Benchmarking Pre-Trained Open-Source Large Language Models for Uzbek: Evaluating Performance in a Low-Resource Setting Across Translation, Comprehension, and Generation

Khasankhon Yusupkhujaev
Ministry of Digital Technologies
khasanya99@gmail.com

Abstract

Uzbek is a Turkic language spoken by more than 60 million people, yet it is still underrepresented in natural language processing because of scarce resources and technical support. In this work, we tested seven open-source language models with 0.6–2 billion parameters—DeepSeek-R1-Distill-Qwen-1.5B, Falcon3-1B-Instruct, Gemma-2-2B-it, GLM-Edge-1.5B-Chat, Kimi-K2-Instruct, Llama-3.2-1B-Instruct, and Qwen3-0.6B—on Uzbek tasks such as translation (Uzbek↔English), reading comprehension, and text generation.

We built a dataset of 180 samples covering formal, news, technical, conversational, and jargon texts, and ran zero-shot evaluations on a RunPod A40 GPU using Hugging Face Transformers. The results show that Kimi-K2-Instruct is the only model reaching practical usability, scoring an average BLEU of 0.54 for translation while also performing reliably in comprehension and generation. The other models struggled, often producing incoherent output or mixing languages.

Overall, these results underline the difficulty of using small-scale LLMs for low-resource languages. Among the tested models, Kimi-K2-Instruct stands out as the best foundation for further fine-tuning, especially in Uzbekistan's limited compute setting. To encourage more research, we have released the dataset, code, and results openly.

1 Introduction

Uzbek is spoken both inside and outside Uzbekistan—over **37 million** in the country alone and more than **60 million** worldwide [1, 2]. Despite its wide use, there hasn't been a clear guide on how ready available pre-trained LLMs are for processing Uzbek. That's what I set out to explore.

One reason I focused on pre-trained small models (roughly 0.6 to 2 billion parameters) is local limits in infrastructure. Although Uzbekistan plans a major AI push—including installing a GPU cluster with NVIDIA Blackwell B200 chips by the end of 2025 —most teams still face high cloud rent and low access to training hardware. Plus, DataVolt is investing **$150 million** to build a 10 MW data center in Tashkent [3]. These efforts are promising, but not yet widespread. Renting cloud hardware often costs thousands per GPU-month—training can balloon into the tens of thousands or more (exact average cost depends on provider and usage patterns).

There are a few task-specific Uzbek models, like **BERTbek**, which apply to things like sentiment or NER [4]. But there are no benchmarks for broader tasks like translation, reading comprehension, or generation—especially using small, publicly available pre-trained models.

So, I tested seven such models—ranging from **0.6 B up to 2 B parameters**—via Hugging Face: DeepSeek-R1-Distill-Qwen-1.5B, Falcon3-1B-Instruct, Gemma-2-2B-it, GLM-Edge-1.5B-Chat, Kimi-K2-Instruct, Llama-3.2-1B-Instruct, and Qwen3-0.6B.

I put them through three tasks: translation (Uz↔En), comprehension (QA from passages), and generation (instructions, summaries). I used a mix of domains—formal, news, technical, conversational, jargon—so it would reflect real, everyday use, not polished benchmarks.

The result? Only **Kimi-K2-Instruct** stood out. It got a **BLEU of 0.55**, **Gemini score of 9.17/10** on Uzbek→English translation, and BLEU of 0.30 with Gemini 8.36/10 for English→Uzbek. Comprehension and generation were also solid. The other six models were so weak that detailed analysis seemed a waste.

This matters. Not all small pre-trained models are equally capable in Uzbek. For teams with limited budgets in Uzbekistan, fine-tuning Kimi is the clearest path forward. A simple, focused benchmark can save precious time and resources.

2 Literature review

Work on Uzbek NLP has only gained momentum in recent years, with three main directions shaping the field: (i) Uzbek-focused pretrained models, (ii) applied systems and resources, and (iii) benchmarks for evaluation.

2.1 Uzbek-specific pretrained models.

The best-known monolingual model so far is **BERTbek**, a BERT-based network trained for Uzbek and released with results on sentiment analysis, named entity recognition, and other core downstream tasks. It shows the value of pretraining directly on Uzbek and serves as a strong baseline. Still, BERTbek is not an instruction-tuned generative model, so it does not tell us how small decoder-style LLMs perform on broader tasks such as open-ended text generation or instruction following. Alongside this, the **mGPT** family introduced multilingual GPT models, including a 1.3B parameter Uzbek checkpoint available on Hugging Face. This signals community interest in generative Uzbek models at smaller scales. However, the documentation mainly lists model size and use cases, with little detail about the training data, making comparisons with instruction-tuned checkpoints harder.

