multi-style NMT (CTRL-NMT), Simultaneous NMT, Grammatical Error Correction (Tag-based GEC), and a BERTScore-based model for semantic evaluation. Training data were collected from WMT20, AI Challenger 2018, TED, and BSD corpora. For grammatical error correction, both human-annotated and pseudo-synthetic data in English, Chinese, and Japanese were used. Additional training data were continuously acquired via user feedback. Experimental results showed that models based on pre-trained language models (e.g., BERT [17] and XLNet) achieved significant improvements in grammatical correction accuracy. Moreover, CTRL-NMT was able to generate multi-style translations using a single model with performance comparable to or better than specialized models. Character-level models outperformed word-level models for Chinese and Japanese, and while simultaneous translation remained slightly inferior to standard translation, it demonstrated strong potential for further improvement.

Saeedi et al. investigated the joint effects of contextual information and morphological tokenization on the translation of Arabic social media conversations [18]. The authors employed the OPUS-mt-ar-en model (a Transformer-based model available on HuggingFace) and fine-tuned it on real-world Twitter data from Oman. The dataset comprised 228,155 messages from 2,700 authentic social media conversations collected between 2010 and 2022, with the goal of translating Omani Arabic into English. Preprocessing steps included text cleaning, removal of emojis and irrelevant messages, and length- and content-based filtering. Tokenization experiments were conducted using various schemes from CAMeL Tools, including ATB and D3. Evaluation was performed using BLEU, ChrF, NIST, and COMET metrics. Three configurations were tested: no context, full conversational context, and partial context including a limited number of previous responses up to a minimum of n words. The results showed that incorporating conversational context increased translation quality from 14.1 to 25.4 BLEU, corresponding to an improvement of approximately 80%. Morphological tokenization did not have a statistically significant impact on translation quality. The ChrF and NIST metrics also demonstrated substantial improvements in context-aware configurations.

2.2. Machine Translation of Low-Resource Texts

Danish Khaleeq et al. employed the Transformer-based MarianMT model for neural machine translation of Qur'anic verses from Arabic to Urdu [14]. Their study focused on the translation of Surah Al-Fātiḥah and Surah Yā-Sīn. Evaluation was conducted using automatic metrics such as BLEU, ROUGE, and cosine similarity, complemented by assessments from Islamic scholars. The results demonstrated that domain-specific fine-tuning on a Qur'anic parallel corpus, combined with transfer learning, improves linguistic accuracy and preserves theological meaning. Furthermore, MarianMT outperformed general-purpose translation systems. The authors identified semantic ambiguity, the structural complexity of Arabic, and the scarcity of specialized data as key challenges, and suggested data expansion, human-in-the-loop review, and multimodal learning as promising future directions.

Alabdullah et al. investigated the translation of Dialectal Arabic (DA) into Modern Standard Arabic (MSA), particularly under low-resource conditions [19]. The authors explored two main approaches. The first employed training-free prompting strategies using different large language models (LLMs). The second proposed a cost-efficient fine-tuning framework for limited resources, incorporating model quantization and joint multilingual training. Experimental results showed that, within the prompting paradigm, few-shot prompting consistently outperformed zero-shot, chain-of-thought, and the proposed Ara-TEaR method. Among the prompting-based approaches, GPT-4o achieved the best overall performance. In the fine-tuning setting, a quantized Gemma2-9B model achieved a CHrF++ score of 49.88, outperforming GPT-4o in the zero-shot configuration (CHrF++=44.58). Additionally, jointly trained multilingual models yielded more than a 10% improvement in CHrF++ compared to single-dialect models. The 4-bit quantization reduced memory usage by up to 60%, while incurring less than a 1% performance degradation.

Bamoki et al. constructed a parallel corpus for Qur'anic translation from Arabic into Kurdish (Sorani dialect) to support NLP research, particularly for low-resource languages [13]. The corpus was developed over several years. This work involved collaboration with bilingual translators, linguists, and Islamic scholars to ensure semantic fidelity and theological correctness. In addition, the authors presented detailed statistical and linguistic analyses, such as word frequency distributions, verse length statistics, and lexical diversity. This corpus serves as a valuable resource for neural machine translation, computational linguistics, and Kurdish language studies.

Faheem et al. focused on improving machine translation for low-resource languages, specifically targeting Egyptian Arabic to Modern Standard Arabic (MSA) translation [20]. While unsupervised learning is commonly employed to address such scenarios, this study explored semi-supervised learning using a combination of parallel and monolingual data. The monolingual Egyptian Arabic corpus comprised over 15 million sentences sourced from social media, while the MSA corpus included more than 20 million sentences from formal sources. The parallel corpus consisted of 40,000 Egyptian Arabic sentences manually translated into MSA. Three primary models were implemented and compared: a supervised attention-based Seq2Seq model leveraging shared vocabulary between Egyptian Arabic and MSA; an unsupervised Transformer-based MT (UMT) model trained exclusively on monolingual data and relying on cross-lingual word embedding alignment; and a semi-supervised hybrid model initially trained on parallel data and subsequently refined using monolingual data. Results showed that the semi-supervised hybrid model achieved the highest BLEU score, outperforming both the fully supervised and unsupervised models.

2.3. Natural Language Processing of Qur'anic Text

Hidayat and Menati evaluated the performance of four algorithms—SVM, Naïve Bayes, kNN, and J48—for classifying translated Qur'anic verses into the thematic categories of faith, worship, and ethics [21]. Data preprocessing was conducted in Python; tokenization was performed using StringToWordVector in WEKA; and feature weighting was based on TF–IDF. The WEKA platform was used to implement the classification algorithms, and evaluation was carried out using Accuracy, Precision, Recall, F-measure, and AUC. The experiments involved two translations, Indonesian and English, of Surah Al-Baqarah (286 verses). The highest accuracy was achieved for the Indonesian translation by the SVM model (81.4%), and the best F-measure was obtained by Naïve Bayes. Naïve Bayes achieved both the highest accuracy (78.35%) and the best F-measure for the English translation.

Touati-Hamad et al. developed an automatic system for detecting Qur'anic verses within general Arabic texts using deep learning models [22]. The study employed pre-trained word

embeddings trained with CBOW and Skip-Gram architectures on three different corpora: Wikipedia, Twitter, and the general web. For text classification, both a Convolutional Neural Network (CNN) and a hybrid CNN–LSTM model were designed and evaluated [23]. The dataset consisted of more than 12,000 text samples. Experimental results showed that the hybrid CNN–LSTM model with CBOW embeddings achieved an accuracy of 98.33% in identifying Qur'anic verses. Moreover, the average accuracy, recall, and F1-score [24] of this model reached 97.86%, representing an improvement of approximately 3% over the standalone CNN model.

Antoun et al. introduced AraBERT, a BERT-based Arabic language representation model that leverages a bidirectional Transformer architecture [25]. AraBERT was evaluated on three NLP tasks: sentiment analysis [26], named entity recognition (NER), and question answering (QA), and was compared against multilingual BERT (mBERT) and other state-of-the-art models. The model was pre-trained on a large 24 GB corpus comprising news articles as well as both Modern Standard Arabic and colloquial Arabic texts. Arabic-specific preprocessing included word segmentation using Farasa and tokenization with SentencePiece, aimed at reducing morphological complexity and improving model performance. In sentiment analysis, AraBERT outperformed mBERT and competing approaches on most datasets. For NER, AraBERTv0.1 achieved a 2.53-point improvement in F1-score over Bi-LSTM–CRF (F1 = 84.2), delivering the best performance on the ANERcorp dataset. In QA tasks, F1-scores improved, although exact match scores were slightly lower. Overall, AraBERT established a new state of the art for Arabic NLP tasks, surpassing multilingual models due to its large-scale pre-training data, expanded vocabulary, and Arabic-specific preprocessing.

