PahGen: Generating Ancient Pahlavi Text via Grammar-guided Zero-shot Translation

Farhan Farsi[1]*, Parnian Fazel[2]*, Farzaneh Goshtasb[3], Nadia Hajipour[3], Sadra Sabouri[4], Ehsaneddin Asgari[5], Hossein Sameti[6]

[1]Amirkabir University of Technology, [2]University of Tehran,
[3]Institute for Humanities and Cultural Studies, [4]Open Science Laboratory,
[5]Qatar Computing Research Institute, [6]Sharif University of Technology

farhan1379@aut.ac.ir, parnian.fazel@ut.ac.ir, f.goshtasb@ihcs.ac.ir, n.hajipour@ihcs.ac.ir,
sadra@openscilab.com, easgari@hbku.edu.qa, sameti@sharif.edu

Abstract

The Pahlavi language, aka Middle Persian, is a critical part of Persian cultural and historical heritage which bridges the Old Persian and Modern Persian (Farsi). However, due to its limited digital presence and the scarcity of comprehensive linguistic resources, Pahlavi is at risk of extinction. As an early attempt to preserve this language, this study introduces a framework to translate English text into Pahlavi. Our approach combines grammar-guided term extraction with zero-shot translation, leveraging large language models (LLMs) to generate syntactically and semantically accurate Pahlavi sentences. This framework aims to preserve the Pahlavi language and serves as a model for reviving other endangered languages with similar characteristics. Finally using our framework, we generate a novel dataset of 360 expert-validated parallel English-Pahlavi texts.

1 Introduction

The Pahlavi language, or Middle Persian, was the official language of the Sasanian Empire, bridging Old and Modern Persian (Boyce, 1984). It was historically used across regions of present-day Iran, Afghanistan, Pakistan, and beyond. Despite its significance, Pahlavi faces extinction due to limited digital representation and scarce linguistic resources (Namiranian and Assadi, 2021). Efforts to digitize its scripts via OCR exist (Alirezaee et al., 2005; Sadri et al., 2007), but most materials stem from incomplete transcriptions of ancient manuscripts and Zoroastrian texts, lacking standardized character sets.

The preservation and revitalization of endangered languages have garnered attention in the field of Natural Language Processing (NLP). Recent works (Basit et al., 2024; Chimoto and Bassett, 2022) have highlighted challenges and advancements in this area. Large Language Models (LLMs) have demonstrated adaptability to new tasks (Fayyazi et al., 2025; Patil and Gudivada, 2024), especially translating into an unseen or low-resource languages (Liao et al., 2024; Mao and Yu, 2024; Park et al., 2024); However, their application as end-to-end translators for languages like Pahlavi is hindered due to the lack of digital resources in their original form on the internet, making it difficult for LLMs to train effectively (Farsi et al., 2025).

Translation requires an understanding of the semantic and syntactic structures of both languages. Language learners use translation as a means for learning a new language (Widdowson, 2014). Therefore, exploring the translation mechanism from a well-known language like English to an ancient language like Pahlavi can lay the foundation for further preservation efforts. This approach not only helps safeguard the language but also raises awareness within communities by enabling a deeper understanding of its nuances. Developing a method by which new Pahlavi text is generated also serves as a new resource with modern concepts. However, translating from English to the Pahlavi language has its challenges. It involves adapting tenses to align with aspectual distinctions, accommodating the lack of gendered pronouns, and the absence of terms for contemporary concepts such as "website," or "tennis racket."

In this paper, we present PahGen, a framework designed to translate English text into Pahlavi. Our approach integrates grammar-guided term extraction with zero-shot translation techniques, utilizing LLMs to generate Pahlavi sentences that are both syntactically and semantically accurate. We evaluated our framework on a set of manually crafted parallel sentences between English and Pahlavi.

Additionally, we constructed a parallel English-Pahlavi dataset using our framework, which was manually validated by language experts. As the first endeavor in this domain, we publicly release

*These authors contributed equally to this work

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 171–182

our dataset[1], providing a foundation for researchers to build upon and advance this work.

For end-to-end translation from English to Pahlavi, we initially experimented with few-shot LLM prompting, both with and without providing vocabulary. However, the models often hallucinated nonsensical characters resembling the Pahlavi script, revealing their lack of understanding of its semantic and syntactic structure. To address this, we combined handcrafted rules for grammatical and semantic roles with LLMs to compile sentence components. This approach, common in low-data scenarios, is supported by heuristics and feature handcrafting (Charoenpornsawat et al., 2002; Abdollahi et al., 2024; Ashrafi et al., 2024). Section 3 presents our proposed methodology in detail. Section 4 describes the dataset validation process, followed by comprehensive performance analysis in Section 5. We introduce our novel dataset in Section 6 and present preliminary experiments with LLMs along with their implications for the machine translation community in Section 7. Section 8 discusses the limitations and future directions of our work, and finally, Section 9 concludes the paper.

We hope that our approach—preserving Pahlavi texts through translation by generating Pahlavi text from English—facilitates a deeper understanding of the Pahlavi language and serves as a model for similarly presenting other languages. This work not only aids in the preservation of Pahlavi but also provides a replicable framework for revitalizing other endangered languages with comparable structural features.

2 Related Work

Researchers have explored machine translation for low-resource languages, such as Urdu (Basit et al., 2024), in the modern NLP. While some of them tried to leverage existing data sources like verses and prose (Cadotte et al., 2024) to enhance the translation accuracy, others equipped state-of-the-art models such as transformers (Varanasi et al., 2024; Roy et al., 2024a) and LLMs (Liao et al., 2024; Mao and Yu, 2024) to do so. In addition, two research directions that are more closely related to our work are presented below.

