

DOI: 10.37943/G0JE5782

Bekarys Baibolat

Master's Student, Big Data and Blockchain Technologies Research and Innovation Center
255265@astanait.edu.kz, orcid.org/0009-0002-2739-8779
Astana IT University, Kazakhstan

Aigerim Mansurova

Master's Degree, Researcher,
Big Data and Blockchain Technologies Research and Innovation Center
a.mansurova@astanait.edu.kz, orcid.org/0009-0003-1978-9574
Astana IT University, Kazakhstan

Mukhammadsaid Mamasaidov

Bachelor's Degree, Founder at Tilmo.ch.ai
m.mamasaidov@tilmo.ch.ai, orcid.org/0009-0003-3292-2472
Tilmo.ch.ai, Uzbekistan

Almas Imangaliyev

PhD, Assistant Professor, School of General Education Disciplines
a.imangaliyev@astanait.edu.kz, orcid.org/0000-0001-8175-9745
Astana IT University, Kazakhstan

Aliya Nugumanova

PhD, Director, Big Data and Blockchain Technologies Research and Innovation Center
a.nugumanova@astanait.edu.kz, orcid.org/0000-0001-5522-4421
Astana IT University, Kazakhstan

**MULTILINGUAL NEURAL MODELS FOR TRANSLATING CHAGATAI
HISTORICAL TEXTS**

Abstract: Recent advances in neural machine translation have improved automatic translation for high-resource languages; however, historical and extremely low-resource languages such as Chagatai remain underserved because of scarce parallel data, complex agglutinative morphology, and orthographic variation. Chagatai served as a literary language of Central Asia from the Timurid period until the early twentieth century, yet much of its textual heritage remains inaccessible to non-specialists. This study investigates whether multilingual translation architectures can be adapted to Chagatai under severe data constraints. A parallel corpus of 10,712 aligned Chagatai–English and Chagatai–Kazakh entries were constructed from academic sources using automated extraction, normalization, multilingual alignment, and manual validation of target-language correspondences by three linguists. Two architectures were evaluated: the encoder–decoder NLLB-200 distilled 600M model, extended with a Chagatai-specific language token and fine-tuned in a Seq2Seq framework, and the decoder-only TranslateGemma-4b-t model, adapted using parameter-efficient DoRA fine-tuning. Zero-shot experiments showed near-zero translation performance for both models, demonstrating the necessity of task-specific adaptation. After fine-tuning, NLLB-200 achieved stronger performance across BLEU, chrF++, Exact Match, and XCOMET. For Kazakh, it obtained an XCOMET score of 74.31 compared with 69.75 for TranslateGemma, while also achieving higher BLEU, chrF++, and Exact Match scores. For English, TranslateGemma produced a slightly higher BLEU score, whereas NLLB-200 achieved stronger semantic and exact-match performance. Error analysis showed that TranslateGemma generated fewer omissions but more hallucinations, while NLLB-200 exhibited more conservative source-anchored errors. These findings suggest that translation-specialized multilingual encoder–decoder models provide a more reliable foundation for machine translation of extremely low-resource historical Turkic languages.

Keywords: NLP; neural machine translation; multilingual models; low-resource languages; historical text translation; Chagatai language.

Introduction

The preservation of cultural heritage plays a crucial role in protecting collective memory, identity, and cultural continuity across generations. In Central Asia, the disruption of cultural and linguistic traditions during the Soviet period substantially undermined the transmission of historical knowledge, making its recovery and preservation an ongoing scholarly and societal imperative [1]. In the digital era, however, preservation requires more than the simple digitization of historical materials. Making historical knowledge accessible also requires the interpretation and translation of historical texts, which often remain unavailable to broader scholarly and public audiences.

The intellectual history of Central Asia, particularly from the Timurid period to the early twentieth century, was largely conducted in the Chagatai language, a common literary language of the Turkic-speaking peoples of the region. Systematic scholarly interest in Chagatai lexicographic resources dates to the first quarter of the nineteenth century and continues to develop as an active field of Turcological research [2]. While many earlier Turkic literary monuments such as “*Dīwān Lughāt al-Turk*” by Mahmud al-Kashgari and “*Kutadgu Bilig*” by Yusuf Balasaghuni belong to earlier stages of the Turkic literary tradition, they form part of the intellectual and linguistic heritage that later influenced the development of Chagatai literature [3]. Today, much of this textual heritage remains inaccessible because the majority of Chagatai manuscripts have not been translated into modern languages and can only be interpreted by a small number of specialists. As a result, large parts of the written cultural history of Central Asia risk entering a form of “digital dark age,” where materials may be digitally preserved but remain intellectually inaccessible.

Artificial Intelligence (AI) has the potential to automate translation, transforming the field from time-consuming manual work into scalable, data-driven systems. The development of Neural Machine Translation (NMT) marked a major shift from rule-based and statistical approaches to fully neural models. Early Recurrent Neural Networks (RNNs) [4] with Long Short-Term Memory (LSTM) [5] units improved translation fluency but faced difficulties with long-range dependencies, which are common in Turkic languages with Subject-Object-Verb word order. The introduction of Transformer architecture greatly advanced the field by using self-attention mechanisms, allowing models to process sequences in parallel and capture context across entire sentences [6].

Despite these advances, the translation of Chagatai texts presents significant challenges for neural models. Chagatai is an agglutinative Turkic language with complex morphology, where grammatical relations are expressed through chains of suffixes attached to lexical stems. A single stem may produce a large number of surface word forms depending on tense, aspect, mood, and person markers [7]. When only limited parallel data are available, neural models struggle to learn these morphological patterns and generalize across unseen word forms. Sparse training data also limits the ability of traditional statistical translation systems to construct reliable phrase tables, while neural sequence models often fail to capture the full range of morphological variation present in agglutinative languages [8-9].

To address the problem of limited data, researchers have explored Multilingual Neural Machine Translation (MNMT). Models such as mBART [10] and M2M-100 [11] support the idea that training a single model on multiple languages enables transfer learning. A model trained on high-resource Turkic languages, like Turkish and Uzbek, can learn patterns of agglutinative morphology and Subject-Object-Verb word order that generalize to related low-resource languages such as Chagatai, through shared subword vocabularies and cross-lingual parameter transfer [8]. However, early dense MNMT models faced a “curse of capacity,” where adding more languages reduced performance on individual languages. The No Language Left Behind (NLLB) project [12] addressed this issue using a Conditional Compute model with a Mixture-of-Experts

(MoE) architecture, successfully applied to low-resource Asian languages such as Hindi, Malay, Thai, and Vietnamese [13]. In this design, only a subset of parameters (experts) is activated for each token, enabling the model to scale to 200 languages, including closely related Turkic dialects, without a proportional increase in computational cost [12].

