Large Language Model-Based Post-OCR Correction for Low-Resource Kazakh Scripts

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Kazakh is written in the Arabic, Cyrillic, and Latin script which present unique challenges for OCR and post-OCR correction research. Despite this complexity, NLP research on Kazakh and its low-resource scripts remains extremely scarce. We analyze common OCR error patterns in all three Kazakh scripts using Tesseract and evaluate four large language models (LLMs) for post-OCR correction using minimal, confusion-aware, and few-shot prompting strategies. Our results reveal three systematic, writing-system-driven failure modes in LLM-based post-OCR correction: script switching, hallucination, and instruction-following breakdown. Arabic script post-OCR correction remains unsuccessful across all setups. In the Cyrillic script, post-OCR correction improvements are minimal due to the high baseline OCR performance on Cyrillic. For the Latin script, few-shot prompting with Gemini 2.5 Flash yields substantial improvements, reducing CER by 8.58 points and WER by 32.49 points to levels better than high-resource Kazakh Cyrillic script OCR. These findings demonstrate that LLM post-OCR correction failure modes are predictable from writing system properties such as script resource asymmetry and co-existing script dominance and demonstrate the need for typology-aware evaluation frameworks for multi-script and under-resourced languages.

Keywords: OCR, Kazakh, post-OCR correction, low-resource scripts, writing system typology

1. Introduction

Kazakh is a rare language in the sense that three different scripts are used for Kazakh writing, making optical character recognition (OCR) challenging. Arabic, Cyrillic, and Latin scripts are each used in Kazakh in different geographic regions (Honkasalo and Temirbekova, 2024). Due to this script diversity, natural language processing (NLP) tools and research have focused predominantly on Kazakh Cyrillic, leaving the Arabic and Latin scripts significantly underrepresented.

Kazakh has undergone multiple script transitions driven by political and cultural forces. Before Soviet rule, Kazakh was written in the Arabic script (Honkasalo and Temirbekova, 2024; Batyrbekkyzy et al., 2018). Under Soviet policy, in 1928, Kazakh was changed to the Latin script for the first time, and in 1940, Kazakh switched from the Latin to the Cyrillic script, which is currently the dominant script in practice for Kazakh (Batyrbekkyzy et al., 2018). Currently, in China, Kazakh is written using a modified Arabic alphabet, which is taught in schools. People also use an informal Latin script in casual contexts, differing from official Latin orthographies. In 2017, the first proposal for a new Latinized alphabet for Kazakh was created (Honkasalo and Temirbekova, 2024). Initially, a switch to the Latin script in Kazakhstan was designated for completion in 2025, but this timeline has been postponed to 2031. The majority of Kazakhstan residents believe that the Latin script is the best script for writing Kazakh, although most residents do not use the Latin script in practice (Honkasalo and Temirbekova, 2024). The sociolinguistic history of Kazakh shapes the asymmetry in script NLP resources, with current work and resources being predominantly focused on the Cyrillic script (Gagnier et al., 2026). This complex digraphic state of Kazakh must be acknowledged and further researched in the context of NLP.

Optical character recognition (OCR) is the process of digitizing a document image into its constituent characters (Bunke and Wang, 1997). Many texts resulting from OCR are noisy and need to be post-corrected. Post-correction can be done manually, using isolated-word approaches, feature-based machine learning models, sequence-to-sequence models, and language models (Nguyen et al., 2021). Prior work has shown that general-purpose OCR tools are not robust to data-scarce endangered language settings, and that post-correction methods must be tailored accordingly (Rijhwani et al., 2020). Large language models (LLMs) have recently emerged for post-OCR error correction in high-resource languages (Koynov and Doan, 2025). Danilova and Aangenendt (2025) found that German Mistral enhanced the OCR output, decreasing word error rate (WER) and character error rate (CER) remained unchanged or decreased. Kanerva et al. (2025) found that LLMs show promise for reducing CER in English, while in Finnish, a practically useful performance was not reached. LLM-based

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 79–88
12 May, 2026.

post-OCR correction is an emerging and important task, but even in well-resourced languages like Finnish, it is difficult, suggesting great challenges in under-resourced settings such as low-resource Kazakh scripts.

Work in Kazakh OCR has primarily focused on the Cyrillic script, reflecting the abundance of digitized Kazakh Cyrillic script resources compared to the Arabic and Latin scripts. Yeleussinov et al. (2023) worked on improving Kazakh OCR accuracy in the Cyrillic script using generative adversarial network (GAN) models. Nurseitov et al. (2021) and Toiganbayeva et al. (2022), respectively, created HKR and KOHTD, which are both large datasets in the Cyrillic script for handwritten text recognition. Kamshat et al. (2024) surveyed OCR, NLP, and speech recognition techniques for Kazakh as a low-resource language, reporting an OCR model achieving 85% accuracy. The Kazakh Arabic script and the Kazakh Latin script both use additional letters, modified graphemes, and different orthographic conventions, making it different from high-resource languages that use the Arabic and Latin scripts. Recently, the low-resource Latin and Arabic have been studied. (Gagnier et al., 2026) constructed KazakhOCR, a synthetic benchmark for OCR in Kazakh in all three Kazakh scripts, and found that models have high error rates in the Kazakh Arabic and Latin scripts. To our knowledge, this is the first work on post-OCR correction for Kazakh in any script, and work in OCR for low-resource Kazakh scripts is emerging.