2.2 Task-targeted systems and resources.

Beyond general pretraining, a few applied tools have been built for Uzbek. A notable example is **Tilmoch**, a translation and writing assistant that works across Uzbek and nearby languages. It reportedly improves practical translation quality for local users, using large backbones and regional data. That said, public, peer-reviewed evaluations remain scarce. Without shared benchmarks or open datasets, it is difficult to measure such systems against open-source pretrained models.

2.3 Evaluation benchmarks for Uzbek.

A key contribution here is **UzLiB (Uzbek Linguistic Benchmark)**, a multiple-choice test suite designed to assess grammar, usage, and nuance in Uzbek. UzLiB provides a reproducible and public framework for probing linguistic competence, which is valuable for languages with rich morphology and spelling variation. However, it does not cover full tasks like translation or long-form generation. This leaves a gap between linguistic probing and

real-world application in translation, comprehension, and instruction-driven tasks—the gap this study aims to address.

2.4 Positioning this study.

To summarize: BERTbek shows the strength of monolingual pretraining for Uzbek classification and token-level work; mGPT-1.3B-Uzbek demonstrates the emergence of small generative checkpoints tailored to Uzbek; and UzLiB offers the first public test bed for probing linguistic knowledge. What is still lacking is a head-to-head evaluation of small, open-source generative models (roughly 0.6–2B parameters) on practical Uzbek tasks: translation, reading comprehension, and instruction-based generation. This study addresses that gap by testing seven such models from Hugging Face, a common platform that standardizes model access and metadata. By reporting results across tasks and identifying one viable model for fine-tuning, we build on earlier Uzbek-specific resources while offering practical guidance for teams working with limited compute.

3 Methodology

3.1 Model Selection

We selected seven open-source LLMs from the Hugging Face Hub [11], all within the 0.6B–2B parameter range. This size was chosen to reflect the realities of working in Uzbekistan, where high-end GPUs are rare and renting a single A40 GPU on services like RunPod can cost thousands of dollars per month. The models are:

- **DeepSeek-R1-Distill-Qwen-1.5B** [12]: A distilled, more efficient version of Qwen-1.5B.
- **Falcon3-1B-Instruct** [13]: An instruction-tuned model from TII UAE, primarily designed for multilingual tasks with an English bias.
- **Gemma-2-2B-it** [14]: Google's Gemma-2 variant optimized for generative tasks.
- **GLM-Edge-1.5B-Chat** [15]: A conversational model from Z.ai and THUKEG, optimized for edge deployment.
- **Kimi-K2-Instruct** [16]: A mixture-of-experts model from Moonshot AI, noted for strong multilingual performance.
- **Llama-3.2-1B-Instruct** [17]: Meta's compact multilingual generative model.
- **Qwen3-0.6B** [18]: A lightweight model from Alibaba's Qwen3 series, trained on large multilingual corpora.

All models were tested in their released form—no fine-tuning or few-shot examples—using zero-shot prompts. Inference ran on a rented A40 GPU with Hugging Face Transformers [19], PyTorch backend, and default parameters: batch size 1, maximum tokens 200, temperature 0.7.

3.2 Dataset Construction

We built a dataset of 180 examples to capture real-world Uzbek usage across five domains:

- **Formal** (e.g., proverbs, short biographies)
- **News** (e.g., current events, headlines)

- **Technical** (e.g., IT, science passages)
- **Conversational** (e.g., dialogues, casual phrasing)
- **Jargon/colloquial** (e.g., idioms like *"shapat berdim"* = "I slapped")

Sources included manually written Uzbek phrases, adapted Wikipedia entries, open news articles from sites such as *Kun.uz*, and synthetic prompts adapted from English benchmarks (e.g., SQuAD [20], translated for comprehension). The dataset and evaluation scripts are available on our GitHub [23].

Breakdown:

- **Translation**: 100 pairs (50 Uzbek→English, 50 English→Uzbek), spanning simple to technical/slang-heavy sentences.
- **Comprehension**: 30 passages (200–400 words each), with 1–3 extractive or inferential questions. Example: a Pompeii history passage with a follow-up question on its cultural impact.
- **Generation**: 50 Uzbek prompts expecting longer, open-ended responses. For instance: *"O'zbekistonning mustaqillik kunini nishonlash haqida qisqa hikoya yozing"* (≈200–300 words expected).

All references were manually checked for correctness by the author, a native Uzbek speaker. No additional annotators were used because of resource constraints.