In parallel with developments in NMT, Large Language Models (LLMs) based on decoder-only architectures, such as GPT-4 [14], PaLM [15], and Gemma [16], have emerged as powerful general-purpose models. Unlike the encoder-decoder architecture of NLLB, which is specifically designed for sequence-to-sequence translation, decoder-only LLMs generate translations autoregressively, based on instructions and input text prompts [17]. Recent research indicates that while LLMs may perform slightly worse than specialized NMT models in zero-shot translation for low-resource languages, they show advantages in semantic coherence and handling ambiguity when fine-tuned [18-19]. The release of the Gemma 3 family, particularly TranslateGemma [20], provides an opportunity to study this potential. TranslateGemma is instruction-tuned for cross-lingual tasks, potentially combining generative fluency with improved translation accuracy.

Chagatai presents specific challenges for both NMT and LLMs because of its agglutinative morphology and extremely limited data. Even state-of-the-art multilingual models do not have enough parallel data to fully learn Chagatai's complex suffix patterns, while decoder-only LLMs struggle with zero-shot translation due to the lack of digital resources [21]. This study examines domain-specific fine-tuning of both TranslateGemma and NLLB using a specialized Chagatai-English parallel dataset, mainly compiled from historical texts and scholarly sources. The goal is to determine whether targeted adaptation on this small, morphologically rich corpus can take advantage of similarities with related Turkic languages and overcome these challenges.

Despite advances in NMT and large language models, comparative research on extinct or historical languages remains limited. Most studies on low-resource languages focus on African languages, such as Nguni and Sotho-Tswana [22], or Indic languages, including Hindi, Urdu, and Nepali [23]. In contrast, Chagatai has been unexamined in computational translation studies, despite its importance as a major literary and administrative language of Central Asia. As a result, it remains unclear how modern multilingual translation architectures behave when applied to historical Turkic texts characterized by sparse data, complex agglutinative morphology, and significant orthographic variation. This study addresses this gap by comparing two neural translation paradigms for Chagatai: the encoder-decoder NLLB-200 model and the decoder-only TranslateGemma model.

The contribution of this work is threefold. First, it develops a linguistically verified parallel corpus of Chagatai entries aligned with English and Kazakh translations, which serves as the foundation for fine-tuning and evaluation in this study. Second, it evaluates the performance of encoder-decoder and decoder-only architectures under conditions of extreme data scarcity, showing that source-anchored multilingual models provide more reliable translations than generative alternatives. Third, rather than only reporting general translation quality scores, the study examines architecture-specific translation behavior, including semantic fidelity, omission, and hallucination tendencies.

The practical significance of the study lies in its relevance for digital humanities, historical linguistics, and cultural heritage preservation. Reliable machine translation of Chagatai texts can support scholars in accessing manuscripts and historical sources that currently require highly specialized expertise. At the same time, the findings provide methodological guidance for researchers working with historical low-resource Turkic languages.

Methods and Materials

To provide a clear and consistent description of the experimental design, this section is structured around four interrelated components: 1) data source and description; 2) data preprocessing and quality control; 3) model architectures and fine-tuning; and 4) evaluation metrics.

1. Data source and description

To address the lack of digital resources for Chagatai, an extinct Turkic literary language, an experimental parallel corpus was created by combining two main sources, as shown in Fig. 1.

1. The main source for grammatical structures was “An Introduction to Chagatai: A Graded Textbook of Classical Central Asian Turkic” by Eric Schluessel [24]. Sentences and phrases were extracted along with their English translations and scholarly Latin transliterations. This part of the dataset ensures that the model can learn Chagatai’s morphological rules and sentence structures.
2. To expand vocabulary coverage, we included the Thesaurus Dictionary of Abulgazi Bahadur Khan's “Genealogy of the Turks” by Mamyrbekova and Seitbekova [25]. Accessed through the Sozdikqor national corpus platform, this dataset contains individual lexemes and terms from the historical chronicle Shajara-i Turk. It provides Chagatai entries with corresponding Kazakh translations and dual transliterations in Cyrillic and Latin.

Chaghatay	Transliteration	Definition
كون	kün	N. day, sun
آی	ay	N. month, moon (Turkic)
ماه	māh	N. month, moon (Persian) (Before a month, it has an <i>ezāfe</i> : <i>māh-i šawwāl</i> “the month of Shawwāl.” [See 10.3])
شوال	Šawwāl	Islamic month of Shawwāl
صفر	Safar	Islamic month of Safar
یل	yil	N. year
ای-ایر-	e-/er-	V. to be
طرفان	Turpan	PN. place name, Turpan/Turfan
ممت	Mämät	PN. man’s name
ابن	ibn	“son of;” usually abbreviated as “b.” in English translation
علی	‘Alī	PN. man’s name

a)

МҰРАДЫМЫЗ ҰА МАҚСҰДЫМЫЗ

lat: MURADYMYZ ҰА MAQSUDYMYZ زینو صدو ۰ مینار

0:00 / 0:02

Мұрат-мақсатымыз, арман-тілегіміз

Оң хан Чынгыз ханға кіші иберді үлгі ұғлы Жошыға қызымды берейін өзі біздің үйге келсін екіміздің һәм мұрадымыз Ұа мақсұдымыз түрүр той қыпалы Ұа тақы той баһане қылыб бірнеше күн ғиш Ұа ғишратға мәшғүл болалық (33, 1-2). Оң хан Шыңғыс ханға кісі жіберді: “Үлкен ұлы Жошыға қызымды берейін, өзі біздің үйге келсін, енеуіміздің де мұратымыз бен мақсатымыз бір болар, той жасайық, (тойды) себеп қылып бірнеше күн қызық пен қуанышқа көнелейік”.

Г. Мамырбекова, А. Сейтбекова: «Аблғазы баһадүр ханның «Түркі шежіресінің» тезаурусы сөздігі. Алматы: «Фирма Орман», 2013. - 480 б.

b)

Figure 1. Examples of source data used for dataset construction: (a) Example of a grammatical entry with English translation and Latin transliteration from Eric Schluessel Textbook and (b) An example of a lexical article with a Kazakh translation and transliterations from the Thesaurus Dictionary of "Genealogy of the Turks"

2. Data preprocessing and quality control

Before these materials could serve as training data, the raw extractions required systematic normalization and alignment to resolve the orthographic inconsistencies inherent in historical Arabic-script manuscripts. Fig. 2 illustrates the technical workflow encompassing extraction, multilingual alignment, linguistic validation, and data partitioning.