Work on Kazakh OCR and post-OCR correction is necessary but very scarce, especially for low-resource Kazakh scripts. We present four main contributions in this paper: (1) the first study of post-OCR correction across all three Kazakh writing systems, (2) LLM-based post-OCR correction is unreliable for low-resource and multi-script settings, even with confusion-aware prompting, (3) script switching and hallucination are writing-system-driven failure modes not captured by standard OCR metrics, and (4) few-shot prompting can overcome challenges in Latin Kazakh, but does not generalize across scripts. The purpose of this paper is to (1) identify issues and challenges in Kazakh OCR and (2) evaluate LLMs for Kazakh post-OCR correction. We also aim to broaden script coverage in Kazakh NLP and encourage future work on the computational modeling of under-resourced writing systems.

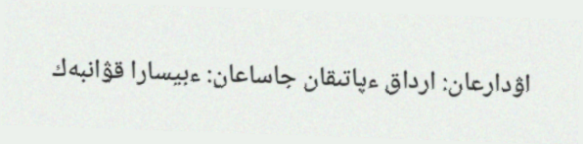

Figure 1: Example image from the KazakhOCR benchmark in the Arabic script

2. KazakhOCR

We use the KazakhOCR benchmark[1] (Gagnier et al., 2026), which consists of 7,219 synthetically created images for Kazakh OCR in all three Kazakh scripts with variations to increase authenticity and representativeness of tasks. Texts are typically medium-length, with average lengths of all texts between 700 and 750 characters in each script. In this benchmark, multimodal large language models have CERs of 0.355 to 0.725, 0.216 to 0.310, and 0.053 to 0.257 in Arabic, Latin, and Cyrillic scripts, respectively, while Tesseract has CERs of 0.150, 0.114, and 0.043 in Arabic, Latin, and Cyrillic scripts, respectively. As traditional OCR approaches perform significantly better than current multimodal large language models or MLLMs (particularly Gemma-3-12B-it, Qwen2.5-VL-7B-Instruct, and Llama-3.2-11B-Vision-Instruct) for Kazakh OCR (Gagnier et al., 2026), we choose to use Tesseract for experiments in this paper.

The three Kazakh scripts differ in visual form and typological properties, which make OCR challenging for Kazakh. The Kazakh Arabic script is an abjad script in which short vowels are often omitted and represented by diacritic marks (Daniels and Bright, 1996), creating character-level ambiguity that increases OCR errors and limits post-correction models. The Kazakh Cyrillic script is an alphabet adapted to the Kazakh phonology, including letters not present in Russian. The Kazakh Latin script is a script that encodes Kazakh phonemes through dense diacritics, creating a high density of marks that are fragile with OCR and poorly represented in NLP data. These three scripts are not equally represented in digital text, as Cyrillic Kazakh makes up the majority of text and LLM training data, while Latin and Arabic Kazakh are under-resourced. This asymmetry created LLM biases across Kazakh scripts and is an essential context for work on Kazakh in NLP and interpreting post-OCR correction results.

[1] https://huggingface.co/datasets/henrygagnier/kazakh-ocr

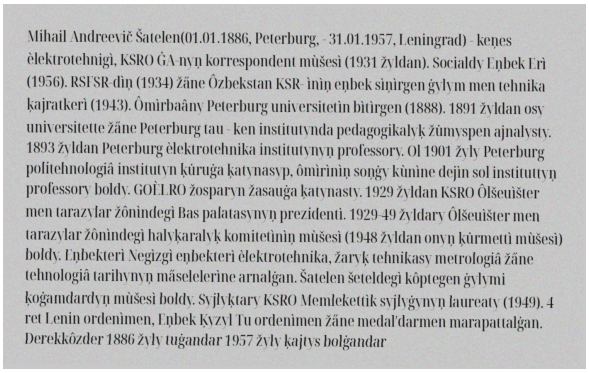
Mihail Andreevič Šatelen(01.01.1886, Peterburg, - 31.01.1957, Leningrad) - keŋes èlektrotehnigì, KSRO GA-nyŋ korrespondent mùšesì (1931 žyldan). Socialdy Eŋbek Erì (1956). RSFSR dìŋ (1934) žäne Özbekstan KSR- inìŋ eŋbek sìŋirgen ġylym men tehnika ķajratkerì (1943). Ômìrbaâny Peterburg universitetìn bitìrgen (1888). 1891 žyldan osy universitette žäne Peterburg tau - ken institutynda pedagogikalyķ žùmyspen ajnalysty. 1893 žyldan Peterburg èlektrotehnika institutynyŋ professory. Ol 1901 žyly Peterburg politehnologiâ institutyŋ ķùruğa ķatynasyp, ômìriniŋ soŋġy kùnìne deјìn sol instituttyŋ professory boldy. GOÈLRO žosparyn žasauğa ķatynasty. 1929 žyldan KSRO Ölšeuišter men tarazylar žônindegì Bas palatasynyŋ prezidenti. 1929-49 žyldary Ölšeuišter men tarazylar žônindegi halyķaralyķ komitetìniŋ mùšesì (1948 žyldan onyŋ ķùrmetti mùšesi) boldy. Eŋbekterì Negizgì eŋbekterì èlektrotehnika, žaryķ tehnikasy metrologiâ žäne tehnologiâ tarihynyŋ mäselelerìne arnalğan. Šatelen šeteldegi kòptegen ġylymi ķoğamdardyŋ mùšesì boldy. Svjlyķtary KSRO Memlekettik svjlyġynyŋ laureaty (1949). 4 ret Lenin ordenimen, Eŋbek Kyzyl Tu ordenimen žäne medal'darmen marapattalğan. Derekközder 1886 žyly tuğandar 1957 žyly ķajtys bolġandar