Translating to Preserve Endangered Languages. Preserving a language through translation extends beyond linguistic conversion to include cultural and contextual nuances. As Rout (2020) highlights, translation is key to preserving language and culture, yet challenges arise when region-specific words lack direct equivalents in the target language. Comparative studies on Odisha and Assam's folk literature reveal lexical disparities and sociocultural differences that complicate translation. Similarly, Do Phuong et al. (2022) emphasizes that cultural awareness is essential to maintaining linguistic integrity. Machine translation, as discussed by Mahmud et al. (2019), offers a tool for documenting endangered languages, though the absence of structured lexicons remains a hurdle. Researchers have explored methods for translating texts from high-resource languages to endangered ones, such as Shu et al. (2024), who introduced a retrieval module to LLMs for Cherokee, Tibetan, and Manchu, and Merx et al. (2024), who examined personalized Tetun translation. Beyond language preservation, NLP techniques also address societal issues, as demonstrated by Soumah et al. (2023), who developed an automated system to analyze gender representation in French-language media. While most translation-based efforts focus on spoken languages (Haddow et al., 2022), some, like Iyer et al. (2024), have instructed LLMs to translate Spanish into 11 indigenous American languages to aid in preservation. Literary translations further embed cultural values, ensuring linguistic heritage is maintained across generations and geographies.

Rule-based Translation. The nuances behind low-resource languages, especially ancient ones, make their end-to-end translation hard. In addition, end-to-end translators usually have less interpretability. Therefore using rule-based methods for their translation is preferred. For example, Khanna et al. (2021) presented an open-source rule-based tool for Apertium translation. Some researchers utilized the rule-based method in their framework for better accuracy when dealing with low-resource languages (Torregrosa et al., 2019; Singh et al., 2021), even with the existence of LLMs to yield better translations (Coleman et al., 2024).

Our work bridges the gap between these two niche research directions by proposing a translation framework and an expert-validated automatically generated dataset that produces insights into the Pahlavi language for the machine translation community and language learners.

[1] https://huggingface.co/datasets/PahGen/Parsig-English

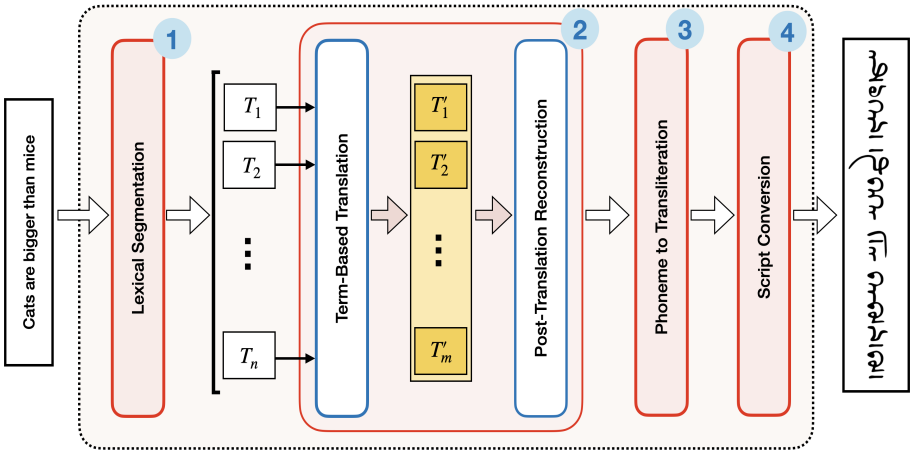

Figure 1: This figure outlines the three-step architecture of our dataset preparation model: (1) lexical segmentation, (2) Translator, (3) phoneme to Transliteration converter, and (4) script converter.

3 Method

For translating English text to Pahlavi with Book Pahlavi script, we employed a structured pipeline which has 4 steps. The steps in our approach are as follows:

- **Lexical Segmentation**: Divide the input sentence into individual terms based on grammatical structure (Section 3.1).
- **Translation**: The translation process follows a two-stage approach: (1) individual translation of segmented terms into Pahlavi language, and (2) construction of the final sentence by arranging the previously translated terms and integrating various linguistic components such as word order, kasreh-ezafeh, and other morphosyntactic features. The output of this process is a Pahlavi language sentence represented in phonemes (Section 3.2).
- **Phoneme to Transliteration**: In this step, the phoneme representation obtained from the previous step is converted to transliteration. This is achieved by using a dictionary that maps each phoneme representation to its corresponding transliteration, along with applying specific linguistic rules (Section 3.3).
- **Script Conversion**: The final step involves converting the transliterated output into the Book Pahlavi script. This process also includes addressing ligature challenges associated with the current font version to ensure accuracy and readability (Section 3.4).

An overview of our method is shown in Figure 1, and the following sections provide detailed explanations of each step.

3.1 Lexical Segmentation

We first segment the input into individual terms to facilitate their translation using an existing open-source English–Pahlavi dictionary[2]. We used the Stanza (Qi et al., 2020) library to extract syntactic features, such as part-of-speech tags, lemma, recognition of named entities, and other structural attributes of the input sentence.

Additionally, we employed GPT-4o in a few-shot setting to extract morphosyntactic features as metadata to capture useful grammatical information such as object count, grammatical person, verb transitivity, and related properties. These features enriched each segment with syntactic and morphological structure, enabling more accurate translation within our rule-based framework.