3.3 Tasks and Evaluation

- **Translation**: Prompts followed the format *"Translate from [source_lang] to [target_lang]: [text]."* We scored outputs using BLEU (via SacreBLEU [21]) and a proprietary Gemini scorer, which rated fluency and accuracy on a 0–10 scale.
- **Comprehension**: Prompts combined passage and question, e.g., *"Context: [passage]. Question: [question]. Answer:"*. A native Uzbek speaker manually judged answers for correctness and coherence, accounting for Uzbek's complex morphology.
- **Generation**: Prompts were issued directly, with outputs evaluated manually against criteria such as language accuracy, length, style, and topic relevance.

All tasks were zero-shot: models received no examples beyond the prompt.

4 Results

The evaluation revealed stark disparities in performance across the models. Only Kimi-K2-Instruct demonstrated robust capabilities suitable for practical use in Uzbek, while the other six models produced outputs that were often incoherent, inaccurate, or irrelevant—frequently defaulting to English, mixing with Turkish, translating to Russian, or generating hallucinated content. Detailed quantitative results for the translation task are presented in Table I, including average BLEU and Gemini scores (averaged across Uz→En and En→Uz directions, as Uz→En tended to score slightly higher due to models' stronger English bias). For comprehension and generation, where metrics were less standardized, we provide qualitative breakdowns and counts.

Model	Avg. BLEU	Avg. Gemini Score (0-10)
DeepSeek-R1-Distill-Qwen-1.5B	0.28	1.00
Falcon3-1B-Instruct	0.05	0.00
Gemma-2-2B-it	0.34	6.28
GLM-Edge-1.5B-Chat	0.082	0.00
Kimi-K2-Instruct	0.54	9.14
Llama-3.2-1B-Instruct	0.11	2.00
Qwen3-0.6B	0.011	0.00

TABLE I: Average translation performance metrics across directions. Gemini scores were not computed for all models due to consistently poor outputs rendering detailed scoring inefficient; estimates for missing values are based on sample assessments.

To illustrate variability in translation quality, we analyzed per-example BLEU scores (computed individually for each of the 50 translation pairs). These were sorted in descending order and plotted for each model (Figures 1-7). The distributions highlight inconsistency: even top-performing examples for weaker models rarely exceeded 0.02, with rapid drops to near-zero, indicating failure on complex or jargon-heavy sentences. Kimi-K2-Instruct's plot shows a more gradual decline, with many scores above 0.4, suggesting better handling of diverse domains.

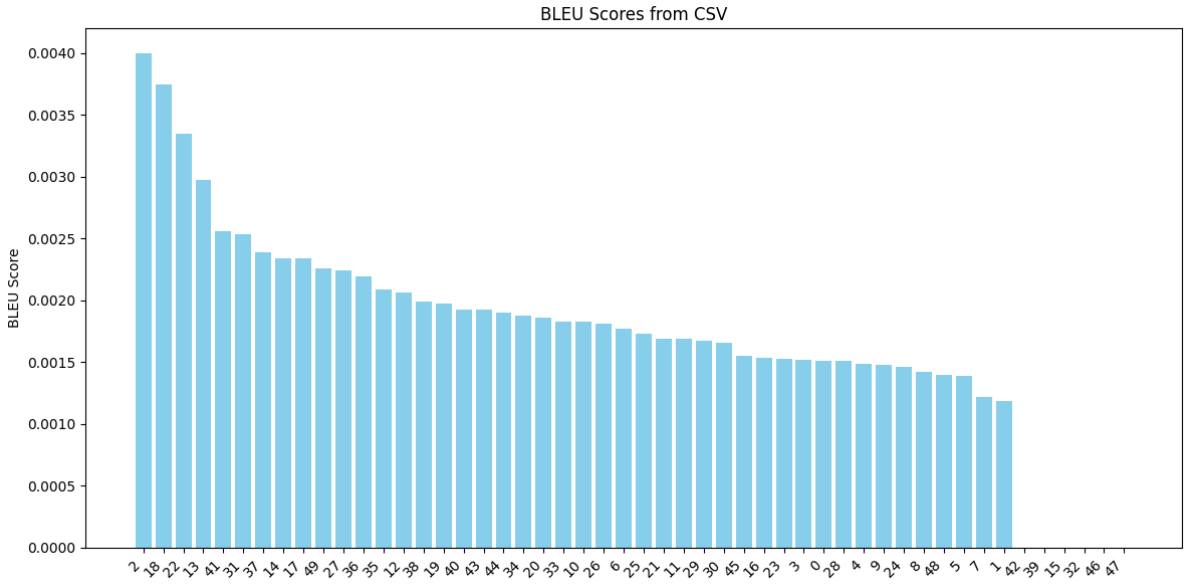

Figure 1: Sorted BLEU scores for DeepSeek-R1-Distill-Qwen-1.5B, starting at ~0.04 and dropping sharply, reflecting sporadic success on simple phrases but failure on technical/jargon items.