Figure 2: Example image from the KazakhOCR benchmark in the Latin script

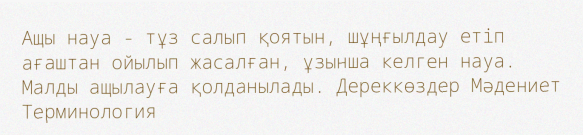
Ащы наура - тұз салып қоятын, шұңғылдау етіп ағаштан ойылып жасалған, ұзынша келген наура. Малды ащылауға қолданылады. Дереккөздер Мәдениет Терминология

Figure 3: Example image from the KazakhOCR benchmark in the Cyrillic script

3. Methodology

We perform OCR using Tesseract, analyze script-specific OCR error patterns, and evaluate large language model (LLM)-based post-OCR correction using multiple prompting strategies.

3.1. Optical Character Recognition

We use Tesseract (Smith, 2007) for OCR in this paper. While Tesseract supports Arabic, Cyrillic, and Latin script recognition through separate trained models, it provides a Kazakh-specific language model only for the Cyrillic script, with no Kazakh-specific support for the Arabic or Latin scripts. To enable consistent comparison across all three scripts under equivalent conditions, we conduct OCR using the generic script-level configurations throughout on 200 images in each script (600 images total). Tesseract adds new lines based on the position of the text in the image, so we remove all new lines from the text before further experiments. Prior to all metric computation, we verified that all ground truth and Tesseract output strings are already in NFC form, as Unicode normalization had no effect on reported metrics. We use 150 images in each script for detecting common errors in Kazakh OCR and developing confusion-aware prompts, and we use 50 images in each script as a test set for LLM-based post-OCR correction.

3.2. Post-OCR Correction

Recently, LLMs have shown promise for post-OCR error correction in high-resource languages (Thomas et al., 2024). We evaluate four LLMs for post-OCR correction: GPT-4o-mini (OpenAI, 2024), DeepSeek v3.2 (DeepSeek-AI, 2025), Claude 3.5 Haiku (Anthropic, 2024), and Gemini 2.5 Flash (Google DeepMind, 2025). In all models, we set the `temperature` to 0.0. All 50 test images per script are submitted for post-correction regardless of OCR output quality, reflecting a realistic deployment scenario. We evaluate all models on post-OCR correction using three different prompt types. We provide the exact prompt text below.

- **Minimal Prompt:** In the minimal prompt, we provide a short instruction to correct OCR errors. We also provide the language and script of the text, and instruct the model to return the text in the original script.

- **Confusion-Aware Prompt:** Using our error analysis, we develop a prompt informing the model of common OCR errors, including deletions, substitutions, and insertions. We again provide the language and script of the text and instruct the model to return the text in the original script.

- **Few-shot Prompt:** We give the model three randomly chosen examples of an OCR output in the same script and the corresponding ground truth correction and instruct the model to correct OCR errors.

Minimal Prompt

You are correcting OCR errors in Kazakh text written in {script} script.
Original OCR output: {ocr_output}
Fix spelling and OCR recognition errors while preserving the original meaning. Return only the corrected text in the {script} script.

Arabic Script Confusion-Aware Prompt

You are correcting OCR errors in Kazakh text written in Arabic script.

Task: Correct OCR errors while preserving the original meaning, word boundaries, and punctuation usage appropriate for Kazakh.

Observed OCR error patterns in this dataset: Character substitutions: - ’،’ → ’.’, ’,’, ’»’, ’›’ - ’ى’ → ’ی’ or deleted - ’ي’ → ’ی’ - ’ك’ → ’ک’ - ’ہ’ → ’ه’ - ’پ’ → ’ب’

Frequently deleted characters: - ,’ى’ ’ ’, ,’ا’ ,’ن’ ’ه’

Frequently inserted characters: - ’ ’, ,’ه’ ,’ى’ ’ى’

Original OCR output: {ocr_output}

Return only the corrected Kazakh text in the Arabic script. Do not include explanations or metadata.

Latin Script Confusion-Aware Prompt

You are correcting OCR errors in Kazakh text written in Latin script.

Task: Correct character-level OCR errors while preserving the original meaning, word boundaries, and punctuation usage appropriate for Kazakh.

