









Article

Enhancing Post-Editing of Kazakh Translations Using Fine-Tuned Large Language Models

Akbayan Bekarystankyzy ^{1,2}, Diana Rakhimova ^{3,*}, Aliya Zhiger ^{2,3,4,*}, Assel Sakatay ⁵,
Nazym Zhumakhan ⁵, Aigerim Yerimbetova ^{1,6,*}, Dina Oralbekova ¹ and Mussa Turdalyuly ^{1,6}

- ¹ Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies, Almaty 050000, Kazakhstan; akbayan.bekaristankyzy@narxoz.kz (A.B.); dinaoral@mail.ru (D.O.); m.turdalyuly@gmail.com (M.T.)
- ² School of Digital Technologies, Narxoz University, Almaty 050035, Kazakhstan
- ³ Faculty of Information Technology and Artificial Intelligence, Farabi University, Almaty 050040, Kazakhstan
- ⁴ Faculty of Computer Technologies and Cybersecurity, International Information Technology University, Almaty 050040, Kazakhstan
- ⁵ Department of IT Engineering and Artificial Intelligence, Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty 050013, Kazakhstan; asel.sakatay@mail.ru (A.S.); nazym.sembekkyzy@gmail.com (N.Z.)
- ⁶ School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
- * Correspondence: di.diva@mail.ru (D.R.); aliya.zhunussova.zh@narxoz.kz (A.Z.); aigerian8888@gmail.com (A.Y.); Tel.: +7-777-242-0775 (D.R.); +7-708-949-4020 (A.Z.); +7-701-583-3302 (A.Y.)

Abstract

Machine translation for low-resource languages such as Kazakh remains a complex task due to the scarcity of training data, intricate morphological structures, and culturally specific linguistic characteristics. This study presents the first extensive exploration of fine-tuning large language models for automated post-editing of Kazakh translations. We introduce KazPE, a carefully curated and annotated dataset that includes 10,008 training sentences and 311 test sentences spanning six domains: the medical, scientific, journalistic, oral, fiction, and legal. The dataset features detailed error classifications across 9 linguistic categories. Our method fine-tunes GPT-4.1 mini using supervised learning to enhance translation quality by systematically correcting targeted errors. According to human evaluations, conducted on a continuous 0–1 scale, the fine-tuned model achieves an average quality score of 0.84, surpassing the baseline score of 0.80, corresponding to a 5% relative improvement. The greatest improvements are observed in handling morphological and lexical errors, as well as in domain-specific texts—particularly in legal (+17%) and medical (+22%) domains. In addition, the translations were evaluated using the automatic metrics: BLEU, TER and METEOR. The fine-tuned model shows improvements across all automatic metrics (BLEU, TER, METEOR), which confirms better n-gram overlap with reference texts, fewer edits needed, and enhanced lexical and semantic alignment with the reference texts. Comprehensive error analysis shows that the fine-tuning process effectively mitigates challenges related to Kazakh’s agglutinative morphology and specialized terminology, while preserving accuracy on already correct sentences. This research establishes the first structured evaluation framework for Kazakh translation post-editing and offers valuable guidance for enhancing machine translation in morphologically rich, low-resource languages. To facilitate further progress in Turkic language processing, we publicly release the KazPE dataset, trained models, and evaluation framework.

Keywords: Kazakh; low-resource languages; machine translation; post-editing; large language models; fine-tuning; NLP; morphologically rich languages



Academic Editor: Vicente García-Díaz

Received: 12 November 2025

Revised: 10 February 2026

Accepted: 19 February 2026

Published: 6 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Kazakh, a member of the Kipchak branch of the Turkic language family, is spoken by an estimated 13 million people worldwide [1]. Although it holds official status in Kazakhstan, the language remains significantly underrepresented in natural language processing (NLP) research and resources when compared to widely studied languages such as English, Chinese, or Spanish [2]. This lack of linguistic data and computational tools creates major obstacles for developing effective machine translation (MT) systems. The challenges are especially pronounced due to Kazakh's agglutinative morphology, flexible syntactic structure, extensive case marking, and productive derivational patterns [3].

The structural properties of Kazakh make it particularly difficult for automated translation models to process. Being an agglutinative language, Kazakh forms words by attaching multiple suffixes to a base morpheme, resulting in a vast number of possible word forms. A single Kazakh word can often express concepts that require several words in English. For instance, the word оқушылардікі (oqushylardiki) translates to “belonging to the students.” This high degree of morphological complexity, along with relatively free word order and vowel harmony constraints, poses substantial challenges for both statistical and neural MT systems [4].

Recent progress with large language models (LLMs) such as GPT-3 [5], GPT-4 [6], and PaLM [7] has revealed impressive zero-shot and few-shot translation abilities across a wide range of language pairs. These models have proven especially effective for low-resource languages by leveraging multilingual training and cross-lingual transfer learning techniques [8]. Nevertheless, despite these achievements, LLMs continue to face considerable challenges when dealing with translations involving under-resourced languages. Their outputs frequently contain semantic inaccuracies, morphological mistakes, unnatural phrasing, and culturally inappropriate expressions [9]. These issues are particularly evident in morphologically complex languages such as Kazakh, where even minor morphological variations can lead to significant changes in meaning.

Post-editing—the process of refining machine-generated translations through human or automated corrections has emerged as an effective strategy to narrow the quality gap in MT systems [10]. In contrast to methods that aim to enhance the translation model itself, post-editing focuses on analyzing and correcting errors within the produced translations. This makes it especially useful for low-resource languages, where creating robust translation models from the ground up is often infeasible due to limited data availability. Earlier post-editing systems relied predominantly on rule-based or statistical approaches [11]; however, more recent research increasingly adopts neural and transformer-based architectures for automated post-editing [1,12]. Fine-tuning large language models for post-editing represents a promising direction, combining the extensive multilingual knowledge of pre-trained LLMs with task-specific adaptation.

Fine-tuning enables large language models to capture specific patterns of translation errors and corresponding correction strategies while retaining their broad linguistic competence [13]. This capability is especially valuable for languages such as Kazakh, where the intricate interaction between morphology, syntax, and semantics demands a deep linguistic understanding that benefits from both multilingual pre-training and language-specific adaptation.

Existing research in Kazakh natural language processing has largely concentrated on foundational tools, including morphological analyzers [3], text-to-speech systems [14], and early-stage machine translation prototypes [15]. However, systematic studies on translation quality assessment and post-editing remain scarce. One of the main limitations hindering progress in Kazakh NLP is the absence of standardized, richly annotated datasets that capture detailed error types—making it difficult to evaluate and compare different translation approaches effectively.

In this study, we present a comprehensive framework for enhancing Kazakh machine translation through fine-tuned post-editing based on GPT-4.1 mini. Our methodology addresses several key shortcomings in existing research by introducing both new datasets and evaluation protocols designed to systematically improve translation quality. The primary contributions of this work are as follows:

Dataset development. We introduce KazPE, a publicly available, large-scale, and systematically annotated English–Kazakh translation dataset that includes an extensive typology of translation errors. The resource consists of 10,008 training and 311 test sentences, carefully selected from six heterogeneous domains: the medical, scientific, journalistic, oral, fiction, and legal. Each sample in the corpus is tagged across 9 distinct error types, covering semantics, lexical choice, morphology, terminology, stylistic adequacy, word order, grammatical accuracy, orthography, idiomaticity.

Methodological innovation. We propose a supervised fine-tuning approach specifically tailored to improve post-editing performance for Kazakh. This method utilizes a prompt–response interaction framework, which directs the model to locate and revise translation inaccuracies in a structured and interpretable way, ensuring that the original semantic intent of the source text remains intact.

Comprehensive evaluation framework. To ensure robustness, we carry out the first extensive evaluation of large language models in the context of Kazakh translation post-editing. The framework integrates human assessments and an in-depth linguistic error taxonomy, enabling a detailed examination of both overall translation quality and the model’s precision in resolving particular linguistic challenges.

Empirical analysis. Our study offers the first detailed assessment of GPT-4.1 mini on English–Kazakh translation post-editing tasks. The findings reveal clear trends in how the model processes Kazakh morphological patterns, syntactic diversity, and semantic subtleties, shedding light on its strengths, limitations, and potential areas for enhancement in handling low-resource, morphologically rich languages.

Practical impact. Our study establishes baseline performance benchmarks and offers practical insights for the development of effective translation post-editing systems for Kazakh and other Turkic languages.

Our experimental results show that fine-tuning GPT-4.1 mini on the annotated KazPE dataset leads to substantial improvements in translation quality, with the model achieving an 84% accuracy rate compared to 80% for the base model. These gains are observed consistently across various text domains and levels of linguistic complexity, indicating the robustness of the proposed methodology. The most pronounced improvements are seen in managing morphological–lexical interactions, where the fine-tuned model reaches 100% accuracy relative to 70% for the baseline, highlighting the effectiveness of targeted training for Kazakh’s highly agglutinative morphology.

The uniformity of performance gains across different domains, including specialized areas such as legal and medical texts demonstrates both the practical utility and real-world applicability of our approach. These findings offer valuable guidance for researchers working on low-resource, morphologically rich languages, revealing both the potential and the current limitations of fine-tuning strategies in improving machine translation quality under resource-constrained conditions.

2. Related Works

2.1. Machine Translation for Low-Resource Languages

Research on machine translation for languages with limited resources has consistently drawn attention in computational linguistics. These languages, which lack extensive parallel corpora, often challenge standard statistical and neural methods.

Koehn and Knowles [16] noted that neural models typically need far more data than statistical approaches to reach comparable accuracy, making it especially difficult to support under-resourced languages.

Recent developments in multilingual machine translation have opened new possibilities for languages with limited resources, particularly through cross-lingual transfer learning. Johnson et al. [17] showed that Google's multilingual neural MT system can exploit training data from multiple language pairs, allowing low-resource languages to benefit from knowledge learned in high-resource languages. Similarly, architectures like mT5 [18] and multilingual pre-training of mBERT [19] provide foundational cross-lingual capabilities that improve the processing of under-resourced languages. Nevertheless, most research has concentrated on languages with a strong online presence or those closely related to high-resource languages. Morphologically complex languages, especially within the Turkic family, continue to present challenges due to their agglutinative nature, extensive case systems, and intricate morphophonological patterns, which are often poorly handled by conventional subword tokenization techniques in neural models [20].

2.2. Turkic Language Processing and Kazakh NLP

Studies on Turkic language processing have consistently pointed out the difficulties arising from the morphological complexity inherent to this language family. Washington et al. [3], for instance, created finite-state morphological transducers for Kazakh, providing essential tools for computational exploration of its agglutinative characteristics. Their research underscored the intricacies of Kazakh morphology, such as extensive case systems, possessive forms, and productive derivational patterns, all of which pose significant challenges for automated processing.