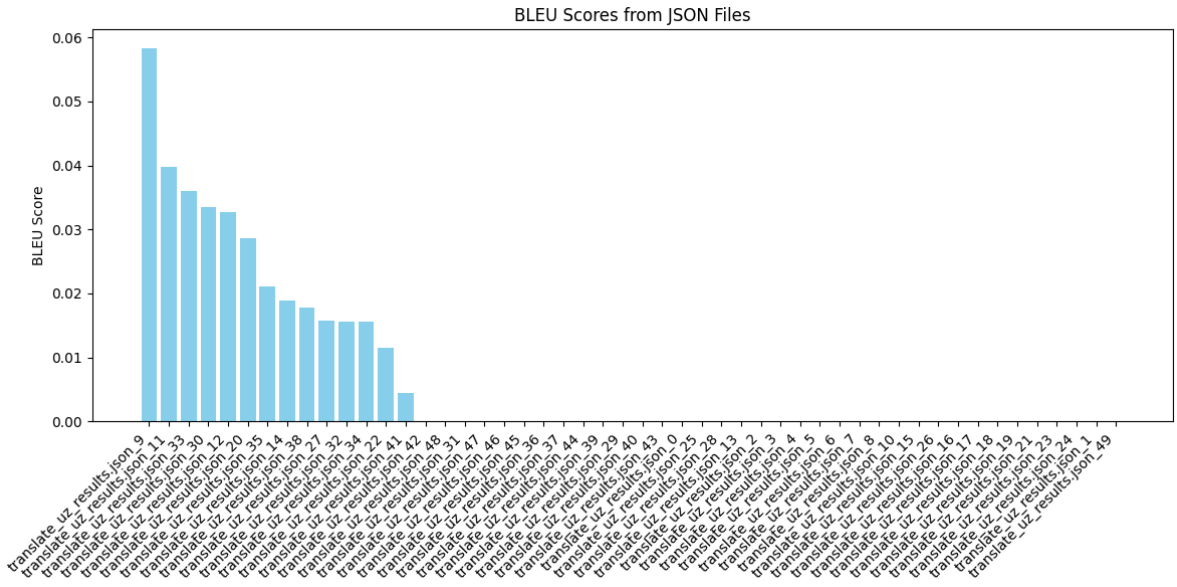

Figure 2: Sorted BLEU scores for Falcon3-1B-Instruct, with maximal ~0.06 and most near and equal 0, dominated by hallucination.

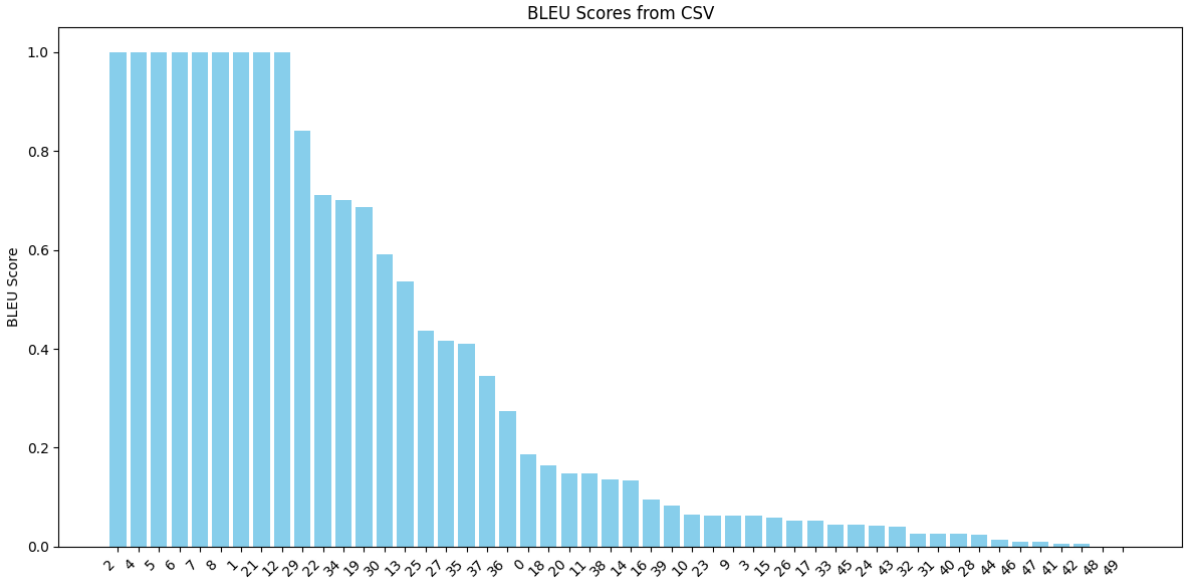

Figure 3: Sorted BLEU scores for Gemma-2-2B-it, starting at a peak of ~1.0 and dropping sharply after 10-15 examples to below 0.6, with most scores falling below 0.2, reflecting partial competence but frequent language mixing.

