Evaluating the Performance of Transformers in Translating Low-Resource Languages through Akkadian

Daniel Jones
Lancaster University, United Kingdom
d.jones15@lancaster.ac.uk

Ruslan Mitkov
University of Alicante, Spain
ruslan.mitkov@ua.es

Abstract

In this paper, we evaluate the performance of various fine-tuned, transformer-based models in translating Akkadian into English. Using annotated Akkadian data, we seek to establish potential considerations when developing models for other low-resource languages, which do not yet have as robust data. The results of this study show the potency, but also cost ineffeciency of Large Language Models compared to smaller Neural Machine Translation models. Significant evidence was also found demonstrating the importance of fine-tuning machine translation models from related languages.

Keywords: transformer, neural machine translation, low-resource, Akkadian

1 Rationale

Ancient languages serve as vital links to our cultural and historical heritage. Akkadian, once the lingua franca of Mesopotamia (George, 2007), exemplifies this connection. Massive digitisation initiatives such as ORACC and the CDLI projects have generated extensive corpora of transliterated cuneiform texts (Tinney et al., 2025; CDLI Contributors, 2025); for instance, the Ur III corpus comprises over 72,000 transcribed texts, yet only 2.2% have been translated into modern languages (Punia et al., 2020). This stark bottleneck underscores the need for robust machine translation (MT) tools that can democratise access to these historical records for Assyriologists and scholars alike.

Since the introduction of the Transformer architecture (Vaswani et al., 2017), MT solutions have swiftly moved away from statistical models in favour of neural approaches. The surge in academic interest in low-resource languages – characterised by their limited digital presence and sparse representation in training data – has further highlighted the challenges faced by contemporary Large Language Models such as ChatGPT or Gemini in low-resource scenarios (Hasan et al., 2024). These languages necessitate tailored strategies to achieve robust translations, whether through fine-tuning on curated sentence pairs or via methods like Retrieval Augmented Generation (RAG), as explored by Shu et al. (2024). However, this study focuses exclusively on fine-tuning.

Akkadian itself, though extinct and largely confined to the realm of Assyriology, benefits from a uniquely rich, highly annotated corpus[1] with many bidirectional translations – an advantage seldom seen in low-resource languages. This abundance of quality data obviates the need for extensive data augmentation techniques often employed in projects for under-documented languages (e.g., as described in NLLB, 2022). Instead, Akkadian offers an ideal testbed for evaluating the performance of different pre-trained transformer architectures – both sequence-to-sequence (seq2seq) models and causal (decoder-only) models – in a low-resource, morphologically distinct setting. Additionally, the relative simplicity of the cuneiform transliteration system, in which wedge clusters represent syllables (Schmandt-Besserat, 2014), enables a straightforward conversion into the Latin alphabet, making Akkadian particularly amenable to phonetic-like translation tasks.

Our investigation evaluates how model architecture, parameter count, and the nature of pre-training data (including exposure to related Semitic languages) influence translation quality. By establishing a performance baseline for Akkadian-to-English translation, our study not only addresses a critical gap in the digitisation and translation of ancient texts but also lays the groundwork for broader applications of MT for low-resource languages.

[1] Entries in Oracc contain descriptions such as line rulings, which are not necessary regarding this paper, but may be useful for future research into Akkadian OCR

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 39–47, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-005

The remainder of this paper is structured as follows. Section 2 provides a background on related work, including an overview of Akkadian, the digitisation efforts of cuneiform texts, and research on MT for low-resource languages. Section 3 details the methodologies used to fine-tune and train the models, along with the compromises made during data preparation, and the metrics used for evaluation. Section 4 offers a comprehensive evaluation of the results, and Section 5 and 6 conclude with a discussion of the study's implications and potential directions for future research.

2 Preliminaries

2.1 Akkadian

Akkadian is an extinct East Semitic language that was spoken in Mesopotamia – roughly corresponding to modern-day Iraq. Historically, it was written in cuneiform on clay tablets using a wedge-based script. Each cuneiform symbol represents a syllable; for example, to write the word "cat," the script would use two distinct symbols, representing "ca" and "at" respectively (as illustrated by examples from the British Museum). The transliteration of these cuneiform texts into the Latin script forms the source for our machine translation (MT) task.

