# A Comparative Evaluation of Open-Source Models for Russian-Kazakh Translation

**Gleb Shanshin**
ITMO University
Saint Petersburg, Russia
gleb.shanshin@niuitmo.ru

## Abstract

We describe an evaluation of several open-source models under identical inference conditions without task-specific training. Despite covering a wide range of available models, including both multilingual systems and models specifically designed for Russian–Kazakh translation, the results indicate that the highest performance is achieved by the language-specific approach.

## 1 Introduction

Kazakh is a low-resource language for machine translation, characterized by limited availability of high-quality parallel corpora and linguistic tools, despite having tens of millions of speakers. This scarcity poses challenges for building robust MT systems, especially given Kazakh's rich morphology and orthographic variability.

To address these issues, the Turkic LoRes MT shared task, organized within the LoResMT workshop series, focuses on machine translation for low-resource Turkic languages, including Russian-Kazakh. The shared task provides a common evaluation framework and standardized test sets, encouraging participants to explore practical system-building strategies under realistic low-resource conditions.

This paper describes approaches developed for the Russian-Kazakh track of the shared task, based on the evaluation of multiple open-source MT models combined with task-specific post-processing techniques aimed at improving translation quality.

## 2 Dataset

The organizers provided only a test dataset consisting of 4,626 sentence pairs, split evenly into public and private subsets for evaluation. During manual inspection, we observed that not all entries contain valid Russian-Kazakh translations.

As shown in Table 1, four types of sentence pairs were identified:

- **Correct**: valid Russian-Kazakh translation pairs.
- **Copied**: cases where the Russian source text is partially or fully copied into the target field.
- **Russian Different**: corrupted pairs where the target contains unrelated Russian text.
- **Kazakh Different**: pairs where the target is written in Kazakh but is semantically unrelated to the source sentence.

To improve dataset reliability, we applied a simple automatic filtering procedure. Both source and target strings were normalized by lowercasing and removing punctuation. If all characters in both fields belonged to the Russian alphabet and the length of the longest common substring between normalized source and target exceeded 30 characters, the pair was removed from the cleaned dataset. Additionally, two sentence pairs (kk_03501 and kk_03995) with severe semantic mismatches were identified and excluded manually.

As a result, 278 sentence pairs were removed from the original dataset. All subsequent experiments were conducted on two versions of the data: the original test set and the cleaned subset.

We note that sentence pairs of the *Kazakh Different* type were not removed from either version of the dataset. While such cases correspond to semantically mismatched translations, they are harder to reliably detect automatically without external semantic models or manual annotation. Therefore, these examples were retained and are reported only to document the presence of this type of noise in the evaluation data.

March 28, 2026 ©2026 Association for Computational Linguistics

| Type | Row Id | Source | Expected Translation |
|---|---|---|---|
| Correct | kk_03310 | Первый компонент - базовая пенсия , которая в настоящее время выплачивается государством из средств республиканского бюджета в одинаковом размере для всех граждан , достигших пенсионного возраста , независимо от их трудового стажа и заработной платы . | Бірінші компонент – зейнеткерлік жасқа жеткен барлық азаматтарға олардың еңбек өтілі мен жалақысына қарамастан , бірдей мөлшерде республикалық бюджет қаражатынан мемлекет төлейтін базалық зейнетақы . |
| Copied | kk_01035 | Таким образом , спецсоцуслуги через неправительственный сектор получают более 2 тыс . человек . | Таким образом , спецсоцуслуги через неправительственный сектор получают более 2 тыс , человек . |
| Russian Different | kk_03995 | Продолжается развитие фармацевтической отрасли , В республике насчитывается 1874 аптечных склада и 9590 объектов розничной аптечной сети . | В рамках развития электронного здравоохранения планируется к 2020 году внедрение электронных паспортов здоровья . |
| Kazakh Different | kk_04424 | В частности , женщинам обеспечивается право покупки пенсионных аннуитетов в возрасте 50 лет . | « Кейбір жағдайларға қатысты еліміздің біраз азаматтарына тиесілі қаражат бұл , Ондай азаматтар мүмкін қайтыс болды , мүмкін шет елге шығып кетті , содан қалып қойғандары да бар . |

Table 1: Examples of different row types in the Russian-Kazakh dataset

## 3 Evaluated Models

The organizers of the shared task proposed ChrF1 as the primary evaluation metric, which we adopt throughout all experiments.

### 3.1 Baseline

As a simple baseline, we use a trivial system that copies the Russian source text as the output. This baseline achieves a ChrF1 score of 22.32 on the private test set before cleaning and 17.23 after cleaning.