3.2 Translation

Our translation pipeline follows a multi-stage process that integrates dictionary-based translation, rule-based morphological generation, and semantic similarity matching. Below, we outline the key steps.

3.2.1 Term-Based Translation

The translation process for each term operates as follows:

Named Entity Recognition. Named entities such as names of people or cities are identified and

[2]https://erman.super.site/hamzin-uzvn-prsg-learning-prsg-language/vznmag-dictionary

Type	Pahlavi Postfixes	Pahlavi Example	English Translation
Adverb of Manner	"-īhā"	rōšnīhā	clearly
Comparative Adjective	"-dar"/"-tar"	abardar/vattar	higher/worse
Superlative Adjective	"-dum"/"-tum"	abardum/vattum	highest/worst
Pluralization	"-an"	draxtan	trees
Agentive Noun	"-gar"/"-kar"	nigargar/ardīkkar	painter/warrior
Possessor Noun	"-wand"/"-awand"	hunarawand	artist
Guardian Noun	"-bān"/"-pān"	marzbān/pāygōspān	margrave/governor
Master Noun	"-bed"	mubed	head priest
Container Nouns	"-dān"/"-yān"	zēndān/pusiyān	prison/womb
Place Nouns	"-estān"	bōyestān	garden

Table 1: Examples of Pahlavi Affixes. This table showcases selected Pahlavi postfixes, corresponding Pahlavi examples, and their English translations.

preserved in their original form to maintain proper noun integrity.

Dictionary-Based Translation. Each term is matched with an English-Pahlavi dictionary. If a term has multiple meanings due to part-of-speech (POS) ambiguity, POS information is used to resolve the correct translation.

Rule-Based Morphological Generation. New words can be formed using suffixes, prefixes, and stems (Goshtasb and Hajipour, 2022) in the Pahlavi language (similar to English, where "player" stems from "play" + "-er" or "chemist" from "chem"+ "-ist"). Therefore, our framework employs predefined Pahlavi morphological rules to construct appropriate translations for terms not found in the dictionary, where applicable. These rules are detailed in Appendix A.

In cases where the rule-based approach fails to produce a translation, we employ BERT (Devlin et al., 2019) embeddings to compute the cosine similarity between the target term and existing dictionary keys. If a sufficiently close match is found, determined by a predefined similarity threshold, the closest matching term is selected as the translation. For terms where no similar key exists in the dictionary, we flag them as untranslatable. This is particularly evident for contemporary invented words that did not exist in the historical period of the language, such as "tennis rackets" or "television."

3.2.2 Post-Translation Reconstruction

After individual term translation, we reconstruct coherent sentences considering relationships between translated elements. This stage is crucial for preserving the semantic and syntactic integrity of the original text. The reconstruction process incorporates key linguistic features of the Pahlavi language, including word order patterns specific to Pahlavi language syntax, historical punctuation conventions, and the kasreh-ezāfeh construction. The kasreh-ezāfeh serves as a crucial grammatical particle that establishes relationships between nouns, expressing possession, attribution, or association (Samvelian, 2018).

We employ GPT-4o in a zero-shot learning configuration for sentence reconstruction. The model receives paired input of the original term from the source text and its corresponding semantic interpretation. The model then generates reconstructed sentences in phonemic representation format, maintaining the linguistic characteristics of the Pahlavi language while ensuring semantic accuracy. This approach allows for contextually appropriate integration of translated terms without requiring explicit rules for grammatical construction.

3.3 Phoneme to Transliteration

To convert phoneme representations into transliterations, we first developed a dictionary using the Pahlavi database (Goshtasb et al., 2021)[3], which includes Pahlavi words and their transliterations, encompassing both stems and affixes. Using this dictionary, we employ a straightforward rule-based method: each word is broken down into its stem and affixes, and we then construct the transliteration by combining the transliterations of these parts. Given that Pahlavi is an extinct language with a limited, fixed vocabulary, this approach effectively meets our needs.

[3]available at https://www.parsigdatabase.com/

For handling out-of-vocabulary (OOV) words like named entities, geographical locations, and numerical expressions, which do not exist in our dictionary, we extend our method with a character-level transliteration model. This model applies systematic Pahlavi phoneme-to-transliteration conversion rules, as outlined by (Tafazzoli and Amoozgar, 1996). These rules are explained in more detail in Appendix B, complementing our morphological approach, ensuring comprehensive coverage and accuracy in the transliteration process.

3.4 Script Conversion

The final stage of our pipeline converts the transliterated output into authentic Book Pahlavi script characters through systematic linguistic and typographical rules. Currently, the *Ham-dibrih* represents the sole available digital typeface for Book Pahlavi script rendering. However, the script's complexity, particularly its extensive use of ligatures for character connections, poses significant challenges in digital representation.

In our work, we address several issues present in the current version of the *Ham-dibrih* typeface and have developed an improved version 2 of this font. The enhancements include: (1) modifying the shapes of certain characters that appeared overly pointed, (2) resolving connectivity issues between specific characters, and (3) adjusting characters that were positioned above the baseline. These improvements were achieved using FontLab.

4 Evaluation Dataset

To assess our model's performance, we developed a parallel Pahlavi-English test dataset. This dataset comprises 200 sentence pairs, including genuine excerpts from historical books and inscriptions, alongside sentences from various domains such as art, fruits, and animals.