Akkadian qualifies as a low-resource language due to its limited online presence (Magueresse et al., 2020). For the majority of low-resource languages, this results in difficulty obtaining good training data – ideally a dataset of parallel sentences. Recent efforts to leverage technology for preserving cultural heritage and enhancing digital inclusivity (Galla, 2018; Joshi et al., 2020) have elevated interest in such languages. Notably, collaborative projects like CDLI and Oracc have been instrumental in digitising vast collections of clay tablets, thereby providing a rich bilingual corpus that is rarely available for other low-resource languages.

There are a few important caveats when using Akkadian as a benchmark. Although Akkadian was originally written in cuneiform, its digitised representation in CDLI and Oracc is transliterated into the Latin script. Consequently, the performance of our MT models specifically reflects translation challenges for low-resource languages presented in this format. Additionally, Akkadian occasionally incorporates Sumerian elements – sumerograms or logograms – into its script. For instance, the Akkadian word for "king" is pronounced "sharum" yet may be rendered with the Sumerian term "Lugal." These instances are statistically infrequent and unlikely to significantly affect overall model performance.

This study leverages Akkadian's unique position as a well-annotated yet low-resource language to evaluate and refine neural machine translation techniques, ultimately contributing to both the preservation of ancient cultural heritage and the advancement of MT for underrepresented languages.

2.2 Low-Resource Languages

The digital preservation of ancient languages is vital not only for maintaining the cultural heritage of communities but also for tapping into a vast potential market – after all, there are nearly 3 billion speakers of low-resource languages worldwide (Kshetri, 2024). Initiatives such as Meta's No Language Left Behind (NLLB Team, 2022) have shown that investment in multilingual translation goes beyond charity; it opens up entire emerging markets while preserving unique cultural identities. In this context, Akkadian stands out as a particularly interesting case study. Its status as a low-resource language is compounded by the fact that our source material is transliterated text – the conversion of ancient cuneiform (originally inscribed on clay tablets using wedge impressions) into the Latin alphabet. An example of this transliteration process is evident in the texts provided by Oracc (2025).

2.3 Low-Resource Comparisons and Cuneiform Translation

Solutions to translate Akkadian have been explored previously. Krueger (2023a) details the development of an AI Cuneiform Corpus – a resource for Assyriology that leverages a fine-tuned T5 transformer model to generate translations of both Sumerian and Akkadian texts. His work employs bidirectional translation training, whereby the model is also trained to translate back from English to the source language. This strategy helps stabilise convergence across epochs, even though the work does not report modern metrics, such as BLEU scores, to benchmark its performance. The model's availability on HuggingFace (Krueger, 2023b) enables direct, side-by-side comparisons with other approaches, making it a valuable reference point for our own T5 experiments.

During a period of heightened interest in the machine translation of ancient languages, Punia

et al. (2020) evaluated multiple architectures for translating Sumerian to English. Their study compared a Base Translator – an LSTM-based model without pre-trained embeddings – against an Extended Translator that incorporated pre-trained embeddings from the Wikipedia corpus (Pennington et al., 2014), as well as a Transformer-based model. Despite Transformers' well-known advantage in handling long sequences through self-attention, the brevity of cuneiform inscriptions (with an average of just 2.8 tokens per phrase in Punia's dataset) appears to limit the benefits of this architectural choice. As a result, the Extended Translator achieved a slightly higher BLEU score (21.6) compared to the Transformer (20.9), though the Transformer still outperformed the Base Translator. These findings underscore that while self-attention offers robust performance overall, fine-tuning specifics – such as access to pre-trained embeddings – can be particularly crucial in scenarios where input sequences are very short.