Observed OCR error patterns in this dataset: Character substitutions: - ’ì’ → ’i’ - ’k’ → ’ķ’ - ’n’ → ’ŋ’ - ’a’ → ’á’ - ’ģ’ → ’g’ - ’u’ → ’ú’, ’ů’ - Combining marks (◌̵, ◌̋, ◌̃) removed or altered

Frequently deleted characters: - ◌̵, ’ ’, ◌̋, ◌̃, ’u’

Frequently inserted characters: - ’i’, ’ ’, ’ķ’, ’ŋ’, ’g’

Original OCR output: {ocr_output}

Return only the corrected Kazakh text in the Latin script. Do not include explanations or metadata.

Cyrillic Script Confusion-Aware Prompt

You are correcting OCR errors in Kazakh text written in Cyrillic script.

Task: Correct OCR errors while preserving the original meaning, word boundaries, and punctuation usage appropriate for Kazakh.

Observed OCR error patterns in this dataset: Character substitutions: - ’ғ’ → ’г’ - ’ұ’ → ’ү’, ’у’ - ’ө’ → ’о’ - ’ң’ → ’ц’, ’н’ - ’н’ → ’п’ - ’ж’ → ’с’ - ’—’ → ’–’

Frequently deleted characters: - ’ ’, ’З’, ’—’, ’.’, ’Г’

Frequently inserted characters: - ’ ’, ’.’, ’ə’, ’с’, ’,’

Original OCR output: {ocr_output}

Return only the corrected Kazakh text in the Cyrillic script. Do not include explanations or metadata.

Few-Shot Prompt

Correct OCR errors in Kazakh {script} text. Here are examples:

{examples}

Now correct this text and return it in the {script} script: OCR: {ocr_output} Correct:

We apply the Wilcoxon signed-rank test to determine if differences between the baseline OCR text and the post-corrected text are significant, as this test is appropriate for small paired samples. We also compute the 95% bootstrap confidence interval on the mean improvement to quantify effect size and uncertainty. Each bootstrap resample draws 50 pairs with replacement from these observations and computes the mean difference, and repeating this 10,000 times yields an empirical distribution from which the 2.5th and 97.5th percentiles form the confidence interval.

4. Results

4.1. Optical Character Recognition

4.1.1. Tesseract Performance

We first look at the performance of Tesseract (Table 1), which presents the baseline OCR performance across all three Kazakh scripts. Tesseract achieves the best metrics in the Cyrillic script with a CER of 0.029 and a WER of 0.090. Performance is substantially worse in the Latin and Arabic scripts. In the Latin script, Tesseract has a CER of 0.102, and in the Arabic script, Tesseract has a CER of 0.142. These results highlight that the Arabic and Latin scripts require post-OCR correction for usability and that significant challenges remain in Kazakh OCR. All experiments use script-level rather than language-specific Tesseract configurations to enable consistent comparison across all three scripts, as no Kazakh-specific models exist for the Arabic and Latin scripts.

Script	CER	WER
Arabic	0.142	0.376
Latin	0.102	0.399
Cyrillic	0.029	0.090

Table 1: OCR Performance metrics Tesseract across different low-resource Kazakh scripts. CER = Character Error Rate, WER = Word Error Rate

4.1.2. Common OCR Errors

We now look at the most common substitutions, insertions, and deletions in Kazakh OCR by Tesseract generated from a subset of 150 images in each script (Tables 2 and 3). We use this analysis to inform models in confusion-aware prompts.

In the Kazakh Arabic script, errors are primarily from confusion between visually similar characters and punctuations, such as comma variants and substitutions involving ى (Alef Maksura, U+0649), ي, and ى. Deletions were frequent for ى (Alef Maksura, U+0649) and spaces, indicating that there is difficulty in preserving word segmentation and character identity. Deletions were also much more common than insertions.

Kazakh Cyrillic script errors involve confusion between Kazakh-specific letters and Russian letters (e.g., ғ → г, ө → о, ң → н), likely due to our use of a non-language-specific Cyrillic OCR model. While insertion and deletion counts are similar to the other two scripts, substitution counts in Cyrillic are substantially lower than both Arabic and Latin. Spaces are the most common character to be inserted and deleted in Cyrillic.

For the Latin Kazakh script, the most prominent errors involve diacritics and combining characters. Characters such as ◌̧ (combining comma below, U+0326), ◌̇ (combining dot above, U+0307), and accented vowels are frequently deleted or substituted, often collapsing distinct Kazakh letters into their unaccented forms. This leads to a high WER despite moderate CER, as diacritic loss frequently changes word identity.

4.2. Post-OCR Correction

We now look at the result of the LLM-based post-OCR correction (Table 4). We look at changes in CER and WER after applying post-OCR correction. Negative values indicate degradation or a worse OCR output, and positive values indicate reductions in errors.

4.2.1. Arabic Script

For the Arabic script, LLM-based post-OCR correction generally degrades OCR output across models and prompting strategies. Almost all setups resulted in a degradation in performance, with no setup achieving an improvement in both CER and WER.

A notable improvement of 6.79 points in WER was observed using few-shot prompting with Gemini; however, the associated change in CER was a degradation of 6.40 points, making the overall result inconsistent. Improvements in performance were very inconsistent, and despite four different models and three different prompting strategies, we were unable to reach a practically useful performance. In some cases, the resulting text was substantially worse than the input. CER degraded 73.73 points, and WER degraded 64.20 points using GPT-4o-mini with the confusion-aware prompt.