Alsaleh et al. applied AraBERT to semantic similarity detection between pairs of Qur'anic verses using the QurSim dataset [27]. Due to AraBERT's limitations in handling Classical Arabic, verse pairs labeled as "1" (weakly related verses) were removed. After data cleaning, additional unrelated verse pairs (labeled "1−") were generated to balance the dataset for binary classification. Three dataset variants containing labels "2", "0", and "-1" were evaluated. Experimental results indicated that AraBERTv0.2 outperformed version v2, achieving an accuracy of 92%. Identified limitations included inadequate recognition of Classical Arabic synonyms and insufficient modeling of the religious context of verses. The authors suggested data enhancement and domain-specific fine-tuning as future research directions. This study represents the first application of AraBERT to semantic similarity analysis of Qur'anic text and provides a foundation for further research in Arabic NLP and religious text processing.

To evaluate the performance of Qur'anic question answering (QA) systems, reliable and reusable test collections are required [28]. The scarcity of such datasets in Arabic and English motivated the creation of AyaTEC, the first reusable, verse-centric test collection for evaluating Arabic Qur'an-based QA systems. AyaTEC features comprehensive answer coverage, broad topical scope, support for diverse user types, varied question formats, and rigorous evaluation protocols.

To construct AyaTEC, 145 questions were collected from curious users and research sources, 62 questions from skeptical users, and 1,762 question–answer pairs from online and printed sources. Questions with excessively long answers were removed to facilitate comprehensive annotation. For answer extraction, two Qur'an-knowledgeable freelancers identified relevant verses, following guidelines that required extracting all answer instances and contiguous verse spans. Answers were labeled as direct, indirect, or incorrect. Indirect answers were retained to distinguish questions lacking direct answers from those with zero answers. Final labels were determined via majority voting. Duplicate answers were removed using different rules for single-answer and multi-answer questions, and mappings between verses and answer instances were created to support accurate multi-answer evaluation. For single-answer questions, evaluation metrics included F1 over verse identifiers, Matching@1 (M@1), Partial Reciprocal Rank (pRR), pRecall, and pPrecision. For multi-answer questions, iRecall and pPrecision were used, and unanswered questions were assigned a score of 1. AyaTEC enables standardized and comparative evaluation of Qur'anic QA systems.

Putra et al. proposed a semi-supervised machine learning approach for Qur'anic interpretation [29]. By conducting semantic analysis of verses and employing the principle of Qur'an-by-Qur'an interpretation, the authors designed a system capable of automatically processing data and generating objective interpretations. Semantic modeling was based on a Qur'anic ontology comprising 1,050 concepts and over 2,700 relations. Additionally, a dataset named QurSim, consisting of 7,600 thematically similar verse

pairs, was developed to support interpretative data mining. The results suggest that artificial intelligence technologies can shift the human role from interpreter to scientific supervisor, potentially reducing cultural and personal biases in Qur'anic interpretation. This study represents a novel step toward automated Qur'anic exegesis and the development of intelligent, data-driven religious systems grounded in ontology-based representations.

3. Construction of a Multi-Reference Parallel Corpus for Qur'anic Translation from Arabic to Persian (QAP-MRT)

Data constitute the most critical component in the design of a machine translation model. In this study, we construct a multi-reference parallel corpus for Qur'anic verse translation from Arabic to Persian, referred to as QAP-MRT (Qur'anic Arabic–Persian Multi-Reference Translation), comprising 37,416 sentence pairs. For each Arabic verse, six independent Persian reference translations are paired with the source verse as separate parallel sentence pairs, resulting in a multi-reference training corpus. By fine-tuning Transformer-based models on a Qur'an-specific dataset, we aim to improve translation quality and generate semantically faithful and stylistically diverse outputs. The corpus is built from the complete text of the Holy Qur'an, consisting of 114 surahs and 6,236 verses. In constructing the multi-reference Qur'anic corpus, six distinct reference translations (formal, exegetical, and literary) were collected from tanzil.net, including well-established human translations such as those by Mohammad Mahdi Fooladvand, Mostafa Khorramdel, Abolfazl Bahrampour, Hussain Ansarian, Mohsen Gharaati, and Naser Makarem Shirazi. The inclusion of these multiple authoritative references enables the model to generate more diverse, fluent, and natural translations.

Since each Persian translation reflects slightly different lexical, stylistic, and interpretive choices, pairing each Arabic verse with multiple independent references exposes the model to a broader range of valid translation alternatives. Rather than forcing the model to imitate a single translation style, the multi-reference corpus encourages learning semantically equivalent expressions and improves generalization. Although stylistic and interpretive variations may introduce additional diversity, all selected translations are recognized Qur'anic translations that preserve the original meaning, making the diversity beneficial rather than detrimental for model learning.

The six reference translations exhibit differences in lexical choice, stylistic expression, and interpretive nuance. While such variation may introduce potential inconsistencies, it also provides diverse valid renderings of the same verse and allows the model to learn beyond the style of a single reference translation. These differences should therefore be considered both a source of diversity and a potential limitation of multi-reference training and evaluation.

The Holy Qur'an comprises 6,236 verses, of which 181 are repeated and 6,055 are unique. For example, the verse "فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبَانِ" occurs 31 times in Sura al-Rahman. These repetitions are part of the original Qur'anic text and contribute to its rhetorical structure and natural distribution; therefore, they do not constitute artificial data duplication. From the perspective of neural machine translation (NMT), removing such verses during training would entail discarding part of the linguistic and rhetorical knowledge of the sacred text and would disrupt verse–context dependencies. Given the relatively small proportion of repeated verses within the entire corpus (181 out of 6,236 verses), the risk of overfitting in Transformer-based models—especially when using shuffling and dropout—is minimal. Consequently, repeated verses are retained in the training data, as their removal would distort the natural distribution of the source text and weaken context-aware translation learning. To construct a multi-reference parallel corpus suitable for Transformer-based models, each Qur'anic verse is paired independently with each reference translation, forming a distinct parallel sentence pair, as follows:

$$(A_i, PT_{i,j}), i \in \{1,...,6236\}, j \in \{1,...,6\} \quad (1)$$

To construct a Transformer-compatible multi-reference parallel corpus, each Qur'anic verse is paired independently with each reference translation, forming a separate parallel sentence pair. Formally, let A_i denote the *i*-th Arabic Qur'anic verse and $PT_{i,j}$ denote its Persian translation provided by the *j*-th reference translator. Each verse is associated with a unique surah–verse identifier. Each reference translation is aligned independently, without merging or reordering the references. This design preserves precise verse alignment and enables both single-reference and multi-reference evaluation.