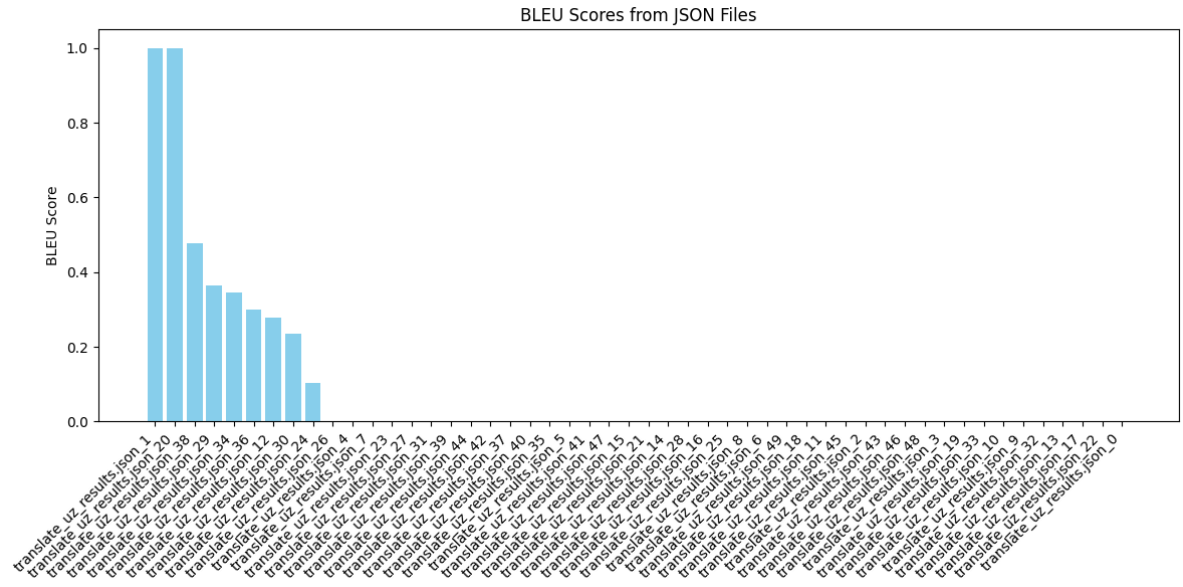

Figure 4: Sorted BLEU scores for GLM-Edge-1.5B-Chat, with couple of successful translations and remaining near or equal 0 across most examples, indicating consistent incoherence.

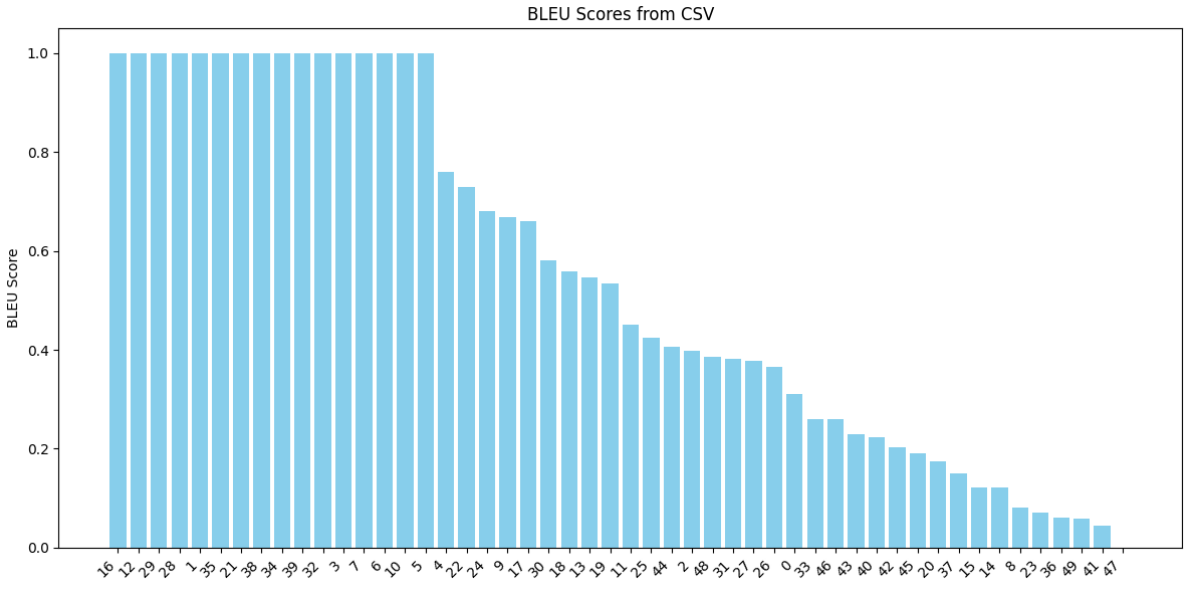

Figure 5: Sorted BLEU scores for Kimi-K2-Instruct on Uzbek translation tasks. The chart shows several high-performing translations with BLEU scores near 1.0, while the rest gradually decrease, reflecting strong overall performance with some variability across examples.