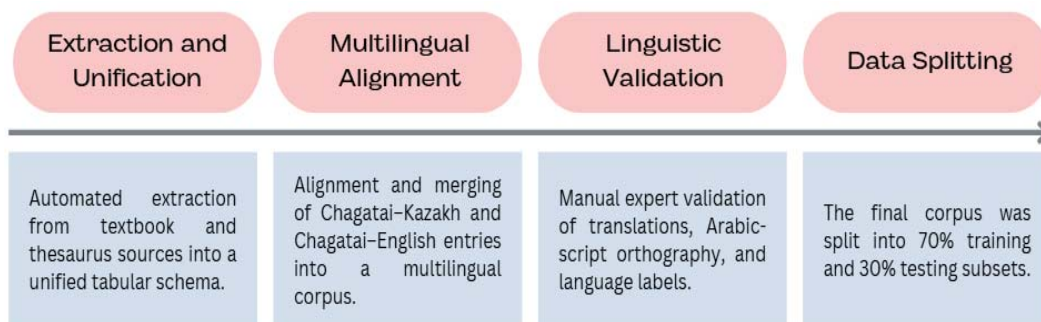


Figure 2. Technical workflow of dataset preparation

The data curation pipeline consisted of the following steps:

1. **Extraction and Structuring:** Raw data from the textbook and thesaurus sources were extracted using custom scripts and normalized into a unified schema with four attributes: `src_text` (original Chagatai), `tgt_text` (translation or transliteration), `src_lang` (Chagatai), and `tgt_lang` (English, Kazakh, or Chagatai transliteration). This ensured consistent structure across different sources.
2. **Multilingual Alignment:** Chagatai–Kazakh thesaurus pairs were aligned with Chagatai–English textbook entries using rule-based lexical and phrasal alignment. This enabled cross-lingual transfer learning through the genealogical similarity of Chagatai and Kazakh, as well as semantic bridging via English.
3. **Linguistic Validation:** Due to the scarcity of domain experts with specialized proficiency in the extinct Chagatai language, a full manual re-verification of the historical Chagatai texts was unfeasible. Consequently, the source Chagatai entries and their baseline translations were directly sourced from authoritative academic publications [24-25]. Manual linguistic validation was explicitly focused on the target language alignments. Specifically, three independent linguists reviewed the translation pairs to ensure semantic fidelity and grammatical consistency across the corpus. Disagreements among the experts were resolved through collaborative consensus rather than formal inter-annotator agreement (IAA) metrics, as individual annotation logs were not recorded separately. This process ensured a unified, high-quality translation standard prior to model training.
4. **Train-Test Partitioning:** The dataset was split into training (70%) and testing (30%) sets using a stratified approach based on the target language. This ensured that the test set included a balanced representation of both the closely related language (Kazakh) and the more distant language (English), avoiding biased evaluation results. Dataset statistics are provided in Table 1.

Table 1. Dataset Distribution

Split	Chagatai-Kazakh Entries	Chagatai-English Entries	Total Entries
Train	3742	3756	7498
Test	1610	1604	3214
Total	5352	5360	10712

Given the orthographic variability inherent in historical manuscripts, a rigorous preprocessing pipeline was implemented to ensure input consistency.

1. **Unicode Normalization:** All text entries were normalized using Normalization Form Compatibility Composition (NFKC) [26] to resolve encoding inconsistencies commonly found in digitized historical scripts.
2. **Script Standardization:** A rule-based standardization protocol was applied to the Arabic-script Chagatai inputs. Variants of specific characters, such as “kaf” (from [ك] to [ڪ]), “ha” (from [ه] to [ه]), “nun” (from [ن] to [ن]), “fa” (from [ف] to [ف]), “qaf” (from [ق] to [ق]), “ta” (from [ت] and [ت] to [ت]), and “yeh” (from [ي] to [ي]), were algorithmically mapped to a unified Persian-style representation. This step is methodologically critical for low-resource agglutinative languages, as it reduces vocabulary sparsity by preventing the tokenizer from treating orthographic variants of the same lexeme as distinct entities [27].

3. Model Architectures and fine-tuning

In this study, two translation models were evaluated: (a) the encoder-decoder NLLB-200 model and (b) the decoder-only TranslateGemma model. The general architectural differences between these models are illustrated in Fig. 3.

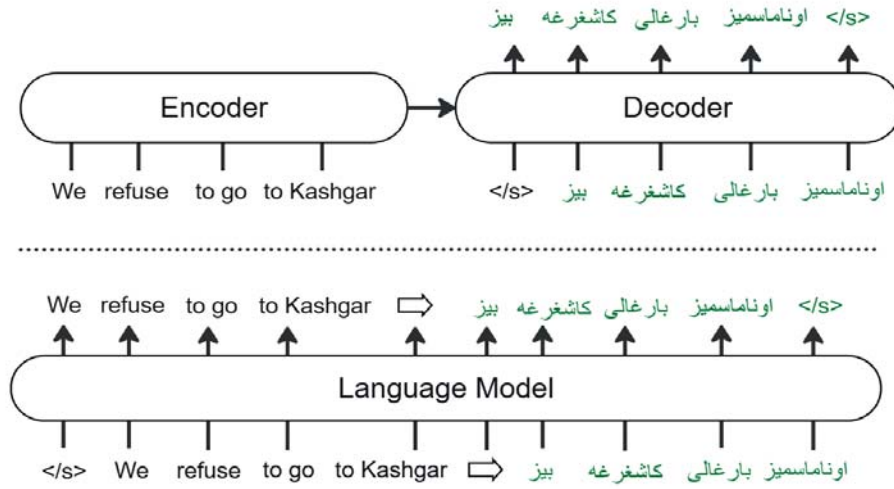


Figure 3. Encoder-Decoder vs. Decoder-Only Architectures for Translation

Machine translation across both evaluated architectures is modeled as a conditional sequence generation problem. Given a source Chagatai text represented as a token sequence $X = (x_1, x_2, \dots, x_N)$, the model computes the conditional probability of predicting the target translation sequence $Y = (y_1, y_2, \dots, y_M)$. Using the chain rule of probability, this autoregressive process is factorized as:

$$P(Y|X; \theta) = \prod_{t=1}^M P(y_t | y_{<t}, X; \theta) \quad (1)$$

where $y_{<t} = (y_1, \dots, y_{t-1})$ represents the sequence of previously generated target tokens, and θ denotes the trainable model parameters. At each decoding step t , the conditional probability distribution over the target vocabulary V is computed by applying the Softmax function to the decoder's output logit vector z_t :

$$P(y_t | y_{<t}, X; \theta) = \text{Softmax}(z_t) = \frac{\exp(z_{t,y_t})}{\sum_{v \in V} \exp(z_{t,v})} \quad (2)$$