3.1. Data Preprocessing

After constructing the parallel corpus, several preprocessing steps were applied to ensure data quality and suitability for Transformer-based models:

- **Removal of Diacritics from Qur'anic Arabic**

Arabic consists of 28 letters and exhibits a level of complexity comparable to English. In Arabic, diacritics (ḥarakāt)—small marks placed above or below letters—are used to resolve lexical and semantic ambiguities [1]. One of the key challenges in Arabic machine translation is the presence of diacritics, as Arabic texts may be fully diacritized, partially diacritized, or entirely undiacritized (vocalized vs. unvocalized text) [30, 31]. A common preprocessing strategy in Arabic NLP is to remove all diacritics and generate undiacritized text. This practice reduces the number of surface word forms (i.e., vocabulary size), which is generally beneficial for machine translation systems [32]. Accordingly, all diacritics were removed from the Arabic Qur'anic text.

- **Removal of Extraneous Symbols**

The presence of extraneous symbols in Persian translations of the Qur'an constitutes a major text preprocessing challenge. Translations produced in different styles often contain various non-linguistic markers, such as surah and verse indices (e.g., 113|4), Persian punctuation marks (e.g., ؟ ؛ ،), explanatory glosses in square brackets (e.g., [در ذات و صفات]), and parenthetical annotations. Since the goal of the translation model is to generate natural Persian translations while preserving the semantic content of the source text, non-linguistic markers and explanatory annotations were removed from the data. Transformer-based translation models such as M2M100 learn translation patterns from bilingual statistics and alignments; the presence of brackets, parentheses, and inline explanations introduces noise into these patterns. Moreover, because the model does not inherently distinguish between verse text and commentary, such artifacts may be reproduced in the output. Therefore, these elements were removed from the input data. Standard punctuation marks (e.g., ! . ? ؛ and Persian quotation marks «…») were retained, as they contribute to sentence structure and naturalness and can be learned appropriately by the model. All preprocessing steps were implemented in Python using the pandas and re (regular expressions) libraries.

- **Variable Verse Length**

A further challenge in Qur'anic translation is the high variability in verse length. Verse 282 of Surah Al-Baqarah, known as Āyat ad-Dayn (the "Debt Verse"), which addresses financial transactions and testimony, is the longest verse in the Qur'an. Depending on orthography, it contains approximately 129 words and over 500 characters. Verse length has significant linguistic and technical implications. From a linguistic and human translation perspective, long verses often correspond to multiple sentences in the target language, requiring translators to segment meaningfully rather than translate word by word. From a technical standpoint, in language models and embedding-based representations, longer verses yield longer sequences and higher-dimensional representations. Figure 1 illustrates the distribution of word counts per surah, highlighting the uneven length distribution across the Qur'anic text.

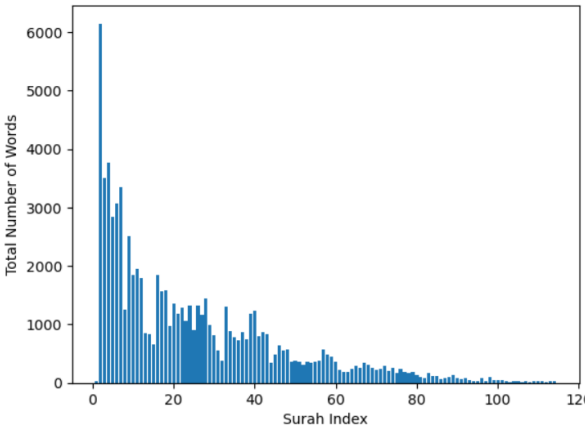

Figure 1. Distribution of Word Counts per Surah in the Quran.

As shown in Figure 1, the length distribution of Qur'anic verses is highly skewed. Most surahs consist predominantly of short verses (fewer than 20 words), while verses with 70 words or more are exceedingly rare. To avoid information loss while maintaining computational efficiency, the sequence lengths were configured such that all Qur'anic verses could be processed without truncation.

- **Disjoint Letters (Ḥurūf al-Muqattaʿāt)**

The ḥurūf al-muqattaʿāt (disjoint or mysterious letters) are a set of isolated Arabic letters in the Qur'an with distinctive properties. They appear in 14 combinations across 29 surahs, yet their precise meaning remains unknown [33]. These letters include combinations such as "الم" (Alif–Lām–Mīm), "یس" (Yāʾ–Sīn), "طه" (Ṭāʾ–Hāʾ), "حم" (Ḥāʾ–Mīm), and "کهیعص" (Kāf–Hāʾ–Yāʾ–ʿAyn–Ṣād), and have long been a subject of debate among Qur'anic exegetes. Many classical scholars maintain that their full meaning is known only to

God, while other interpretations view them as manifestations of the Qur'an's linguistic inimitability, devices for capturing the audience's attention, or symbolic references to the structure of revelation. Across Persian translations, these letters are rendered inconsistently, for example as "الف. لام. میم", "الم", or "الف، لام، میم". For the purpose of fluent and consistent machine translation of the Qur'an, it is essential to define a uniform representation for these tokens in the training data so that the model learns that no semantic transformation is intended. Therefore, in the multi-reference corpus, all ḥurūf al-muqattaʿāt were normalized to a single, consistent Persian rendering. For instance, all occurrences of "الم" were replaced uniformly with "الف، لام، میم". This normalization enables the model to learn a deterministic translation behavior for these religiously specific tokens. In addition, in the raw text files, the opening verse of each surah is preceded by "بسم الله الرحمن الرحیم" (Bismillāh al-Raḥmān al-Raḥīm), which was manually removed from the beginning of all surahs to ensure consistency across the dataset.

4. Models and Training Setup

The objective of this study is to propose a Transformer-based framework for training and optimizing a neural machine translation (NMT) model for translating Qur'anic verses from Arabic into Persian. The proposed approach aims to enhance the translation quality of religious texts with complex linguistic structures by leveraging a pre-trained multilingual model and fine-tuning it on the proposed multi-reference parallel Qur'anic corpus (QAP-MRT). Within the multi-reference framework, each Qur'anic verse is paired with multiple valid Persian translations, as described in the preprocessing section, thereby exposing the model to legitimate semantic and stylistic variations. This strategy is intended to improve both semantic robustness and linguistic naturalness.

Since the repeated verses constitute only a small fraction of the entire corpus (181 out of 6,236 verses), their impact on overfitting is expected to be limited. During training, the data were randomly shuffled at the beginning of each epoch, and dropout regularization was applied to further reduce the risk of overfitting. Consequently, the likelihood of the model memorizing these repeated verses is expected to be low. Furthermore, retaining these verses preserves the natural distribution and rhetorical structure of the Qur'anic text rather than introducing artificial duplication.

To prevent potential data leakage, the dataset was first partitioned at the level of unique Arabic verses into training, validation, and test sets using the Hugging Face Datasets library (train_test_split). Afterward, each verse within a split was paired with its six corresponding Persian reference translations. Consequently, no Arabic verse or any of its reference translations appeared in more than one data split, ensuring a fair and leakage-free evaluation.

For model development and evaluation, the unique Arabic verses were divided into training, validation, and test subsets, with 80% used for training, 10% for validation, and 10% for testing. This partitioning was performed before associating the six Persian reference translations with each verse. Thus, all six translations of a given verse remained within the same subset. The validation set was used for monitoring model performance and selecting the final training configuration, whereas the independent test set was reserved for the final evaluation.

In this work, two Transformer-based NMT models—M2M100 and mBART-50—are employed. Both models are used in their publicly available pre-trained forms (baseline models) and further fine-tuned on the proposed Qur'anic dataset. The following sections describe the architectures and configurations of the mBART-50 and M2M100 models used in this study.