4.2.2. Cyrillic Script

Given the strong baseline OCR performance for Cyrillic Kazakh, most LLM configurations either fail to improve results or slightly degrade performance. Some models were found to have slight increases in performance.

In select cases, performance in CER and WER increased from the baseline. Using Gemini and few-shot prompting, an improvement of 0.40 in CER and 1.54 in WER was observed. In Gemini 2.5 Flash, performance improved in all three prompting setups for Cyrillic, with few-shot prompting being the most effective, and the minimal prompt being the least effective. This suggests that confusion-aware prompting may provide marginal improvements to minimal prompts. This finding also suggests that some models have much greater capabilities for Kazakh post-OCR

Arabic			Cyrillic			Latin		
#	Substitution	Count	#	Substitution	Count	#	Substitution	Count
1	، → .	474	1	ғ → г	103	1	ì → i	969
2	ى → ی	363	2	ұ → у	91	2	k → ķ	618
3	ي → ی	223	3	ө → о	83	3	n → ņ	387
4	ك → ک	211	4	— → –	54	4	a → á	233
5	ہ → ه	202	5	н → п	48	5	ģ → g	177
6	، → ,	166	6	ж → с	47	6	u → ú	130
7	، → »	110	7	ң → ц	42	7	◌̦ (U+0326) → ú	124
8	، → ›	104	8	ң → н	39	8	u → ů	101
9	ى → □	83	9	ы → Ы	31	9	□ → ’	93
10	پ → ب	78	10	ұ → у	28	10	◌̦ (U+0326) → ů	91

Table 2: Top 10 character substitutions by Tesseract OCR across Kazakh scripts. The arrow indicates the ground-truth character (left) incorrectly recognized as the OCR output (right). Combining diacritics are identified by Unicode name: ◌̦ = combining comma below (U+0326); ◌̇ = combining dot above (U+0307); ◌̋ = combining double acute accent (U+030B).

Arabic			Cyrillic			Latin		
Top 5 Deleted Characters								
#	Char	Count	#	Char	Count	#	Char	Count
1	ى (U+0649)	1,252	1	□	1,299	1	◌̦ (U+0326)	2,960
2	□	1,046	2	З	17	2	□	1,236
3	ا	541	3	—	17	3	◌̇ (U+0307)	625
4	ن	488	4	.	14	4	◌̋ (U+030B)	529
5	ه	451	5	Г	11	5	u	275
Top 5 Inserted Characters								
#	Char	Count	#	Char	Count	#	Char	Count
1	□	153	1	□	127	1	i	216
2	ه	40	2	.	24	2	□	55
3	ى (U+0649)	30	3	ə	23	3	ķ	48
4	U+200E (LRM)	18	4	c	15	4	ņ	32
5	ی	18	5	,	14	5	g	12

Table 3: Top 5 deleted and inserted characters by Tesseract OCR across Kazakh scripts. Deleted characters are missing from OCR output; inserted characters are spuriously added. Combining diacritics are identified by Unicode name: ◌̦ = combining comma below (U+0326); ◌̇ = combining dot above (U+0307); ◌̋ = combining double acute accent (U+030B).

correction, given the consistent improvement by Gemini, and the consistent degradation by GPT-4o-mini. In some cases, such as Claude with few-shot and minimal prompting, performance decreased dramatically despite Claude’s success in confusion-aware prompting. This suggests that models may be highly sensitive to different prompts.

4.2.3. Latin Script

Results for the Latin script are mixed but predominantly negative. Some setups have significant improvements in WER and CER from the baseline of 0.102 and 0.399.

Performance decreased in most models, with the decrease often being extremely substantial. Using the minimal and confusion-aware prompts, we find that CER performance degrades between 2.20 and 26.21 points, and WER performance degrades between 9.45 and 30.71 points. Conversely, using the few-shot prompts, performance increases dramatically in almost all models. Using GPT-4o-mini, Gemini 2.5 Flash, and DeepSeek v3.2, CER performance increased between 1.80 and 8.58 points, and WER performance increased between 7.39 and 32.49 points. Ultimately, the improvements of Gemini 2.5 Flash improve CER and WER to 0.016 and 0.074, to levels even lower than the initial Kazakh Cyrillic OCR. This few-shot setup makes the Latin script OCR outputs usable for downstream tasks, but is highly sensitive to the use of few-shot prompts.