To construct this dataset, we followed a structured process: (a) We initially generated text covering several domains, utilizing diverse grammatical structures in English, with sentences composed of 5 to 12 words. (b) We filtered these sentences to include only those with words that are present in our dictionary or are translatable into Pahlavi, deliberately excluding terms that would not have existed during the Sassanian era, such as "airplane." (c) These selected sentences underwent expert translation and thorough quality checks by additional experts to ensure high accuracy.

Additionally, we enriched our dataset with data from historical books and inscriptions, using the same selection criteria and processes used for our synthetically generated content.

Our dataset consists of 200 sentences, with an average length of 7.5 words per sentence. It includes two categories: 167 sentences from created texts and 33 sentences from historic texts. The dataset contains a total of 402 unique words.

5 Results

To evaluate the effectiveness of our proposed English-to-Pahlavi machine translation framework, we assess two baselines: (1) Zero-shot and (2) Zero-shot learning with vocabulary. We evaluated these baselines using two types of assessment. First, we applied well-known translation metrics, such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), to measure translation quality, which is particularly useful for languages with complex sentence structures. Additionally, since word order in Pahlavi can vary while preserving meaning and these metrics have false negatives, we conducted human evaluations following prior studies (Chi et al., 2020; Maurya et al., 2021; Roy et al., 2024a) to assess our model's output qualitatively as well. This allowed us to gain a more comprehensive understanding of our model's output.

5.1 Automatic Evaluation

To assess the correctness of model-generated translations at the word level, we used the BLEU score. Additionally, since Pahlavi is a morphologically-rich language with a complex system of affixes, the CHrF metric, which is sensitive to character-level translation, helps evaluate character-level accuracy. The results are available in Table 2. Our model outperformed both baselines by a significant margin across all metrics, demonstrating its effectiveness.

Model	**ChrF**	**BLEU-Score**	
		BLEU-1	BLEU-2
Zero-shot	21.87	13.5	0.2
Zero-shot + vocab	25.06	30.8	0.95
Our model	**83.1**	**65.3**	**51.2**

Table 2: The proposed model outperforms zero-shot and vocabulary-augmented approaches in BLEU and ChrF.

5.2 Human Evaluation

To qualitatively evaluate the translated sentences from English to Pahlavi, we randomly selected 20

test data points for each language pair. We used three key criteria for assessment: fluency, relatedness, and correctness.

Fluency refers to the smoothness and coherence of the translated text, assessing how well the sentences flow and adhere to grammatical rules.

Relatedness measures how effectively the translations connect to the original ground truth sentences, ensuring they capture the essential information.

Correctness evaluates the accuracy and appropriateness of the translations in terms of meaning and semantics.

These criteria allow for a comprehensive assessment of the translation quality, ensuring that translations are not only grammatically correct but also semantically accurate and closely related to the source material.

With two experts, each holding a PhD in the field of language studies and with extensive research experience on the Pahlavi language, including several published papers, informed about the task and evaluated the sentences using a 5-point scale, where 1 denotes very poor quality and 5 denotes excellent quality, across each of the three metrics. They divided the 20 sentences and finalized the results through negotiated agreement. The average value of sentences generated by each model is presented in Table 3. This demonstrates that our model not only achieves better results in automatic metrics but also shows qualitative improvements.

Model	Fluency	Relatedness	Correctness
0-shot	1.5	1	1
0-shot + vocab	3	2.5	3
our model	**4**	**4**	**4**

Table 3: Average Likert scale assessment (1: very poor, 5: excellent) of our approach and its variations across Fluency, Relatedness, and Correctness.

6 Parallel Dataset Creation

We utilized our framework to generate a larger parallel English-Pahlavi dataset. This dataset is intended to facilitate further research in Pahlavi language processing by offering a dataset that can serve as a useful resource for this low-resource language. The dataset includes 360 sentence pairs spanning diverse domains. Sentences were selected and filtered based on linguistic diversity, domain coverage, and grammatical complexity.

6.1 Thematic Distribution

The dataset encompasses a broad range of themes to ensure linguistic diversity across various domains. The primary thematic categories and representative examples are summarized in Table 4.

6.2 Syntactic Variety

The dataset includes diverse syntactic constructions to capture structural variability in sentence formation. We incorporated various sentence types to reflect both conversational and narrative styles. These types include simple sentences (English: She teaches her skills with signs.; Pahlavi: ōy hunarān-š hamōzed ped daxšagān.), compound sentences (English: He placed the letters on the table and waited for her to read them.; Pahlavi: ōy fravardagān abar mīzd nihād hend u-š pād dā ōy-išān xvāned.), complex sentences (English: She baked fresh bread in the kitchen while the children played in the garden.; Pahlavi: ōy nān ī tāzag andar xvardpazxānag puxt ka kōdakān ped bāv vāzīg kird.), and imperative sentences (English: Be friend to everyone.; Pahlavi: kas dōst bāš.). These examples highlight the range of syntactic variety present in the dataset.

6.3 Lexical Variety

Lexical variety in the dataset is achieved by selecting words from a range of semantic fields. This diversity facilitates the model's acquisition of a broad vocabulary to accurately translate terms across different contexts. The lexical selection spans several grammatical categories. For instance, nouns such as teacher, merchant, and doctor, represent a variety of everyday items and professions. Verbs like run, eat, think, learn, paint, and write capture different actions and processes. Adjectives such as red, green, bright, strong, and careful describe various attributes, while adverbs like quickly, usually, brightly, and carefully modify actions in terms of frequency or manner. Temporal expressions, including phrases such as every morning, before sunset, and at noon, indicate specific times or recurring events. The inclusion of such diverse lexical items ensures the dataset has the necessary variety to support accurate translation across different domains.