4.1. Base mBART-50 Model

mBART-50 is a supervised multilingual sequence-to-sequence model introduced by Tang et al. [34]. It is designed for text generation tasks and is particularly well suited for low-resource machine translation. The model supports direct translation between any pair of 50 languages without relying on English as a pivot language; translation is controlled by prepending language-specific tokens to the source and target sequences [35].

The model contains approximately 610 million parameters and is pre-trained using a multilingual denoising pretraining objective. Let denote a set of monolingual $T = \{T_1, ..., T_L\}$ corpora across L languages, where each T_i represents a collection of documents in language *i*. During pretraining, the input documents are corrupted using two noise functions: (1) random sentence permutation and (2) a span in-filling strategy. The model is trained to reconstruct the original text, with the target sequence shifted and fed into the decoder [30]. The training objective is the cross-entropy loss defined as follows:

$$\ell_\theta = \sum_{T_i \in T} \sum_{y \in T_i} \log P(y \mid g(y); \theta) \quad (2)$$

where y is a training example in language i, and P denotes the probability distribution defined by the sequence-to-sequence model.

In this study, we use the publicly available many-to-many multilingual mBART-50 model (facebook/mbart-large-50-many-to-many-mmt) as the baseline model, referred to as mBART-50-ZS. This model consists of 12 encoder layers, 12 decoder layers with cross-attention, 16 attention heads, a hidden size of 1024, and a linear softmax output layer.

4.2. Fine-Tuned mBART-50 Model

The model is obtained by fully fine-tuning the multilingual many-to-many mBART-50 model (facebook/mbart-large-50-many-to-many-mmt) on the proposed QAP-MRT corpus for Arabic-to-Persian Qur'anic translation. Fine-tuning is performed using the optimized hyperparameters listed in Table 1. The overall workflow of the proposed framework is illustrated in Figure 2.

Table 1. Hyperparameters of the model.

Parameter	Value
Epochs	4
Batch size (train)	4
Batch size (eval)	4
Learning rate	3e-5
Max sequence length	512
padding	max_length
Optimizer	AdamW
Loss Function	Cross-Entropy Loss

The hyperparameters listed in Table 1 were selected through preliminary validation experiments based on convergence behavior and performance on the validation set. The same hyperparameter configuration was used for the fine-tuned models to ensure a fair comparison. During training, the data were randomly shuffled at each epoch, and dropout regularization and early stopping were employed to improve generalization and reduce the risk of overfitting. The multi-reference training setup also provides a form of data augmentation by exposing the models to six human-authored Persian translations of each Qur'anic verse, thereby increasing the diversity of target-language realizations without introducing synthetically generated translations.

The maximum token length values were selected to fully accommodate all Qur'anic verses without truncation, ensuring complete coverage of the source text during model training and inference.

In the flowchart shown in Figure 2, the input data consist of the complete Arabic text of the Holy Qur'an and the corresponding Persian translations. Each verse is repeated six times, with each repetition aligned to a translation produced by a different human translator, resulting in a total of 37,416 rows. The data are subsequently preprocessed and tokenized using SentencePiece.

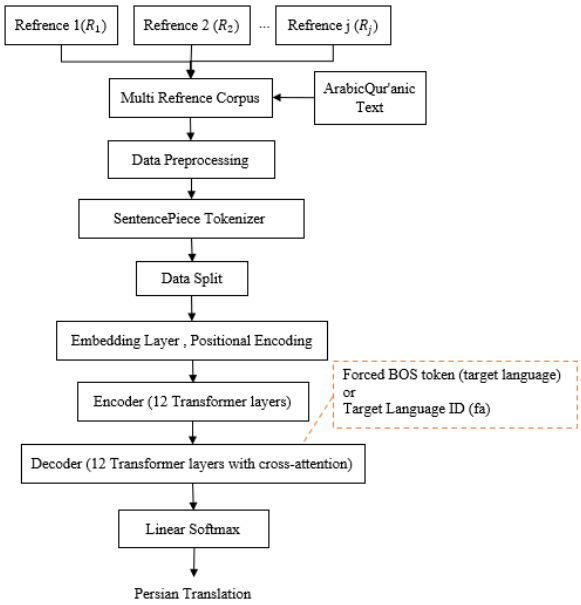

Figure 2. Flowchart of the proposed framework.

Transformer-based models are then employed for training. The architecture conceptually consists of one encoder for the Arabic verse, multiple encoders corresponding to the Persian translations, and a single decoder that generates the final Persian translation. By training on verses with multiple reference translations, the model implicitly learns a one-to-many mapping through exposure to multiple valid targets (supervised learning with multiple references). The output is a single Persian translation, sampled from a learned distribution grounded in multiple high-quality human references, and evaluated using both single-reference and multi-reference metrics.

4.3. Base M2M100 Model

The M2M100 model is a multilingual machine translation model introduced in 2020 by Angela Fan et al. and is publicly available via the Hugging Face library. It is a Transformer-based encoder–decoder architecture designed for direct translation between 100 languages without reliance on English as a pivot language [36]. The model is implemented in PyTorch and uses special language identification tokens at the beginning of the input sequence to specify the source and target languages.

M2M100 comprises 12 encoder layers, 12 decoder layers with cross-attention, 16 attention heads, a hidden dimension of 1024, and a linear softmax output layer. Structurally, it is similar to mBART-50. The M2M100 model (facebook/m2m100_418M) is one of the most suitable options for Arabic-to-Persian translation and contains approximately 418 million parameters. It is capable of jointly modeling multiple languages (418 languages) and generating more semantically accurate translations. The training objective of this model is based on the cross-entropy loss function, and the target language is controlled via a beginning-of-sentence (BOS) token. In this study, the publicly available pre-trained multilingual M2M100 (418M) model is used as the baseline model, denoted as M2M100-ZS.

4.4. Fine-Tuned M2M100 Model

In addition, the multilingual Facebook M2M100 (418M) Transformer model was fine-tuned on the proposed parallel Qur'anic corpus using the optimized hyperparameters listed in Table 1. The hyperparameters were selected through iterative experimentation and were kept consistent with the configuration used for the fine-tuned mBART-50 model to ensure a fair comparison. During inference, beam search decoding with a beam size of 10 was employed, which improves translation quality at the cost of increased computational complexity. The overall architecture of the proposed framework follows the framework illustrated in Figure 2. The abbreviated names of all models used in this study are summarized in Table 2.None of the baseline models listed in Table 2 (mBART-50-ZS and M2M100-ZS) were exposed to Qur'anic Arabic or Persian translations during their pretraining phase; therefore, fine-tuning is required to achieve high-quality performance in the task of Qur'an-to-Persian translation.

Table 2. Abbreviated names of the models.

Model	Training Setup
mBART-50-MRT	Fine-tuned with QAP-MRT
M2M100-MRT	Fine-tuned with QAP-MRT
mBART-50-ZS	Zero-shot pre-trained baseline
M2M100- ZS	Zero-shot pre-trained baseline

5. Machine Translation Evaluation

Automatic evaluation metrics are widely used in machine translation research due to their efficiency and low cost. These metrics can be broadly categorized into n-gram–based (surface-level) metrics and semantic-based metrics.

Classic n-gram–based metrics used include:

- BLEU (Bilingual Evaluation Understudy): A widely adopted metric that computes translation quality by measuring the overlap of n-grams between the generated translation and reference translations. BLEU primarily evaluates adequacy and fluency but is known to have limited sensitivity to deeper linguistic qualities, such as grammatical correctness and semantic faithfulness [37].
- METEOR (Metric for Evaluation of Translation with Explicit ORdering): A metric that extends exact word matching by incorporating synonyms, stems, and morphological variants, providing a more linguistically informed comparison than BLEU.

