KazakhOCR: A Synthetic Benchmark for Evaluating Multimodal Models in Low-Resource Kazakh Script OCR

Henry Gagnier[1], **Sophie Gagnier**[1], **Ashwin Kirubakaran**[2]
[1]Pittsford Sutherland High School, [2]Edison Academy Magnet School,
Correspondence: henrygagnier9@gmail.com

Abstract

Kazakh is a Turkic language using the Arabic, Cyrillic, and Latin scripts, making it unique in terms of optical character recognition (OCR). Work on OCR for low-resource Kazakh scripts is very scarce, and no OCR benchmarks or images exist for the Arabic and Latin scripts. We construct a synthetic OCR dataset of 7,219 images for all three scripts with font, color, and noise variations to imitate real OCR tasks. We evaluated three multimodal large language models (MLLMs) on a subset of the benchmark for OCR and language identification: Gemma-3-12B-it, Qwen2.5-VL-7B-Instruct, and Llama-3.2-11B-Vision-Instruct. All models are unsuccessful with Latin and Arabic script OCR, and fail to recognize the Arabic script as Kazakh text, misclassifying it as Arabic, Farsi, and Kurdish. We further compare MLLMs with a classical OCR baseline and find that while traditional OCR has lower character error rates, MLLMs fail to match this performance. These findings show significant gaps in current MLLM capabilities to process low-resource Abjad-based scripts and demonstrate the need for inclusive models and benchmarks supporting low-resource scripts and languages.

1 Introduction

Kazakh is a Turkic language spoken in Kazakhstan, China, Mongolia, Russia, Kyrgyzstan, and Uzbekistan by 16 million people (McCollum and Chen, 2020). Kazakh is written in different scripts depending on the geographic and political context. In Central Asia, a Cyrillic script is used for Kazakh, while in Europe, America, and Turkey, a Latin script is used, and in China, Afghanistan, Pakistan, and Iran, an Arabic script is used (Honkasalo and Temirbekova, 2024). The use of these different writing scripts makes Kazakh unique and makes optical character recognition (OCR) particularly complex, considering Cyrillic, Latin, and Arabic scripts.

OCR refers to the electronic translation of handwritten, typewritten, or printed text, and its conversion into a machine-readable form (Faizullah et al., 2023). It enables the recognition of text in digital images, scanned documents, and videos, converting images into machine-coded text. OCR has many applications, including document scanning and form processing. Convolutional neural networks (CNNs) and transformer architectures have recently improved OCR performance across many scripts and languages. Multimodal models have been used for OCR, enabling text extraction without segmentation and allowing for greater robustness with noise. Research on these models has primarily focused on high-resource languages and has not been performed with the Kazakh Arabic script.

Prior work for Kazakh OCR has been primarily done on the Cyrillic script. Toiganbayeva et al. (2022) created a KOTHD, a Kazakh offline handwritten text dataset using Cyrillic script. Razaque et al. (2024) developed a handwritten text recognition system for handwritten Cyrillic script using CNNs and RNNs. Pirniyazova et al. (2023) evaluates a neural network on the recognition of Latin letters of the Kazakh alphabet. Yeleussinov et al. (2023) use a Generative Adversarial Network (GAN) to improve Kazakh handwriting recognition using Cyrillic script. The Kazakh Arabic script uses additional letters, modified graphemes, and different orthographic conventions, making it different from high-resource Arabic varieties and an important test case for Abjad-based low-resource NLP. While work has been performed on Kazakh OCR, it does not focus on the Arabic script and scarcely focuses on the Latin script.

Synthetic datasets are emerging as a solution to a lack of images and content in low-resource languages. The creation of a synthetic OCR dataset often consists of the collection of text corpora, fonts, and noise filters (Saini et al., 2022). Haq et al.

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 60–65
March 28, 2026.

(2025) developed a synthetic benchmark for Pashto and the Pashto script. Margner and Pechwitz (2001) proposed a method for the creation of synthetic data for Arabic OCR. Saini et al. (2022) synthetically created a benchmark of 90k images for 23 Indic languages. Synthetic data has also been used for post-OCR correction, and has been shown to have advantages over traditional OCR training in low-resource languages (Guan and Greene, 2024). Synthetic data is particularly useful when limited resources for OCR are available, such as in low-resource Kazakh scripts.