7 Discussion

This section first examines our attempts to use LLM as end-to-end translators and provides preliminary explanations for why we believe it fails. We then

Theme	English	Pahlavi
Daily Activities	The boy runs with the dog for exercise.	rēdak abāg sag daved varzišn kirdan rāy.
Geography, Flora & Fauna	In the garden, the cat chases the butterflies.	ped bāv, gurbag parrānakān xvēhed.
Professions	The merchant trades silk for gold.	vāzāragān abrēšum vahāg kuned zarr rāy.
Family & Social Life	The family gathered around their table for dinner.	dūdag šām xvardan rāy pērāmōn ī-šān mīzd amvašt hend.
Objects & Tools	He paints a picture with brushes.	ōy nigārag-ē nigāred ped garvāšān.
Seasons & Time	Every winter, the snow falls to cover.	harv zimestān, vafr pōšīdan rāy kafed.
Food & drink	They eat apples in the morning.	avēšān ped bāmdād sēb xvarend.
Religion	The satisfaction of the body of good people is joy and worship to the gods.	*šnāyišn ī tan ī wehān rāmišn ud niyāyišn ī yazdān*

Table 4: Thematic Categories in the Parallel dataset. The text in italics represents the ancient Pahlavi script.

Figure 2: GPT-4o (*gpt-4o-2024-08-06*)'s translation for the sentence: "*the warrior walks to battle*" when prompted zero-shot. Lighter colors indicate high entropy, showing the model's uncertainty, while darker colors indicate low entropy, indicating higher certainty in the generated tokens.

discuss the limitations of our work and outline how future research can address these challenges while building upon our findings. Finally, we discuss the implications of our work for the machine translation community.

7.1 LLM End-to-End Translation

We did a series of experiments on zero-shot prompting OpenAI GPT-4o API with and without providing the vocabulary to assess its capability to translate English into Pahlavi.

First, we simply asked ChatGPT to translate some English text into Pahlavi using the following prompt template.

```
Translate the following text from English to
    Pahlavi (middle Persian language) phonetic
    representation: {text}
Only provide me the final output without any
    explanation.
```

Then we provided the vocabulary needed for the translation in the prompt to give it more context.

```
Translate following text from English to Parsig
    (middle Persian language) phonetic
    representation: {text}
Only provide me the final output without any
    explanation here are some vocab that you
    need for translation: {vocabs}
```

We observed that GPT-4 heavily hallucinated the output, generating words that resembled Pahlavi but were meaningless. For example when given the sentence "the warrior walks to battle" as input, GPT-4o would translate it into "zēnīg ō hamāg padēw gāhēd." Here, only the translation of the first two words was somewhat accurate: 'zēnīg' means 'a person with a weapon,' and 'ō' means 'to.' However, from the third word onward, the translation becomes problematic—'padēw' and 'gāhēd' have no known meaning in Pahlavi, while 'hamāg', which means 'everyone,' does not fit the context.

To deepen our analysis, following the method used by Zhou et al. (2023), we looked into the entropy of generated tokens. With this measure, we can probe the uncertainty level of the model for each generated token. For calculating the entropy for tokens, we need their probability distribution over all vocabulary but since the OpenAI API for GPT-4o only returns log probability for the top 20 tokens, we apply a soft-max on those 20 sets and calculate the entropy on that restricted set, as done by Roy et al. (2024b).

Figure 2 represents the example sentence which is color-coded by the entropy values of the tokens. The average entropy over all tokens of this sentence was 1.528 while the least entropy (the most certain token) was for 'ō'. We hypothesize that since it's a stopword in Pahlavi, it should have been frequent in the sparse resources that LLM was trained based on. Therefore model is confident about its presence and uses it. It's noteworthy that while the model is uncertain on completely hallucinated words like 'padēw', it is somehow certain on a word that has a meaning but is not meaningful in this context—'hamāg'. This implies that GPT4o may memorize some words but doesn't know the contextual use of them for the Pahlavi language.

To assess the extent of hallucination in translations to Pahlavi, we compared the average entropy of tokens in sentences of our evaluation dataset for both Persian and Pahlavi translations. Applying a one-way t-test comparing the distribution of average entropy for Pahlavi and Persian translated sentences yielded a statisically significant difference ($t = 24.5549,\ p < 0.05$). The entropy distribution for the generated Persian sentence had a mean of 0.7989 and a standard deviation of 0.2579, while Pahlavi-generated text had a mean of 0.1163 and a

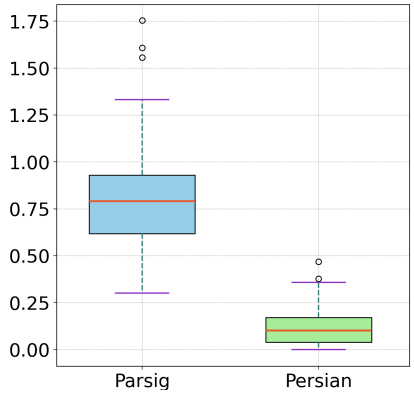

Figure 3: Box plots illustrating the distribution of entropy levels in translations to Persian and Pahlavi.

standard deviation of 0.0960. These findings highlight the model hallucinates more when translating into Pahlavi compared to Persian, as indicated by the lower entropy and higher certainty of incorrect tokens in Pahlavi. Figure 3 presents the box plots illustrating the distribution of entropy for translations into Persian and Pahlavi.