During fine-tuning, model parameters θ are optimized by minimizing the standard Negative Log-Likelihood (Cross-Entropy Loss) over the parallel training dataset $\mathcal{D}$:

$$\mathcal{L}(\theta) = - \sum_{(X,Y) \in \mathcal{D}} \sum_{t=1}^M \log P(y_t | y_{<t}, X; \theta) \quad (3)$$

(a) *NLLB-200 (Distilled 600M)*

The distilled version of the NLLB-200 model (nllb-200-distilled-600M) was employed. While the original NLLB models use a Mixture-of-Experts (MoE) architecture, this smaller variant is a dense Transformer encoder-decoder. Training was performed via knowledge distillation,

where the model learns to replicate the outputs of a larger MoE teacher model. This allows the 600M-parameter model to retain high translation quality and cross-lingual transfer capabilities from high-resource Turkic languages, such as Uzbek and Turkish, while remaining computationally efficient for fine-tuning [28].

To adapt the model to the low-resource target, the tokenizer was expanded to include a `chg_Arab` (Chagatai in Arabic script) language token, enabling the encoder to learn embeddings specific to Chagatai while leveraging pre-trained weights through lexical overlap. Training was performed using the Seq2SeqTrainer framework with AdamW optimization, a learning rate of 2×10^{-5} , weight decay of 0.01, and a per-device batch size of 16. Mixed precision (FP16) was employed to optimize GPU memory usage and training speed. Early stopping with the patience of two epochs was applied, monitoring the BLEU score on the Kazakh validation subset to prevent overfitting and ensure selection of the optimal checkpoint.

(b) *TranslateGemma-4b-t*

The `google/translategemma-4b-t` model, part of the Gemma 3 family (Google DeepMind), was employed to evaluate the performance of generative Large Language Models (LLMs) on historical dialects. This decoder-only Transformer architecture predicts target translations autoregressively, without constructing a separate encoded representation of the source sentence. The 4b-t variant has been instruction-tuned on a large number of translation pairs. Its 4-billion-parameter size enables extensive storage of world knowledge, potentially improving the resolution of historical ambiguities that require context beyond a single sentence [29].

Fine-tuning the 4-billion-parameter model was performed using Parameter-Efficient Fine-Tuning (PEFT) to reduce computational costs. Standard LoRA was insufficient for capturing the morphological complexity of the Chagatai script; therefore, DoRA was applied, which decomposes weight updates into magnitude and direction components to improve training stability and achieve convergence closer to full fine-tuning. Training included the input embedding layers (`embed_tokens`) and the language modeling head (`lm_head`) to align Chagatai orthography with the model’s internal representations, as this language presents a distribution potentially unseen by the base model.

Specifically, DoRA decouples the pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ into a magnitude vector $m \in \mathbb{R}^{1 \times k}$ and a directional matrix $V \in \mathbb{R}^{d \times k}$:

$$W_0 = m \odot \frac{V}{\|V\|_c} \quad (4)$$

where $\|V\|_c$ denotes the vector-wise norm of V across columns, and $\odot$ denotes column-wise multiplication. During fine-tuning, the directional matrix is updated via a scaled low-rank decomposition $\Delta V = \frac{\alpha}{r} BA$, while m is optimized directly:

$$W' = m \odot \frac{V + \frac{\alpha}{r} BA}{\left\| V + \frac{\alpha}{r} BA \right\|_c} \quad (5)$$

where $A \in \mathbb{R}^{r \times k}$ is initialized with Kaiming uniform distribution, $B \in \mathbb{R}^{d \times r}$ is initialized to zero ($B = 0$), and $r \ll \min(d, k)$. The architecture and structural weight decomposition of the DoRA mechanism applied to the model are illustrated in Figure 4.

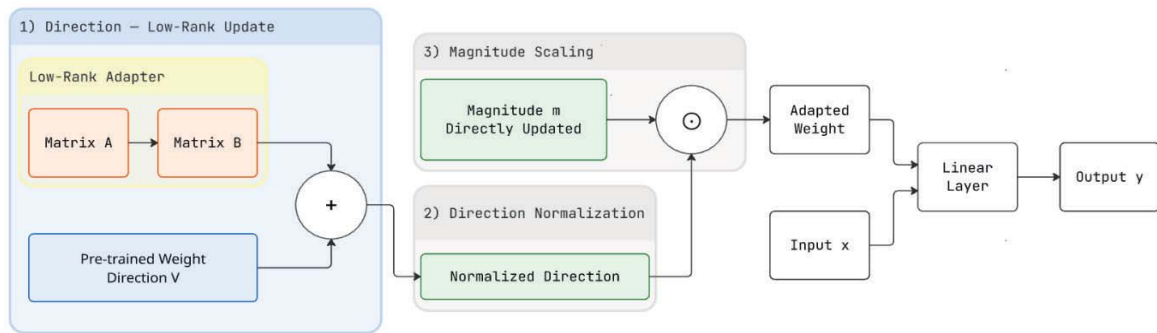


Figure 4. Schematic illustration of Weight-Decomposed Low-Rank Adaptation (DoRA), showing the decomposition of pre-trained weights into magnitude (m) and directional (V) components, with directional updates modeled via low-rank matrices A and B .

The DoRA configuration was defined with a rank of $r = 32$, scaling parameter $\alpha = 64$ (resulting in a constant scaling factor $\frac{\alpha}{r} = 2.0$), and a dropout rate of $p = 0.1$ applied to the adapter inputs to prevent overfitting on the low-resource dataset. Adaptation targeted the query, key, value, and output projections (q_proj , k_proj , v_proj , o_proj) as well as the gate, up, and down projections ($gate_proj$, up_proj , $down_proj$) of the Feed-Forward Networks, while the vision tower and multi-modal projectors remained frozen. Upon completion of training, the decomposed adapter weights were merged directly into the base weights W_0 , incurring zero additional latency during inference.

