KyrgyzLLM-Bench: Benchmarking Kyrgyz Language Understanding

Anonymous ACL submission

Abstract

Evaluating large language models (LLMs) across languages remains challenging, as most multilingual benchmarks rely on translated English datasets, often obscuring linguistic and cultural specificity in the target language. This issue is particularly pronounced for less-resourced languages such as Kyrgyz, where reliable natively authored evaluation data are scarce. Building on previously introduced Kyrgyz-language evaluation datasets, this work reports the first systematic and large-scale evaluation of LLMs in Kyrgyz using the KyrgyzLLM-Bench benchmark suite. KyrgyzLLM-Bench comprises two natively authored datasets—*KyrgyzMMLU* and *KyrgyzRC*—together with carefully translated and manually post-edited versions of *WinoGrande*, *HellaSwag*, *BoolQ*, and *TruthfulQA*. We evaluate 26 open- and closed-source LLMs under zero-shot and few-shot settings, analyzing model performance, cross-lingual transfer, and the impact of translation artifacts on evaluation reliability.

1 Introduction

The rapid progress of large language models (LLMs) has increased the need for robust and diverse evaluation benchmarks. Suites such as MMLU (Hendrycks et al., 2021) have become a de facto standard for assessing reasoning and knowledge capabilities. However, current evaluations remain heavily skewed toward English (Wu et al., 2025), leaving substantial gaps in our understanding of model performance across diverse linguistic and cultural contexts.

To broaden multilingual evaluation, a number of benchmarks have been developed, most commonly by machine-translating English datasets (Lai et al., 2023). While pragmatic, this approach introduces well-documented issues, including subtle translation errors, unnatural phrasing, and "translationese" (Vanmassenhove et al., 2021). More critically, translation often removes cultural and contextual grounding intrinsic to the source material—an essential component of language understanding, particularly in domains such as social sciences, history, and literature (Singh et al., 2025). For less-resourced languages such as Kyrgyz, the lack of high-quality, natively curated evaluation data continues to hinder both systematic research and the development of reliable language technologies.

In this work, we report a systematic and large-scale evaluation of LLMs in Kyrgyz using *KyrgyzLLM-Bench*, a multi-faceted benchmark suite based on previously introduced Kyrgyz-language evaluation datasets. In contrast to evaluations relying solely on translated data, the core components of KyrgyzLLM-Bench are natively authored in Kyrgyz, ensuring linguistic naturalness and cultural relevance.

KyrgyzLLM-Bench comprises: (1) *KyrgyzMMLU*, a large-scale multitask multiple-choice question-answering benchmark with 7,977 items written by curriculum experts and aligned with the Kyrgyz national education registry, covering subjects such as mathematics, physics, literature, and history; (2) *KyrgyzRC*, a native reading-comprehension dataset consisting of 400 questions based on authentic Kyrgyz texts, requiring contextual understanding and multi-sentence reasoning; (3) translated benchmarks: manually post-edited Kyrgyz versions of *WinoGrande*, *HellaSwag*, *BoolQ*, and *TruthfulQA*, ensuring linguistic fidelity and cultural appropriateness. (4) *mainstream model evaluation*: benchmarking a wide range of frontier open- and closed-source LLMs on both native and translated tasks under zero-shot and few-shot settings, providing the first systematic analysis of model performance on complex, culturally grounded Kyrgyz tasks; (5) *open access toolkit*: public release of all datasets, evaluation code, and results.

Based on prior work on translated benchmarks

and cross-lingual evaluation, we put forward the following hypotheses: (1) core reasoning and question-answering capabilities partially transfer from English to Kyrgyz, preserving relative model rankings on structurally similar tasks such as *BoolQ* and *WinoGrande*; (2) event-continuation benchmarks that rely on plausibility judgments, such as *HellaSwag*, are particularly sensitive to translation artifacts, leading to *plausibility shifts*—changes in perceived naturalness and coherence that affect performance in the target language in an unpredictable manner and likely lead to unreliable results.

More broadly, we aim to answer several guiding questions: (i) how well modern LLMs perform on natively authored Kyrgyz-language tasks; (ii) to what extent cross-lingual performance rankings transfer from English to Kyrgyz; (iii) how in-context learning affects performance in a less-resourced, morphologically rich Kyrgyz language; and (iv) how translation artifacts could affect the reliability of measurements.

In the remainder of the paper, we first review the relevant prior research and describe the construction of KyrgyzLLM-Bench, then present extensive evaluations of open-source and proprietary LLMs, and finally discuss cross-lingual transfer, robustness, and implications for low-resource evaluation.