Research on OCR for Kazakh using the Arabic script is scarce, but extremely necessary. The purpose of this study is to (1) construct synthetic OCR benchmarks for low-resource Kazakh scripts, (2) evaluate multimodal models on the OCR of low-resource Kazakh scripts, and (3) identify challenges in the OCR of Kazakh. This study also aims to encourage the inclusion and research of low-resource scripts, particularly languages using the Arabic Kazakh script.

2 Methodology

2.1 Kazakh Text Corpora

While it is possible to obtain perfect transliterations in Kazakh using pre-defined rules, cultural differences in the Cyrillic and Arabic scripts make transliteration from high-resource scripts into low-resource scripts sub-optimal for training (Zhang et al., 2024a). We obtained authentic Kazakh text in the Arabic script from the first Kazakh corpus in the Arabic script, the MC² corpus Zhang et al. (2024a). Wikipedia is also a good resource for low-resource languages such as Kazakh. For Cyrillic text, we obtained authentic data from the Kazakh Wikipedia[1].

We cleaned the Cyrillic script data from Wikipedia using kaznlp (Yessenbayev et al., 2020), and transliterated it to the Latin script using the QazNLTK package[2]. As the number of characters in each sample of the Arabic (3746 characters) script text was much greater than the number of characters in the Cyrillic (689 characters) and Latin (718 characters) script texts, we implemented random balanced subsampling to make the text lengths similar in each of the scripts. After balance subsampling, we had 2,417 Arabic texts, 2,402 Cyrillic texts, and 2,400 Latin texts, with similar text length distributions across scripts, which were all used in the final benchmark.

2.2 Benchmark Construction

In order to make synthetic images representative of authentic OCR tasks, such as document scanning, we implement many variations across images, using the Pillow Python library. Examples of the benchmark and its variation can be seen in Figure 1, 2, and 3.

- **Fonts:** We downloaded 2,491 fonts from Google Fonts and filtered all fonts for compatibility with Kazakh Cyrillic, Latin, and Arabic script. We found 135 fonts compatible with Arabic, 431 with Cyrillic, and 1376 with Latin. These fonts were randomly selected to create synthetic images. We also varied font size randomly from 24 to 56 across images.

- **Noise:** We set the noise level to be random between 4 and 18, and a blur effect to be randomly between 0 and 0.8. Images are all rotated randomly between -3 and 3 degrees randomly.

- **Color:** We designed two color palettes, each with 40 colors. One contained light background colors, and one contained darker text colors. When these colors were randomly selected, a minimum contrast ratio of 4.5 was used as a standard in all images.

To foster reproducibility and encourage future work on low-resource Kazakh scripts, we release the full benchmark, which is available at https://huggingface.co/datasets/henrygagnier/kazakh-ocr.

2.3 Multimodal Large Language Models

Multimodal large language models (MLLMs) preserve the reasoning and decision-making capabilities of large language models (LLMs), but also empower multimodal tasks, such as OCR (Zhang et al., 2024b). We evaluated `gemma-3-12b-it` (Team et al., 2025), `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025), and `Llama-3.2-11B-Vision-Instruct` (Grattafiori et al., 2024) using a subset of the benchmark, with 200 images per script, or a total of 600 images. We chose parameter-efficient variants of recent multimodal models for cost and computational efficiency, to ensure that our methodology remains

[1] https://huggingface.co/datasets/wikimedia/wikipedia
[2] https://pypi.org/project/qaznltk/

