Ensemble Methods for Low-Resource Russian-Kyrgyz Machine Translation: When Diverse Models Beat Better Models

Adilet Metinov
metinovab@kstu.kg

Abstract

We present our submission to the LoResMT 2026 Shared Task on Russian-Kyrgyz machine translation. Our approach demonstrates that ensembling diverse translation models with simple majority voting can significantly outperform individual models, achieving a +1.37 CHRF++ improvement over our best single model. Notably, we find that including "weaker" models in the ensemble improves overall performance, challenging the conventional assumption that ensembles should only combine top-performing systems. Our best submission achieved 49.31 CHRF++, placing 3rd in the Russian-Kyrgyz track, using only open-weight models without any fine-tuning on parallel Kyrgyz data. We also report several counter-intuitive findings: (1) simple voting outperforms quality-weighted selection, (2) more diverse models help even when individually weaker, and (3) post-processing "corrections" can hurt performance when reference translations contain similar artifacts.

1 Introduction

Machine translation for low-resource languages remains challenging due to limited parallel data and linguistic resources. Kyrgyz, a Turkic language spoken by approximately 4.5 million people, falls into this category despite growing digitalization efforts.

The LoResMT 2026 Shared Task on Russian-Kyrgyz translation provides an opportunity to explore effective strategies for this language pair. In this system description paper, we present our ensemble-based approach that achieved competitive results without requiring expensive model fine-tuning or access to large parallel corpora.

Our main contributions are:

- A simple yet effective voting-based ensemble method that combines diverse translation models
- Empirical evidence that including weaker models improves ensemble performance
- Analysis of why simpler selection strategies outperform complex quality filtering
- Practical insights for low-resource MT without fine-tuning

2 System Description

2.1 Base Models

We generated translations using multiple models from two families:

NLLB Models We used three variants of Meta's No Language Left Behind (NLLB) models (NLLB Team et al., 2022):

- NLLB-200 600M (distilled)
- NLLB-200 1.3B
- NLLB-200 3.3B

DeepSeek Models We also included translations from DeepSeek-R1, a large language model with strong multilingual capabilities, running locally on our GPUs with open weights.

Table 1 shows the individual performance of each model.

Model	CHRF++
NLLB 3.3B	47.94
NLLB 1.3B	47.34
NLLB 600M	46.59
DeepSeek (basic)	46.73
DeepSeek-R1	45.80

Table 1: Individual model performance on the test set.

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 235–237

2.2 Ensemble Method

Our ensemble approach uses consensus-based voting to select the best translation for each sentence. Given n candidate translations for a source sentence, we compute pairwise similarity scores and select the translation with highest average similarity to all others.

Similarity Metric We use character-level n-gram similarity (Jaccard coefficient over character trigrams):

$$sim(t_1, t_2) = \frac{|ngrams(t_1) \cap ngrams(t_2)|}{|ngrams(t_1) \cup ngrams(t_2)|} \quad (1)$$

Voting Procedure For each sentence, we select:

$$t^* = \arg\max_{t_i} \frac{1}{n-1} \sum_{j \neq i} sim(t_i, t_j) \quad (2)$$

This simple approach selects the translation that has the highest consensus with other models, effectively implementing a "wisdom of the crowd" strategy.

2.3 Quality Filtering (Ablation)

We also experimented with quality-based filtering before voting, detecting:

- English words in translations
- Russian words that should have been translated
- Encoding artifacts and repeated characters
- Unusual length ratios compared to source

However, as we report in Section 3, quality filtering did not improve results.

3 Experiments and Results

3.1 Main Results

Table 2 shows our ensemble experiments. Our best result (49.31 CHRF++) was achieved with simple voting over 5 models.

3.2 Key Findings

Finding 1: Weaker models improve the ensemble. Counter-intuitively, adding DeepSeek-R1 (45.80 CHRF++ individually) to our NLLB ensemble *improved* overall performance from 48.22 to 49.31. This suggests that model diversity contributes more to ensemble success than individual model quality.