Building on this work, Altenbek and Wang [4] developed segmentation systems for Kazakh inflectional affixes, tackling a crucial preprocessing step for computational analysis of the language. Makazhanov et al. [14] extended these efforts by producing open-source tools for Kazakh speech synthesis, thereby enhancing the overall ecosystem of Kazakh language technologies. Initial machine translation attempts for Kazakh were primarily rule-based or relied on limited statistical approaches. For example, Yeshpanov et al. [21] introduced KazParC, one of the first large-scale parallel corpora for Kazakh machine translation, addressing the long-standing challenge of scarce parallel data and enabling more effective statistical and neural MT systems. Recent studies on neural machine translation for English–Kazakh have demonstrated improvements over earlier systems, while still showing significant quality gaps compared to high-resource languages [15].

Kartbayev [12] carried out a comprehensive investigation of neural sequence-to-sequence models for agglutinative languages, including Kazakh, highlighting the difficulties involved in modeling complex morphological patterns. This research offered important insights into the limitations of conventional neural architectures when applied to languages with rich morphology.

Despite these advancements, Kazakh NLP still faces a significant challenge: the absence of large-scale, consistently annotated datasets. Available resources are generally small and tailored to specific tasks, which restricts the development of robust evaluation frameworks and hinders systematic comparisons between different modeling approaches.

2.3. Automatic Post-Editing

Automatic Post-Editing (APE) has emerged as a practical approach to improving machine Automatic Post-Editing (APE) aims to enhance translation quality without altering the underlying machine translation systems. The field initially relied on rule-based approaches, which applied handcrafted correction patterns to systematically address errors

in MT output. Early studies by Simard et al. [11] demonstrated that statistical phrase-based methods could be effectively adapted for post-editing, framing the task as a monolingual translation problem from “incorrect” to “correct” translations.

The introduction of neural approaches represented a major advancement in APE. Junczys-Dowmunt and Grundkiewicz [22] investigated various neural sequence-to-sequence architectures, showing that encoder–decoder models are capable of learning to correct systematic translation errors. Their work established foundational baselines and highlighted the potential of neural methods for automatic post-editing tasks.

More recent studies in automatic post-editing (APE) have explored the use of additional contextual information and transformer-based architectures. Correia and Martins [23] proposed a method that fine-tunes pre-trained BERT models for both the encoder and decoder of an APE system, achieving strong performance even with a relatively small training corpus. Building on this work, Yang et al. [24] demonstrated that integrating pre-trained language models into post-editing workflows can substantially improve error correction.

The Workshop on Machine Translation (WMT) has served as an important venue for the systematic evaluation of APE methods through shared tasks [25]. Results from these shared tasks consistently show that even incremental improvements in post-editing can be valuable, especially for domain-specific content or languages lacking high-quality MT systems.

Nevertheless, most APE research has concentrated on high-resource language pairs, such as English–German and English–Spanish. Systematic studies of APE for low-resource or morphologically complex languages remain limited, leaving a significant gap in understanding how these methods generalize to languages with different linguistic characteristics [26].

2.4. Large Language Models for Translation

The advent of Large language models (LLMs) has significantly reshaped research in machine translation. Brown et al. [5] demonstrated that GPT-3 is capable of performing zero-shot and few-shot translation across a wide range of language pairs without requiring explicit parallel training data. This ability arises from the model’s extensive multilingual pre-training and its capacity for in-context learning through example prompts.

Subsequent studies have systematically assessed the translation capabilities of LLMs. Hendy et al. [27] conducted extensive evaluations of GPT models, showing that while these models exhibit impressive zero-shot performance, they generally underperform compared to specialized translation systems for high-resource language pairs. Nevertheless, for low-resource languages, the performance gap is often smaller, suggesting that LLMs may offer a competitive alternative to conventional methods.

Jiao et al. [28] specifically analyzed ChatGPT’s translation capabilities, reporting that GPT-4 achieves noticeably higher translation quality than its predecessors, particularly for low-resource languages. Their findings underscore the value of instruction-tuned models and the role of carefully designed prompting strategies in maximizing translation performance.

Research on enhancing LLM translation has explored fine-tuning approaches. Li et al. [29] examined multilingual machine translation with LLMs, showing that targeted fine-tuning can substantially improve results on specific language pairs. Yang et al. [30] demonstrated that LLMs can manage simultaneous multilingual translation, highlighting their flexibility across multiple languages. Vilar et al. [31] investigated prompting strategies for PaLM, revealing that prompt design has a significant influence on translation quality. Their work provides crucial insights into effectively leveraging LLMs for translation tasks and emphasizes the need for evaluation methodologies that account for the unique characteristics of LLM-generated translations [32].

2.5. Fine-Tuning Approaches for Translation

The use of fine-tuning techniques to enhance translation quality has become an increasingly active area of research. Early fine-tuning approaches in neural machine translation primarily focused on domain adaptation, where models trained on general parallel corpora were further refined for specific domains or text genres [33]. The emergence of large pre-trained language models has, however, opened new avenues for fine-tuning strategies.

Xu et al. [13] proposed significant innovations in fine-tuning LLMs for machine translation, showing that parameter-efficient methods can deliver substantial performance gains while requiring minimal computational resources. Such approaches are particularly valuable for low-resource language pairs, where conventional training from scratch is often impractical.

Parameter-efficient fine-tuning techniques, including Low-Rank Adaptation (LoRA) [34] and prefix-tuning [35], have demonstrated the ability to adapt LLMs to specific tasks without the heavy computational cost of full model fine-tuning. These methods are especially relevant in resource-constrained scenarios, which are common in low-resource language processing. Additionally, incorporating human feedback into fine-tuning has proven highly effective. Ouyang et al. [36] showed that reinforcement learning from human feedback (RLHF) can significantly improve the alignment of model outputs with human preferences, enhancing the quality and appropriateness of generated text—a feature highly relevant for translation post-editing.

Despite these advances, the majority of research on fine-tuning for translation has concentrated on high-resource languages or synthetic datasets. Investigations into fine-tuning strategies specifically tailored for low-resource languages with complex morphological structures remain limited, highlighting a critical gap in current research [37].

2.6. Evaluation of Translation Quality

Evaluating machine translation quality, particularly for low-resource languages, involves substantial methodological challenges. Traditional automatic metrics, such as BLEU [38] and METEOR, have well-known limitations when applied to morphologically rich languages, where surface-level similarity does not always reflect true semantic equivalence. More recent neural-based evaluation metrics, including BERTScore [39] and COMET [40], offer improvements by incorporating semantic similarity measures. However, these methods still face difficulties for languages with limited pre-trained representations or complex morphological systems.

Human evaluation remains the gold standard for assessing translation quality, yet conducting systematic evaluations for low-resource languages is often challenging due to the scarcity of qualified annotators and the difficulty of defining consistent evaluation criteria [41]. Freitag et al. [42] provide valuable guidance on best practices for human MT evaluation, but most studies focus on high-resource language pairs.

For morphologically complex languages like Kazakh, effective evaluation frameworks must consider multiple levels of translation accuracy, including morphological correctness, lexical choice, syntactic structure, semantic fidelity, and stylistic appropriateness. Designing comprehensive evaluation methodologies that capture these dimensions while remaining feasible for systematic application represents a major challenge for the field [43].

Research Gaps

Despite significant advancements in machine translation, automatic post-editing (APE), and fine-tuning of large language models, several key gaps remain in the literature:

1. Limited focus on real low-resource languages: most studies have concentrated on high-resource languages or synthetic low-resource scenarios. Systematic investigations

of genuinely low-resource languages, which face authentic data scarcity and unique linguistic challenges, remain scarce.

2. Handling morphological complexity: while some research has considered morphologically rich languages, there is limited systematic analysis of how existing approaches address the agglutinative and highly inflected structures typical of Turkic languages. Examples: For examples:
 - үйлерімізде = үй (house) + -лер (plural) + -іміз (1st person plural possessive) + -де (locative case) = “in our houses”
 - кітаптарымыздан = кітап (book) + -тар (plural) + -ымыз (1st person plural possessive) + -дан (ablative case) = “from our books”
3. Post-editing for morphologically complex languages: APE research has predominantly targeted languages with simpler morphological systems. Applying post-editing techniques to languages with extensive agglutinative morphology remains underexplored.
4. Development of specialized evaluation frameworks: there is a shortage of standardized evaluation methodologies specifically designed for morphologically complex, low-resource languages, which must account for multiple layers of linguistic complexity, including morphology, syntax, semantics, and style.
5. Availability of annotated datasets: few systematically annotated corpora with detailed error categorization exist for most low-resource languages, limiting both the development and evaluation of targeted improvement strategies.

Our work addresses these gaps by providing the first systematic study of fine-tuning for Kazakh translation post-editing, creating comprehensive evaluation methodologies, and developing annotated datasets that will serve as a foundation for future research in this critical area.

3. Dataset

3.1. Train Set

Our dataset, Kazakh Post Editing (KazPE), was compiled from three primary sources with careful consideration of the unique features and challenges of translating into Kazakh. The source materials were originally collected in English under the supervision of linguistics specialists and later translated into Kazakh using ChatGPT 4.1 mini. Additionally, medical corpora and sentences in scientific style were gathered from open data sources.

Under the guidance of Kazakh language experts, the translations were analyzed and several types of errors were identified. Semantic errors occur when the meaning is not conveyed correctly, lexical errors involve issues with word choice such as homonyms, synonyms, or dialect words, and terminological errors arise when domain-specific terms in medicine or science are not properly translated. Word order errors occur when the subject and predicate are incorrectly positioned within a sentence. Orthographic errors involve violations of phonetic harmony rules, for example, using «лер» instead of «лар». Idiomatic errors appear when set expressions are translated literally without preserving their meaning. Grammatical errors affect the sentence structure, for example, «Мен кітап оқып едім» being incorrectly rendered as «Мен оқып едім кітап». Morphological errors occur when suffixes are incorrectly applied, for instance, «Ол тау барды» instead of the correct «Ол тауға барды».

This is an important clarification. In our taxonomy:

- Morphology refers specifically to word-internal structure: incorrect case markers, possessive suffixes, plural markers, and other affixation errors characteristic of Kazakh’s agglutinative system (e.g., using the wrong case ending on a noun).

- Grammar (Grammatical Accuracy) refers to sentence-level syntactic structure: word order violations, agreement mismatches between sentence constituents, improper clause construction, and other phenomena that occur above the word level. This distinction is linguistically principled and particularly relevant for agglutinative languages like Kazakh, where morphological complexity within words is distinct from syntactic relationships between words.

ChatGPT-Generated Sentences (9212 Sentences)

The number of sentences generated by ChatGPT and drawn from each source was determined randomly to ensure an unbiased representation of data across all sources.

The prompt used for translation generation:

“Translate the following English sentence to Kazakh. Provide a natural, fluent translation that preserves the original meaning while following Kazakh grammatical conventions. Do not provide explanations or alternative translations.”

For the domain-specific sentence generation, we provided ChatGPT-4.1 miniGPT-4.1 mini with domain indicators (e.g., “Generate medical terminology sentences” or “Generate legal document sentences”) to produce stylistically appropriate content. The exact prompts for different domains can be included in our Supplementary Materials.