Training was performed for 10 epochs with a learning rate of 2×10^{-4} and a weight decay of 0.1 to regularize the model given the limited dataset size. Flash Attention 2 and BFloat16 precision were employed to maximize computational throughput. Input data followed the chat-template structure required by Gemma 3 (Fig. 5), with custom insertion of empty `token_type_ids` to satisfy multimodal architecture constraints. In this setup, the Chagatai inputs were processed using the existing Uzbek language token, leveraging typological proximity rather than a dedicated Chagatai token.

```
messages = [
  {
    "role": "user",
    "content": [
      {
        "type": "text",
        "source_lang_code": "uz",
        "target_lang_code": "en",
        "text": "ایدی باز اریغه تاش فرقه سی بر موندین",
      }
    ],
  }
]
```

Figure 5. Input structure for TranslateGemma-4b-t showing Chagatai-to-English translation with the Uzbek language code

To quantify the contribution of supervised adaptation, both models were additionally evaluated in a zero-shot configuration on the same test set and with the same evaluation pipeline. Under this condition, NLLB-200 received `chg_Arab` token initialization from `uzn_Latn` (Uzbek in Latin script), consistent with the proxy used during fine-tuning, but no parameter updates on

Chagatai data. TranslateGemma was prompted using the instruction format described above without any adapter weights loaded. The resulting zero-shot scores serve as a reference lower bound for measuring the improvement attributable to fine-tuning on the Chagatai corpus.

4. Evaluation Metrics

To compare these two translation architectures on equal terms, model performance was evaluated using a multi-metric suite designed to capture different aspects of translation quality, including BLEU, chrF++, Exact Match (EM), and XCOMET. All four metrics are reported on a common 0 to 100 scale, where higher values correspond to closer alignment with reference translations and stronger semantic fidelity.

Bilingual Evaluation Understudy (BLEU): Measures n-gram overlap and surface-level fluency [30]. Although limited in capturing deep semantic equivalence, BLEU provides a foundational baseline, ensuring backward compatibility with historical machine translation studies and offering a rigorous measure of literal adherence to reference translations [31].

Character n-gram F-score plus words (chrF++): Designed to address morphologically rich and agglutinative languages, chrF++ combines character-level n-grams with word-level representations, mitigating the penalization of valid suffix variations [32]. By integrating word-level n-grams with character representations, chrF++ successfully balances morphological flexibility with structural syntax constraints, producing higher correlations with human evaluations than BLEU [32].

Exact Match (EM): Evaluates precise retrieval of fixed historical terminology [33]. By calculating strict string-level equivalence between generated outputs and references, EM ensures accurate translation of proper nouns, toponyms, and historically defined glossaries, providing a stringent measure of factual integrity [34].

XCOMET (Unbabel/XCOMET-XL): Provides a semantic evaluation based on neural embeddings, offering a proxy for human judgment of meaning preservation beyond lexical overlap [35]. Built on XLM-RoBERTa XL, XCOMET maps source text, translation, and reference into a shared embedding space, enabling comparison of semantic meaning across languages, including unseen or low-resource languages [35-36]. The Unbabel/XCOMET-XL variant ($\approx 3.5B$ parameters) was selected for this study due to its balance of high explainability, strong zero-shot cross-lingual transfer, and computational efficiency [35]. By predicting both quality scores and error spans, it provides interpretable evaluations of semantic fidelity, capturing deviations even when the target language was not observed during training [35].

In addition to the aggregate quality score, the error span annotations produced by XCOMET-XL enable a finer-grained analysis of model behavior. Four derived indicators are reported in the error analysis. Major errors per sentence denote the average number of critical error spans per output sentence, where a critical span is a contiguous segment, whose mistranslation substantially alters source meaning [35]. The omission rate indicates the proportion of sentences in which semantically significant source content is absent from the translation without a corresponding target segment. The hallucination rate captures the proportion of sentences in which the model generates content that has no correspondence in the source text. The BLEU/XCOMET correlation coefficient measures the degree of agreement between surface-level n-gram overlap scores and neural semantic quality scores at the sentence level, indicating how reliably lexical accuracy approaches semantic accuracy for a given language pair.

Results

This section presents the experimental findings from the evaluation of encoder-decoder NLLB-200 model and decoder-only TranslateGemma model on the Chagatai test corpus. The results are organized in two subsections: 1) quantitative performance, comparing both architectures across English and Kazakh translation directions using standard MT metrics, and 2) qualitative error analysis, examining representative outputs to characterize the failure patterns specific to each architecture.

1. Quantitative Performance

The evaluation reveals notable differences in performance across two experimental conditions and two target languages. Table 2 reports results for both the zero-shot configuration, reflecting model behavior prior to any Chagatai-specific adaptation, and the fine-tuned configuration, following supervised training on the parallel corpus described in the Methods section. This comparison directly quantifies the contribution of fine-tuning for NLLB-200 (Distilled-600M) and TranslateGemma (4B) on the Chagatai test set.

Table 2. Zero-shot baseline and fine-tuned performance of NLLB-200 and TranslateGemma on the Chagatai test set. Best fine-tuned scores per language pair are highlighted in bold. XCOMET was not computed for zero-shot configurations.

Target Language	Model	BLEU (↑)	chrF++ (↑)	Exact Match	XCOMET (↑)
English	NLLB-200 (zero-shot)	0.50	9.41	0.43	—
	NLLB-200 (fine-tuned)	16.89	31.37	17.64	74.97
	TranslateGemma (zero-shot)	1.48	12.92	0.93	—
	TranslateGemma (fine-tuned)	17.48	30.48	13.54	73.8
Kazakh	NLLB-200 (zero-shot)	0.16	8.61	1.06	—
	NLLB-200 (fine-tuned)	18.89	38.72	25.44	74.31
	TranslateGemma (zero-shot)	0.65	10.28	0.81	—
	TranslateGemma (fine-tuned)	17.83	36.96	23.44	69.75

In the zero-shot configuration, both models produced near-zero scores on the English direction (NLLB-200: BLEU 0.50, chrF++ 9.41; TranslateGemma: BLEU 1.48, chrF++ 12.92), confirming that the Chagatai script and vocabulary are outside the effective coverage of both models without task-specific adaptation.