Script	Model	Prompt	△CER	△WER
Arabic	GPT-4	minimal	-0.5321*	-0.4580*
Arabic	GPT-4	confusion-aware	-0.7373*	-0.6420*
Arabic	GPT-4	few-shot	-0.5182*	-0.4254*
Arabic	Claude	minimal	-0.3185*	-0.3326*
Arabic	Claude	confusion-aware	-0.1927*	-0.2098*
Arabic	Claude	few-shot	-0.4411*	-0.4185*
Arabic	Gemini	minimal	-0.0574	+0.0166
Arabic	Gemini	confusion-aware	-0.1974	-0.1052
Arabic	Gemini	few-shot	-0.0640	**+0.0679**
Arabic	DeepSeek	minimal	-0.0569*	-0.2024*
Arabic	DeepSeek	confusion-aware	**-0.0357***	-0.1602*
Arabic	DeepSeek	few-shot	-0.0499*	-0.0815*
Cyrillic	GPT-4	minimal	-0.0028	-0.0389*
Cyrillic	GPT-4	confusion-aware	-0.0020	-0.0248
Cyrillic	GPT-4	few-shot	-0.0001	-0.0179
Cyrillic	Claude	minimal	-0.1818*	-0.2341*
Cyrillic	Claude	confusion-aware	+0.0012	+0.0030
Cyrillic	Claude	few-shot	-0.6464*	-0.7371*
Cyrillic	Gemini	minimal	+0.0019	+0.0033
Cyrillic	Gemini	confusion-aware	+0.0021	+0.0054
Cyrillic	Gemini	few-shot	**+0.0040***	**+0.0152**
Cyrillic	DeepSeek	minimal	-0.0272*	-0.0748
Cyrillic	DeepSeek	confusion-aware	-0.0074	-0.0202
Cyrillic	DeepSeek	few-shot	+0.0031	+0.0074
Latin	GPT-4	minimal	-0.1083*	-0.1710*
Latin	GPT-4	confusion-aware	-0.0220*	-0.0945*
Latin	GPT-4	few-shot	+0.0303*	+0.1163*
Latin	Claude	minimal	-0.1982*	-0.2660*
Latin	Claude	confusion-aware	-0.2621*	-0.3018*
Latin	Claude	few-shot	-0.1855	-0.0514
Latin	Gemini	minimal	-0.2237*	-0.3071*
Latin	Gemini	confusion-aware	-0.2163*	-0.2898*
Latin	Gemini	few-shot	**+0.0858***	**+0.3249***
Latin	DeepSeek	minimal	-0.1179*	-0.2880*
Latin	DeepSeek	confusion-aware	-0.0804*	-0.2620*
Latin	DeepSeek	few-shot	+0.0180	+0.0739

Table 4: Change in character error rate (△CER) and word error rate (△WER) after LLM-based post-OCR correction across scripts. Negative values indicate degradation with respect to the original OCR output. *Statistically significant change (Wilcoxon signed-rank test, $p < 0.05$, with 95% bootstrap confidence intervals not crossing zero).

4.3. Error Analysis

To complement the quantitative analysis and exemplify major errors, we present a short qualitative analysis of errors made by LLMs when correcting Kazakh text (Table 5).

- **Script change:** In cases when CER and WER decreased dramatically, errors often resulted from the LLM not preserving the original script and transliterating the text to Cyrillic Kazakh. In example 1, it can be seen that the Latin OCR output is changed to Cyrillic text. In example 2, it can be seen that the Arabic script is mixed with the Cyrillic script in the LLM correction. This behavior accounts for some of the largest single-setup degradations in our results. Script switching towards the dominant Kazakh script is a predictable consequence of resource asymmetry in low-resource languages, a rare typological configuration when multiple orthographies coexist for a single language at extremely unequal levels of representation. In Kazakh, Cyrillic text dominates online corpora and LLM training data significantly, creating an association with “Kazakh” and the Cyrillic script in current language technologies. When models encounter a low-resource script with an unfamiliar orthography, it often overrides explicit instructions, producing outputs in the better-known script of the language. This is an important writing-system-specific failure that

has no direct comparison in monoscript post-OCR correction research.

- **Hallucination/overcorrection:** In some cases, LLMs made frequent incorrect corrections, even in confusion-aware settings. In examples 3 and 4, k is often corrected to q despite a q → k substitution not mentioned in the Latin script confusion-aware prompt. These rewrites may partly reflect influence from Turkish, which uses a similar Latin-based orthography and is substantially better represented in LLM training data than Kazakh Latin; models may default to Turkish orthographic patterns when Kazakh-specific knowledge is absent. More broadly, this pattern reflects the lack of orthographic knowledge that current LLMs have of Kazakh Latin specifically. Kazakh Latin uses a unique orthography with many unique diacritics differing from other high-resource languages using the Latin script. This means that models cannot use learned orthographic patterns and instead apply superficial substitutions based on high-resource co-script languages. This is a hallucination driven by the lack of script knowledge by LLMs.
- **English Text Insertion**: Surprisingly, Cyrillic OCR performance decreased dramatically in Claude. We found that Claude frequently inserted English text explaining its changes, despite the prompt asking the LLM to "return only the corrected text." We did not find this issue in other models.

Both of these errors are prevalent across the Latin and Arabic scripts and account for the dramatic decreases in CER and WER in some setups.

5. Discussion

We identify common errors in the OCR of all three Kazakh scripts and evaluate four LLMs for post-OCR error correction using three major prompt types. Our evaluation reveals that LLM-based post-OCR correction in multi-script settings fails in structured, characterizable ways that go beyond simple performance degradation. Script switching driven by resource imbalance, hallucination driven by insufficient script knowledge, and instruction-following breakdown under constrained output conditions each map onto distinct properties of the Kazakh writing system context, suggesting that post-OCR correction for multi-script languages and under-resourced writing systems requires evaluation frameworks and prompting strategies that account for script dominance hierarchies. Performance varies dramatically across scripts. In the Arabic script, post-OCR correction was unsuccessful, frequently resulting in great increases in CER and WER, demonstrating that LLM-based post-OCR correction is unreliable when the target language is severely underrepresented in LLM training data relative to other languages sharing the same script. In the Cyrillic script, LLMs provided marginal improvements and, in some cases, introduced new errors. The Latin script presents a promising case for post-OCR correction where models using few-shot prompting achieve substantial reductions in WER and CER, making their output suitable for downstream NLP tasks and usage.