The dataset was organized into six stylistic subdomains to encompass a broad spectrum of register-specific linguistic characteristics:

- Medicine—1476 sentences
- Scientific style—2493 sentences
- Journalistic—2161 sentences
- Oral style—1258 sentences
- Fiction—1323 sentences
- Legal—501 sentences

The average sentence length in the dataset ranged from 4 to 10 words. Each stylistic subdomain was systematically assessed for translation errors across nine linguistic categories: (1) semantics, (2) lexical, (3) morphology, (4) terminology, (5) style, (6) word order, (7) grammar, (8) orthography, (9) idiomatic expressions.

In the medical subset, 613 stylistic errors, 159 lexical errors, and 53 terminology errors were identified, while 539 sentences were completely error-free. By contrast, the legal subset exhibited a higher prevalence of stylistic inconsistencies (473 instances) and terminology issues (17 instances), with only 11 sentences free from errors. Other subdomains showed distinct patterns of error distribution, highlighting the unique linguistic characteristics of each style (see Table 1).

Table 1. Error distribution within ChatGPT-generated test set by style.

Source/Style	Total	Sem	Lex	Morph	Term	Style	WO	Gram	Ortho	Idiom	NoErr
ChatGPT—Medicine	1476	15	159	12	53	613	21	75	1	0	539
ChatGPT—Scientific	2493	249	167	49	123	696	224	3	1	0	980
ChatGPT—Journalistic	2161	21	13	1	21	1313	15	0	0	0	777
ChatGPT—Oral	1258	44	167	40	0	144	49	0	4	9	820
ChatGPT—Fiction	1323	86	139	0	0	401	12	1	0	127	581
ChatGPT—Legal	501	0	0	0	17	473	0	0	0	0	11
Medical Corpus	647	4	41	2	2	3	0	0	0	0	596
Scientific Papers	149	0	0	0	0	0	0	0	0	0	149

Legend: Sem = Semantics, Lex = Lexical, Morph = Morphology, Term = Terminology, WO = Word Order, Gram = Grammar, Ortho = Orthography, Idiom = Idiomatic.

The medical corpus consisted of 647 sentences sourced from specialized medical texts Available at: https://github.com/Franck-Dernoncourt/pubmed-rct?utm_source=chatgpt.com. (accessed on 18 February 2026).

The medical corpus sentences were extracted from publicly available Kazakh medical texts and terminology databases developed by medical institutions in Kazakhstan. The selection process prioritized:

- Terminological diversity (covering various medical specialties)
- Sentence complexity range (4–15 words)
- Authentic medical communication contexts

We will add specific corpus citations in the revision. The 647-sentence sample was selected to represent common medical documentation styles while maintaining manageable annotation scope for our expert.

Most sentences (596) were error-free, while the remainder included errors in semantics (4), lexical (41), morphology (2), terminology (2), and style (3). This subdomain proved particularly useful for evaluating the accuracy of domain-specific terminology translation.

Scientific Papers (149 sentences) Available at: https://github.com/VladislavKaryukin/kk_en_corpora. (accessed on 18 February 2026). This subset was compiled from peer-reviewed scientific articles, ensuring that the original texts were grammatically correct and free of linguistic errors. It serves as a reliable benchmark for scientific style and technical precision in translation tasks.

Scientific papers were selected from peer-reviewed journals in computer science, linguistics, and related fields published in English with potential relevance to Kazakh NLP research. Selection criteria included:

- Clear, well-formed English sentences
- Technical terminology representation
- Grammatical correctness (serving as high-quality references)

Sentences were extracted from abstract, introduction, and methodology sections (2–4 sentences per paper from approximately 40 papers), focusing on complete, self-contained statements rather than specific positional sampling. This approach ensured grammatically correct, professionally written content suitable as error-free reference material.

The complete dataset (see Figure 1) combines domain-specific content, stylistic variety, and varying error patterns, providing a comprehensive foundation for evaluating Kazakh machine translation systems. Its design allows for rigorous assessment of models' ability to handle different registers and accurately translate specialized vocabulary.

The "Total" column in Table 1 reflects the total number of sentences in each subset. It should be noted that individual sentences often contain multiple error types (semantics and idiomatic; style and semantics; morphology and lexical, morphology and terminology, semantics and orthography), which is clearly represented in our data. The total number of errors across categories for a given style often exceeds the difference between "Total" and "NoErr," as all errors present in each sentence were annotated, not just the primary one.

For instance, in the ChatGPT-Medicine subset, which includes 1475 sentences, 540 of which are error-free, the remaining 935 sentences contain 962 errors distributed across categories: 613—style, 160—lexical, 53—terminology, and others. Some sentences contain 2 error types simultaneously, which is characteristic of machine translation output and enhances the value of the dataset for training comprehensive post-editing systems. For ease of comparison, and to ensure that even categories with very low error counts (such as orthographic errors) are visible, the data are presented on a logarithmic scale (see Figure 1).

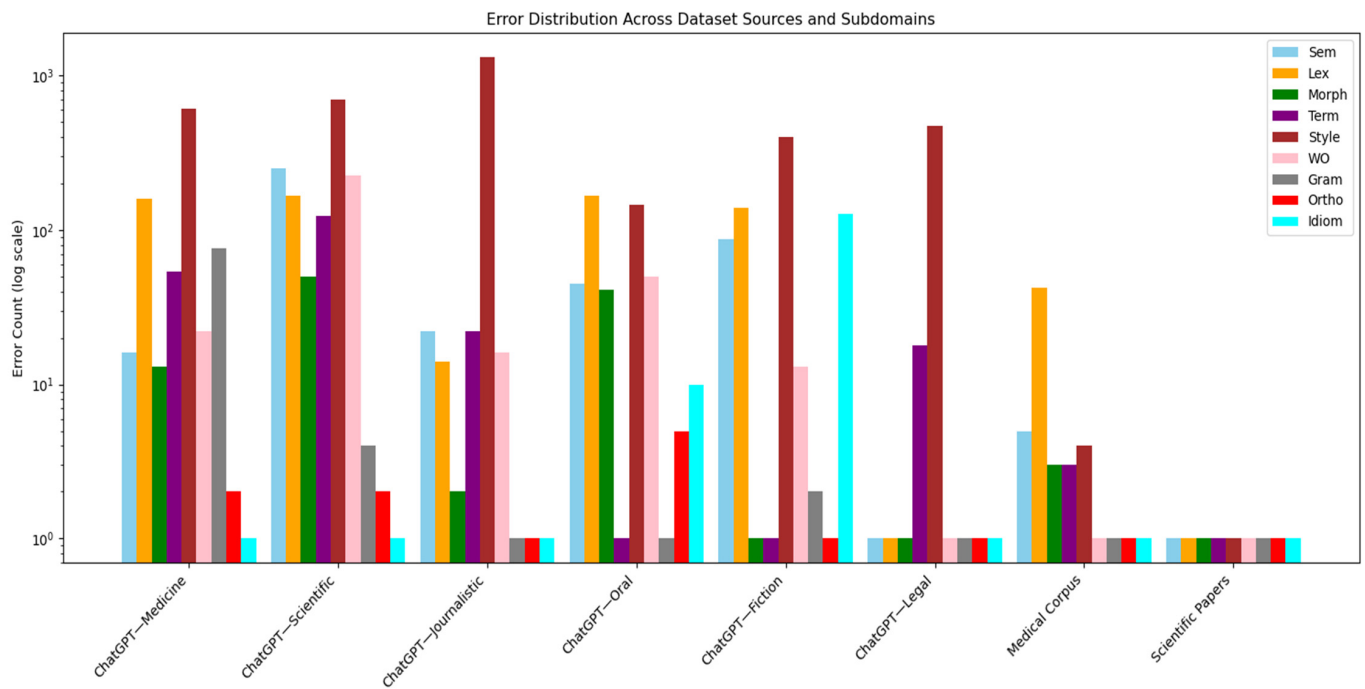


Figure 1. Translation error distribution in the dataset.

Table 2 presents sentences extracted from the Kazakh Post Editing (KazPE) corpus along with the types of errors identified in each sentence.

The analysis of ChatGPT translations from English to Kazakh reveals systematic limitations. The model frequently produces errors across multiple linguistic categories, including word order, lexical choice, idiomatic expressions, terminology, morphology, and orthography. These errors arise primarily because ChatGPT, while powerful in general language understanding, does not fully capture the syntactic, morphological, and idiomatic characteristics of Kazakh. Kazakh is a morphologically rich and agglutinative language with flexible word order and domain-specific terminology. As a result, literal or overly generalized translations often occur, leading to unnatural phrasing, incorrect lexical selection, and misrepresentation of idiomatic or formal expressions. This highlights the need for task-specific fine-tuning, incorporation of linguistic knowledge, and human evaluation to improve translation quality for low-resource languages like Kazakh.

Table 2. Translation errors analysis.

English Text	ChatGPT Translation	Human Translation	Error Types	Explanation
He underwent a physical therapy session to improve his mobility.	Қозғалу қабілетін жақсарту үшін ол физикалық терапия сеансынан өтті.	Ол қозғалу қабілетін жақсарту үшін физикалық терапия сеансынан өтті.	Word Order	In Kazakh, the subject («Ол») is placed at the beginning of the sentence.
He cleaned his room before going out.	Ол сыртқа шықпас бұрын бөлмесін жинады.	Ол далаға шықпас бұрын бөлмесін жинады.	Lexical	The word chosen does not match the context: «далаға» instead of «сыртқа».
They are head over heels in love.	Олар бас-аяқ ғашық.	Олар шынайы ғашық.	Idiomatic	ChatGPT translated literally («Олар бас-аяқ ғашық»), but the natural idiomatic Kazakh translation is «Олар шынайы ғашық» (human translation).

Table 2. Cont.

English Text	ChatGPT Translation	Human Translation	Error Types	Explanation
The authors argue that technological advancements should be regulated to prevent misuse.	Авторлар технологиялық жетістіктердің теріс пайдаланылу болдырмау үшін реттелуі қажет деп пайымдайды.	Авторлар технологиялық жетістіктердің теріс пайдалануын болдырмау үшін реттелуі тиіс деп есептейді.	Grammar	«пайдаланылу»—a participle formed with a passive suffix, which does not match the intended meaning; the correct form «пайдалануын» is a noun, which is grammatically suitable as the object of the action.
The defendant was sentenced to five years in prison for tax evasion.	Айыпталушы салық төлеуден жалтарғаны үшін бес жылға түрмеге қамалды.	Айыпталушы салықтан жалтарғаны үшін бес жылға бас бостандығынан айырылды.	Style	ChatGPT's translation uses a colloquial and less formal phrasing («түрмеге қамалды»—"was put in prison"), which is suitable for everyday speech but not for formal/legal style. The human translation («бас бостандығынан айырылды»—"deprived of liberty") follows official/legal style, which is more appropriate for judicial contexts.
Researchers conducted a non-inferiority European randomized controlled clinical trial (RCT) in 37 centers.	Зерттеушілер 37 орталыққа кемшіліксіздік принципіне негізделген еуропалық рандомизацияланған бақылаулы клиникалық зерттеу өткізді.	Зерттеушілер 37 орталықта кемшіліксіздік принципіне негізделген еуропалық рандомизацияланған бақылаулы емханалық зерттеу өткізді.	Morphology, Terminology	Орталыққа- morphology error. Correct: «Орталықта» (location, not direction). «Клиникалық»—Terminology error. Correct: «емханалық» (matches the context of "clinical trial").
The clock struck exactly on time.	Сағат бұлқымсыз шешімімен соқты.	Сағат өзгермей дәл соқты.	Semantics, Orthography	Orthography: «соқты» → should be «соқты». Semantics: «бұлқымсыз шешімімен» → meaning does not match «exactly, without deviation».