In the English translation task, TranslateGemma achieved a marginally higher BLEU score (17.48) compared to NLLB-200 (16.89). However, NLLB-200 outperformed the LLM-based model in all other metrics. NLLB achieved a higher Exact Match rate (17.64% vs. 13.54%) and a higher XCOMET score (74.97 vs. 73.80), indicating that while Gemma produces n-gram overlaps, NLLB generates translations that are semantically closer to the reference.

The Kazakh direction showed an even more pronounced gap: zero-shot BLEU scores were 0.16 for NLLB-200 and 0.65 for TranslateGemma, representing the lowest baseline performance across all conditions.

For the Kazakh target subset, NLLB-200 demonstrated superior performance across all evaluated metrics. It achieved a BLEU score of 18.89 and a significantly higher chrF++ score of 38.72 (compared to Gemma's 36.96). The semantic evaluation via XCOMET favoured NLLB (74.31) over Gemma (69.75) by a margin of 4.56 points. Additionally, NLLB achieved the highest Exact Match rate in the study (25.44%), confirming its ability to correctly reconstruct full sentences in the related Turkic language.

To examine translation errors beyond surface-level metrics, a detailed analysis was conducted using XCOMET-XL. Table 3 presents the error distribution, highlighting Major Errors per sentence, omission rates, hallucination rates, and the relationship between BLEU and XCOMET scores.

Table 3. Deep Error Analysis of NLLB-200 and TranslateGemma. (Lower values are better for Errors, Omissions, and Hallucinations; Higher values are better for Correlation).

Target Language	Model	Major Errs/Sent (↓)	Omissions (%) (↓)	Hallucinations (%) (↓)	Corr (BLEU/XCOMET) (↑)
English	NLLB-200	0.43	31.2	15.7	0.469
	Translate Gemma	0.41	26.8	23.5	0.414
Kazakh	NLLB-200	0.41	24.4	11.7	0.476
	Translate Gemma	0.40	16.8	25.1	0.500

2. Qualitative Error Analysis and Architectural Artifacts

Reflecting fundamental differences in model architecture, a manual inspection of the generated outputs, supported by XCOMET-XL error logs, reveals a clear trade-off between hallucinations and omissions, highlighting distinct failure patterns for each model: (a) NLLB-200 (Distilled 600M) and (b) TranslateGemma-4b-t.

(a) NLLB-200 (Distilled 600M)

NLLB-200 errors were more conservative and primarily related to phonetic or orthographic confusion rather than generative hallucinations. Observed failure modes include:

- **Phonetic and Embedding Confusion:** The complex Chagatai botanical enumeration “سرو شريان سرو ازاد سرو سه سرو عرعر” (sumac, cypress, elm, and beech trees”) was entirely mistranslated as “staff, a large army, and a few free men.” This error is likely due to the proximity of the Chagatai word “serv” (cypress) to phonetically similar Turkic/Persian words like “sarbaz” (soldier) or “sarvar” (leader/staff) within the model's shared multilingual embedding space.
- **Orthographic/Semantic Shift (Kazakh Target):** In a historical account regarding a conflict, the Chagatai verb “ات يلدی” [atyldy] (was shot) was confused with the Kazakh verb “айтылды” [aytyldy] (was said). This single-letter orthographic and phonetic confusion significantly altered the semantic meaning of the entire passage.
- **Decoding Repetition Loops:** In low-confidence scenarios, particularly with short or ambiguous inputs, NLLB outputs suffered from repetition loops (e.g., generating “ma? ma? ma?” continuously). This is a well-documented artifact of beam-search decoding in Seq2Seq architectures, which was notably absent in TranslateGemma's outputs.

(b) TranslateGemma-4b-t

TranslateGemma exhibited a pronounced tendency to hallucinate, with error rates of 23.5% for English and 25.1% for Kazakh, while omission rates remained comparatively lower (16.8%–26.8%). Its generative nature often produced fluent but factually inaccurate content. Key failure modes include:

- **Severe Semantic Hallucination (English Target):** Numerical phrases such as “مېڭ بېر” (one thousand five hundred households), Gemma generated the disturbing and entirely unrelated phrase “I will make 1,005 people into skeletons.” This indicates a severe decoupling from the source text, where the model prioritizes hallucinated narrative generation over translation.
- **Encyclopedic Intrusion:** For the singular noun “حوض” (reservoir/pool) triggered extended, explanatory outputs “saba, a vessel made of horsehide...” rather than direct lexical equivalents, reflecting a reliance on text completion learned during pre-training.
- **Contextual Noise (Kazakh Target):** Gemma introduced the out-of-context term “орыстарында” (among Russians) in a sentence where the source text referred explicitly only to “مسلمان كاف ردا” (Muslims and Infidels). This demonstrates the injection of

semantic noise based on latent historical or textual associations in the LLM's weights rather than the input prompt.

Discussion

The experimental results demonstrate that multilingual encoder–decoder architectures remain more reliable than generative large language models when applied to extremely low-resource historical languages. In the Kazakh translation task, the NLLB-200 model achieved higher scores across all evaluation metrics, including BLEU, chrF++, Exact Match, and XCOMET. This indicates that the encoder–decoder architecture more effectively preserves both the morphological structure and semantic fidelity of Chagatai sentences when translating into a closely related Turkic language. The higher chrF++ score obtained by NLLB indicates that the model is able to reproduce the complex suffix structures typical of Turkic languages with greater consistency. This result supports the hypothesis that multilingual models benefit from typological proximity between languages when transferring grammatical knowledge from high-resource languages such as Turkish or Uzbek to historically related low-resource languages.

Although TranslateGemma achieved a slightly higher BLEU score for the English translation task, the broader evaluation metrics reveal limitations of the decoder-only generative architecture in this context. BLEU mainly reflects surface n-gram overlap and may favor fluent but partially inaccurate translations. In contrast, the lower XCOMET scores and reduced Exact Match rates indicate that TranslateGemma introduces semantic deviations more frequently than the NLLB model. The qualitative error analysis further confirms that the generative model tends to produce hallucinated or encyclopedic completions that are not grounded in the source text. This is consistent with recent work on low-resource LLM-based translation, which shows that decoder-only systems remain vulnerable to rare words and underrepresented languages even when external linguistic resources are available. In that literature, dictionaries, grammar books, and parallel examples can help, but they do not remove the basic instability of LLM-based translation under severe resource scarcity. The present TranslateGemma errors such as semantic intrusions, encyclopedic completions, and numerically distorted outputs fit that failure mode closely. The problem is therefore not simply that the model is “worse”; it is that it is fluent before it is faithful [37].

