Machine Translation for Low Resource Turkic Languages: English-Tatar

Alexander Dikov
National Research Nuclear
University MEPhI (NRNU MEPhI)
dae007@campus.mephi.ru

Abstract

This paper outlines our winning submission to the English-to-Tatar translation task. We evaluated three strategies: few-shot prompting with **Gemini 3 Pro Preview**, specialized trans-tokenized **Tweeties** models, and the RL-distilled **TranslateGemma** family. Results demonstrate that large commercial models significantly outperform smaller specialized ones in this low-resource setting. Gemini secured first place with a chrF++ score of 56.71, surpassing the open-source baseline of 25.23.

1 Introduction

Machine translation for low-resource languages like Tatar remains a significant challenge, particularly in specific domains such as Natural Language Understanding (NLU) for virtual assistants. The test set for this task consists of short, imperative sentences related to alarms, weather forecasts, and media playback. Our goal was to evaluate the capabilities of modern Large Language Models (LLMs) in zero-shot and few-shot scenarios compared to smaller models fine-tuned specifically for Tatar.

2 Data

The data is provided in CSV format with `id` and `source_en` columns. Table 1 illustrates the structure of the input data. As seen in the examples, the source text often exhibits characteristics of spoken language or ASR (Automatic Speech Recognition) output:

- Sentences may start with lowercase letters (e.g., *"how hot is it..."*).
- There are spaces before question marks (e.g., *"... rain today ?"*).
- The segments are extremely short, lacking context, which makes ambiguity resolution challenging for translation models.

Given the nature of the text, the translation requires maintaining specific named entities and time formats while ensuring natural, conversational phrasing in Tatar.

ID	Source English Text
valid_1	Is it going to rain today ?
valid_2	how hot is it going to get today
valid_3	How hot is it ?
valid_4	Will it be sunny today?
valid_5	Is it cloudy today?

Table 1: Sample entries from the test dataset showing weather-related intents.

We did not use additional parallel corpora for fine-tuning the LLMs, relying instead on their pre-trained knowledge and in-context learning capabilities.

3 System Description

We experimented with four different model architectures, ranging from large proprietary APIs to specialized open-source models utilizing novel tokenization strategies.

3.1 Gemini 3 Pro Preview

To establish a high-resource baseline, we utilized Google's Gemini 3 Pro Preview. Unlike the other models in our experiments, this is a proprietary, closed-source model accessed via API. (DeepMind & Google, 2025)

We employed a few-shot prompting strategy, providing the model with context examples, constructing a prompt that included:

1. **System Instruction:** A role-playing directive defining the persona.
2. **Contextual Examples:** 5 pairs of high-quality English-Tatar translations.
3. **Target Input:** The query sentence to be translated.

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 222–224

System: You are a helpful assistant. Translate the following user commands from English to Tatar. Keep the tone natural and preserve time formats.

User: Set an alarm for 7 am.
Model: Иртәгә сәгать 7-гә сигнал куй.

User: Is it going to rain?
Model: Бүген яңгыр явачакмы?

User: Play some jazz music.
Model: Джаз музыкасын уйнат әле.

... [More examples] ...

User: `{Input_Sentence}`
Model:

Figure 1: Structure of the few-shot prompt used for the Gemini 3 Pro Preview submission.

Figure 1 illustrates the structure of the prompt used for inference.

This strategy proved to be the most effective, securing the **first place** in the shared task leaderboard. The model demonstrated superior handling of the domain nuances and low-resource morphology.

3.2 TranslateGemma

Following the competition conclusion, we extended our evaluation to the TranslateGemma family (specifically the 4B and 12B variants). Consequently, official leaderboard metrics were not recorded for these models. (Finkelstein et al., 2026)

These models are based on the Gemma 3 architecture and were trained on the WMT24++, SMOL, GATITOS and additional language pairs derived from synthetic data. Notably, the official technical report does not explicitly list Tatar, Bashkir, or Chuvash among the supported or tested languages, although it includes related Turkic languages like Kazakh and Kyrgyz.

Our qualitative experiments revealed that due to this "zero-shot" nature regarding Tatar, the models exhibit significant *language confusion*. While TranslateGemma successfully generates text using the correct script and Turkic morphological structure, it frequently hallucinates vocabulary or grammar specific to Kazakh or Turkish rather than producing authentic Tatar, limiting its immediate utility without further fine-tuning.