3.2. Test Set

The test set was designed following the same methodology as the training set, ensuring representation across multiple domains and relevant error categories for Kazakh translation. It includes three main sources (see Table 3):

ChatGPT-Generated Sentences (258 sentences). This subset contains sentences with varied stylistic and structural characteristics, aimed at evaluating the model's ability to handle diverse linguistic phenomena. Error distribution includes semantics (7), lexical (32), style (3), word order (1), idiomatic expressions (2), and morphology (2), while 215 sentences were free from errors.

Medical Corpus (46 sentences). Sourced from specialized medical texts, this subset primarily tests domain-specific vocabulary, with 15 sentences containing vocabulary errors and no other error types observed. A total of 31 sentences were error-free, providing a reliable reference for medical terminology translation.

Scientific Papers (7 sentences). Extracted from peer-reviewed scientific literature, all sentences in this subset are error-free, serving as high-quality benchmarks for technical and scientific translation style.

This test set structure enables systematic evaluation of machine translation systems across general content, specialized domains, and stylistically constrained texts, ensuring comprehensive assessment of translation quality.

Table 3. Error distribution across test set sources. “NoErr” indicates sentences free of any detected issues.

Source/Style	Total	Sem	Lex	Morph	Style	WO	Idiom	NoErr
ChatGPT	258	7	32	2	3	1	2	215
Medical Corpus	46	0	15	0	0	0	0	31
Scientific Papers	7	0	0	0	0	0	0	7

Furthermore, the ChatGPT-generated portion of the test set is subdivided into distinct stylistic categories, with error distributions summarized in Table 4.

Table 4. Error distribution within the ChatGPT-generated test set by style.

Style	Total	Lexical	No Error	Semantic	Morphology	Style	Idiom	WO
Medicine	49	10	39	-	-	-	-	-
Legal/Journalistic	49	3	45	1	-	1	-	-
Fiction	50	3	41	3	-	1	2	1
Oral Style	50	11	35	3	2	1	-	-
Scientific Style	60	5	55	-	-	-	-	-

Note: Dashes (-) indicate zero or unreported errors for those categories in the respective style.

Tables 3 and 4 do not include categories such as orthography, terminology, and grammar, as no errors were observed in these components across all domains of the test set.

Based on the data presented in Table 4, the following conclusion can be drawn. Some categories of errors are absent in the test set, as formal and standardized texts (medical and scientific) do not contain idioms, complex morphological constructions, or non-standard style. Errors are observed predominantly in more flexible styles—fictional and oral texts—where the presence of idioms, variable word order, and less formal constructions creates more opportunities for violations.

This comprehensive breakdown makes it possible to evaluate model performance with greater precision across different linguistic registers and domains in the Kazakh test set.

3.3. Translation Generation

We employed ChatGPT-4.1 mini/GPT-4.1 mini to produce initial Kazakh translations for all sentences in the training and test sets. The translation prompt was carefully designed to generate fluent, natural Kazakh translations that retain the meaning of the original English sentences:

“Translate the following English sentence into Kazakh. Ensure the translation is fluent, natural, and consistent with Kazakh grammatical conventions. Avoid providing explanations or alternative translations.”

Translations were generated in batches of 13 sentences to facilitate efficient processing.

3.4. Annotation Process

Due to limited resources, a single expert annotator reviewed each translation in the test set. The annotator had the following qualifications:

- Native speaker of Kazakh with university-level education in Kazakhstan
- Advanced proficiency in English
- Knowledge of Kazakh computational linguistics research

Each translation was scored on a continuous scale from 0 (completely incorrect) to 1 (perfect translation). In addition, all errors were annotated according to the established error taxonomy.

4. Fine-Tuning ChatGPT-4.1 miniGPT-4.1 mini to Improve Post-Editing of Kazakh Text

4.1. Baseline Configuration

For comparison purposes, we evaluated a zero-shot baseline using the unmodified base ChatGPT-4.1 miniGPT-4.1 mini model. This baseline employed a straightforward prompting strategy: “Review the following Kazakh sentence and correct any grammatical, lexical, stylistic, or typographical errors. Provide only the corrected sentence without explanations.” The baseline used a temperature of 0.1 to minimize randomness and ensure consistent corrections, with a maximum token limit of 256 matching the fine-tuned model. Crucially, the baseline received no fine-tuning whatsoever and no few-shot examples, representing the out-of-the-box performance of the model on Kazakh post-editing tasks.

4.2. Model Fine-Tuning

For our fine-tuning experiments, we selected GPT-4.1 (GPT-4.1 minigt-4.1 mini) as the foundational model. Access and model customization were carried out using the OpenAI fine-tuning API, which offers a streamlined workflow for supervised adaptation.

Figure 2 illustrates the scheme of how the GPT-4.1 model adapts incorrect sentences into correct ones during fine-tuning.

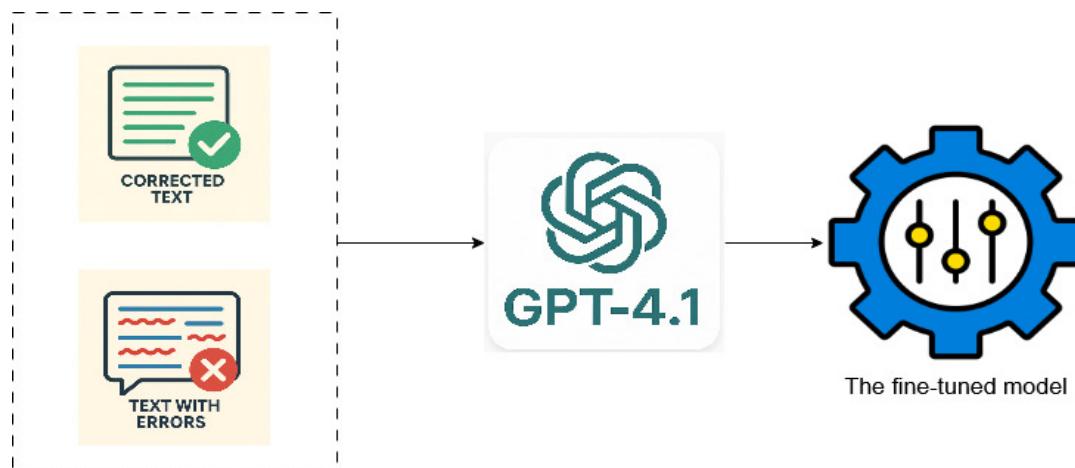


Figure 2. Scheme of the fine-tuning process.

The supervised fine-tuning (SFT) procedure aimed to teach the model to produce corrected Kazakh translations from imperfect inputs. Training samples were structured in a conversational prompt-response format to match the model’s native interaction style:

System: You are a professional editor of Kazakh. Given a sentence that may contain grammatical, lexical, stylistic, or typographical errors, rewrite it accurately in Kazakh while preserving its original meaning.

User: [INCORRECT_KAZAKH_SENTENCE]

Assistant: [CORRECTED_KAZAKH_SENTENCE]

This setup explicitly directed the model to focus on precise error correction while ensuring semantic fidelity to the source text.

The fine-tuning process employed the following hyperparameters:

- Learning rate multiplier: 2
- Batch size: 13 examples
- Training epochs: 2 (selected based on validation performance)

We monitored training progress using held-out validation loss and implemented early stopping to mitigate overfitting. Figures 3 and 4 illustrate the evolution of training accuracy and loss, respectively.

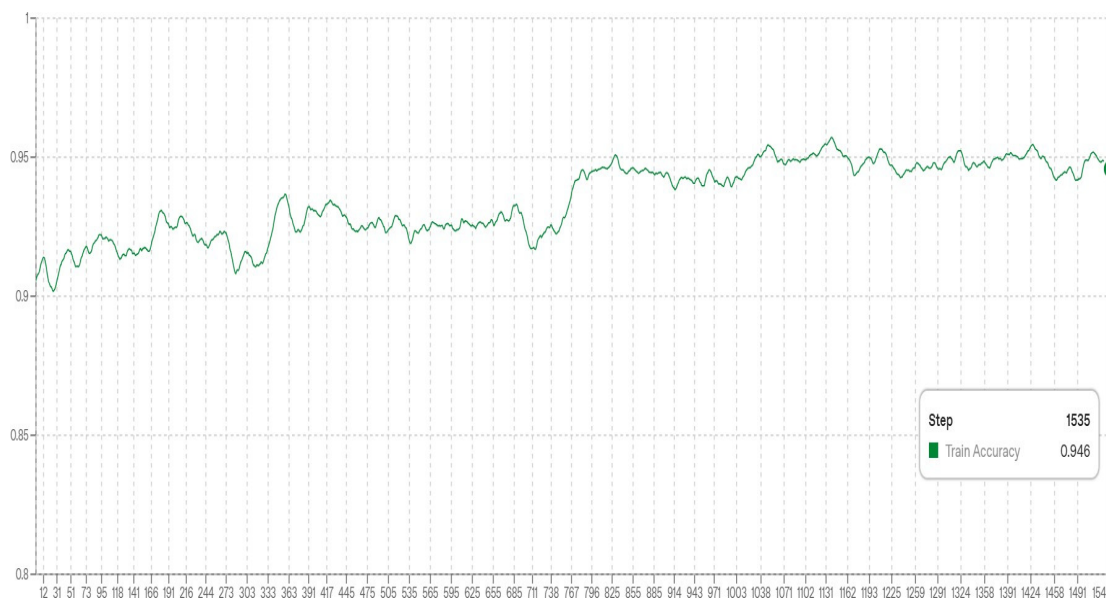


Figure 3. Training accuracy over iterations.

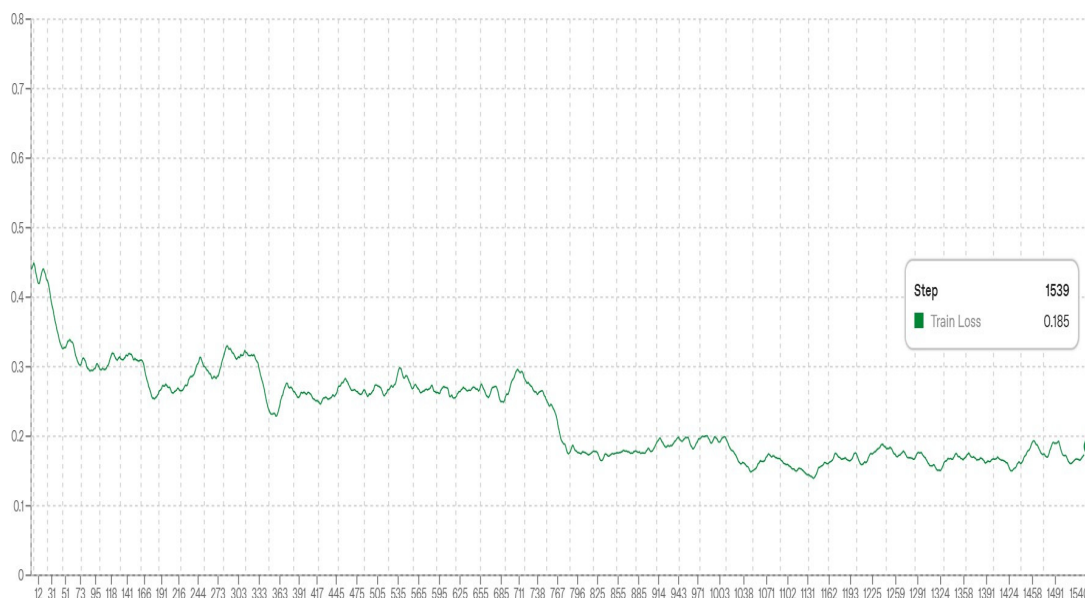


Figure 4. Training loss over iterations.