These findings align with previous work showing that LLM-based post-OCR correction is less reliable in under-resourced scripts and languages compared to high-resource languages (Kanerva et al., 2025). Low-resource Kazakh scripts lack full support in LLMs, limiting model performance on Kazakh tasks. These findings conversely show promise in few-shot methods for the Kazakh Latin script and provide few-shot methods as a future direction for low-resource post-OCR correction.

We find that some models frequently fail to preserve the original script, often transliterating or partially converting text into Cyrillic Kazakh even when instructed not to do so. This behavior accounts for some of the largest degradations in CER and WER and exemplifies a limitation of current LLMs in multi-script settings. This difficulty is rare due to Kazakh's use of three scripts and must be considered in future systems using any low-resource Kazakh script. Hallucination and overcorrection played roles in the degradation of the Latin script, where models introduced corrections that were not supported by error patterns. These errors suggest that current LLMs may rely on incorrect assumptions about Kazakh. The insertion of English text by Claude when instructed not to do so was unexpected and displays the fragility of instruction-following in LLMs for text generation tasks such as post-OCR correction.

Future work should develop post-OCR correction methods to prevent script changes and overcorrection of texts. Future work should develop new post-OCR correction systems, such as hybrid approaches that combine rule-based filtering with LLM generation. Correction systems should be studied with other OCR systems other than Tesseract, as MLLMs have significantly more errors, and other models may also produce outputs with differing amounts of error from Tesseract. Future work could also include supervised or instruction-tuned post-OCR correction models trained specifically on Kazakh scripts. Work should also look into the impact of structured outputs such as JSON, which may result in more adherence to instructions because many models have been instruction-tuned

#	Ground Truth	OCR Output	LLM Correction	Error
1	Orys ůltyķ-mądeni ortalyģy, Taldykorģan ķalasynda 1998 ž. ķúrylģan. Ortalyķtyn baģdarlamasynda orys halķynyn saltdństùrìn, tìlìn, mądenietìn damytyp, saķtau kôzdelgen...	Orys últyk-mőűdeni ortalygy, Taldykorģan kalasynda 1998 Žž kúrylģan. Ortalyktyn baģdarlamasynda orys halkynyn soltdőstùrìn, tiin, mődenietin damytyp, saktau kôzdelgen…	Орыс ұлттық-мәдени орталығы Талдықорған қаласында 1998 жылы құрылған. Орталықтың бағдарламасында орыс халқының салтдәстүрін, тілін, мәдениетін дамытып, сақтау көзделген...	Script change (Latin → Cyrillic)
2	ايدىك 3-كۇنى، انتقا مىنگەن سايا‌حاتشىلار قابا اۋدانىنىڭ جايلاۋىن سايا‌حاتتاۋدا. قابا اۋدانىنىڭ جايلاۋىنىڭ ەكولوگيالىق ورتاسى جاقسى، ...	ايدىك 3-كۇنى. انتقا مىنگەن سايا‌حاتشىلار قابا اۋدانىنىڭ جايلاۋىنىڭ سايا‌حاتتاۋدا. قابا اۋدانىنىڭ جايلاۋىنىڭ ... ەكولوگيالىق ورتاسى جاقسى. ءار	ايدىك 3-كۇنى. انتقا مىنگەن سايا‌حاتشى ىلار قابا‌تاۋ‌داныныңجايلاۋىنداسايا‌حاتتап. قابا‌تاۋ‌داныныңجايلاۋىنىڭ ەكولوگيالىق ۇورتاسى جاقسى...	Partial script change (Arabic → Cyrillic)
3	..Taldykorģan ķalasynda 1998 ž. ķúrylģan...	...Taldykorģan kalasynda 1998 Žž kúrylģan...	...Taldyqorghan qalasynda 1998 jyl qurylgan...	Over correction/hallucination
4	...Abaj oblysy Aâgôz audanyndaģy auyl...	...Abaj oblysy Aâgôz audanyndaģy auyl...	...Abai oblysy Aiağöz audanyndağy auyl...	Over correction/hallucination
5	N/A	N/A	(Note: I did not find any obvious OCR errors in this text that require correction.)	English text insertion

Table 5: Representative examples of LLM errors in Kazakh text correction.

for structured output. MLLMs for post-OCR correction should be explored as they may prevent overcorrection and hallucination, and their output quality may benefit from LLM-based post-OCR correction strategies such as few-shot learning.

LLM-based post-OCR correction for Kazakh is largely unreliable, but under certain setups, it can perform extremely well. Arabic-script correction remains a challenge, and Cyrillic-script correction offers limited benefits due to the strong baseline OCR performance. In the Latin script, we develop a setup to decrease CER by 8.58 points and WER by 32.49 points using a few-shot setup with Gemini. These findings display the promise and the limitations of LLMs for post-OCR correction in underresourced and multi-script languages.