Shu et al. (2024) further contribute to this discussion by demonstrating how Retrieval Augmented Generation (RAG) can be used effectively in low-resource settings – in their case, for Cherokee translation. Their RAG model, although yielding moderate BLEU scores, showed impressive semantic understanding as measured by BERTScore. This suggests that even when lexical overlap is low, models can capture deep semantic meaning if properly contextualised through additional data retrieval. While these findings point to the potential of large language models when fine-tuned or augmented appropriately, the higher computational costs involved also highlight the appeal of achieving strong performance through more focused, fine-tuning methods – exactly the approach taken in this project.

3 Data and methodology

3.1 Corpora

A major reason for the choice of Akkadian as the language of interest is the organisation of data that exists. Assyriologists have worked to digitise the world's discovered cuneiform tablets into organised corpora. Since this digitising process occurred over many years, there are inconsistencies within the standardisation of cuneiform, a problem discussed by Krueger (2023a). Some symbols are translated in ASCII, whereas more modern forms maybe transliterated using accented Unicode characters. For this project, data was gathered from the Oracc (2025) and CDLI (2025) corpora. These corpora are extensive and hold enough translation examples to train a competent translator.

3.1.1 Oracc Corpus

The Oracc corpus grew out of recognition of the limitation present in the Electronic Text Corpus of Sumerian Literature (Black et al., 2002). ETCSL was initiated by Jeremy Black and Graham Cunningham of the University of Oxford, and had the ambitious goal to create an online corpus with Sumerian literary texts, along with their English translations (Ebeling, 2007). Though a valuable resource, it was limited in many aspects. Its focus on Sumerian was problematic when considering cuneiform, a writing system used to write countless, linguistically unrelated languages. Furthermore, ETCSL was largely static, limiting the ability of the community to contribute and stunting its development. Recognising these limitations, the Oracc corpus was developed. It allowed for richer annotation, beyond just translations (Oracc, 2019), and emphasised openness and collaboration. Oracc includes glossaries for each subproject within the overall corpus. These glossaries provide information about all the words used within the subproject. While this is not useful for this project, since the model should be able to learn words from exposure during fine-tuning, it does have potential to be useful when considering a technique like RAG (Shu et al., 2024) discussed earlier. These glossaries could be used to scrape a wordlist, which can be used as context for larger models with very potent few-shot capabilities.

3.1.2 CDLI

The Cuneiform Digital Library Initiative (CDLI), unlike Oracc, focuses on being a digital archive for the objects themselves. As such, the tablets have high quality photos and line art, occasionally with transliterations. It has an emphasis on unpublished materials, allowing researchers to access tablets worldwide for research and study. Both CDLI and Oracc provide a valuable resource for this project. Scraping their data allows for transliterated Akkadian to be gathered en masse. Furthermore, it allows for untranslated, but still transliterated, texts to be gathered. While not vital to this project, it is important to a project such as AICC, which used Krueger's model to translate previously undeciphered texts into English.

3.2 Methodology

3.2.1 Models used

The transformers chosen for comparison primarily differ in their architecture, pre-training scope and parameter size. Broadly, the models chosen can be split into Sequence-to-sequence (encoder-decoder) and causal (decoder-only) architectures. The *seq2seq* models used here are considered Neural Machine Translation (NMT) solutions, whereas the *causal* models are considered LLM models. The models chosen were as follows:

- **T5-base** (Raffel et al., 2020): 250 million parameter encoder-decoder model that was used by Krueger (2023a) to translate Akkadian and Sumerian. Uses C4 corpus with mostly English text scraped from the web, and is trained to perform a variety of tasks, including translation. Since this model was used by Krueger, it serves as a baseline and sense check for the quality of our own experiments.

- **MarianMT** (Junczys-Dowmunt et al., 2018): MarianMT has roughly 75 million parameters, and is an encoder-decoder model specifically designed for translation tasks. It has half the layers of T5-base, and is trained on the OPUS corpus (Tiedemann, 2012), which contains parallel corpora in multiple languages. This model provides insight in the tradeoff between fine-tuning a multi-purpose model as opposed to a model designed specifically for translation.

- **Qwen 0.5B-Instruct** (Yang et al., 2024): A 500 million parameter decoder-only model. It is inherently multilingual, being trained across multiple languages. It also utilises some advancements to the transformer architecture, such as Rotary Positional Embeddings, which may allow it to better understand semantics within sentences. It also uses Grouped-Query Attention, which may allow for faster inference times.