7.2 Implication for MT Researchers

Machine translation researchers working with low-resource ancient languages can adapt our hybrid approach, which combines rule-based methods with neural embeddings, to address similar challenges. Our work demonstrates how these techniques improve translation quality despite data scarcity, offering insights for refining dictionary-based methods and model fine-tuning.

Human validation remains crucial for assessing translation quality. Future research can integrate iterative feedback, active learning, or expert-guided refinement. Expanding datasets through semi-automated translation and verification can further mitigate data limitations and enhance translation reliability.

8 Limitations and Future Work

This work is among the first to contribute to the preservation of the ancient Pahlavi language and like any other work, has specific limitations.

Dataset Size and Diversity Due to the scarcity of experts in this language, manually verifying our model's outputs at scale was a significant challenge. To ensure dataset quality, we collaborated with domain experts, creating a reliable but small gold-standard parallel dataset of 360 sentence pairs. However, the dataset may not include every aspect of the Pahlavi language, limiting its effectiveness as a benchmark and making it unclear whether our approach generalizes beyond handcrafted examples.

Future work can address this limitation by generating new translations using our model and manually verifying them or by incorporating a wider range of texts, such as historical manuscripts and inscriptions, to improve linguistic diversity and enhance generalization.

Analysis of Low-Scoring Sentences While we conducted human evaluations, we did not thoroughly analyze cases where our model performed poorly. A more in-depth examination of low-scoring sentences could provide insights into common failure patterns and areas for improvement of our model or rules which are used in different parts. Future studies could explore these errors systematically to refine the translation framework and better understand its limitations.

9 Conclusion

Our study presented a framework for translating English text into Pahlavi, combining grammar-guided term extraction with zero-shot translation and leveraging large language models (LLMs) to generate syntactically and semantically accurate sentences. This approach not only aids in the preservation of Pahlavi but also serves as a model for reviving other endangered languages. We also created a novel dataset of 360 expert-validated parallel English-Pahlavi texts, which is publicly available to support further research and advancements in this field. By addressing key challenges in low-resource language translation, this work lays the foundation for future efforts in both ancient and modern language preservation.

Acknowledgments

We extend our gratitude to Ario Sedaghat[4], a key member of the Erman website, for his assistance in evaluating our results and instructing team members unfamiliar with the basics of the Pahlavi language. We also thank Hessam Behdani for his contributions to creating an enhanced version of the font. Additionally, we sincerely thank the anonymous reviewers for their insightful suggestions.

References

Armin Abdollahi, Negin Ashrafi, and Maryam Pishgar. 2024. Advanced predictive modeling for enhanced

[4] ario.sedaghat@gmail.com

mortality prediction in icu stroke patients using clinical data. *arXiv preprint arXiv:2407.14211*.

Shahpour Alirezaee, Hassan Aghaeinia, Majid Ahmadi, and Karim Faez. 2005. An efficient restoration algorithm for the historic middle-age Persian (Pahlavi) manuscripts. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2114–2120. IEEE.

Negin Ashrafi, Armin Abdollahi, and Maryam Pishgar. 2024. Enhanced prediction of ventilator-associated pneumonia in patients with traumatic brain injury using advanced machine learning techniques. *arXiv preprint arXiv:2408.01144*.

Abdul Basit, Abdul Hameed Azeemi, and Agha Ali Raza. 2024. Challenges in Urdu machine translation. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 44–49, Bangkok, Thailand. Association for Computational Linguistics.

Mary Boyce. 1984. *Zoroastrians: Their Religious Beliefs and Practices*. Psychology Press.

Antoine Cadotte, Nathalie André, and Fatiha Sadat. 2024. Machine translation through cultural texts: Can verses and prose help low-resource indigenous models? In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 121–127, Bangkok, Thailand. Association for Computational Linguistics.

Paisarn Charoenpornsawat, Virach Sornlertlamvanich, and Thatsanee Charoenporn. 2002. Improving translation quality of rule-based machine translation. In *COLING-02: machine translation in Asia*.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7570–7577.

Everlyn Asiko Chimoto and Bruce A Bassett. 2022. Very low resource sentence alignment: Luhya and swahili. *arXiv preprint arXiv:2211.00046*.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thao Do Phuong et al. 2022. Preserving cultural values in translation to english. *Croatian International Relations Review*, 28(90):252–268.

Farhan Farsi, Parnian Fazel, Sepand Haghighi, Sadra Sabouri, Farzaneh Goshtasb, Nadia Hajipour, Ehsaneddin Asgari, and Hossein Sameti. 2025. Parsipy: NLP toolkit for historical persian texts in Python. *arXiv preprint arXiv:2503.17810*.

Arya Fayyazi, Mehdi Kamal, and Massoud Pedram. 2025. Facter: Fairness-aware conformal thresholding and prompt engineering for enabling fair LLM-based recommender systems. *arXiv preprint arXiv:2502.02966*.

Farzaneh Goshtasb, Masood Ghayoomi, and Nadia Hajipour Artarani. 2021. Corpus-based analysis of middle persian texts based on the pārsīg database. *Language Studies*, 12(1):255–280.

Farzaneh Goshtasb and Nadia Hajipour. 2022. Denominative verbs in Zoroastrian Middle Persian texts (statistical and corpus-based analysis). *Language and Linguistics*, 18(35):1–19.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch-Mayne. 2024. Exploring very low-resource translation with llms: The university of Edinburgh's submission to americasnlp 2024 translation task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220. Association for Computational Linguistics (ACL), Association for Computational Linguistics (ACL).