Configuration	Models	CHRF++
Best single model	1	47.94
3 NLLB models (vote)	3	48.22
5 models (vote)	5	**49.31**
7 models (vote)	7	49.17
5 models (qual. filter, t=50)	5	49.23
5 models (qual. filter, t=40)	5	49.23
5 models (qual. filter, t=60)	5	48.60

Table 2: Ensemble results. Simple voting with 5 models achieves the best performance.

Finding 2: Simple voting beats quality filtering. Our quality-weighted ensemble (49.23) performed worse than pure voting (49.31). We hypothesize that the quality heuristics filtered out translations that, while appearing "incorrect," actually matched the reference translation style.

Finding 3: There is a sweet spot for ensemble size. Performance improved from 3 models (48.22) to 5 models (49.31), but degraded slightly with 7 models (49.17). Too many models may introduce noise that dilutes the consensus signal.

Finding 4: Post-processing hurts performance. We attempted to "clean" our outputs by:

- Removing double spaces
- Normalizing punctuation
- Adding missing end punctuation

This reduced our score from 49.31 to 49.27, suggesting that reference translations contain similar artifacts.

3.3 Source Distribution Analysis

Table 3 shows how often each model was selected in our best ensemble.

Model	Selected	%
NLLB 1.3B	736	31.8%
NLLB 3.3B	642	27.8%
NLLB 600M	544	23.5%
DeepSeek-R1	253	10.9%
DeepSeek (basic)	136	5.9%

Table 3: Source distribution in the final ensemble. All models contribute, with NLLB variants selected most frequently.

Interestingly, even the worst-performing model (DeepSeek basic, 46.73) was selected for 136 sen-

tences (5.9%), indicating it provided the best consensus translation for those cases.

4 Analysis

4.1 Why Does Diversity Help?

We hypothesize that different model architectures make different types of errors. NLLB models, trained specifically for translation, may handle common patterns well but struggle with unusual constructions. LLMs like DeepSeek, while potentially less consistent, may handle creative or contextual translations better.

When these diverse models agree, the consensus is likely correct. When they disagree, the voting mechanism tends to select translations that share common elements, which often correlates with correctness.

4.2 Why Does Quality Filtering Hurt?

Our quality heuristics were designed to detect:

- Untranslated Russian words
- English contamination
- Formatting issues

However, we found that reference translations in the test set may contain similar "issues" — for example, some technical terms left untranslated, or inconsistent spacing. By filtering these out, we moved away from the reference distribution, hurting our CHRF++ score.

This highlights an important consideration for MT evaluation: optimizing for automatic metrics may require matching reference artifacts, not just producing "clean" translations.

4.3 Comparison to Top Systems

The top two systems achieved 51.03 and 51.02 CHRF++, approximately 1.7 points above our best result. Given their minimal submission counts (2–3 submissions), we hypothesize they may have used:

- Fine-tuned models on Kyrgyz parallel data
- Different model architectures better suited for Turkic languages
- Access to additional training resources

Our approach, using only off-the-shelf open-weight models without fine-tuning, represents a strong baseline for resource-constrained settings.

5 Conclusion

We presented a simple ensemble approach for Russian-Kyrgyz machine translation that achieves competitive results without model fine-tuning. Our key findings — that diversity matters more than individual quality, and that simple voting beats complex filtering — provide practical insights for low-resource MT.

Future work could explore:

- Learned combination weights instead of uniform voting
- Quality estimation models trained on this language pair
- Fine-tuning base models on available Kyrgyz parallel data

Limitations

Our work has several limitations. First, our ensemble method was only evaluated on one language pair (Russian-Kyrgyz), and the findings may not generalize to other low-resource pairs. Second, we relied entirely on automatic metrics (CHRF++) without human evaluation, which may not fully capture translation quality. Third, our approach requires running inference with multiple models, which increases computational cost compared to a single model. Finally, we did not explore fine-tuning on available parallel Kyrgyz data, which could potentially improve base model quality and ensemble performance.

Acknowledgments

We thank the LoResMT 2026 organizers for providing this shared task and the Selectel company for sponsoring the competition.

References

NLLB Team, Marta R. Costa-jussà, James Cross, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.