- **Mistral 7B** (Lachaux et al., 2023): A 7 billion parameter decoder-only model developed by Mistral AI that leverages Grouped-Query Attention for faster inference. With a parameter count 14 times that of Qwen 0.5B, it offers enhanced few-shot translation performance. The model employs a Byte-Pair Encoding (BPE) tokeniser adapted for translation, ensuring robust handling of both ASCII and Unicode inputs. Training on consumer hardware is made feasible by using quantisation techniques (Gholami et al., 2022) and LoRA (Hu et al., 2021), with most parameters remaining unchanged during fine-tuning.

We seek to establish the performances of these various transformer-based NMT and LLM models – each with different architectures, parameter counts and specifity. As stated, the models chosen for this experiment are MarianMT, T5, Qwen 0.5B instruct, and Mistral 7B. These models are advanced enough to provide worthwhile translations, but can still be trained on consumer hardware, and run cheaply. T5 and MarianMT are encoder-decoder transformers, whereas Qwen and Mistral are decoder-only transformers. The decoder-only architecture not only reduces model size, but also does not need labelled input that a encoder-decoder model might (Fu et al., 2023). This means it can be more readily trained with available data on the internet, hence why it is favoured by the larger language models. This might be of benefit in few-shot capabilities when translating Akkadian to English. Architecturally, Qwen and Mistral use more modern techniques when embedding and using multiple attention heads. Qwen, for example, uses Rotary Positional Embeddings (RoPE). This gives the model an advantage in understanding relative word relations by reducing the influence words have on one another with distance, as opposed to T5's relative positional embeddings, which cannot decay dependencies as effectively (Su et al., 2024). It also benefits faster inference, because of improvements such as Grouped-Query Attention, and has a significantly higher context window then MarianMT and T5.

An important compromise was made in order to train Mistal 7B. Given the limited VRAM on consumer hardware, methods were used to lessen this burden. Namely, through the Unsloth library, Low-Rank Adaptation (LoRA) was used to freeze the original model weights, and only train newly injected matrices. In its introductory paper by Hu et al. (2021), it was shown that LoRA drastically lowers the resources required to train. GPT 175B's VRAM consumption during training using LoRA reduced VRAM consumption form 1.2TB to 350GB. Despite this, empirical evidence has also shown that the fine-tuned capability of LoRA trained models equals, or occasionally outperforms fully-trained models (Agiza et al., 2024; Peters

et al., 2019). As such, the outcome can still be considered alongside the other models.

3.2.2 Evaluation Metrics

Translation quality will be assessed using the following metrics:

- **BLEU (Papineni et al., 2002)**: A widely used metric for evaluating machine translation quality, BLEU measures the overlap between n-grams in the generated translation and reference translations, and includes a brevity penalty. It is particularly effective for assessing lexical similarity.
- **BERTScore** (Zhang et al., 2019): This metric evaluates semantic similarity by comparing contextual embeddings of words in the generated translation against those in reference translations. BERTScore measures the accuracy of semantic meaning, making it useful for assessing whether the model captures intended meanings of text, even if the exact words differ.
- **ROUGE2** (Lin, 2004): This metric measures bigram overlap between the generated and reference translations. By capturing both precision and recall of contiguous word sequences, it assesses whether key multi-word expressions and local phrasal structures are preserved. In doing so, ROUGE2 complements BLEU by providing insight into the preservation of phrase-level content.

3.2.3 Gathering

Oracc provides API's to access lists of projects hosted, given in JSON format. Each project has its own collected corpus of cuneiform transliterations that can be navigated through. Each project provides different types of documents, and it is important to appreciate the difference between these when considering a translator. For example, a project such as *akklove* (2025), contains Akkadian love literature. Within this corpus are lengthy poems that incorporate descriptive vocabulary. On the other hand, a corpus such as The Royal Inscriptions of Assyria online (Grayson et al., 2025) contains royal inscriptions, which are often shorter and much more repetitive than *akklove*. It is important to appreciate the bias and diversity of the dataset in order to make best use of it. Data scraped from these projects are in the form of ATF files, a format specifically designed to digitise cuneiform tablets, whilst maintaining metadata about its format.