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

You-Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun, Yen-Hsiang Wang, Hsiao-Min Li, and Yao-Chung Fan. 2024. Learning-from-mistakes prompting for indigenous language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 146–158, Bangkok, Thailand. Association for Computational Linguistics.

Zulkifli Erlina Mahmud et al. 2019. Preserving culture through literary works and their translations. In *Proceedings Conference on the Environmental Conservation through Language, Arts, Culture and Education*, volume 1, pages 11–15.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Raphael Merx, Hanna Suominen, Adérito José Guterres Correia, Trevor Cohn, and Ekaterina Vylomova. 2024. Low-resource machine translation: what for? who for? an observational study on a dedicated tetun language translation service. *arXiv preprint arXiv:2411.12262*.

Katayoun Namiranian and Sheida Assadi. 2021. Exploring translation strategies of middle persian texts: The case of the 19th chapter of vandidad. *Journal of Foreign Language Teaching and Translation Studies*, 6(1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gyutae Park, Seojin Hwang, and Hwanhee Lee. 2024. Low-resource cross-lingual summarization through few-shot learning with large language models. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 57–63, Bangkok, Thailand. Association for Computational Linguistics.

Rajvardhan Patil and Venkat Gudivada. 2024. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Jyotsna KB Rout. 2020. Importance of translation in preservation of language and culture. *LANGUAGE, LITERATURE, CULTURE & INTEGRITY*, page 1.

Aniruddha Roy, Pretam Ray, Ayush Maheshwari, Sudeshna Sarkar, and Pawan Goyal. 2024a. Enhancing low-resource NMT with a multilingual encoder and knowledge distillation: A case study. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 64–73, Bangkok, Thailand. Association for Computational Linguistics.

Shamik Roy, Sailik Sengupta, Daniele Bonadiman, Saab Mansour, and Arshit Gupta. 2024b. FLAP: Flow adhering planning with constrained decoding in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 517–539, Mexico City, Mexico. Association for Computational Linguistics.

Javad Sadri, Sara Izadi, Farshid Solimanpour, Ching Y Suen, and Tien D Bui. 2007. State-of-the-art in Farsi script recognition. In *2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–6. IEEE.

Pollet Samvelian. 2018. Specific features of persian syntax: The ezâfe construction, differential object marking and complex predictaes. *Oxford handbook of Persian linguistics*, pages 226–269.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, et al. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *arXiv preprint arXiv:2411.11295*.

Muskaan Singh, Ravinder Kumar, and Inderveer Chana. 2021. Improving neural machine translation for low-resource indian languages using rule-based feature extraction. *Neural Computing and Applications*, 33(4):1103–1122.

Valentin-Gabriel Soumah, Prashanth Rao, Philipp Eibl, and Maite Taboada. 2023. Radar de parité: An NLP system to measure gender representation in french news stories. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Ahmad Tafazzoli and Zhaleh Amoozgar. 1996. *Pahlavi Language, Literature. Grammatical Sketch, Texts and Glossary*. Moin. [in Persian].

Daniel Torregrosa, Nivranshu Pasricha, Maraim Masoud, Bharathi Raja Chakravarthi, Juan Alonso, Noe Casas, and Mihael Arcan. 2019. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland. European Association for Machine Translation.

Abhishek Varanasi, Manjira Sinha, and Tirthankar Dasgupta. 2024. Linguistically informed transformers for text to American Sign Language translation. In

Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 50–56, Bangkok, Thailand. Association for Computational Linguistics.

Henry G Widdowson. 2014. *The role of translation in language learning and teaching*, pages 222–240. Springer, London.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Rules

A.1 Verbs

To generate Middle Persian (Pahlavi) verbal forms from English verbs, our rule-based system synthesizes inflected verbs using morphosyntactic features such as tense, aspect, mood, polarity, transitivity, person, and number. This design is grounded in historical Pahlavi language grammars (Tafazzoli and Amoozgar, 1996).

Verbs in the Pahlavi language are conjugated based on their present and past stems, both of which are lexically determined rather than generated through a fixed derivational pattern. We extract these stems from the English–Pahlavi dictionary, which provides both present and past stems for each verb.

Person	Singular	Plural
First	-ēm	-ēm
Second	-ē	-ēd
Third	-ēd	-ēnd

Table 5: Present tense suffixes in Pahlavi language

Person	Singular	Plural
First	hēm	hēm
Second	hē	hēd
Third	-	hēnd

Table 6: Conjugation of the auxiliary verb used in the past tense in Pahlavi Language

A.1.1 Present Tense

The present tense is constructed by attaching subject-agreeing suffixes to the present stem of the verb. The suffixes for the present tense are presented in Table 5. For example, "She sees" is *wēn-ēd* in Pahlavi.

Person	Singular	Plural
First	ēst-ēm	ēst-ēm
Second	ēst-ē	ēst-ēd
Third	ēst-ēd	ēst-ēnd

Table 7: Conjugation of the auxiliary verb used in the present perfect tense in Pahlavi Language

Person	Singular	Plural
First	ēstād hēm	ēstād hēm
Second	ēstād hē	ēstād hēd
Third	ēstād	ēstād hēnd

Table 8: Conjugation of the auxiliary verb used in the past perfect tense in Pahlavi Language

A.1.2 Past Tense

We use the present tense conjugation of the verb *h-* ("to be") as an auxiliary to construct compound past tenses which are presented in Table 6. In the Pahlavi language, verb agreement in the past tense depends on transitivity:

Past Tense of Intransitive Verbs: For intransitive verbs, the past tense agrees with the subject. So, the structure follows: past stem + auxiliary verb(subject). For example, "We went" is *raft hēm* in Pahlavi.