Semantic-based metrics used include:

- chrF (character n-gram F-score): Particularly effective for morphologically rich languages such as Arabic and Persian, chrF operates at the character level, allowing it to better capture inflectional and derivational variations than word-based metrics.
- BERTScore (Bidirectional Encoder Representations from Transformers Score): Utilizes contextual embeddings from pretrained language models such as BERT to compute semantic similarity between machine translations and reference translations [38].
- COMET (Crosslingual Optimized Metric for Evaluation of Translation): A learned evaluation metric based on XLM-R and BERT architectures, designed to assess semantic adequacy. COMET has consistently achieved state-of-the-art performance in multiple WMT shared tasks [39].

5.1. Evaluation Results of the Models

This section reports the evaluation results of the models with fine-tuning on the proposed multi-reference Arabic–Persian Qur'anic parallel corpus (QAP-MRT) as well as without fine-tuning (baseline models). Both single-reference and multi-reference evaluation settings are employed. Additionally, two human evaluators participated in the assessment for human evaluation. In the experimental setup, 90% of the dataset is used for training and 10% for validation. Since the Qur'an constitutes a closed, low-resource corpus (i.e., its content is fixed and cannot be expanded), no separate test set is required. Instead, translation outputs are evaluated using multiple reference-based metrics. Furthermore, we report both the average COMET score and the COMET Max-over-References score across the six human

Persian translations. Over three epochs of training, the M2M100-MRT model showed a clear and steady reduction in loss, decreasing from 1.15 to 0.193, which indicates reliable convergence behavior. A similar trend was observed for the mBART-50-MRT model, whose loss declined from 0.070 to 0.038 over the same training period. The quality of the resulting translations was primarily evaluated using the COMET and BLEU metrics.

5.2. Single-Reference Evaluation

Here, the model output is compared separately with each reference translation using the evaluation metrics BLEU, chrF, METEOR, and BERTScore, in order to explicitly account for the diversity of human translations. The comparison of single-reference evaluation metrics for the M2M100-MRT model is illustrated in Figure 3, while the corresponding results for the mBART-50-MRT model are shown in Figure 4.

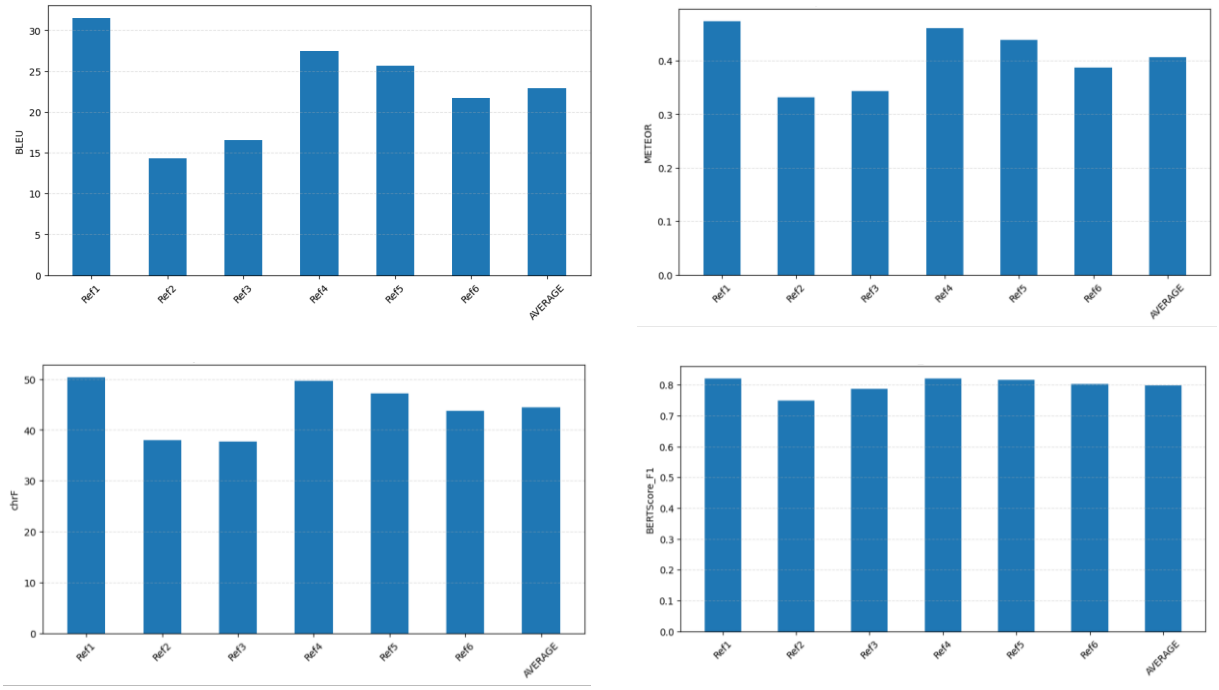

Figure 3. Comparison of the single-reference evaluation metrics for the M2M100-MRT model.

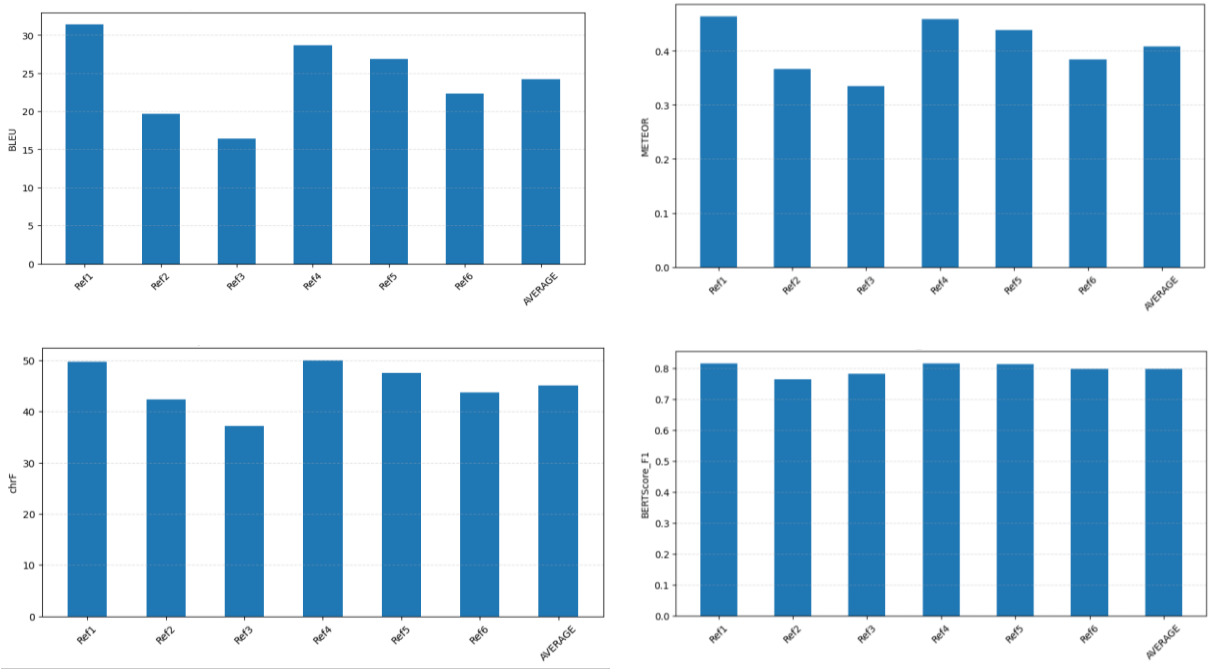

Figure 4. Comparison of the single-reference evaluation metrics for the mBART-50-MRT model.