### 3.2 issai/LLama-3.1-KazLLM-1.0-8B

LLama-3.1-KazLLM-1.0-8B (ISSAI, 2024) is an open-source large language model based on Meta's LLaMA-3.1 architecture, fine-tuned for multilingual use with a particular focus on Kazakh, Russian, and English. The model was released by the Institute of Smart Systems and Artificial Intelligence (ISSAI) under a CC-BY-NC license.

For inference, we use the prompt shown in Figure 1.

The model achieves a ChrF1 score of 51.12 on the original test set and 53.39 on the cleaned version.

**System prompt:**
Сіз орыс тілінен қазақ тіліне кәсіби аудармашысыз. Сізге орыс тіліндегі мәтін ұсынылады және қосымша түсініктемелерсіз қазақ тіліне аударма жазу қажет болады.

**User prompt:**
Орыс мәтіні: {text}
Қазақша аударма:

Figure 1: Prompt used for inference with LLama-3.1-KazLLM-1.0-8B. *Note:* The system prompt translates into English as: *"You are a professional translator from Russian to Kazakh. You will be given a Russian text and are required to produce a Kazakh translation without any additional explanations."*

### 3.3 PolynomeAI/Llama-3.1-8B-kkru

Llama-3.1-8B-kkru (PolynomeAI, 2025) is a fine-tuned variant of Meta's Llama-3.1-8B model, specifically adapted for Russian–Kazakh and Kazakh–Russian machine translation. The model was trained on a mixture of parallel and synthetic data and is optimized for direct translation tasks rather than general-purpose text generation.

For inference, we use the default Alpaca-style prompt provided by the model configuration, shown in Figure 2.

The model achieves a ChrF1 score of 49.64 on the original test set and 51.80 on the cleaned version.

### 3.4 google/translategemma-{4|12|27}b-it

TranslateGemma (Finkelstein et al., 2026) is an open suite of multilingual machine translation models released by Google Translate Research, built on the Gemma 3 foundation models and fine-tuned through a two-stage process of supervised fine-tuning on synthetic and human-translated

| Model | Public | Private | Public (clean) | Private (clean) |
|---|---|---|---|---|
| deepvk/kazRush-ru-kk | **76.77** | **76.24** | **80.05** | **80.28** |
| facebook/nllb-200-3.3B | 54.56 | 54.08 | 56.61 | 56.57 |
| facebook/nllb-200-1.3B | 53.97 | 53.38 | 55.98 | 55.81 |
| facebook/nllb-200-distilled-1.3B | 53.88 | 53.56 | 55.91 | 56.00 |
| facebook/nllb-200-distilled-600M | 53.00 | 52.77 | 54.96 | 55.19 |
| google/translategemma-27b-it | 51.48 | 51.08 | 53.35 | 53.36 |
| issai/LLama-3.1-KazLLM-1.0-8B | 51.38 | 51.12 | 53.23 | 53.39 |
| PolynomeAI/Llama-3.1-8B-kkru | 50.35 | 49.64 | 52.14 | 51.80 |
| google/translategemma-12b-it | 48.02 | 47.60 | 49.67 | 49.61 |
| google/translategemma-4b-it | 39.59 | 39.36 | 40.76 | 40.78 |
| tencent/HY-MT1.5-7B transcripted + spellcheck | 33.23 | 33.00 | 34.21 | 34.17 |
| tencent/HY-MT1.5-7B transcripted | 29.56 | 29.41 | 30.33 | 30.33 |
| Russian text | 21.61 | 22.32 | 17.47 | 17.23 |

Table 2: ChrF scores on public and private test sets, evaluated on the original and cleaned data variants.

Below is an instruction that describes a task.

**### Instruction:**
Translate from Russian to Kazakh.

**### Input:**
{text}

**### Response:**

Figure 2: Alpaca-style prompt used for inference with Llama-3.1-8B-kkru.

parallel corpora followed by reinforcement learning. The family includes 4B, 12B, and 27B parameter variants optimized for efficient, high-quality translation across 55 languages, where the mid-sized model often outperforms larger baselines on standard benchmarks. Our tests show substantial quality increase with size of model rising up to 51.08 private on original data and 53.36 on cleaned data.

### 3.5 facebook/nllb-200-*

We also include results for Meta's No Language Left Behind (NLLB) family of models (Fan et al., 2022), which are pretrained massively multilingual machine translation systems covering hundreds of languages. In our evaluation, the full-size NLLB model achieves the strongest performance among the multilingual baselines. Smaller configurations, such as nllb-200-1.3B and its distilled variants, show competitive results with substantially fewer parameters, while the distilled 600M model provides a lightweight alternative. Metric differences across NLLB are relatively small, with the best scores reaching 54.08 on the original test set and 56.57 on the cleaned one.