Past Tense of Transitive Verbs (Ergative Alignment): For transitive verbs, Pahlavi follows an ergative alignment, meaning 1) the verb no longer agrees with the subject, 2) instead, it agrees with the object in person and number. So, the structure follows: past stem + auxiliary verb(object). For example, "you saw the men" is *tō mardān dīd hēnd* in Pahlavi.

A.1.3 Perfect Tenses

The perfect tense in the Pahlavi language is formed by combining the past stem of the verb with the auxiliary verb *ēst* ("to be"). Intransitive verbs agree with the subject, while transitive verbs follow ergative alignment, with the verb agreeing with the direct object. This structure applies to both present perfect and past perfect, with the only difference being the tense marking on the auxiliary verb.

For intransitive verbs, the auxiliary *ēstādan* carries subject agreement, meaning the verb form agrees with the person and number of the subject. For transitive verbs, the auxiliary verb *ēst* instead agrees with the direct object, reflecting the ergative nature of Pahlavi language past-tense transitive verbs.

Present Perfect. We use the present tense conjugation of the verb *ēstādan* as an auxiliary to construct the present perfect tense, as shown in Table 7. For example, "I have gone" (intransitive verb) is *raft ēst-ēm*, where *raft* is the past stem of "to go", *ēst* is the auxiliary, and *-ēm* is the first-person singular present suffix.

Past Perfect. We use the past tense conjugation of the verb *ēstādan* as an auxiliary to construct the past perfect tense, as shown in Table 8. For example, "I had gone" (intransitive verb) is *man raft ēstād hēm*, where *ēstād hēm* is the first-person singular past tense of *ēstādan*.

A.1.4 Continuous Tense

Pahlavi language marks continuous aspect using the preverbal particle *hamē*, which precedes the conjugated verb. This construction applies to both present and past tenses and expresses ongoing actions.

A.1.5 Negation

Negation in the Pahlavi language is expressed using *nē*, which precedes the verb phrase. This rule applies across simple, perfect, and continuous tenses while maintaining the same alignment properties as affirmative sentences.

A.2 Constructable Words

Pahlavi language features a productive system for deriving new words through regular morphological processes, including adjectives, adverbs, agentive nouns, locative nouns, possessive forms, plural forms, and various other noun types. We implement these rules by extracting root words from the dictionary and applying affix-based transformations to generate new forms.

A.2.1 Adjectival and Adverbial Derivations

Comparative adjectives are typically formed by adding *-tar* or *-dar* to the base adjective, while superlative adjectives take *-tom* or *-dom*. The selection of suffixes depends on the final sound of the base adjective. If the base ends in a voiceless consonant (f, k, p, s, t, x, š, or d), the comparative suffix *-tar* and the superlative suffix *-tom* are applied. For example, *wuzurg* meaning "big" forms *wuzurg-tar* for "bigger", while *xwaš* meaning "good" forms *xwaš-tom* for "best". If the base ends in any other sound, the comparative suffix *-dar* and the superlative suffix *-dom* are used, as seen in *abar* meaning "high", which forms *abar-dar* for "higher" and *abar-dom* for "highest".

Some adjectives do not follow these suffixation rules and instead have lexicalized or irregular forms. For example, *weh* is used to mean "better". These irregular forms are stored in the dictionary.

Adverbs of manner are systematically derived from adjectives by appending *-īhā*. For example, rōšn-īhā means "clearly".

A.3 Nominal Derivations

Pahlavi language supports systematic noun formation through a range of suffixes, including those that denote pluralization, agency, possession, guardianship, mastery, and locative functions, as detailed in Table 1.

B Conversion between phoneme, transliteration and book Pahlavi script

The Figure 4 illustrates the letters of the Book Pahlavi script, accompanied by their corresponding phonemic representations and transliterations. These are the basis for our rule creation.

Pahlavi Letter	Transliteration	Transcription	Description
𐭠	ʾ h	a/ ā h/ x	This letter, when representing the symbol ʾ, is transcribed in combination with the letter *w* as *aw, āw, ō, o, u, ū*, and in combination with the letter *y* as *ay, āy, ē, e, ī, i*.
𐭡	b	b	Sometimes, this letter is written in place of the Pahlavi letter *y*. It can also represent the number 1.
𐭢 / 𐭣	y d g	j/ y/ e/ ē/ i/ ī d g/ y	Sometimes this letter is written in place of *b* and *k*. Sometimes it is a shortened form of the letter *z*.
𐭤	w r n	w/ u/ ū/ o/ ō r n	In Pahlavi, at the end of words, after letters that do not connect to the following letter, this letter appears as an extra (redundant) character. When this letter is combined with *b*: • If written as *wb*, it is transcribed as *w*. • If written as *nb*, it is transcribed as *m*.
𐭥	z	z	
𐭦	k	k/ g/ γ	This letter can sometimes replace *g*.
𐭧	l	l/ r	
𐭨	m	m	
𐭩 / 𐭪	s	s/ h	The 𐭪 form is used at the beginning of a word and in connection with the letter *w*.
𐭫	p	p/ f/ b	
𐭬	c	č/ z/ j	
𐭭	š	š	
𐭮	t	t/ d	

Figure 4: This image displays the Pahlavi letters (in Book Pahlavi script) along with their corresponding phonemic representations and transliterations.