Furthermore, Figures 5 and 6 present heatmap visualizations of single-reference evaluation metrics (BLEU, METEOR, BERTScore, TER, and chrF) for the M2M100-MRT and mBART-50-MRT models, respectively, using different reference translations. As shown in Figure 3, the highest BLEU score is obtained with Ref1, reaching a value of 31.5076, which indicates that the model's translation exhibits the greatest n-gram overlap with this reference among the evaluated translations. The higher BLEU score for Ref1 suggests that this reference's choice of vocabulary and surface-level syntactic structures is more closely aligned with the patterns produced by the model. Additionally, the BERTScore for this reference is 0.8206. The METEOR score for Ref1 is 0.4738, reflecting lexical correspondence, semantic alignment, and synonym matches between the model-generated translation and this reference. The chrF score is 50.3778, indicating a relatively strong character-level match between the model translation and Ref1, including the reproduction of morphological and orthographic patterns.

According to Figure 4, the highest BLEU score for the model translation was again obtained for Ref1, with a value of 31.4033. The BERTScore for this reference is 0.8155, the METEOR score is 0.4638, and the chrF score is 49.7604.

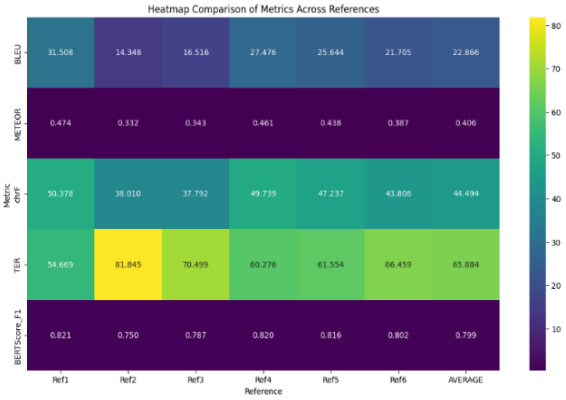

Figure 5. Heatmap of Single-Reference Evaluation Metrics for M2M100-MRT.

Figure 5 illustrates the model translation quality of M2M100 under single-reference evaluation. Each row corresponds to a specific evaluation metric, and each column represents a human reference translation. The color intensity reflects the relative values of each metric: darker shades indicate better model performance in terms of lexical similarity, semantic alignment, and character-level correspondence with the given reference, while lighter shades indicate weaker performance. It should be noted that a high alignment of the model translation with Ref1 does not necessarily imply absolute semantic superiority of this reference; rather, it reflects the sensitivity of automatic metrics to surface-level and stylistic similarities. This underscores the importance of multi-reference evaluation and human analysis when translating sacred texts.

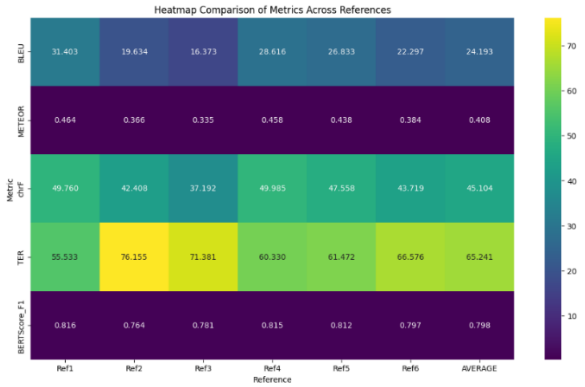

Figure 6. Heatmap of Single-Reference Evaluation Metrics for mBART-50-MRT.

The average scores across different single-reference metrics for the M2M100-MRT and mBART-50-MRT models are reported in Table 3.

Table 3. Average scores of various single-reference evaluation metrics for the M2M100-MRT and mBART-50-MRT models.

Model/Metric	BLEU	METEOR	chrF	BERTScore
M2M100-MRT	22.8661	0.4060	44.4935	0.7990
mBART-50-MRT	24.1928	0.4076	45.1037	0.7977

Table 3 presents the average scores of various single-reference evaluation metrics for the M2M100-MRT and mBART-50-MRT models. The single-reference evaluation results demonstrate reasonable lexical-semantic alignment and strong correspondence, particularly as indicated by METEOR and BERTScore.

5.3. Multi-Reference Evaluation

Given the inherently interpretive nature of Quranic translation and the availability of multiple valid human translations for each verse in the corpus, a multi-reference evaluation was conducted. In addition to reporting the average COMET score across multiple references—which measures the semantic similarity of the model output to all human translations—the COMET Max-over-References metric was also reported, selecting the best match between the model output and any single human reference for each verse. In both cases, the final score was calculated as the mean across all verses. For DA versions of COMET (e.g., wmt22-comet-da), scores above 0.7 are typically interpreted as excellent for general-

domain texts (news, conversations). However, for sacred texts like the Quran—which are more challenging due to classical language, interpretive variability, and literary style—scores above 0.65 indicate near-human-level translation quality. Table 4 presents the multi-reference evaluation results for the proposed framework.

Table 4. Multi-reference evaluation results of the models.

Model	BLEU	METEOR	COMET (avg)	COMET (max)
mBART-50-ZS	0.0002	3.1656	0.3122	0.3156
mBART-50-MRT	63.3153	64.0547	**0.7209**	**0.7347**
M2M100-ZS	1.3943	14.5470	0.3913	0.3947
M2M100-MRT	57.1715	58.8789	0.7119	0.7279

Based on Table 4, the fine-tuned models outperform the base models across all lexical and semantic metrics. For the M2M100-MRT model, the average COMET score is 0.7119, reflecting the mean quality of the output relative to all human translations. The COMET Max-over-References score is 0.7279, indicating that the model is not merely copying a single reference but produces interpretable and flexible translations. Similarly, the mBART-50-MRT model achieves an average COMET score of 0.7209 and a COMET Max-over-References score of 0.7347, demonstrating strong semantic adequacy and interpretive flexibility.

Also, comparisons with the base (pre-trained) models confirm that the fine-tuned models leverage multiple human references and avoid replicating a single translation. The low variance among different references further indicates interpretive consistency. The closeness of the average and maximum COMET scores suggests that the model output is semantically aligned with most human translations and is not overly dependent on any single reference.

While COMET scores above 0.7 are generally considered indicative of high translation quality in multilingual machine translation benchmarks, no universally accepted threshold exists for defining near-human performance in Qur'anic translation. Also, Although COMET and other semantic metrics correlate better with human judgment than traditional lexical metrics, they cannot fully capture the theological fidelity and interpretive nuances required for sacred texts. Therefore, automatic metrics should be interpreted together with human evaluation.

5.4. Human Evaluation

The automatic metrics provide a quantitative evaluation of translation quality; therefore, human evaluation was also conducted to assess the quality of the generated translations from a human perspective. For human evaluation, 60 verses were randomly selected from the full set of machine-translated verses. The same verses were used to evaluate the outputs of both models. The evaluation was conducted independently by two human evaluators with expertise in Qur'anic studies and Persian linguistics. The evaluators were blinded to the model identities and were not informed which model generated each translation. Each evaluator independently assessed the translations according to four criteria: adequacy, fluency, faithfulness, and terminology, using a 1–5 scale. The evaluation criteria are presented in Table 5.