The performance of the fine-tuned models was compared to the Base GPT-4.1 miniGPT-4.1 mini, which carried out zero-shot post-editing with standard prompts. This baseline allowed us to measure the gains achieved through supervised fine-tuning.

4.3. Evaluation Methodology

Human evaluation. A single expert annotator reviewed all post-edited sentences, scoring them according to:

- Overall quality: a continuous rating from 0 to 1 reflecting both fluency and fidelity to the source.
- Error categorization: identification and classification of specific errors present in each sentence.

4.3.1. Evaluation Metrics

Our evaluation framework combines human expert assessment as the primary metric with complementary automatic metrics to provide comprehensive model evaluation. The primary metric consists of human quality assessment performed by the same expert annotator who created the dataset, ensuring consistency in evaluation standards. Each post-edited sentence is scored on a continuous scale from 0 (completely incorrect) to 1 (perfect translation), combining assessment of both adequacy (whether the original meaning is preserved) and fluency (whether the output reads naturally in Kazakh).

4.3.2. Evaluation Procedure and Blinding Protocol

For each sentence in the test set, we generated post-edited versions using both the fine-tuned model and the baseline model with their respective prompting strategies. The annotator then assessed both outputs according to a blinding protocol designed to eliminate bias. Model outputs were presented in randomized order without any indication of which system produced which output, and the annotator was instructed to evaluate each output independently without comparing them directly. To further minimize memory effects and bias, the evaluation phase occurred 2 months after the dataset creation phase, reducing the likelihood that the annotator would recall specific sentence corrections from the annotation process.

For each output, the annotator assigned a quality score from 0 to 1 and identified any remaining errors using the 9-category taxonomy. Both outputs were then compared against the gold-standard correction established during annotation to enable automatic metric calculation. This procedure ensured that all evaluation measures—both human judgments and automatic metrics—were computed consistently across both systems.

4.3.3. Reproducibility Statement

To facilitate independent replication of our results, we provide comprehensive documentation of all experimental procedures and commit to releasing materials upon publication. The annotation guidelines document will be made publicly available, along with training scripts and data processing code. The fine-tuned model checkpoint can be accessed through the OpenAI fine-tuning API using the training job identifier that will be provided.

All experiments were conducted using OpenAI API python client version 1.12.0 with Python 3.10.12. Data processing employed pandas 2.0.0 and numpy 1.24.3. Random seeds were set to 42 for all data splitting procedures to ensure reproducibility, and wherever possible, deterministic settings were used in API calls.

4.4. Automatic Evaluation Metrics

In addition to human evaluation, we assessed the quality of post-edited Kazakh text using three widely adopted automatic metrics—BLEU, METEOR, and TER. These metrics provide complementary perspectives on translation quality, particularly in terms of surface similarity and edit distance.

Bilingual Evaluation Understudy (BLEU) score evaluates the n-gram overlap between the model output and reference translations. It is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N \omega_n \log p_n\right) \quad (1)$$

where p_n is the precision of n-grams, ω_n are weights (commonly uniform), and BP is the brevity penalty to penalize short outputs:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (2)$$

with c the length of the candidate translation and r the length of the reference.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) aligns output and reference text based on exact, stem, synonym, and paraphrase matches. The score is computed as:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - P_{\text{frag}}) \quad (3)$$

where $F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$ is the harmonic mean of precision P and recall R (weighted toward recall), and P_{frag} is a penalty for fragmented matches. Translation Error Rate (TER) measures the minimum number of edits (insertions, deletions, substitutions, and shifts) required to change the system output into the reference translation:

$$\text{TER} = \frac{\text{Number of edits}}{\text{Number of words in reference}} \quad (4)$$

Lower TER values indicate better translation quality, as fewer edits are needed to match the reference. These metrics together provide a quantitative evaluation of translation fidelity, fluency, and surface-level accuracy for Kazakh post-editing.

5. Results and Analysis

This section reports the evaluation findings for our fine-tuned ChatGPT-4.1 miniGPT-4.1 mini model applied to Kazakh post-editing tasks. We assess its performance by benchmarking it against the original, non-fine-tuned ChatGPT-4.1 miniGPT-4.1 mini model.

Examples of translations evaluated by linguists using fine-tuned models are as follows. In the medical domain, the baseline translation was «Трансплантация операциясы сәтті өтті деп танылды», and the fine-tuned and post-edited version was «Трансплантация отасы сәтті өтті деп танылды». Here, a lexical error was corrected. Specifically, the word «операция» was not used correctly in the Kazakh medical context; the correct term is «ота».

In the fiction domain, the baseline translation was «Олар трендке секіріп кетті». The fine-tuned and post-edited version was «Олар трендке еріп кетті». This correction addressed an idiomatic error because the English phrase They jumped on the bandwagon, when translated literally as «секіріп», changed the meaning. The post-editing provided an appropriate idiomatic equivalent in Kazakh («еріп кетті»).

In the legal domain, the baseline translation was «Тараптар соттан тыс шешімге келді», and the fine-tuned and post-edited version was «Тараптар сотқа дейін келісімге келді». Here, a semantic error was corrected, as the original translation was unclear in meaning, and the post-edited version provided a contextually accurate translation.

Table 5 provides a summary of the performance metrics for both models based on a test set containing 311 sentences. The evaluation was carried out by a human annotator using a continuous rating scale ranging from 0 (indicating a completely incorrect translation) to 1 (indicating a flawless translation), with intermediate scores reflecting varying degrees of translation quality.

Table 5. Overall performance comparison between baseline and fine-tuned models based on human evaluation.

Model	Mean Quality Score
ChatGPT-4.1 miniGPT-4.1 mini	0.80
Fine-tuned ChatGPT-4.1 miniGPT-4.1 mini	0.84
Absolute improvement	+0.04
Relative improvement	5.0%

Human evaluation scores range from 0 (completely incorrect) to 1 (perfect translation).

The fine-tuned model obtained an average quality score of 0.84, compared to 0.80 for the baseline ChatGPT-4.1 miniGPT-4.1 mini, indicating a measurable improvement in post-editing performance based on human evaluation. Human evaluation scores improved from 0.80 to 0.84, a 5.0% relative gain.

Table 6 presents the BLEU, TER, and METEOR scores obtained from the ChatGPT-4.1 miniGPT-4.1 mini and Fine-tuned ChatGPT-4.1 miniGPT-4.1 mini models on the 311-sentence test set.

Table 6. Evaluation of translation quality using automated metrics.

Model	Bleu ↑	Ter ↓	Meteor ↑
ChatGPT-4.1 miniGPT-4.1 mini	0.7002	0.1155	0.7590
Fine-tuned ChatGPT-4.1 miniGPT-4.1 mini	0.8682	0.0751	0.8500
Absolute improvement	+0.1680	−0.0404	+0.0910
Relative improvement	+24%	−35%	+12%

Note: ↑ indicates higher is better; ↓ indicates lower is better.

As shown in Table 6, the Fine-tuned ChatGPT-4.1 miniGPT-4.1 mini model demonstrates a clear improvement in translation quality compared to the baseline ChatGPT-4.1 miniGPT-4.1 mini. The fine-tuned model achieved notably higher BLEU (0.8682) and METEOR (0.8500), TER (0.0751) scores, indicating better lexical and semantic correspondence with the reference translations. The overall results confirm that fine-tuning led to a measurable enhancement in translation performance.

The fine-tuned model demonstrates consistent improvements across all evaluation metrics. The BLEU score increased from 0.7002 to 0.8682 (+24% relative improvement), while METEOR improved from 0.7590 to 0.8500 (+12%). The TER decreased from 0.1155 to 0.0751, representing an 35% relative reduction in post-editing effort. These consistent improvements across both automatic and human metrics indicate meaningful enhancement in post-editing quality.

The authors also conducted additional experiments on the 311-sentence test set using the LLaMA 3, Gemma, and KazLLM models. The results, evaluated with BLEU, METEOR, and TER metrics, are presented in Table 7.

Table 7. Evaluation metrics of fine-tuned LLaMA 3, KazLLM, and Gemma models.

Model	BLEU ↑	METEOR ↑	TER ↓
Fine-tuned LLaMA 3	0.5262	0.7180	0.1828
Fine-tuned Gemma	0.7769	0.8012	0.1100
Fine-tuned KazLLM	0.7715	0.7890	0.1142

Note: ↑ indicates higher is better; ↓ indicates lower is better.

Compared to the baseline ChatGPT-4.1 miniGPT-4.1 mini, the fine-tuned KazLLM and Gemma models showed improvements across all evaluation metrics, while LLaMA 3 lagged behind. Specifically, Gemma achieved a BLEU score of 0.7769 (+10.95%), METEOR of

0.8012 (+5.56%), and TER of 0.1100 (−4.76%) relative to the baseline, slightly outperforming KazLLM (BLEU 0.7715, METEOR 0.7890, TER 0.1142).

Among all models, the fine-tuned ChatGPT-4.1 miniGPT-4.1 mini delivered the best overall performance, with a BLEU of 0.8682 (+11.8% vs. Gemma, +12.5% vs. KazLLM), METEOR of 0.8500, and TER of 0.0751, demonstrating the highest accuracy and the lowest post-editing effort.

5.1. Domain-Specific Performance

We assessed both the fine-tuned and baseline models across different stylistic and domain-specific categories using the same quality metric. Table 8 presents the scores for each domain, with bolded rows highlighting areas where the fine-tuned model shows superior performance.

Table 8. Domain-specific performance scores for fine-tuned and baseline ChatGPT-4.1 miniGPT-4.1 mini models. (Bolded rows indicate domains where the fine-tuned model outperforms the baseline).

Domain/Style	Fine-Tuned Model	Baseline ChatGPT
Oral Style	0.828	0.740
Fiction	0.906	0.870
Medical Corpus	0.913	0.913
Medical Domain	0.869	0.714
Legal	0.837	0.714
Scientific Style	0.900	0.867
Scientific Corpus	0.857	0.857

The analysis demonstrates that fine-tuning provides notable gains in domains that require precise terminology and formal language use. The most substantial improvements are observed in the legal and medical domains, with score increases of 0.123 and 0.155 respectively. The oral style category also shows moderate improvement, suggesting better handling of conversational and colloquial language. Gains in the scientific style domain are minimal, while the medical and scientific corpora maintain consistent high performance, reflecting their specialized and largely error-free content.