The comparison with other historical-language studies sharpens this point. Recent work on 16th-century Latin and Early New High German reports substantially stronger outcomes for LLM-based historical translation [38], including cases where LLM-based translation outperforms previous approaches and cases where expert-written summaries improve translation quality by more than two BLEU/chrF points. The present Chagatai results are clearly less favorable to the LLM. The most plausible interpretation is not that historical languages in general are unsuitable for decoder-only models, but that Chagatai occupies a harsher point on the resource spectrum. Latin benefits from a denser editorial tradition, richer bilingual scaffolding, and more prompt-ready scholarly support. Chagatai does not yet have that ecosystem at comparable scale. Under those conditions, the latent prior of a generative model is too weak to prevent semantic guessing, whereas a translation-specialized encoder–decoder model remains more tightly anchored to the source text [38].

The corpus design also deserves a more precise interpretation. The use of both sentence-level material and lexical resources is not merely an ad hoc workaround; recent low-resource MT research supports this strategy. Zheng et al. [39] show that bilingual lexical resources and lexicon-derived pseudo-parallel data produce consistent BLEU and CHRF gains in endangered-language MT, while Pei et al. [37] report that for Manchu, dictionary entries, suffix explanations, collocations, and retrieved grammar excerpts substantially improve translation quality over direct prompting. This makes the present Chagatai corpus construction methodologically defensible. However, the same literature also clarifies its limit: lexical resources improve terminology and morpheme-level grounding, but they do not substitute for sentence-level evidence of discourse structure, stylistic regularity, or long-range dependencies. For that reason, the current dataset is

likely stronger for lexical fidelity than for full sentence adequacy, and that asymmetry should be kept in mind when interpreting Exact Match and terminology-sensitive outputs.

The entity and terminology issue is especially important in this task and should not be treated as a side problem. Historical prose is saturated with toponyms, titles, ethnonyms, dynastic names, and culturally specific references whose correct rendering depends on historical knowledge, not just language fluency. The first Entity-Aware Machine Translation shared task was organized precisely because modern MT systems, including LLM-based ones, still struggle with proper nouns and culture-specific entities, and because standard global metrics do not adequately capture entity correctness. This is directly relevant to the observed Chagatai errors. In practical terms, the literature points to an obvious next step: if a decoder-only model is to remain competitive for Chagatai, it will likely need entity-aware support through retrieval, lexica, gazetteers, or knowledge base grounding rather than stronger generic prompting alone [40].

Finally, the hallucination profile points to a specific technical agenda for future work. Hallucination in MT is no longer treated as an incidental artifact; recent studies target it directly at both decoding time and training time. Waldendorf et al. [41] show that contrastive decoding reduces hallucinations in large multilingual MT models, while Tang et al. [42] show that hallucination-focused preference optimization can reduce hallucination rates in LLM-based translation dramatically while preserving overall translation quality. In the present study, TranslateGemma exhibits lower omission but substantially higher hallucination rates than NLLB-200. Under these conditions, anti-hallucination methods are not optional refinements; they are the obvious minimum requirement for any future decoder-only Chagatai system, because for historical translation fabricated content is more damaging than conservative under-translation.

Conclusion

The experimental results demonstrate that multilingual encoder-decoder architectures remain more reliable than generative large language models when applied to extremely low-resource historical languages. In the Kazakh translation task, the NLLB-200 model achieved higher scores across all evaluation metrics, including BLEU, chrF++, Exact Match, and XCOMET. This indicates that the encoder-decoder architecture more effectively preserves both the morphological structure and semantic fidelity of Chagatai sentences when translating into a closely related Turkic language. Therefore, for Chagatai, the bottleneck is not surface fluency but source grounding. The model that remains closer to the source and transfers morphology through multilingual conditioning is currently more useful than the model that writes more smoothly but guesses more often.

Acknowledgments

This research has been/was/is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR28712621). The authors express their sincere gratitude to linguists Almas Imangaliyev, Kutlugjon Sultanbek, and Dameli Bilalova for their expert contribution to the quality assessment and manual review of the Chagatai parallel texts across Kazakh and English.

References

- [1] Abylkhozhin, Zh.B., Akulov, M.L., & Tsai, A.V. (Eds.). (2019). *Zhivaya pamyat'. Stalinizm v Kazakhstane: Proshloe, Pamyat', Preodolenie*. Daik-Press. ISBN: 978-601-290-110-8.
- [2] Kairanbayeva, N.N., & Shadkam, Z. (2019). Shag'ataj tilindegi sozdikterdi zertteulerge sholu. *Journal of Oriental Studies*, 91(4), 61-67. <https://doi.org/10.26577/JOS-2019-4-o6>
- [3] Tangsykbay, A. T. (2024). Qazaq tildi auditorijag'a shag'ataj tili leksikasyn oqytu zholdary («Babynama» mysalynda). *Bulletin of LN Gumilyov Eurasian National University. PHILOLOGY Series*, 149(4), 261-271. <https://doi.org/10.32523/2616-678X-2024-149-4-261-271>