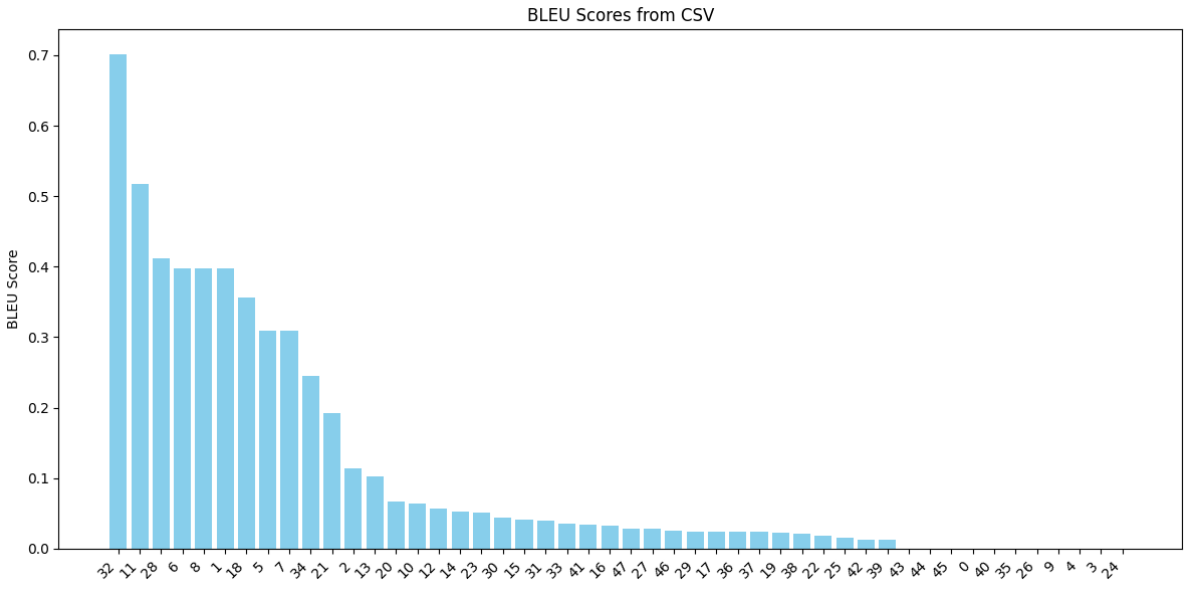

Figure 6: Sorted BLEU scores for Llama-3.2-1B-Instruct, with a couple of successful translations and the highest score reaching up to 0.7, and the remaining being near or equal to 0 across most examples, indicating inconsistent and generally low-quality performance.