6. Conclusion

This study presents the first evaluation of LLMs for post-OCR correction in all three Kazakh scripts. We evaluate Gemini 2.5 Flash, GPT-4o-mini, DeepSeek v3.2, and Claude 3.5 Haiku using minimal, confusion-aware, and few-shot prompting.

We first identify common OCR errors and causes in Kazakh, which we use to inform prompt development. We find that LLMs are largely unsuccessful under most setups, often switching to incorrect scripts, overcorrecting, or incorrectly following instructions. In the Arabic script, we are unable to find a successful model, and in the Cyrillic script, the best models make marginal improvements due to the initial high accuracy of Kazakh Cyrillic OCR. In the Latin script, we find that few-shot prompting is highly effective, decreasing the CER by 8.58 points and WER by 32.49 points, to levels lower than the high-resource Cyrillic script, when using Gemini 2.5 Flash.

These findings display the promise and variability of LLMs for post-OCR correction in low-resource languages and scripts such as Kazakh. These findings contribute to advancing low-resource and Kazakh NLP research, building post-OCR correction models, and supporting the inclusion of low-resource writing systems in NLP.

7. Limitations

There are several limitations that should be considered in this study. First, authentic OCR data does not exist for the Latin and Arabic Kazakh scripts, so all OCR experiments rely on the synthetically generated KazakhOCR benchmark, which may not capture the full diversity of layouts, fonts, and handwriting styles found in real-world documents. Synthetic data is also more controlled, allowing for results to reflect script, model, and prompt differences rather than document-level variation. Second, the evaluation is limited to Tesseract due to its strong previous performance on the benchmark. Different OCR systems may have different error levels, which could affect post-OCR correction performance. Third, we evaluate only zero-shot and few-shot prompting without any model fine-tuning to reflect a realistic and accessible use case. This approach may not capture the potential gains achievable through supervised or instruction-tuned post-OCR correction models. Finally, the test set used consisted of 50 images per script, which is smaller than ideal for evaluation. This was constrained by API costs of running four models across three scripts and three prompting strategies. We applied two statistical tests, and we marked results that were significant in the results to mitigate incorrect findings.

8. Bibliographical References

References

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku.

G. Batyrbekkyzy, T.K. Mahanuly, M.M. Tastanbekov, L.S. Dinasheva, B.K. Issabek, and G.D.

Sugirbayeva. 2018. Latinisation of kazakh alphabet history and prospects. *European Journal of Science and Theology*, 14:125–134.

H Bunke and P S P Wang. 1997. Handbook of character recognition and document image analysis.

Peter T Daniels and William Bright. 1996. *The World's Writing Systems*. Oxford University Press.

Vera Danilova and Gijs Aangenendt. 2025. Post-OCR correction of historical German periodicals using LLMs. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 120–129, Tallinn, Estonia. University of Tartu Library, Estonia.

DeepSeek-AI. 2025. DeepSeek-V3 technical report.

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Sami Honkasalo and Tansulu Temirbekova. 2024. The writing reform and 'Latinization' of written Kazakh: A sociolinguistic survey. *International Journal of Eurasian Linguistics*, 6(1):48–80.

Asmaganbetova Kamshat, Ulanbek Auyeskhan, Nurzhanova Zarina, Serikbayev Alen, and Malika Yeskazina. 2024. Integration ai techniques in low-resource language: The case of kazakh language. In *2024 IEEE AITU: Digital Generation*, pages 7–13.

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. 2025. OCR error post-correction with LLMs in historical documents: No free lunches. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 38–47, Tallinn, Estonia. University of Tartu Library, Estonia.

Radoslav Koynov and Triet Ho Anh Doan. 2025. Opportunities and challenges of LLMs as post-OCR correctors. volume 45, pages 111–118. Polish Information Processing Society.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-OCR processing approaches. *ACM Computing Surveys*, 54(6):1–37.

Daniyar Nurseitov, Kairat Bostanbekov, Daniyar Kurmankhojayev, Anel Alimova, Abdelrahman Abdallah, and Rassul Tolegenov. 2021. Handwritten Kazakh and Russian (HKR) database for text recognition. *Multimedia Tools and Applications*, 80(21–23):33075–33097.

OpenAI. 2024. GPT-4o system card.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

R. Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Nazgul Toiganbayeva, Mahmoud Kasem, Galymzhan Abdimanap, Kairat Bostanbekov, Abdelrahman Abdallah, Anel Alimova, and Daniyar Nurseitov. 2022. KOHTD: Kazakh offline handwritten text dataset. *Signal Processing: Image Communication*, 108:116827.

Arman Yeleussinov, Yedilkhan Amirgaliyev, and Lyailya Cherikbayeva. 2023. Improving OCR accuracy for Kazakh handwriting recognition using GAN models. *Applied Sciences*, 13(9):5677.

9. Language Resource References

Gagnier, Henry and Gagnier, Sophie and Kirubakaran, Ashwin. 2026. *KazakhOCR: A Synthetic Benchmark for Evaluating Multimodal Models in Low-Resource Kazakh Script OCR*. Association for Computational Linguistics.