- [4] Dieng, A. B., Wang, C., Gao, J., & Paisley, J. (2016). Topicrnn: A recurrent neural network with long-range semantic dependency. arXiv preprint arXiv:1611.01702. <https://doi.org/10.48550/arXiv.1611.01702>
- [5] Mirzakhlov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., ... & Chellappan, S. (2021, November). A large-scale study of machine translation in Turkic languages. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5876-5890). <https://doi.org/10.18653/v1/2021.emnlp-main.475>
- [6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762>
- [7] Kornfilt, J. (2018). Turkish and the Turkic languages. In *The world's major languages* (pp. 536-561). Routledge. <https://www.taylorfrancis.com/chapters/edit/10.4324/9780203216149-12/turkish-turkic-languages-jaklin-kornfilt>
- [8] Kocmi, T., & Bojar, O. (2018, October). Trivial transfer learning for low-resource neural machine translation. *In Proceedings of the third conference on machine translation: research papers* (pp. 244-252). <https://doi.org/10.18653/v1/W18-6325>
- [9] Aliyu, Y., Sarlan, A., Danyaro, K. U., Rahman, A. S. B., & Abdullahi, M. (2024). Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources. *IEEE Access*, 12, 66883-66909. <https://doi.org/10.1109/ACCESS.2024.3398635>
- [10] Anfaresi, M. S. T. (2023). Developing multi translation chat application using Django frameworks and M2M100 model (Doctoral dissertation, Universitas Islam Indonesia).
- [11] Lee, E. S. A., Thillainathan, S., Nayak, S., Ranathunga, S., Adelani, D. I., Su, R., & McCarthy, A. D. (2022, May). Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?. *In Findings of the Association for Computational Linguistics: ACL 2022* (pp. 58-67). <https://doi.org/10.18653/v1/2022.findings-acl.6>
- [12] Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672. <https://doi.org/10.48550/arXiv.2207.04672>
- [13] Liang, X., Khaw, Y. M. J., Liew, S. Y., Tan, T. P., & Qin, D. (2025). Towards low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2025.3549795>
- [14] Jiao, W., Wang, W., Huang, J. T., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745. <https://doi.org/10.48550/arXiv.2301.08745>
- [15] Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2023, July). Prompting palm for translation: Assessing strategies and performance. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15406-15427). <https://doi.org/10.18653/v1/2023.acl-long.859>
- [16] Nguyen, H. N., Do, D., Nguyen, B., Nguyen, L., Nguyen, D., & Tran, V. A. (2024, December). Advanced Language Model-Based Translator for English-Vietnamese Translation. In 2024 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 398-402). *IEEE*. <https://doi.org/10.1109/RIVF64335.2024.11009075>
- [17] Chen, S., Shi, X., Li, P., Li, Y., & Liu, J. (2024). Refining translations with llms: A constraint-aware iterative prompting approach. arXiv preprint arXiv:2411.08348. <https://doi.org/10.48550/arXiv.2411.08348>
- [18] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210. <https://doi.org/10.48550/arXiv.2302.09210>
- [19] Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Johnson, M., & Firat, O. (2023, July). The unreasonable effectiveness of few-shot learning for machine translation. *In*

International Conference on Machine Learning (pp. 10867-10878). PMLR. <https://proceedings.mlr.press/v202/garcia23a.html>

[20] Finkelstein, M., Caswell, I., Domhan, T., Peter, J. T., Juraska, J., Riley, P., ... & Vilar, D. (2026). TranslateGemma Technical Report. arXiv preprint arXiv:2601.09012. <https://doi.org/10.48550/arXiv.2601.09012>

[21] Buscaldi, D., & Rosso, P. (2023, November). How Good is NLLB for Low-resource Languages? A Study on the Genoese Language. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)* (pp. 490-493). <https://aclanthology.org/2023.clicit-1.61/>

[22] Mesham, S., Hayward, L., Shapiro, J., & Buys, J. (2021). Low-resource language modelling of south african languages. arXiv preprint arXiv:2104.00772. <https://doi.org/10.48550/arXiv.2104.00772>

[23] Ko, W., El-Kishky, A., Renduchintala, A., Chaudhary, V., Goyal, N., Guzmán, F., Fung, P., Koehn, P., & Diab, M. (2021). Adapting High-resource NMT Models to Translate Low-resource Related Languages without Parallel Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 1, 802–812. <https://doi.org/10.18653/v1/2021.acl-long.66>

[24] Schluessel, E. (2018). An introduction to Chaghatay: A graded textbook for reading Central Asian sources. Michigan Publishing. USA. <https://doi.org/10.3998/MPUB.10110094>

[25] Mamyrbekova, G., & Seitbekova, A. (2013). Abilgazy bakhadur khannyn «Turki shezhiresinin» tezaurus sozdigi [Thesaurus dictionary of the “Turkic chronicle” by Abylgazy Bahadur Khan]. Almaty:«Firma Ornak»[in Kazakh].

[26] Moran, S., & Cysouw, M. (2018). The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles (Vol. 10). *Language Science Press*. <https://doi.org/10.5281/zenodo.1296780>

[27] Asoka Chakravarthi, B. R. (2020). Leveraging orthographic information to improve machine translation of under-resourced languages (Doctoral dissertation, NUI Galway). <https://doi.org/10.13025/17483>

[28] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1-39. <https://doi.org/10.48550/arXiv.2101.03961>

[29] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403. <https://doi.org/10.48550/arXiv.2305.10403>

[30] Post, M. (2018, October). A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers* (pp. 186-191). <https://doi.org/10.18653/v1/W18-6319>

[31] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). <https://doi.org/10.3115/1073083.1073135>

[32] Popović, M. (2017, September). chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation* (pp. 612-618). <https://doi.org/10.18653/v1/W17-4770>

[33] Mozer, R., Miratrix, L., Kaufman, A. R., & Anastasopoulos, L. J. (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4), 445-468. <https://doi.org/10.48550/arXiv.1801.00644>

[34] Alam, M. M. I., Kvapilíková, I., Anastasopoulos, A., Besacier, L., Dinu, G., Federico, M., ... & Nikoulina, V. (2021, November). Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation* (pp. 652-663). <https://aclanthology.org/2021.wmt-1.69/>

- [35] Guerreiro, N. M., Rei, R., Stigt, D. V., Coheur, L., Colombo, P., & Martins, A. F. (2024). xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12, 979-995. https://doi.org/10.1162/tacl_a_00683
- [36] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, November). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2685-2702). <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- [37] Pei, R., Liu, Y., Lin, P., Yvon, F., & Schütze, H. (2025, July). Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 8767-8788). <https://doi.org/10.18653/v1/2025.acl-long.429>
- [38] Volk, M., Fischer, D. P., Fischer, L., Scheurer, P., & Ströbel, P. B. (2024, May). LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024* (pp. 122-128). <https://aclanthology.org/2024.lt4hala-1.15/>
- [39] Zheng, F., Marrese-Taylor, E., & Matsuo, Y. (2024, August). Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11248-11259). <https://doi.org/10.18653/v1/2024.findings-acl.670>
- [40] Conia, S., Li, M., Navigli, R., & Potdar, S. (2025, July). SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* (pp. 2535-2557). <https://aclanthology.org/2025.semeval-1.326/>
- [41] Waldendorf, J., Haddow, B., & Birch, A. (2024, March). Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 2526-2539). <https://doi.org/10.18653/v1/2024.eacl-long.155>
- [42] Tang, Z., Chatterjee, R., & Garg, S. (2025, April). Mitigating hallucinated translations in large language models with hallucination-focused preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers) (pp. 3410-3433). <https://doi.org/10.18653/v1/2025.naacl-long.175>