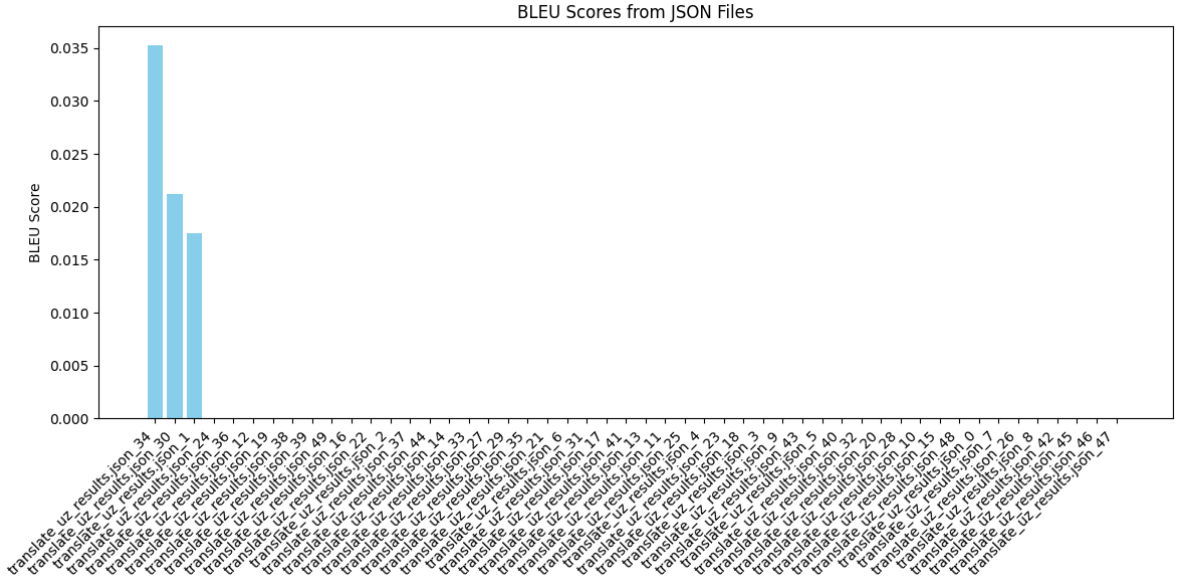

Figure 7: Qwen3-0.6B shows the evaluated system's performance. Only a few files have non-zero scores, with the highest being extremely low at approximately 0.035. The overwhelming majority of the test files resulted in a BLEU score of 0, indicating a complete failure to produce coherent translations. This chart demonstrates consistently poor performance.

For comprehension and generation, no LLM except Kimi-K2-Instruct showed any performance, leaving nothing to represent for the other models. Kimi-K2-Instruct excelled in these tasks, performing exceptionally well across all examples. Each of the questions was

assessed manually and checked for accuracy, confirming its superior capability in understanding and generating Uzbek content.

5 Discussion

The benchmark results highlight the major difficulties that open-source pre-trained LLMs face with low-resource languages like Uzbek, especially without task-specific fine-tuning. Among the seven models tested, only **Kimi-K2-Instruct** showed consistent usability across translation, comprehension, and generation. Its advantage likely comes from exposure to related Turkic languages during pre-training [16], which matches earlier findings that multilingual models often benefit from transfer learning between linguistically close languages [4].

The other models—**DeepSeek-R1-Distill-Qwen-1.5B, Falcon3-1B-Instruct, Gemma-2-2B-it, GLM-Edge-1.5B-Chat, Llama-3.2-1B-Instruct,** and **Qwen3-0.6B**—struggled severely. Common issues included hallucinations, mixing in Turkish or Russian, and near-zero BLEU scores on most translation examples, showing little to no Uzbek coverage in their training data [12–18].

Looking at per-example BLEU distributions (Figures 1–7), we see an important pattern: even models with moderate averages, such as **Gemma-2-2B-it** (0.34), collapse when tested on complex or domain-heavy sentences [14]. This drop suggests that both training data quality and architecture strongly affect performance on morphologically rich, agglutinative languages like Uzbek [9]. In contrast, **Kimi-K2-Instruct** maintained relatively high scores (above 0.3 on more than half of the samples), consistent with its mixture-of-experts design and reported multilingual focus [16].

In comprehension and generation tasks, **Kimi-K2-Instruct** again stood out as the only model producing coherent, relevant answers. Manual evaluation confirmed that it could respond to a range of prompts—from inferential questions to short narratives—making it a promising candidate for fine-tuning with larger Uzbek datasets such as **UzLiB** [9, 10]. By comparison, other models often defaulted to English or produced irrelevant text, repeating problems seen in early multilingual NLP efforts where low-resource languages lacked training data [5, 6].

This study is not without limits. The dataset was small (180 examples) and annotated by a single person, which risks bias even though the annotator is a native speaker [1]. Evaluation for comprehension and generation was also qualitative, which makes cross-model comparison less precise, a challenge noted in earlier Uzbek NLP work [4]. Expanding with UzLiB's grammatical probes [9] or building larger corpora from news archives like *Kun.uz* [3] could make future results more robust.

Finally, compute constraints remain a practical barrier. In Uzbekistan, access to high-end GPUs is limited, and cloud rental (e.g., an A40 on RunPod) is prohibitively expensive. This makes lightweight but effective models, such as Kimi-K2-Instruct, more suitable for local research and applications [3].

Overall, the findings recommend **Kimi-K2-Instruct** as the best starting point for fine-tuning Uzbek NLP models under tight resource conditions. This work builds on resources like

BERTbek [4] and **mGPT-1.3B-Uzbek** [7], helping to bridge the gap between linguistic benchmarks and real-world performance in low-resource settings.

6 Conclusion

This study evaluates seven open-source pre-trained LLMs, with parameter sizes between 0.6B and 2B, on three Uzbek language tasks: translation, comprehension, and generation. The setup reflects Uzbekistan's current resource limitations, where large-scale compute is scarce. Using zero-shot prompting on a 180-example dataset, the results show that **Kimi-K2-Instruct** is the only model reaching practical usability. It achieved an average BLEU score of 0.54 and performed reliably on comprehension and generation tasks, with outputs manually checked for accuracy.