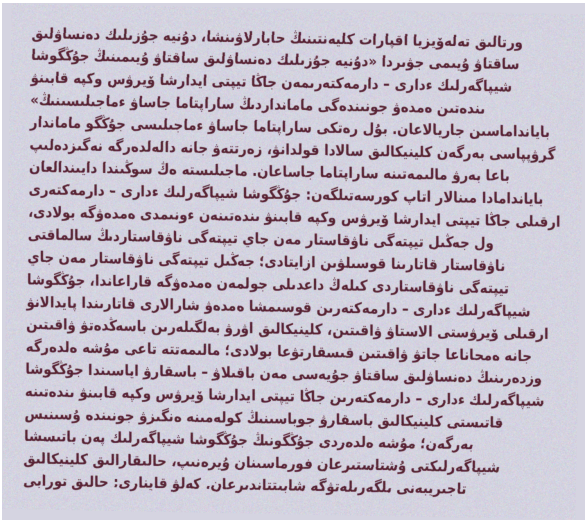

Figure 1: Example of an OCR task in the Kazakh Arabic script

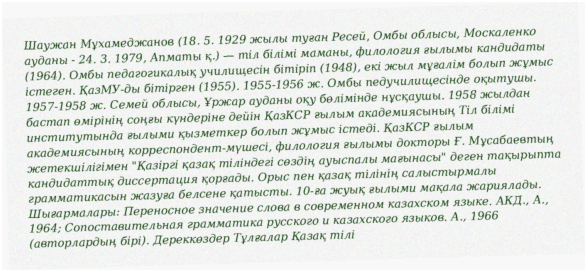

Figure 2: Example of an OCR task in the Kazakh Cyrillic script

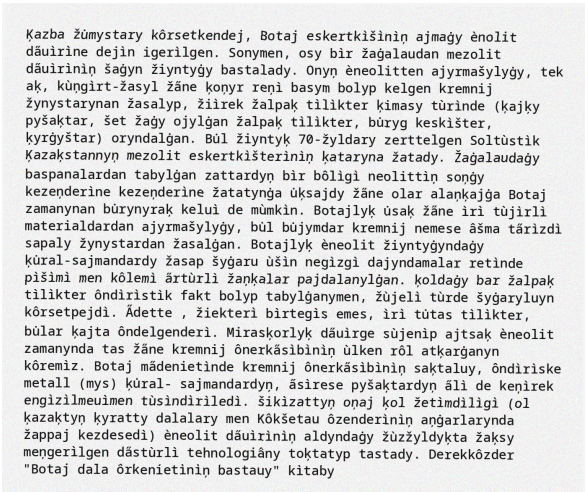

Figure 3: Example of an OCR task in the Kazakh Latin script

accessible under low-computational budgets, while still covering diverse model architectures in our evaluation. A temperature of 0.0 was used in all models with the prompt in Appendix A.

2.4 Classical OCR Baseline

We evaluate a classical OCR baseline using Tesseract as a baseline. We use the Tesseract OCR engine (Smith, 2007) with its pretrained models on the Arabic, Latin, and Cyrillic scripts for each of the Kazakh scripts. No fine-tuning or adaptations were made, and the models were evaluated on an identical subset of 200 images per script. As Tesseract does not perform language identification, we only report character error rate and word error rate.

3 Results

We now look at the performance of the three models in the Arabic, Latin, and Cyrillic Kazakh scripts to find which scripts multimodal models experience challenges in (Table 1). In all models, character error rate (CER) and word error rate (WER) were best in Cyrillic script, and worst in Arabic script. Generally, Qwen performed the best across the three publicly available models, with a CER of 5.3% in the Cyrillic script and a WER of 13.4%. In the Latin and Arabic scripts, CER and WER are considerably worse, with a CER of 26.4% and 35.5% in the Latin and Arabic scripts, respectively.

All models are extremely unsuccessful at recognizing the Arabic script as Kazakh, with the Arabic script most commonly being misclassified as Arabic, Farsi, and Kurdish. The percent of samples correctly classified as Kazakh ranged from 0.0% to 1.1%. Qwen was highly successful at identifying the Cyrillic text as Kazakh, while also having the best OCR accuracy in the Cyrillic script. Gemma had an accuracy of 70.1%, most commonly misclassifying Kazakh as Kyrgyz. Llama had an accuracy of 11.3%, also most commonly misclassifying Kazakh as Kyrgyz. In the Latin script, Qwen was the most successful with an accuracy of 99.4%, while Llama and Gemma had lower accuracies of 13.5% to 31.1%, with the Kazakh being most commonly misclassified as Kyrgyz and Tatar. Seen in Qwen and Gemma, models struggled in identifying low-resource scripts, while being successful with the Cyrillic script.