Table 5. Human Evaluation Criteria.

Criterion	Description	Scale
Adequacy	Does the translation convey the intended meaning?	1-5
Fluency	Is the translation natural and fluent?	1-5
Faithfulness	Does the translation preserve the intended interpretation without distortion or unwarranted additions?	1-5
Terminology	Are religious terms translated appropriately?	1-5

The average human evaluation scores are presented in Table 6.

Table 6. Human evaluation results.

Translation	Terminology	Faithfulness	Fluency	Adequacy
M2M100-MRT	2.906	2.698	2.811	2.925
mBART-50-MRT	3.925	3.868	3.679	3.925

As shown in Table 6, mBART-50-MRT achieved higher scores than M2M100-MRT across all four evaluation criteria. The largest differences were observed in faithfulness and fluency, while mBART-50-MRT also achieved higher scores in adequacy and terminology. Overall, the human evaluation results indicate better translation quality for mBART-50-MRT and provide additional evidence consistent with the automatic evaluation results.

Given the limited number of human evaluators and evaluated verses, these results should be interpreted as a complementary assessment of translation quality rather than a comprehensive measure of theological and interpretive fidelity.

5.5. Ablation Study: Single-Reference vs. Multi-Reference Training

To investigate the contribution of multi-reference training, an ablation study was conducted by comparing single-reference and multi-reference training strategies. In the single-reference setting, each Arabic verse was paired with one Persian

reference translation, whereas in the multi-reference setting, each verse was paired with all six Persian reference translations. The same model architectures and training settings were used in both settings to ensure a fair comparison. The results are presented in Table 7.

Table 7. Ablation results comparing single-reference and multi-reference training.

Reference Setting	Model	BLEU	METEOR	chrF	BERT Score
Single-reference	M2M100-MRT	16.52	0.343	37.79	0.787
	mBART-50-MRT	16.37	0.335	37.19	0.781
Multi-reference	M2M100-MRT	22.78	0.406	44.49	0.799
	mBART-50-MRT	24.19	0.408	45.10	0.798

The results show that multi-reference training improves the performance of both models across all evaluation metrics. BLEU for M2M100 increased from 16.52 to 22.78, while for mBART-50 it increased from 16.37 to 24.19. Also, similar improvements were observed in METEOR, chrF, and BERTScore, which indicate that using multiple reference translations can enhance translation quality by exposing the models to different valid ways of expressing the same content.

5.6. Statistical Significance Analysis

To determine whether the performance difference between the two fine-tuned models was statistically significant, paired bootstrap resampling with 1,000 iterations was performed on the multi-reference BLEU scores. The results indicated that the mBART-50-MRT model significantly outperformed the M2M100-MRT model (63.32 vs. 57.18 BLEU, $p < 0.001$), demonstrating that the observed improvement is statistically significant and unlikely to be due to random variation.

6. Discussion and Conclusion

In this study, a multi-reference parallel corpus of Quranic translations from Arabic to Persian (QAP-MRT), comprising approximately 37,000 sentence pairs, was prepared. Translation was performed using pre-trained Transformer-based models, both in their general form and after fine-tuning on the proposed parallel corpus. Evaluation was conducted using a range of single- and multi-reference metrics. The results indicate that multilingual models such as mBART-50 and M2M100, when domain-adapted for low-resource translation, yield promising performance. Specifically, for the Quran—a sacred text with classical Arabic—the M2M100-MRT model achieved an average COMET score of 0.7119 and a maximum COMET score of 0.7279. The small difference between the average and maximum COMET scores suggests relatively consistent performance across the evaluated data. The diversity of the reference translations may also help the model learn different valid semantic and stylistic realizations rather than relying on a single reference style. Overall, these findings suggest that fine-tuning multilingual models on multi-reference data can improve the translation of religious and doctrinal texts. The Quran serves as a particularly valuable low-resource domain for this investigation and provides a foundation for further research in machine translation of classical and religious texts.

The diversity among the six reference translations is considered a strength of the proposed corpus, as it enables the model to learn multiple acceptable renderings of the same verse while preserving semantic fidelity.

Although automatic metrics such as BLEU, METEOR, BERTScore, and COMET provide objective and reproducible evaluation, they cannot fully capture the theological, interpretive, and cultural nuances of Qur'anic translation. Therefore, automatic evaluation should be complemented by human expert assessment, which was also conducted in this study.

Although this study focused on the Arabic Quran—a low-resource and closed dataset—the proposed framework combining multi-reference training and model fine-tuning can potentially be extended to other low-resource texts and language pairs, such as Kurdish, Pashto, or Urdu. In this study, a diacritics-free version of the Quran was used. Integration of these models with language models such as AraBERT could further enhance the semantic understanding of the verses. Additionally, the multi-reference parallel corpus could support style-based translation or the incorporation of multiple interpretations and translations into a unified model for future research.

A limitation of this study is that automatic evaluation metrics cannot fully assess the cultural, theological, and interpretive fidelity required for Qur'anic translation. Although the proposed approach improves translation quality according to both automatic metrics and human evaluation, these metrics may not fully capture the religious nuances and interpretive fidelity that are particularly important in sacred texts such as the Qur'an. Therefore, the results of automatic evaluation should be interpreted with caution and complemented by expert human judgment.

Ethical and Cultural Considerations

Given the ethical and semantic sensitivity of translating religious texts, this study employed multiple authoritative reference translations and diverse automatic evaluation metrics to reduce the risk of misrepresentation or superficial interpretation. Human evaluation was also conducted to provide an additional assessment of translation quality.

The QAP-MRT corpus and trained models will be made publicly available for research purposes upon acceptance of this article.

References

[1] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019.

[2] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proc. 2021 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies*, pp. 2545–2568, 2021.

[3] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[4] J. Raiman and J. Miller, "Globally normalized reader," *arXiv preprint arXiv:1709.02828, 2017.*

[5] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," in *Proc. 28th International Conference on Computational Linguistics*, pp. 3861–3867, Barcelona, Spain, 2020.

[6] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni, "Colorless green recurrent networks dream hierarchically," in *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1195–1205, New Orleans, LA, 2018.

[7] C. Vania, Y. Kementchedjhieva, A. Søgaard, and A. Lopez, "A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages," *arXiv preprint arXiv:1909.02857*, 2019.

[8] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 567–573, Vancouver, Canada, 2017.

[9] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, pp. 452–457, New Orleans, LA, 2018.

[10] Z. Li, K. Parnow, M. Utiyama, E. Sumita, and H. Zhao, "MiSS: An assistant for multi-style simultaneous translation," in *Proc. 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 1–10, 2021.

[11] A. M. Mutawa and A. Alrumaih, "Determining the meter of classical Arabic poetry using deep learning: A performance analysis," *Frontiers in Artificial Intelligence*, vol. 8, p. 1523336, 2025.

[12] S. Altammami, E. Atwell, and A. Alsalka, "Towards a joint ontology of Quran and Hadith," *International Journal on Islamic Applications in Computer Science and Technology*, vol. 10, no. 2, pp. 01–12, 2022.

[13] M. Bamoki, S. H. Wady, and S. Badawi, "Holy Quran Kurdish Sorani translation dataset for language modelling," *Data in Brief*, vol. 60, p. 111533, 2025.