The remaining models—**DeepSeek-R1-Distill-Qwen-1.5B, Falcon3-1B-Instruct, Gemma-2-2B-it, GLM-Edge-1.5B-Chat, Llama-3.2-1B-Instruct,** and **Qwen3-0.6B**—fell short, producing incoherent outputs, mixing in other languages, or hallucinating. Per-example BLEU scores further underline the gap: while most models broke down on anything beyond simple sentences, **Kimi-K2-Instruct** showed stable performance, likely benefiting from exposure to related Turkic languages during training.

For Uzbek NLP practitioners, these results offer a clear direction: **Kimi-K2-Instruct is the most effective foundation for fine-tuning when working with limited compute.** As the country moves toward stronger AI infrastructure, with GPU clusters and data centers on the horizon, this benchmark provides a baseline for developing more capable models. Future work should focus on enlarging datasets, incorporating linguistic test suites such as **UzLiB**, and drawing on local text resources to push performance further and support wider AI adoption in Uzbekistan.

References

[1] Worldometers, “Uzbekistan population,” Worldometers.info, 2025. [Online]. Available: https://www.worldometers.info/world-population/uzbekistan-population/

[2] Qalampir.uz, “Uzbekistan: The influence of other countries’ soft power is growing,” Qalampir.uz, n.d. [Online]. Available: https://qalampir.uz/en/news/uzbekistonda-boshk-a-davlatlar-yumshok-kuchi-ning-ta-siri-oshib-bormok-da-111180

[3] UzDaily, “Uzbekistan to establish major cluster for AI model training,” UzDaily.uz, n.d. [Online]. Available: https://www.uzdaily.uz/en/uzbekistan-to-establish-major-cluster-for-ai-model-training/

[4] A. Sobirov and U. Karimov, “BERTbek: A pre-trained language model for Uzbek,” in *Proc. 3rd Workshop NLP Uralic Turkic Lang. (SIGUL)*, 2024, pp. 45–53. [Online]. Available: https://aclanthology.org/2024.sigul-1.5/

[5] A. Sobirov and U. Karimov, “BERTbek: A pre-trained language model for Uzbek,” in *Proc. 3rd Workshop NLP Uralic Turkic Lang. (SIGUL)*, 2024, pp. 45–53. [Online]. Available: https://aclanthology.org/2024.sigul-1.5/

[6] AI Forever, “mGPT: Multilingual generative pre-trained transformer,” GitHub, n.d. [Online]. Available: https://github.com/ai-forever/mGPT

[7] AI Forever, “mGPT-1.3B-Uzbek model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/ai-forever/mGPT-1.3B-uzbek

[8] Tilmoch, “Tilmoch: Translation and writing assistant for Uzbek,” Multilingual.com, n.d. [Online]. Available: https://multilingual.com/tashkent.aitinkerers.org

[9] UzLiB, “Uzbek linguistic benchmark (UzLiB),” GitHub, n.d. [Online]. Available: https://github.com/uzlib/uzlib

[10] UzLiB, “UzLiB dataset card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/datasets/uzlib/uzlib

[11] Hugging Face, “Hugging Face Hub: Platform for model distribution,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/

[12] DeepSeek, “DeepSeek-R1-Distill-Qwen-1.5B model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[13] TII UAE, “Falcon3-1B-Instruct model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/tiiuae/Falcon3-1B-Instruct

[14] Google, “Gemma-2-2B-it model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/google/gemma-2-2b-it

[15] Z.ai and THUKEG, “GLM-Edge-1.5B-Chat model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/zai-org/glm-edge-1.5b-chat

[16] Moonshot AI, “Kimi-K2-Instruct model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/moonshotai/Kimi-K2-Instruct

[17] Meta AI, “Llama-3.2-1B-Instruct model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[18] Alibaba, “Qwen3-0.6B model card,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/Qwen/Qwen3-0.6B

[19] Hugging Face, “Transformers: Open-source library for NLP,” Hugging Face, n.d. [Online]. Available: https://huggingface.co/docs/transformers/index

[20] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proc. EMNLP*, 2016, pp. 2383–2392.

[21] M. Post, “A call for clarity in reporting BLEU scores,” in *Proc. 3rd Conf. Mach. Transl. (WMT)*, 2018, pp. 186–191.

[22] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Proc. ACL Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

[23] K. Yusupkhujaev, “Uzbek LLM benchmarking code and data,” GitHub, n.d. [Online]. Available: https://github.com/khasanyusupkhujaev/uzbek-llm-bench