In contrast, the fiction domain shows a slight decline in performance following fine-tuning, indicating that improvements in domain-specific accuracy may come at a minor cost to handling creative or literary language. Overall, these results suggest that fine-tuning substantially improves the model's capacity to generalize and adapt to a variety of linguistic phenomena, particularly in areas that require precise terminology and formal conventions, while generally maintaining performance in other styles.

Table 9 presents the results of the automatic evaluation of text quality generated by the ChatGPT-4.1 miniGPT-4.1 mini model and its fine-tuned version (a model additionally trained on specialized data) across various domains and styles. The evaluation employed standard metrics: BLEU, TER, and METEOR, which reflect the degree of similarity to reference texts.

The results demonstrate that the fine-tuned model consistently outperforms the base model across all metrics and in all examined domains.

It increases significantly after fine-tuning, especially in specialized domains such as medical and scientific. For example, in the medical domain, the BLEU score rises from 0.6073 to 0.9019, while in the scientific style it increases from 0.6410 to 0.9608. The lowest TER values are observed in the scientific and legal domains (0.0166 and 0.0379, respectively). METEOR also shows consistent growth across all domains, indicating an improvement in the semantic correspondence of the generated texts to the reference.

Table 9. Quality evaluation using automatic metrics for fine-tuned and baseline ChatGPT-4.1 miniGPT-4.1 mini models.

Automatic Metric	Oral Style		Fiction		Medical Corpus		Domain/Style Medical Domain		Legal/Journalistic		Scientific Style		Scientific Corpus	
	ChatGP T-4.1 miniGPT-4.1 mini	Fine-Tuned ChatGP T-4.1 miniGPT-4.1 mini	ChatGP T-4.1 miniGPT-4.1 mini	Fine-Tuned ChatGP T-4.1 miniGPT-4.1 mini	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini
BLEU	0.7861	0.8035	0.6742	0.7704	0.8017	0.8066	0.6073	0.9019	0.3190	0.9281	0.6410	0.9608	0.9297	0.9297
TER	0.2012	0.1258	0.2254	0.1036	0.2002	0.1221	0.3213	0.0369	0.5407	0.0379	0.2792	0.0166	0	0
METEOR	0.7871	0.8220	0.7578	0.8383	0.7908	0.8297	0.7093	0.9217	0.4916	0.9089	0.7184	0.9434	0.6944	0.6944

5.2. Analysis by Error Type

We further examined how each model performs across different categories of translation errors. Table 10 reports the average quality scores for each error type as determined through human evaluation.

Table 10. Mean quality scores by error type for baseline and fine-tuned models.

Error Type	ChatGPT-4.1 mini	Fine-Tuned ChatGPT-4.1 mini
Idiomatic Expression	0.700	0.700
Lexical	0.829	0.863
Morphology, Lexical	0.700	1.000
Semantic	0.880	0.880
Semantic, Idiomatic	0.700	0.700
Style	0.950	0.950
Style, Semantic	1.000	1.000
Word Order	0.900	0.900
No Error	0.982	0.985

The findings show that the fine-tuned model demonstrates improvements across multiple error categories. The most significant enhancement occurs in handling morphological and lexical errors, where the fine-tuned model achieves a perfect score of 1.000, compared to 0.700 for the baseline. Improvements are also observed in Lexical errors (0.863 vs. 0.829) and in the No Error category (0.985 vs. 0.982), indicating an overall boost in translation quality.

5.3. Analysis by Sentence Length

We investigated how sentence length influences translation quality for both models. Table 11 reports the correlation coefficients between sentence length and the corresponding quality scores.

Table 11. Correlation between sentence length and quality scores.

Model	Correlation Coefficient
ChatGPT-4.1 mini	−0.064
Fine-tuned ChatGPT-4.1 mini	−0.066

Both models show a slight negative correlation between sentence length and translation quality, suggesting that longer sentences tend to receive marginally lower scores. The both models show a slight negative correlation between sentence length and translation quality, suggesting that longer sentences tend to receive marginally lower scores. The correlation coefficients are nearly identical for the two models (−0.064 for the baseline and −0.066 for the fine-tuned version), indicating that fine-tuning had little effect on how

sentence length influences performance. Overall, the low magnitude of these correlations implies that sentence length has a minimal impact on translation quality for either model.

6. Discussion

6.1. Key Findings Summary

Our experimental evaluation demonstrates that targeted supervised fine-tuning meaningfully improves large language model performance on Kazakh translation post-editing. The fine-tuned ChatGPT-4.1 mini model achieved consistent improvements across all evaluation metrics:

Overall performance: 5.0% relative improvement in human evaluation (0.80 → 0.84), 24% BLEU improvement (0.7002 → 0.8682), and 35% TER reduction (0.1155 → 0.0751)

- Morphological accuracy: 43% relative improvement in morphology/lexical error handling (0.700 → 1.000)
- Domain specialization: Strongest gains in legal (+0.123) and medical (+0.155) domains
- Lexical precision: 4.1% relative improvement in lexical error correction (0.829 → 0.863)

These improvements, while modest in absolute terms, represent meaningful progress given the already strong baseline performance (0.80 human evaluation score). The consistency across automatic and human metrics validates the robustness of our fine-tuning approach.

6.2. Why Fine-Tuning Improves Kazakh Post-Editing

The observed improvements can be attributed to specific mechanisms that address Kazakh's unique linguistic characteristics:

6.2.1. Agglutinative Morphology Handling

Kazakh's agglutinative structure poses significant challenges for translation models. Words are formed by concatenating multiple suffixes to roots, with each suffix encoding grammatical information (case, number, possession, person, tense).

The baseline model frequently produced errors in suffix ordering, vowel harmony violations, or inappropriate suffix selection. Our fine-tuned model, trained on 10,008 corrected examples with explicit morphological error annotations, learned to:

- **Preserve suffix order constraints:** Kazakh requires strict ordering (plural → possessive → case)
- **Apply vowel harmony rules:** Suffixes must harmonize with the root's vowel characteristics (front/back, rounded/unrounded)
- **Select contextually appropriate allomorphs:** Many suffixes have multiple phonologically conditioned variants

This explains the dramatic 43% improvement in morphology/lexical errors, where the model must simultaneously handle correct lexical choice and complex morphological inflection.

6.2.2. Domain-Specific Terminology and Register

The substantial improvements in legal (+17%) and medical (22%) domains reflect successful adaptation to specialized vocabulary and formal registers. These domains in Kazakh feature:

- **Technical borrowings:** Many legal and medical terms are borrowed from Russian or Arabic, requiring knowledge of integration patterns (e.g., Russian loanwords typically take native Kazakh suffixes but may resist certain phonological processes)
- **Formal register conventions:** Professional domains use specific verb forms, honorifics, and sentence structures that differ from colloquial Kazakh

- **Terminology consistency:** Technical domains require precise term selection from multiple near-synonyms

Our domain-diverse training data (covering six distinct styles) exposed the model to these specialized patterns, enabling it to generalize beyond simple word-for-word post-editing to contextually appropriate formal language production.

6.2.3. Lexical Selection in Context

The 4.1% improvement in lexical errors indicates enhanced sensitivity to contextual word choice. Kazakh has multiple words for many concepts, with selection depending on:

- **Semantic nuance:** Subtle meaning differences between near-synonyms
- **Collocational preferences:** Certain words co-occur more naturally
- **Register appropriateness:** Formal vs. informal vocabulary choices
- **Cultural connotations:** Words may carry cultural associations affecting appropriateness

The systematic error-focused training enabled the model to learn these contextual dependencies, reducing inappropriate lexical substitutions that preserved literal meaning but violated natural Kazakh usage patterns.

6.3. Linguistic Insights from Error Analysis

6.3.1. Morphological Processing Success

The perfect performance (1.000) on morphology/lexical errors represents the most significant finding of our evaluation. Analysis of corrected examples reveals the model learned to handle:

- **Case-marking patterns:** Correct application of nominative, accusative, genitive, dative, locative, ablative, and instrumental cases based on syntactic and semantic context
- **Possessive suffix selection:** Appropriate choice among six possessive markers (1st/2nd/3rd person \times singular/plural) considering both semantic possession and syntactic agreement
- **Derivational morphology:** Proper use of productive derivational suffixes that create new lexical items while maintaining semantic coherence

This success contrasts with the baseline's struggles, where morphological errors often cascaded—an incorrect case suffix would propagate through dependent words, creating multiply malformed constructions.

6.3.2. Persistent Challenges in Idiomatic and Stylistic Phenomena

Several error categories showed no improvement, revealing the limitations of our current approach:

- **Idiomatic expressions (0.700 for both models):** Kazakh idioms often involve culturally specific metaphors, livestock-related imagery (reflecting nomadic heritage), or Islamic cultural references. These require cultural knowledge beyond linguistic pattern recognition.
- **Semantic-idiomatic combinations (0.700 for both models):** Cases where literal semantic accuracy conflicts with idiomatic naturalness remain challenging, suggesting the need for specialized training data or explicit cultural annotation.
- **Style errors (0.950 for both models):** Already high baseline performance leaves little room for improvement, but the stability suggests fine-tuning did not enhance stylistic nuance detection.

These persistent challenges indicate that while morphosyntactic phenomena respond well to supervised fine-tuning, cultural and stylistic dimensions may require alternative approaches such as:

- Larger-scale data with explicit cultural annotation
- Integration of cultural knowledge bases
- Preference-based learning methods that capture subjective stylistic judgments

6.3.3. Sentence Complexity Handling

The minimal correlation between sentence length and quality (-0.066) suggests both models handle Kazakh's relatively flexible word order reasonably well. Unlike languages with strict word order constraints, Kazakh allows considerable syntactic variation for information structure and emphasis. The weak negative correlation likely reflects:

- **Increased ambiguity:** Longer sentences introduce more potential attachment sites for modifiers and more complex discourse structure
- **Compound constructions:** Extended sentences often involve clause chaining and multiple embedded structures
- **Computational limits:** Very long sequences may exceed the effective attention span of transformer models

Importantly, fine-tuning did not degrade this capability, indicating our training approach preserves the model's existing strengths while adding new capabilities.

6.4. Methodological Strengths and Limitations

6.4.1. Strengths

Comprehensive evaluation framework: Our multi-metric evaluation (BLEU, METEOR, TER, human assessment) provides robust validation across different quality dimensions, addressing common criticisms of single-metric APE studies.

1. **Systematic error taxonomy:** The seven-category error classification (semantic, lexical, syntactic, morphological, fluency, idiomatic, style) enabled targeted analysis and training, going beyond generic quality improvement.
2. **Domain diversity:** Training data spanning six distinct styles (medical, scientific, journalistic, oral, fiction, legal) promotes robust generalization and prevents overfitting to narrow domain characteristics.
3. **Native speaker expertise:** Annotation by qualified native Kazakh speakers with computational linguistics training ensured high-quality, theoretically grounded corrections.
4. **Consistency in evaluation:** Using a single expert annotator for test set evaluation eliminates inter-annotator variability, though this comes with trade-offs discussed below.