3.3 Tweeties

We evaluated two distinct models from the `Tweeties` family, both designed to address the scarcity of Tatar language data in standard open-source LLMs through vocabulary adaptation. (Remy et al., 2024)

The first, `tweety-7b-tatar-v24a`, is derived from `Mistral-7B-Instruct-v0.2` using a trans-tokenization approach. This method replaces the original vocabulary with a tokenizer explicitly trained on Tatar corpora, ensuring that the agglutinative morphology of the target language is encoded into meaningful units rather than fragmented bytes.

The second model, `tweety-tatar-hydra-base-7b`, builds upon `Unbabel/TowerInstruct-7B-v0.1` using the hydra architecture. Unlike standard trans-tokenization, the hydra approach employs a dual-tokenizer mechanism: it retains the original tokenizer for encoding source (English) input while utilizing a dedicated Tatar tokenizer for the target output. This necessitates an embedding alignment strategy where source tokens are mapped into a shared vector space, allowing the model to leverage its pre-trained multilingual knowledge while generating syntactically correct Tatar.

The fundamental difference between these approaches lies in their handling of the cross-lingual gap. The Mistral-based model acts primarily as a monolingual Tatar specialist, having overwritten its original embeddings to maximize generation efficiency in the target language. In contrast, the Hydra model preserves the source language understanding of the TowerInstruct base through its hybrid input processing. This makes the Hydra architecture theoretically more robust for translation tasks, as it maintains access to the rich English semantic representations of the base model while utilizing the native Tatar tokenizer for high-quality surface realization.

Despite the architectural advantages, our experimental results yielded a modest chrF++ score of 25.23. In our zero-shot inference settings, the base trans-tokenized models successfully produced grammatically coherent Tatar but struggled with the specific NLU domain terminology and strict formatting constraints, confirming that while vocabulary adaptation is crucial, it must be paired with robust instruction tuning to match the reasoning capabilities of larger commercial models.

4 Conclusion

In this work, we presented a comparative analysis of diverse approaches for English-to-Tatar translation within the specific domain of NLU commands.

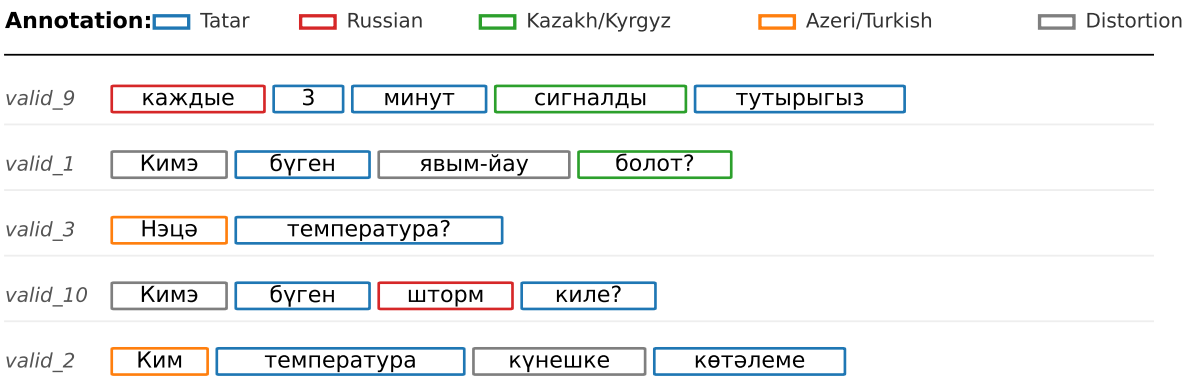

Figure 2: Visual error analysis of the specialized model outputs.

Gemini 3 Pro Preview demonstrated superior performance, achieving a chrF++ score of 56.71. This proprietary model benefits from massive scale and extensive multilingual pre-training. Its few-shot reasoning capabilities allowed it to grasp the NLU domain nuances effectively without specific fine-tuning, significantly outperforming the smaller models.

The **Tweeties (tweety-7b-tatar-v24a)** model achieved a score of 25.23. As a trans-tokenized 7B parameter model, it serves as a robust open-source baseline. While efficient, it struggles to match the generalist reasoning and vocabulary coverage of the large commercial model in this specific zero/few-shot setting.

Our evaluation of the **TranslateGemma** family revealed significant limitations. Despite utilizing state-of-the-art architecture and knowledge distillation from Gemini, these models proved **unsuitable** for high-quality English-to-Tatar translation in a zero-shot setting. The lack of explicit Tatar language representation in the training data leads to severe language confusion and hallucinations from related Turkic languages, rendering the model ineffective for this specific pair without substantial fine-tuning.

References

DeepMind & Google. 2025. Approach, methodology & results: Gemini 3 pro. Technical report (PDF).

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *Preprint*, arXiv:2408.04303.