Krueger (2023b) had already scraped the corpora, and using the same training data provides a foundational benchmark to assess our models against existing solutions. Overall, 95,629 samples were used, with a split of 90% training (86,066) and 10% testing (9,563). A validation set was not used in this case, since tuning of hyperparameters was not performed. Instead, a learning rate of 2e-5 was used for all models, with varying numbers of epochs. This is a common learning rate for fine-tuning transformer models, and was used by Krueger (2023a).

4 Experiments and Results

The models were each trained with a different number of epochs. MarianMT models, along with T5, were trained for 15 epochs, while Qwen 0.5B was trained for 3. Mistral7B was only trained for 1 epoch. This is mostly due to limitations in compute, but provides insight into the few-shot capabilities of the larger models.

4.1 All Sentences

Table 1 reports BLEU, ROUGE-2 and BERTScore over all test sentences.

- Prec - ROUGE2 Precision
- Recall - ROUGE2 Recall
- F1 - ROUGE2 F1

It shows 6 fine-tuned models. The LLM's, Mistral 7B and Qwen 0.5B, as opposed to the NMT models, MarianAr (Arabic→English), MarianEs (Spanish→English), Krueger (T5) and T5. The models are ordered by BLEU score, with Mistral 7B achieving the highest BLEU of 0.478, followed by MarianAr at 0.453, Krueger at 0.416, Qwen at 0.403, T5 at 0.376 and MarianEs at 0.122.

Table 1: Evaluation on all sentences

Model	BLEU	Prec	Recall	F1	BERT
Mistral	0.478	0.527	0.494	0.501	0.930
MarianAr	0.453	0.541	0.508	0.512	0.931
Krueger	0.416	0.530	0.484	0.493	0.930
Qwen	0.403	0.516	0.487	0.491	0.929
T5	0.376	0.420	0.397	0.399	0.914
MarianEs	0.122	0.198	0.303	0.209	0.842

Mistral's 7 billion-parameter model achieved the highest BLEU of 0.478 (3 sf), indicating strong

n-gram overlap with the reference. MarianAr followed at 0.453 – a 5.2% deficit – while Krueger's and Qwen trailed at 0.416 and 0.403 respectively. Our T5, trained for only 15 epochs, reached 0.376, and MarianEs (Spanish→English) lagged at 0.122, a 74.5% drop relative to Mistral.

In ROUGE-2 Precision, MarianAr led with 0.541 (54.1% of generated bigrams in the reference), followed by Krueger (0.530), Mistral (0.527) and Qwen (0.516), all within a 4.6% band. T5 and MarianEs fell to 0.420 and 0.198. For Recall, MarianAr attained 0.508, Mistral 0.494, Qwen 0.487 and Krueger 0.484. Combined F1 placed MarianAr at 0.512, Mistral at 0.501 and the next best systems within 4.1%.

On BERTScore most models clustered around 0.930-0.931, effectively within margin of error. MarianAr scored 0.931, with Mistral, Krueger and Qwen at 0.930. T5 scored 0.914 and MarianEs 0.842.

4.2 Long Sentences (Reference $\geq$ 4 words)

Table 2 shows metrics when restricting to sentences with four or more reference tokens.

Table 2: Evaluation on long sentences

Model	BLEU	Prec	Recall	F1	BERT
Mistral	0.473	0.549	0.510	0.519	0.928
MarianAr	0.446	0.562	0.522	0.529	0.929
Krueger	0.407	0.554	0.499	0.510	0.928
Qwen	0.395	0.532	0.496	0.503	0.927
T5	0.370	0.434	0.407	0.409	0.911
MarianEs	0.149	0.230	0.340	0.242	0.854

BLEU dipped slightly for all models (e.g. Mistral from 0.478→0.473, MarianAr 0.453→0.446), while MarianEs rose from 0.122→0.149. Precision increased across the board – MarianAr reached 0.562, Krueger 0.554 – preserving the rank order. Recall and F1 mirrored this improvement (MarianAr recall 0.522, F1 0.529). BERTScore remained effectively unchanged.