Across all three scripts, Tesseract achieves lower CERs than the evaluated MLLMs. This gap is the biggest in the Arabic script when CER is 15.0%

when using Tesseract, compared to 35.5%–72.5% when using the MLLMs. While there are large gaps in Arabic and Latin scripts, Qwen achieves a low CER similar to Tesseract. These results indicate that MLLMs, despite their general multimodal capabilities, do not match the OCR accuracy of traditional approaches on low-resource Kazakh scripts.

4 Discussion

We construct KazakhOCR, the first OCR benchmark for low-resource Kazakh scripts. Evaluating multimodal large language models on the benchmark, we find significant disparities in OCR performance and language identification across the three scripts. The Cyrillic script, or the primary script in which Kazakh is written, achieved the best results across models, with Qwen2.5-VL-7B-Instruct achieving a CER of 5.3% and a WER of 13.4%. The Latin script had CERs ranging from 21.6 to 31.0%, and the Arabic script had CERs ranging from 35.5% to 72.5%.

Comparing MLLMs with Tesseract reveals that traditional approaches have much greater accuracy than MLLMs, with lower CERs in all three scripts. The gap in performance is more pronounced in the Latin and Arabic scripts, demonstrating that current MLLM capabilities do not match OCR system capabilities for low-resource scripts. All MLLMs failed at identifying the Kazakh Arabic script, with correct identification rates of 0.0% to 1.1%. Most commonly misidentified as Arabic, Farsi, and Kurdish, this incapability of the model displays a lack in current MLLM understanding of low-resource Kazakh scripts. We theorize that this error is especially important in MLLM-based OCR because MLLMs often hallucinate (He et al., 2025), and when hallucinating in OCR for low-resource scripts, they may use words, characters, and diacritics found in languages that predict the text uses, leading to low OCR CER and WER. Qwen almost perfectly identified the Latin and Kazakh scripts, with accuracies of 99.4% and 100.0%, respectively, displaying that the Kazakh Arabic script is the least recognized by current MLLMs. These results are especially concerning for Kazakh speakers in China, Afghanistan, Pakistan, and Iran, where MLLMs do not adequately support the Kazakh Arabic script.

Future work with multimodal models for OCR should further experiment on prompt sensitivity and engineering for optimal OCR results. Work should also use authentic Latin script Kazakh text to properly represent cultural differences in the use of the script. Future LLMs and MLLMs should strive to support low-resource Kazakh scripts, considering the use of the Arabic script and the growing use of the Latin script. Future work should also use authentic low-resource Kazakh script texts, as opposed to synthetic benchmarks, when available.

This study demonstrates that multimodal large language models have extreme disparities in processing low-resource Kazakh scripts, particularly for the Kazakh Arabic script. The KazakhOCR benchmark reveals significant gaps in current model capabilities for low-resource NLP, and highlights the need for inclusive training processes that support the Kazakh Latin and Arabic scripts. By making this benchmark publicly accessible, we aim to address these disparities in Kazakh NLP and encourage the development of more equitable multimodal models for the diverse Kazakh scripts.

5 Conclusion

This study synthetically constructs the first benchmark for low-resource Kazakh script OCR and identification. We evaluate Qwen2.5-VL-7B-Instruct, Llama-3.2-11B-Vision-Instruct, and Gemma-3-12B-it, three publicly available multimodal large language models, on our benchmark.

We find that all models are unsuccessful with Latin and Arabic script Kazakh OCR. Models have CERs of 35.5% to 72.5% for the Arabic script, and CERs of 26.4% to 31.0% for the Latin script. All models were unable to identify the Arabic script Kazakh as Kazakh, mistaking it for Arabic, Farsi, and Kurdish. Models were also largely unable to identify the Latin script. In all models, the best performance was on the Cyrillic script, with Qwen achieving a CER of 5.3% and a WER of 13.4%,

These findings display that low-resource Kazakh scripts are largely unsupported in MLLMs, with models failing at OCR and language identification, and demonstrate the necessity for the inclusion of the Kazakh Arabic and Kazakh Latin scripts in NLP.