6.4.2. Limitations

1. **Single-annotator evaluation:** While ensuring consistency, reliance on one expert annotator limits evaluation perspectives. Post-editing quality assessment involves subjective judgments about naturalness, style appropriateness, and cultural fit that would benefit from multiple perspectives. Inter-annotator agreement studies would strengthen confidence in our human evaluation metrics.
2. **Dataset scale constraints:** Our training dataset of 10,008 sentences is substantial for Kazakh NLP but may not capture the full range of linguistic phenomena and cultural contexts. The test set of 311 sentences, while carefully stratified across domains and error types, represents a limited sample. Larger-scale evaluation would provide more robust estimates of domain-specific performance.
3. **Domain coverage gaps:** While we include six domains, certain important genres remain underrepresented:
 - a. Government/administrative documents
 - b. Technical manuals and user guides

- c. Social media and informal digital communication
 - d. Literary translation requiring creative adaptation
 - e. These gaps may limit generalization to real-world translation scenarios.
4. **Limited baseline comparisons:** We compared only against the base ChatGPT-4.1 mini model. Comparisons with other large language models (e.g., other GPT variants, Claude, Gemini) or specialized MT systems would provide broader context for our results.
 5. **Error category imbalance:** Some error types (abbreviation expansion, style-semantic combinations) have few examples in the test set, limiting statistical confidence in category-specific conclusions.
 6. **Computational accessibility:** Fine-tuning GPT-4.1 mini requires significant computational resources (GPU memory, training time, API costs), potentially limiting reproducibility and accessibility for resource-constrained research environments.
 7. **Absence of COMET scores:** While we report BLEU, METEOR, and TER, we did not include COMET (Crosslingual Optimized Metric for Evaluation of Translation), which has shown superior correlation with human judgments in recent MT evaluation research. Future work should incorporate this metric.
 8. **Domain-specific bias potential:** Strong performance on legal and medical domains may reflect greater availability of these text types in the training data rather than inherent model capabilities. Analyzing the training data distribution would clarify this concern.

6.5. Implications for Low-Resource Language Processing

Our findings carry several important implications for the broader field of low-resource language NLP:

6.5.1. Fine-Tuning Viability for Low-Resource Languages

The consistent 4–9% improvements across metrics demonstrate that supervised fine-tuning remains effective for enhancing LLM performance on low-resource languages, even when baseline performance is already substantial. This challenges the assumption that large pre-trained models have reached performance plateaus that cannot be meaningfully improved through specialized training.

For low-resource language communities, this suggests that investing in high-quality post-editing annotation (even at relatively modest scales of ~10,000 examples) can yield measurable improvements in translation quality. The cost–benefit trade-off favors this approach compared to training models from scratch or waiting for organic improvements through general pre-training.

6.5.2. Error-Focused Training Value

Our error-categorized training approach proved more effective than undifferentiated quality improvement methods. By explicitly annotating error types (morphological, lexical, semantic, etc.), we enabled the model to learn specific patterns of common mistakes and their corrections rather than relying on implicit quality signals.

This has practical implications for annotation guidelines in low-resource settings: annotation efforts should prioritize systematic error identification and categorization over simple quality rating or minimal edits. The marginal cost of detailed error annotation appears justified by the resulting performance gains, particularly for morphologically complex languages.

6.5.3. Domain Specialization Benefits

The notable improvements in legal (+17%) and medical (+22%) domains underscore the value of domain-specific fine-tuning. These specialized domains feature technical terminology, formal registers, and discourse conventions that may be underrepresented in general pre-training corpora.

For practical applications in low-resource language contexts (legal document translation, medical information access, technical documentation), domain-adapted models trained on even modest amounts of specialized data can significantly enhance usability. This is particularly important for Kazakh and similar languages where domain-specific translation services are limited.

6.5.4. Morphological Complexity Addressable Through Fine-Tuning

The 43% improvement in morphology/lexical errors demonstrates that current fine-tuning approaches can effectively address complex agglutinative structures characteristic of Turkic languages. This is encouraging for other morphologically rich, low-resource languages (e.g., other Turkic languages, Uralic languages, many Indigenous languages).

The success suggests that:

- Morphological complexity, while challenging, is not an insurmountable barrier for LLM adaptation
- Systematic exposure to corrected morphological patterns enables models to internalize productive morphological rules
- Agglutinative languages may benefit more from fine-tuning than isolating languages, as morphological errors are both more common and more amenable to systematic correction

6.5.5. Multilingual Transfer Potential

While not directly tested in this work, our results suggest promising directions for multilingual transfer learning. Kazakh shares typological features with other Turkic languages (Kyrgyz, Uzbek, Turkish, Azerbaijani, Tatar, Turkmen). Fine-tuning approaches that successfully address Kazakh morphology may transfer to these related languages, potentially requiring less language-specific data through shared morphological patterns.

6.5.6. Contextualizing Improvements in Post-Editing Research

Although the 5% relative improvement observed in our study may appear modest, it represents a meaningful advancement given the high baseline performance of large language models. The baseline ChatGPT-4.1 mini achieved a mean quality score of 0.80, indicating a strong existing post-editing capability, which makes additional gains inherently difficult. When compared to previous work in automatic post-editing (APE), our final mean score of 0.84 is competitive, particularly for a low-resource language like Kazakh.

Additionally, in our study the translations were evaluated using the automatic metrics (BLEU, TER and METEOR) the baseline ChatGPT-4.1 mini achieved a BLEU of 0.7002, TER of 0.1155 and METEOR of 0.7590, while the fine-tuned model reached a BLEU of 0.8682, TER of 0.0751 and METEOR of 0.8500. These results reinforce that fine-tuning not only improves overall average quality, but also delivers substantial gains on standard reference-based evaluation metrics.

Prior large-scale APE benchmarks, such as the WMT shared tasks, highlight the difficulty of achieving improvements on strong baselines. For instance, Chatterjee et al. [25] reported absolute Translation Edit Rate (TER) reductions of 1–3% for English–German APE in WMT 2020, while Grundkiewicz et al. [44] achieved similar gains using transformer-based approaches. Negri et al. [26] also observed that TER improvements beyond 2–4% are

uncommon when starting from high-quality machine translation outputs. Studies on APE for low-resource or morphologically rich settings report more modest, but meaningful improvements. Lee et al. [45] employed a multi-source transformer and achieved -0.73 TER и $+1.49$ BLEU. Similarly, Lee [46] used a cross-lingual transformer with word-level and sentence-level quality estimation, improving the baseline by -3.95 TER and $+4.50$ BLEU.

Our findings align with these, suggesting that incremental gains of 3–5% on relevant metrics are typical when fine-tuning on morphologically complex, specialized post-editing datasets.

Lankford et al. [47] demonstrated that in English-to-Irish translation (a low-resource language), the adapted adaptMLLM model outperformed the baseline LoResMT2021 system by 5.2 BLEU, which corresponds to a 14% improvement. Lankford et al. [47] used the EN \rightarrow GA language pair to train their models with a small corpus of 13,171 parallel sentences, similar to our work.

Fine-tuning a GPT model for low-resource language pairs significantly improves translation quality according to both automatic and human-centered metrics. For the English \rightarrow Kazakh direction, the fine-tuned models achieved a +24% increase in BLEU and a -35% reduction in TER, while human evaluation on a 0–1 scale improved by approximately 5%. Compared to previous studies, the results show the largest TER reduction, competitive BLEU gains, and measurable improvements in human-perceived translation quality, confirming the effectiveness of specialized fine-tuning for low-resource languages.

6.6. Future Research Directions

6.6.1. Scale and Data Expansion

1. **Increased training scale:** Expanding the training dataset to 50,000+ examples across more diverse domains could yield additional improvements. Key priorities include:
 - a. Government and administrative texts
 - b. Informal social media language
 - c. Technical documentation and user manuals
 - d. Conversational dialog from various registers
2. **Active learning integration:** Implementing active learning strategies to identify the most informative examples for annotation could improve training efficiency and reduce annotation costs. The model could flag uncertain cases or systematically sample from underrepresented linguistic phenomena.

6.6.2. Multilingual Turkic Transfer Learning

Applying similar fine-tuning approaches to related Turkic languages could:

- Validate the generalizability of our approach across the Turkic language family
- Enable cross-linguistic transfer learning that reduces data requirements for each individual language
- Provide insights into which morphological and syntactic patterns transfer across related languages vs. require language-specific training

Promising candidates include:

- Kyrgyz: Very close to Kazakh, with high mutual intelligibility
- Uzbek: Significant structural similarity but different writing systems (Latin vs. Cyrillic)
- Turkish: More distantly related but with substantial resources for comparison

Expected benefits for Kazakh: Preference-based learning is particularly relevant because linguistic quality cannot be fully captured by accuracy metrics alone. Subtleties such as case harmony, agglutinative suffix selection, and culturally appropriate lexical choice

often require nuanced native speaker judgments. Training on preference data could help the model internalize these subtleties and produce outputs that are not only grammatically correct but also idiomatically natural and culturally appropriate. For example, similar to how Proximal Policy Optimization (PPO) stabilizes incremental improvements in reinforcement learning, preference-based fine-tuning can enable gradual, reliable enhancement of model outputs without destabilizing previously learned structures [48].

Challenges: Implementation must address high annotation costs for Kazakh (limited pool of qualified annotators), risk of reward model bias from limited preference data, and careful balancing between supervised objectives and preference optimization to avoid catastrophic forgetting of rare grammatical structures.

6.6.3. Multimodal Context Integration

Incorporating visual or cultural context information could help address culturally specific translation challenges:

- **Image-grounded translation:** For documents with accompanying images, visual context could disambiguate cultural references or provide clarification for metaphorical language
- **Cultural knowledge bases:** Integrating structured knowledge about Kazakh culture, history, and social practices could inform translation choices involving culturally laden terminology
- **Cross-cultural adaptation:** Developing methods to explicitly model source-target cultural differences and adapt translations accordingly

6.6.4. Explanation and Interpretability

Developing methods to explain why the model makes specific post-editing decisions would:

- Increase trust and adoption in professional translation workflows
- Enable iterative improvement through targeted feedback on decision-making processes
- Help identify systematic biases or errors that aggregate metrics might miss
- Facilitate training of human post-editors by demonstrating the reasoning behind corrections

6.6.5. Real-World Application Development

- Translating research findings into practical applications requires:
- Integration with CAT tools: Incorporating the fine-tuned model into Computer-Assisted Translation (CAT) tools used by professional translators
- Interactive post-editing interfaces: Developing user interfaces that present post-editing suggestions with confidence scores and allow efficient human review
- Domain adaptation workflows: Creating streamlined processes for adapting the model to new specialized domains with minimal additional training data
- Quality estimation: Building automatic quality estimation systems that predict when post-editing is needed vs. when machine translation can be used directly

6.6.6. Linguistic Documentation and Analysis

Our work could contribute to broader Kazakh language documentation:

- **Error corpus creation:** The annotated errors represent a valuable resource for understanding common translation difficulties and could inform Kazakh language pedagogy
- **Morphological pattern analysis:** Systematic analysis of the morphological corrections could reveal productive patterns in contemporary Kazakh morphology
- **Cross-dialectal variation:** Extending the approach to different Kazakh dialects (Standard Kazakh, Southern dialects, Northeastern variants) could document dialectal differences

7. Conclusions

This study provides a systematic examination of fine-tuning large language models for post-editing Kazakh machine translations, addressing a significant gap in NLP for low-resource and morphologically complex languages. We introduced KazPE, a comprehensive dataset comprising 10,008 training sentences and 311 test sentences, all annotated with detailed error information. This resource represents the first large-scale benchmark for evaluating Kazakh translation quality systematically.