4.3 Short Sentences (Reference $<$ 4 words)

Table 3 isolates sentences shorter than four words.

Table 3: Evaluation on short sentences

Model	BLEU	Prec	Recall	F1	BERT
Mistral	0.602	0.408	0.407	0.404	0.938
Qwen	0.593	0.430	0.435	0.428	0.942
Krueger	0.586	0.404	0.406	0.401	0.940
MarianAr	0.541	0.426	0.429	0.423	0.941
T5	0.520	0.346	0.348	0.344	0.927
MarianEs	0.016	0.025	0.103	0.031	0.778

Short sentences boosted BLEU markedly for all except MarianEs: Mistral rose to 0.602, Qwen to 0.593, Krueger to 0.586, each surpassing MarianAr's 0.541. Qwen led ROUGE-2 precision (0.430), recall (0.435) and F1 (0.428). BERTScore peaked at 0.942 for Qwen, with MarianAr at 0.941 and Krueger 0.940.

4.4 Significance Testing

To determine the statistical significance of the differences in BLEU scores between models, we conducted significance tests. Table 4 presents the p-values and confidence intervals for BLEU delta between pairs of models.

Table 4: Significance Test Results (BLEU Delta)

Model 1	Model 2	BLEU Δ Lower CI	BLEU Δ Upper CI	p-value
Krueger	**MarianAr**	–4.10	–3.31	0.00
Krueger	**Mistral7b**	–7.00	–5.68	0.00
Krueger	**MarianEs**	27.72	30.08	0.00
Krueger	**Qwen05**	0.87	1.63	0.00
Krueger	**T5**	3.58	4.32	0.00
MarianEs	**MarianAr**	–33.65	–31.50	0.00
MarianEs	**Mistral7b**	–36.05	–34.31	0.00
MarianEs	**Qwen05**	–28.77	–26.45	0.00
MarianEs	**T5**	–25.98	–23.91	0.00
MarianAr	**Mistral7b**	–3.15	–2.14	0.00
MarianAr	**Qwen05**	4.49	5.39	0.00
MarianAr	**T5**	7.24	8.06	0.00
Mistral7b	**Qwen05**	6.94	8.23	0.00
Mistral7b	**T5**	9.75	10.86	0.00
Qwen05	**T5**	2.28	3.13	0.00

The significance testing results in Table 4 provide a detailed statistical analysis of the differences in BLEU scores between pairs of models. The table includes the lower and upper confidence intervals for the BLEU delta, along with the corresponding p-values. All p-values are less than 0.05, indicating that the differences in BLEU scores between the models are statistically significant.

4.5 Inference Timings

Table 5 compares wall-clock times to translate 500 sentences[2].

Table 5: Inference time for 500 sentences

Model	Total Time (s)	Per Sentence (s)
MarianAr	108	0.22
T5	242	0.48
Qwen	348	0.70
Mistral	1786	3.57

MarianAr was fastest at 108s (0.22s/sentence), over four times quicker than Mistral's 1786s (3.57 s/sentence). T5 and Qwen required 242s and 348s, respectively. Inference speed is an important metric when considering practical applications, and the tradeoff between speed and quality needs to be considered.

4.6 Summary

Overall, MarianAr and Mistral are the top performers: MarianAr leads in all metrics except BLEU (where Mistral narrowly wins). Krueger's T5 surpasses our 15-epoch T5 and Qwen in semantic scores, highlighting the impact of extended training and transfer. Short sentences favour few-shot LLM generalisation (Mistral, Qwen), while longer contexts modestly reduce recall. Finally, smaller specialised models (MarianMT) offer the best tradeoff of speed and quality for practical low-resource language translation.

Shown in Figure 1 are translations between transliterated Akkadian and English, using the MarianAr model.