[14] D. M. A. A. Danish Khaleeq, M. Tariq, and J. Iqbal, "Machine translation of Quranic verses: A transformer-based approach to Urdu rendering," *International Journal of Innovations in Science and Technology*, vol. 7, no. 2, pp. 702–717, 2025.

[15] A. AlSajri, "Challenges in translating Arabic literary texts using artificial intelligence techniques," *EDRAAK*, vol. 2023, pp. 5–10, 2023.

[16] A. R. Ghasemi and J. Salimi Sartakhti, "Multilingual Language Models in Persian NLP Tasks: A Performance Comparison of Fine-Tuning Techniques," *Journal of AI and Data Mining*, vol. 13, no. 1, pp. 107-117, 2025.

[17] F. Moodi, A. Jahangard-Rafsanjani, and F. Sadri, "Aspect-based sentiment analysis based on users' comments in an online marketplace," *Journal of Modeling in Engineering*, vol. 24, no. 84, pp. 27–46, 2026.

[18] F. Saeedi, G. Al Hinai, K. Al Kharusi, and A. A. Abdulsalam, "Effect of context and tokenization on machine translation of Arabic conversations on social media," *Procedia Computer Science*, vol. 258, pp. 1757–1763, 2025.

[19] A. Alabdullah, L. Han, and C. Lin, "Advancing dialectal Arabic to modern standard Arabic machine translation," *arXiv preprint arXiv:2507.20301*, 2025.

[20] M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, "Improving neural machine translation for low resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation," *Scientific Reports*, vol. 14, no. 1, p. 2265, 2024.

[21] R. Hidayat and S. Minati, "Comparative analysis of text mining classification algorithms for English and Indonesian Qur'an translation," *International Journal on Informatics for Development*, vol. 8, no. 1, pp. 47–51, 2019.

[22] Z. Touati-Hamad, M. R. Laouar, I. Bendib, and S. Hakak, "Arabic Quran verses authentication using deep learning and word embeddings," *The International Arab Journal of Information Technology*, vol. 19, no. 4, pp. 681–688, 2022.

[23] F. Moodi, A. J. Rafsanjani, S. Zarifzadeh, and M. A. Z. Chahooki, "Fusion of technical indicators and sentiment analysis in a hybrid framework of deep learning models for stock price movement prediction, " *IEEE Access*, vol. 12, pp. 195696–195709, 2024.

[24] F. Moodi, A. Jahangard Rafsanjani, S. Zarifzadeh, and M. A. Zare Chahooki, "Advanced stock price forecasting using a 1D-CNN-GRU-LSTM model, " *Journal of AI and Data Mining*, vol. 12, no. 3, pp. 393–408, 2024.

[25] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[26] F. Moodi, A. Jahangard-Rafsanjani, and S. Zarifzadeh, "Improving stock price prediction using technical indicators and sentiment analysis, " *Tabriz Journal of Electrical Engineering*, vol. 55, no. 2, pp. 357–370, 2025.

[27] A. Alsaleh, E. Atwell, and A. Altahhan, "Quranic verses semantic relatedness using AraBERT," in *Proc. Sixth Arabic Natural Language Processing Workshop*, pp. 185–190, 2021.

[28] R. Malhas and T. Elsayed, "Ayatec: Building a reusable verse-based test collection for Arabic question answering on the Holy Qur'an," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 6, pp. 1–21, 2020.

[29] D. I. A. Putra and M. Yusuf, "Proposing machine learning of Tafsir al-Quran: In search of objectivity with semantic analysis and natural language processing," in *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 2, p. 022101, 2021.

[30] M. S. Hadj Ameur, Y. Moulahoum, and A. Guessoum, "Restoration of Arabic diacritics using a multilevel statistical model," in *Proc. IFIP International Conference on Computer Science and Its Applications*, pp. 181–192, Cham, Switzerland, 2015.

[31] N. Habash and O. Rambow, "Arabic diacritization through full morphological tagging," in *Proc. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume*, *Short Papers*, pp. 53–56, 2007.

[32] M. Diab, M. Ghoneim, and N. Habash, "Arabic diacritization in the context of statistical machine translation," in *Proc. Machine Translation Summit XI*, 2007.

[33] S. A. Almaaytah and S. A. Alzobidy, "Challenges in rendering Arabic text to English using machine translation: A systematic literature review," *IEEE Access*, vol. 11, pp. 94772–94779, 2023.

[34] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

[35] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[36] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond English-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021.

[37] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," *arXiv preprint arXiv:1904.09675*, 2019.

[39] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," *arXiv preprint arXiv:2009.09025, 2020.*

PerQ: چارچوب ترجمه معنایی چندمرجعی مبتنی بر ترنسفورمر برای متون قرآنی عربی کم‌منبع به فارسی

فاطمه مودی و حسن مجیدی*

گروه زبان و ادبیات عربی، دانشگاه حکیم سبزواری، سبزوار، ایران.

ارسال ۲۰۲۶/۰۴/۱۶؛ بازنگری ۲۰۲۶/۰۸/۰۱؛ پذیرش ۲۰۲۶/۰۸/۱۷

چکیده:

ترجمه ماشینی متون کم‌منبع و حوزه‌محور همچنان چالشی مهم به شمار می‌رود، به‌ویژه در شرایطی که روش‌های ارزیابی مناسب در دسترس نباشند. در این پژوهش، یک چارچوب افزایش داده چندمرجعی برای ترجمه متون کم‌منبع پیشنهاد می‌کنیم. در این چارچوب، دو مدل ترجمه ماشینی مبتنی بر معماری Transformer و ازپیش‌آموزش‌دیده، یعنی mBART-50 و M2M100، به کار گرفته شده و با استفاده از داده‌های چندمرجعی و سازگارشده با حوزه‌خاص، ریزتنظیم می‌شوند. برای ارزیابی کیفیت ترجمه، مجموعه‌ای جامع از معیارهای ارزیابی خودکار، شامل BLEU، chrF، BERTScore، METEOR و COMET، در هر دو حالت تک‌مرجعی و چندمرجعی مورد استفاده قرار می‌گیرد. افزون بر این، ترجمه‌ها توسط دو ارزیاب انسانی متخصص در مطالعات قرآنی و زبان‌شناسی فارسی نیز ارزیابی می‌شوند. نتایج آزمایش‌ها به‌ویژه از نظر امتیازهای BLEU و COMET بهبود قابل‌توجهی را نشان می‌دهند، که بیانگر بهبود عملکرد ترجمه در این وظیفه کم‌منبع و حوزه‌محور است. نتایج همچنین نشان می‌دهند که آموزش چندمرجعی می‌تولند کیفیت ترجمه را بهبود بخشد و ارزیابی چندمرجعی نیز ارزیابی جامع‌تری از عملکرد مدل ارائه می‌دهد. بااین‌حال، معیارهای خودکار قادر نیستند به‌طور کامل میزان وفاداری الهیاتی، فرهنگی و تفسیری موردنیاز در ترجمه قرآن را ارزیابی کنند؛ ازاین‌رو، نتایج حاصل از این معیارها با ارزیابی متخصصان انسانی تکمیل می‌شود.

کلمات کلیدی: ترجمه ماشینی کم‌منبع، ترجمه چندمرجعی، ترجمه قرآن، ترجمه ماشینی عربی- فارسی، ترجمه ماشینی عصبی.