### 3.6 tencent/HY-MT1.5-{1.8|7}B

Tencent's HY-MT1.5 (Zheng et al., 2025) is a recently released family of multilingual machine translation models available in two sizes: a 1.8B parameter variant optimized for on-device and real-time translation, and a 7B parameter variant targeting high-quality server and cloud-based scenarios. Both models support bidirectional translation across 33 languages and several dialectal variants, and are trained using a holistic pipeline combining MT-oriented pre-training, supervised fine-tuning, on-policy distillation, and reinforcement learning.

For inference, we used the default prompt suggested by the authors: *"Translate the following segment into Kazakh, without additional explanation."* followed by the source sentence.

The 1.8B model failed to produce adequate results, frequently generating outputs in unrelated languages such as Ukrainian, Hindi, or Russian instead of Kazakh. As a result, we exclude this configuration from further analysis.

The 7B model consistently produced Kazakh translations; however, the output was written in Arabic script rather than Cyrillic. After direct transcription, the model achieved ChrF1 scores of 28.89 on the original dataset and 29.78 on the cleaned dataset. A detailed inspection revealed systematic orthographic issues.

To address these issues, we trained a lightweight spell correction model in a self-supervised manner. We extracted 50,000 sentences from the Kazakh Wikipedia dump (20231101.kk slice from `wikimedia/wikipedia` (Foundation)). Each sentence was split into 7-word segments and

| Type | Sentence | ChrF1 |
|---|---|---|
| Original | Казахстан является многонациональным и многоконфессиональным государством . | 22.37 |
| HY-MT1.5-7B output | قازاقستان كوپ ۇلتتى جانه كوپ ٴدىندى مەملەكەت. | – |
| Direct transcription | қазақстан коп ұлтты жане коп дінді мемлекет. | 46.90 |
| Spell Corrected | Қазақстан көп ұлтты және көп дінді мемлекет. | 69.53 |
| Correct Target | Қазақстан көпұлтты және көпконфессиялы мемлекет . | – |

Table 3: Example of post-processing stages for sentence pair `kk_02849`

artificially corrupted by introducing character-level noise, including common confusions (e.g., *і↔ы*, *қ↔к*), random insertions, and deletions, with an overall corruption probability of 0.35. A `google/byt5-small` model (Xue et al., 2021) was then fine-tuned to recover the original sentences from their corrupted versions.

During inference, the generated translations were similarly split into fixed-length segments, corrected independently, and merged back. This post-processing step improved ChrF1 scores to 33.00 on the original dataset and 34.17 on the cleaned dataset.

Table 3 illustrates the effect of the post-processing pipeline. Most recoverable errors are corrected (e.g., *жане→және*), while some lexical or stylistic differences remain unavoidable without additional supervision.

### 3.7 deepvk/kazRush-ru-kk

deepvk/kazRush-ru-kk (Lebedeva and Sokolov, 2024) is a Russian–Kazakh neural machine translation model released by DeepVK and trained specifically for direct Russian-Kazakh translation. The model significantly outperforms all other evaluated systems, achieving ChrF1 scores of 76.24 on the private test set and 80.28 on the cleaned private subset.

## 4 Conclusion

Our results demonstrate that language-specific machine translation systems trained explicitly for a single language pair consistently outperform general-purpose multilingual models that are not specialized for the target direction. While large multilingual models provide strong baselines and broad coverage, their performance on Russian-Kazakh translation remains inferior to that of dedicated systems optimized for this specific pair. These findings highlight the importance of task- and language-pair-specific training in low-resource settings and suggest that, when sufficient parallel data is available, specialized models remain the most effective approach for achieving high translation quality.

## References

Angela Fan, Mike Lewis, Tomas Kocisky, and et al. 2022. Beyond english–centric multilingual machine translation. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 339–351.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report. *arXiv preprint arXiv:2601.09012*.

Wikimedia Foundation. Wikimedia downloads.

ISSAI. 2024. LLama-3.1-KazLLM-1.0-8B. `https://huggingface.co/issai/LLama-3.1-KazLLM-1.0-8B`.

Anna Lebedeva and Andrey Sokolov. 2024. kazRush-ru-kk: translation model from Russian to Kazakh.

PolynomeAI. 2025. Llama-3.1-8B-kkru. `https://huggingface.co/PolynomeAI/Llama-3.1-8B-kkru`.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*. ArXiv:2105.13626.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025. HY-MT1.5 Technical Report. *arXiv preprint arXiv:2512.24092*.