Limitations

Several limitations should be considered in this study. Due to the computational costs and time of MLLMs, we chose a sample of 200 images per script. This sample may not be large enough to represent the diversity in the Kazakh scripts. We also

Model	Script	CER (%)	WER (%)	% Kazakh
Qwen2.5-VL-7B-Instruct	Arabic	35.5	90.0	1.1%
	Latin	26.4	57.4	99.4%
	Cyrillic	5.3	13.4	100.0%
Llama-3.2-11B-Vision-Instruct	Arabic	37.3	90.2	0.0%
	Latin	21.6	66.3	31.1%
	Cyrillic	18.2	49.8	11.3%
Gemma-3-12B-it	Arabic	72.5	103.0	0.0%
	Latin	31.0	79.2	13.5%
	Cyrillic	25.7	42.9	70.1%
Tesseract	Arabic	15.0	47.5	—
	Latin	11.4	51.6	—
	Cyrillic	4.3	30.3	—

Table 1: OCR Performance metrics of models across different scripts. CER = Character Error Rate, WER = Word Error Rate, and the percentage of samples correctly identified as Kazakh

used transliterated Latin texts, which may have different uses and cultural differences from the Cyrillic script, although this has not been studied. This study also constructs a synthetic benchmark, due to the lack of real OCR data for low-resource Kazakh scripts, which may not capture all characteristics of real-world documents.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Safiullah Faizullah, Muhammad Sohaib Ayub, Sajid Hussain, and Muhammad Asad Khan. 2023. A survey of ocr in arabic language: Applications, techniques, and challenges. *Applied Sciences*, 13(7):4584.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shuhao Guan and Derek Greene. 2024. Advancing post-ocr correction: A comparative study of synthetic data. *Preprint*, arXiv:2408.02253.

Ijazul Haq, Yingjie Zhang, and Irfan Ali Khan. 2025. Psocr: Benchmarking large multimodal models for optical character recognition in low-resource pashto language.

Zhentao He, Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, and Minghui Qiu. 2025. Seeing is believing? mitigating ocr hallucinations in multimodal large language models. *Preprint*, arXiv:2506.20168.

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.

V. Margner and M. Pechwitz. 2001. Synthetic data for arabic ocr system development. In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pages 1159–1163.

Adam G. McCollum and Si Chen. 2020. Kazakh. *Journal of the International Phonetic Association*, 51(2):276–298.

P. Pirniyazova, E.Yu. Son, and D.Zh. Kulzhan. 2023. Recognition of latin letters of the kazakh alphabet based on a neural network. *Computing amp;amp; Engineering*, 1(3):36–41.

A. Razaque, B. Makezhanuly, O. Alimseitov, Zh. Kalpeyeva, and A. Ayapbergenova. 2024. Development of handwritten text recognition system for the kazakh language. *Computing amp;amp; Engineering*, 2(4):1–7.

Naresh Saini, Promodh Pinto, Aravinth Bheemaraj, Deepak Kumar, Dhiraj Daga, Saurabh Yadav, and Srihari Nagaraj. 2022. Ocr synthetic benchmark dataset for indic languages. *Preprint*, arXiv:2205.02543.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and*

Recognition, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Zhandos Yessenbayev, Zhanibek Kozhirbayev, and Aibek Makazhanov. 2020. Kaznlp: A pipeline for automated processing of texts written in kazakh language. In *Speech and Computer*, pages 657–666, Cham. Springer International Publishing.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024a. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024b. Mm-llms: Recent advances in multimodal large language models. *Preprint*, arXiv:2401.13601.

A MLLM Prompt

We used the following prompt for all MLLM evaluations:

```
You are an expert OCR system.

Read and transcribe all text visible in
the image.

Preserve exact formatting, spacing, and
punctuation.

Identify the language using ISO 639
codes.

Return only valid minified JSON in this
exact format:

{"language":"iso_code","text":"exact
transcription"}
```