Example 1 - Astronomical Text:

Source:
(_mul2_-e)-sza2-_sag_-_gir2-tab 20 si 6 zi ir kur gin ge6 7 sag ge6_ sin ina _igi mul2#_-_kur_-sza2-_kir4_-szil-_pa 3 kusz3_ beta Scorpii
Translation:
The 6th, ZI IR, the east wind blew. Night of the 7th, beginning of the night, the moon was 3 cubits in front of theta Ophiuchi
Example 2 - Royal Inscription:

Source:
(d#)na3#-ku-du-ur2-[ri-uri3] _lugal_ ba-bila2#[(ki)] za-ni-in e2-sag-il2# u3 e2-zi-da# _ibila_ a-sza-re-du sza (d)na3-ibila-uri3 _lugal_ babila2(ki)
Translation:
Nebuchadnezzar, king of Babylon, who provides for the E-sagil and the Ezida, foremost son of Nabopolassar, kingship of Babylonia

Figure 1: Example translations from the MarianAr model showing transliterated Akkadian to English

5 Discussion

When interpreting these results, it's important to remember that our cuneiform corpus is highly repetitive – many near-identical phrases appear in both training and test splits, inflating absolute scores. While this doesn't undermine comparisons between models, it does caution against assuming similar performance on a more varied low-resource dataset.

Firstly, our T5 baseline (15 epochs) underperforms Krueger's T5 (30 epochs) across both BLEU and BERTScore, despite identical hyperparameters. Krueger's additional epochs – and his bidirectional training (English→Akkadian/Sumerian) – helped convergence. This technique is not viable for MarianMT's single-direction architecture, but it signals that smaller models without LLM-style few-shot strength benefit substantially from extended fine-tuning.

Overall, MarianAr (Arabic→English) delivers the best balanced performance. It leads in ROUGE2 and BERTScore, and only narrowly trails Mistral-7B on BLEU. By contrast, MarianEs (Spanish→English) lags dramatically, confirming that even a distantly related Semitic language imbues the model with useful implicit grammatical and lexical knowledge – despite script mismatches and millennia of divergence.

The few-shot prowess of large causal LLMs also shines through. Mistral-7B achieves the top BLEU after a single epoch, and Qwen-0.5B, with three epochs, matches or bests others on very short sentences (<4 words). These results suggest their vast pre-training mitigates sparse data, particularly for lexical matching.

Inference speed highlights practical trade-offs. Although Mistral-7B excels in raw BLEU, its 7 billion parameters slow throughput severely. In contrast, MarianMT variants – especially MarianAr – combine strong quality with sub-second latency, making them better suited for real-world, resource-constrained deployment.

[2]MarianAr was used for timing MarianMT, it can be assumed MarianAr will be roughly equivalent.

6 Conclusions and Future Work

In this paper we have demonstrated that careful adaptation of existing NMT architectures can unlock high-quality Akkadian→English translation even under severe data scarcity. Our experiments show that fine-tuning a MarianMT model pretrained on Arabic (MarianAr) delivers the best balance of surface accuracy (BLEU), semantic fidelity (ROUGE2, BERTScore), and inference efficiency. Despite the millennia that separate Arabic and Akkadian – and the mismatch between Arabic script and romanised transliteration – MarianAr's internalised Semitic grammar and vocabulary proved remarkably transferrable. At the same time, we observed that large causal LLMs such as Mistral-7B and Qwen-0.5B require only one to three epochs to rival or exceed other models on shorter sentences, underlining their potent few-shot adaptation. Yet the hefty parameter counts of these LLMs incur a tangible latency penalty, reaffirming the practical importance of lightweight, specialised NMT when deployment speed and resource budgets are at a premium.

Looking ahead, several avenues promise to extend and deepen these findings. First, pushing MarianAr and our T5 baseline through additional epochs and systematic hyperparameter sweeps will clarify the point of diminishing returns and guard against overfitting. Second, a mixture-of-experts framework – where a fast NMT core handles routine or formulaic passages while a heavyweight LLM tackles longer or more ambiguous sentences – could marry speed with versatility. Third, augmenting our pipeline to ingest raw cuneiform images and output English translations would bridge OCR/transliteration and MT, yielding a seamless toolchain for Assyriologists. Finally, applying this comparative lens to other under-documented Semitic and ancient languages will test the generality of "pre-train on related language + fine-tune" and few-shot paradigms across diverse scripts, dialects, and time periods. By pursuing these threads, we aim to push the frontier of low-resource, historical-language translation ever closer to full academic and cultural utility.