Our experiments show that targeted fine-tuning of GPT-4.1 mini leads to meaningful gains in translation quality, achieving an average human-evaluated score of 0.84 compared to 0.80 for the baseline model. Improvements were observed across multiple domains, with especially notable gains in specialized fields, such as legal texts (+0.123) and medical texts (+0.155), highlighting both the robustness and practical applicability of our approach.

Error analysis provides insight into how fine-tuning addresses specific Kazakh linguistic phenomena. The most pronounced improvement occurred in the morphology/lexical category, where the fine-tuned model reached perfect performance, compared to the baseline's 0.700. This demonstrates enhanced capability in managing Kazakh's complex agglutinative morphology—a result that is particularly relevant for other Turkic languages with similar morphological structures.

In addition to assessments by linguistic specialists, the translation was also evaluated using the automatic metrics—BLEU, TER, METEOR. In the result, the largest gains were observed in specialized domains: the BLEU score increased from 0.6073 to 0.9019 in the medical domain and from 0.6410 to 0.9608 in the scientific style. Meanwhile, the lowest TER values (0.0166 and 0.0379 in the scientific and legal domains respectively) further underscore the enhanced precision of the fine-tuned model. The substantial increase in BLEU and METEOR, combined with the reduction in TER, demonstrate that the fine-tuned model delivers significantly higher accuracy relative to reference texts especially in medical and scientific topics. In summary, the fine-tuning process yields noticeable improvements in the model's performance within specific domains, reaching near-optimal BLEU values in the scientific field and minimal TER values, which indicate minimal editing effort.

Our methodology introduces several innovations for low-resource language NLP. The comprehensive error taxonomy, covering semantic, lexical, syntactic, morphological, and fluency dimensions, offers a replicable framework for systematic translation evaluation. Additionally, the inclusion of diverse domains (medical, scientific, journalistic, oral, fiction, and legal) ensures broad linguistic coverage while retaining real-world relevance.

This work makes several key contributions to computational linguistics:

1. The creation and public release of the first systematically annotated Kazakh post-editing dataset with detailed error categorization.
2. Demonstration of effective fine-tuning strategies for morphologically rich, low-resource languages.
3. Establishment of baseline performance metrics to support future Kazakh NLP research.
4. Provision of practical insights for improving machine translation systems for Turkic languages.

Looking ahead, this study lays the groundwork for continued progress in Kazakh and related language processing. The consistent improvements across domains and error categories indicate that the approach can be adapted to other low-resource languages. Success in managing morphological complexity suggests promising directions for tackling similar challenges in agglutinative language families globally.

By releasing our annotated dataset, trained models, and evaluation frameworks, we aim to empower the research community to build on these findings and advance innovation in low-resource language technology. As large language models continue to evolve, our

systematic fine-tuning approach serves as a valuable template for extending state-of-the-art capabilities to underrepresented languages, promoting more inclusive and equitable NLP solutions.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/a19030199/s1>.

Author Contributions: Conceptualization, D.R. and A.B.; methodology, A.Z. and A.S.; software, A.Z.; validation, N.Z.; formal analysis, A.Y.; investigation, D.O.; resources, M.T.; data curation, A.Z.; writing—original draft preparation, A.B.; writing—review and editing, D.R.; visualization, A.S.; supervision, A.Y.; project administration, N.Z.; funding acquisition, D.O. and M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992875).

Data Availability Statement: The datasets related to this research are available on GitHub by the following link: Available at: <https://drive.google.com/drive/folders/11BZ6lyFOTew0pc3AAx9Jib92w28CCfBZ?usp=sharing> (accessed on 18 February 2026). or Available at: <https://github.com/Aliya19942204/DATASET-.git> (accessed on 18 February 2026).

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

GPT	Generative Pre-trained Transformer
LLM	Large Language Model
TER	Translation Edit Rate
METEOR	Metric for Evaluation of Translation with Explicit Ordering
BLEU	Bilingual Evaluation Understudy
Sem	Semantics
Lex	Lexical
Morph	Morphology
Term	Terminology
WO	Word Order
Gram	Grammar
Ortho	Orthography
Idiom	Idiomatic
NLP	Natural Language Processing
FFN	Feed-Forward Network
MT	Machine Translation

References

1. Ethnologue. Languages of the World. 2023. Available online: <https://www.ethnologue.com/> (accessed on 18 February 2026).
2. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The state and fate of linguistic diversity and inclusion in the nlp world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6282–6293.
3. Washington, J.N.; Ipasov, M.; Tyers, F.M. A finite-state morphological transducer for Kazakh. In Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, Donostia, Spain, 23–25 July 2012; pp. 49–53.
4. Altenbek, G.; Wang, X.-L. Kazakh segmentation system of inflective affixes. In Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, 28–29 August 2010; pp. 183–190.
5. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
6. OpenAI. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]

7. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **2022**, *24*, 1–113.
8. Zhang, W.; Xu, H.; Tu, H.; Murray, K.; Koehn, P.; Murray, K.; Koehn, P. How multilingual is multilingual bert? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 4191–4197.
9. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Bali, Indonesia, 14 October 2023; pp. 675–718.
10. Allen, J.; Hogan, C. Toward the development of a post-editing module for raw machine translation output: A controlled language perspective. In Proceedings of the Machine Translation Summit IX, New Orleans, LA, USA, 18–22 September 2003; pp. 17–24.
11. Simard, M.; Goutte, C.; Isabelle, P. Statistical phrase-based post-editing. In Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, USA, 23–25 April 2007; pp. 508–515.
12. Kartbayev, A. Machine translation of agglutinative languages using neural networks. In Proceedings of the International Conference on Computer Processing of Turkic Languages, Kazan, Russia, 18–21 October 2017; pp. 87–96.
13. Xu, H.; Kim, Y.J.; Sharaf, A.; Awadalla, H.H. A paradigm shift: Efficient fine-tuning of large language models for machine translation. *arXiv* **2023**, arXiv:2305.13105.
14. Makazhanov, A.; Sultangazina, A.; Makhambetov, O.; Sabyrgaliyev, I.-L. An open source Kazakh speech synthesis system. In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 1491–1495.
15. Toral, A.; Edman, L.; Yeshmagambetova, G.; Spenader, J. Neural machine translation for English–Kazakh with morphological segmentation and synthetic data. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Florence, Italy, 1–2 August 2019; pp. 386–392, Association for Computational Linguistics.
16. Koehn, P.; Knowles, R. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 4 August 2017; pp. 28–39.
17. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Vi'egas, F.; Wattenberg, M.; Corrado, G.; et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [[CrossRef](#)]
18. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021; pp. 483–498.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Aji, A.F.; Bogoychev, N.; Heafield, K.; Sennrich, R. In neural machine translation, what does transfer learning transfer? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7701–7710.
21. Yeshpanov, R.; Polonskaya, A.; Varol, H. KazParC: Kazakh Parallel Corpus for Machine Translation. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 9633–9644.
22. Junczys-Dowmunt, M.; Grundkiewicz, R. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Stockholm, Sweden, 13–19 July 2018; pp. 120–129.
23. Correia, G.M.; Martins, F.T. A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3050–3056.
24. Yang, H.; Wang, M.; Wei, D.; Shang, H.; Guo, J.; Li, Z.; Lei, L.; Qin, Y.; Tao, S.; Sun, S.; et al. HW-TSC's Participation at WMT 2020 Automatic Post Editing Shared Task. In Proceedings of the Fifth Conference on Machine Translation, Online, 19–20 November 2020; pp. 797–802.
25. Chatterjee, R.; Negri, M.; Rubino, R.; Turchi, M. Findings of the WMT 2020 shared task on automatic post-editing. In Proceedings of the Fifth Conference on Machine Translation, Online, 19–20 November 2020; pp. 646–659.
26. Negri, M.; Turchi, M.; Chatterjee, R.; Bertoldi, N. ESCAPE: A large-scale synthetic corpus for automatic post-editing. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 4784–4788.
27. Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y.J.; Afify, M.; Awadalla, H.H. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv* **2023**, arXiv:2302.09210. [[CrossRef](#)]
28. Jiao, W.; Wang, W.; Huang, J.-T.; Wang, X.; Tu, Z. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv* **2023**, arXiv:2301.08745. [[CrossRef](#)]

29. Li, Z.; Li, X.; Du, M.; Jiang, J. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv* **2023**, arXiv:2304.04675. [[CrossRef](#)]
30. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. Exploring the use of large language models for multilingual machine translation. *arXiv* **2023**, arXiv:2305.04662.
31. Vilar, D.; Freitag, M.; Cherry, C.; Luo, J.; Ratnakar, V.; Foster, G. Prompting palm for translation: Assessing strategies and performance. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 7264–7278.
32. Rafailov, E.; Shi, P.; Zelikman, E.; Ouyang, L.; Chris-tiano, P.; Amodei, D. Direct preference optimization: Your language model is secretly a reward model. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023.
33. Kocmi, T.; Bojar, O. Fine-tuning on clean data for end-to-end speech translation: FBK@ IWSLT 2018. *Nat. Lang. Eng.* **2020**, *26*, 727–737.
34. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.* **2021**, *1*, 3.
35. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 4582–4597.
36. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
37. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. *arXiv* **2021**, arXiv:2109.01652. [[CrossRef](#)]
38. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
39. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2020**, arXiv:1904.09675. [[CrossRef](#)]
40. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. Comet: A neural framework for mt evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 2685–2702.
41. Christiano, P.F.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
42. Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1460–1474. [[CrossRef](#)]
43. Krippendorff, K. Computing Krippendorff’s Alpha-Reliability. In *Departmental Papers*; University of Pennsylvania: Philadelphia, PA, USA, 2011; p. 43.
44. Grundkiewicz, R.; Junczys-Downmunt, M.; Heafield, K. Neural automatic post-editing of machine translation output. In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1–2 August 2019; pp. 188–199.
45. Lee, W.; Shin, J.; Lee, J.-H. Transformer-based Automatic Post-Editing Model with Joint Encoder and Multi-source Attention of Decoder. In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1–2 August 2019; pp. 112–117.
46. Lee, D. Cross-Lingual Transformers for Neural Automatic Post-Editing. In Proceedings of the Fifth Conference on Machine Translation, Online, 19–20 November 2020; pp. 772–776.
47. Lankford, S.; Afli, H.; Way, A. adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds. *Information* **2023**, *14*, 638. [[CrossRef](#)]
48. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.