Acknowledgments

This work has been partially supported by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana.

References

Ahmed M. Agiza, Kai Zhu, Tianlong Zhang, and Shijie Han. 2024. Mtlora: Low-rank adaptation approach for efficient multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16196–16205.

Jeremy Black, Graham Cunningham, Eleanor Robson, and Gábor Zólyomi. 2002. The electronic text corpus of sumerian literature. https://etcsl.orinst.ox.ac.uk/index1.htm. Accessed: 2025-03-17.

CDLI Contributors. 2025. Cuneiform digital library initiative: Home. https://cdli.mpiwg-berlin.mpg.de/. Accessed: 2025-01-21.

Jarle Ebeling. 2007. The electronic text corpus of sumerian literature. *Corpora*, 2(1):111–120.

Zhengluo Fu, Guangxiang Zhou, Xiaogang Li, Jin Wang, Aiyun Zeng, Liang Lyu, and Xing Zhou. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*.

Candace Kaleimamoowahinekapu Galla. 2018. Digital realities of indigenous language revitalization: A look at hawaiian language technology in the modern world. *Language and Literacy*, 20(3):100–120.

Andrew George. 2007. *Babylonian and Assyrian: A history of Akkadian*, pages 31–71. British School of Archaeology in Iraq.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2022. *A survey of quantization methods for efficient neural network inference*, pages 291–326. Chapman and Hall/CRC.

Kirk Grayson, Jamie Novotny, and Poppy Tushingham. 2025. The royal inscriptions of assyria online (riao) project. https://oracc.museum.upenn.edu/riao/. Accessed: 2025-03-17.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shengyuan Wang, and Weizeng Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Pratik S Joshi, Chathurika Welivita, Pubudu Liyanapathirana, Ameya Budhiraja, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Sevcik, Roman Vólner,

Anthony Aue, Alexandra Birch, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Frank. A. Krueger. 2023a. I built the world's largest translated cuneiform corpus using ai. https://praeclarum.org/2023/06/09/cuneiform.html. Accessed: 2025-03-17.

Frank A. (praeclarum) Krueger. 2023b. Dataset for cuneiform language translation. Accessed [2025-03-19].

Nir Kshetri. 2024. Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management*, 27(2):95–99.

Arthur Lachaux, Baptiste Lucic, Thomas Kirchner, Arthur Mensch, Sander Lowenthal, Joseph Rouzé, Jean-Baptiste Mensch, Quentin Yu, Thibaut Batigne, Clara Stone, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Oracc. 2019. ATF Primer. https://oracc.museum.upenn.edu/doc/help/editinginatf/primer/index.html. Accessed: 2025-03-17.

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. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Ravneet Punia, Niko Schenk, Christian Chiarcos, and Émilie Pagé-Perron. 2020. Towards the first machine translation system for sumerian transliterations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3454–3460. International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Mihir Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21:1–67.

Denise Schmandt-Besserat. 2014. *The Evolution of Writing*, pages 1–15. Elsevier.

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*.

Michael P. Streck and Nathan Wasserman. 2025. Sources of early akkadian literature: Love literature. https://oracc.museum.upenn.edu/akklove/. Accessed: 2025-03-17.

Jianlin Su, Yu Lu, Shengfeng Xu, and Ahmed Murtadha. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Steve Tinney, Eleanor Robson, Niek Veldhuis, and Jamie Novotny. 2025. Open richly annotated cuneiform corpus. https://oracc.museum.upenn.edu/. Accessed: 2025-01-25.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30.

Aiying Yang, Xiaotian Jin, Rui Zhang, Yibo Zhang, An Li, Jiaming Wu, Wenbin Wang, Chen Xu, Qian Wang, Xing Zhou, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.