2 Related Work

Large language models (LLMs) have achieved strong performance on a wide range of natural language understanding and reasoning benchmarks, including *GLUE* (Wang et al., 2018), *SuperGLUE* (Wang et al., 2019), and *MMLU* (Hendrycks et al., 2021), as well as commonsense reasoning datasets such as *WinoGrande* (Sakaguchi et al., 2020), *HellaSwag* (Zellers et al., 2019), and *GSM8K* (Cobbe et al., 2021). However, these benchmarks are primarily available in English and other high-resource languages, with limited coverage for low-resource languages such as Kyrgyz. To address this gap, multilingual benchmarks such as *XTREME* (Siddhant et al., 2020) and *FLORES* (Goyal et al., 2022) have been developed to evaluate cross-lingual transfer across typologically diverse languages. Yet many low-resource languages, particularly Central Asian and Turkic languages such as Kyrgyz, remain excluded. One practical approach to creating benchmarks for low-resource languages is to translate existing English datasets. For example, *MMLU* and *COPA* have been translated into Latvian (Bakanovs, 2024), although these translations often lack manual post-editing and may introduce additional noise. The *OKAPI* framework includes Kyrgyz via machine-translated evaluations but does not validate them with native speakers, leading to linguistic "noise" as noted by Jumashev et al. (2025). In this study, four widely used benchmarks—*WinoGrande*, *HellaSwag*, *BoolQ*, and *TruthfulQA*—are translated into Kyrgyz and manually reviewed to ensure cultural and linguistic appropriateness. In addition, the benchmark includes original Kyrgyz-language tasks, including a localized *MMLU* and reading comprehension tests based on official Kyrgyz school exams, whose construction and annotation procedures are described in detail in this paper. This approach is consistent with that of Dargis et al. (2024), who used centralized Latvian high school exams to evaluate LLM performance. Original materials offer high relevance and validity for evaluating real-world LLM performance in low-resource language settings.

This study contributes to the growing body of work on evaluating LLMs in underrepresented languages. By combining translated datasets with culturally grounded, original benchmarks, this work provides new insights into multilingual generalization and the capabilities of open-source LLMs for Kyrgyz, a Turkic language spoken by approximately four million people (about 80% of the population) in the Kyrgyz Republic (Salmorbekova et al., 2023). Linguistic research has explored various facets of the Kyrgyz language; see, for example, the overview by Alekseev and Turatali (2024). However, despite increasing interest from Kyrgyz-speaking communities and a growing number of Kyrgyz language-related projects, both commercial (Kan, 2024) and non-commercial (UNESCO-IITE, 2022), there remains a notable lack of manually annotated datasets for Kyrgyz language processing tasks (Mirzakhalov et al., 2021; Veitsman and Hartmann, 2025; Alekseev and Turatali, 2024). This work aims to address this gap.

3 KyrgyzLLM-Bench Construction

The datasets used in this study were previously announced in a conference talk (Turatali et al., 2025). In contrast to the brief introduction given in the talk by Turatali et al. (2025), this paper provides a detailed description of the datasets and reports large-scale evaluation results that offer actionable

insights for the community.

3.1 Original benchmark: KyrgyzMMLU

KyrgyzMMLU is based on materials from the General Republican Testing (GRT), conducted in Kyrgyzstan since 2002. The GRT is administered by the Center for Educational Assessment and Teaching Methods (CEATM) in cooperation with the Ministry of Education of the Kyrgyz Republic. The test set was officially sourced from the Department for Development of Education Quality under the Ministry of Education and Science and comprises 9 school subjects taught from 6th to 11th grade, as well as specialized categories such as Medicine. The subject distribution is summarized in Table 1. The GRT aims to ensure equal access to higher education through fair and independent testing. The test assesses applicants' ability to successfully continue their studies at a higher education institution and is conducted in Kyrgyz and Russian. The assessment targets reasoning skills and the application of school knowledge.

Regarding test structure, the rejection of tests focused entirely on school subjects was largely due to structural challenges in Kyrgyzstani education. While applicants nationwide must be measured using a single tool, teaching quality across the country remains heterogeneous due to shortages of qualified teachers (mainly in rural areas), lack of educational materials in many schools, and unequal access to technical resources and mass media.

Obviously, applicants aspiring to higher education start from different positions in terms of school-acquired knowledge and skills. Given unequal living and learning conditions, the GRT was designed to mitigate these disparities. This was a strong argument for the chosen main test structure (i.e., the test taken by all applicants).

The Kyrgyzstani GRT uses a multiple-choice format with one correct option (Listing 1 shows a sample question). The primary metric for *KyrgyzMMLU* and *KyrgyzRC* is accuracy.

3.2 Original content benchmark: KyrgyzRC

KyrgyzRC is a natively authored reading-comprehension dataset designed to evaluate understanding and reasoning in Kyrgyz. It consists of 400 manually curated multiple-choice questions drawn from diverse sources, including Kyrgyz Wikipedia, national news articles, literary excerpts, and school-level math problems. Each item consists of a 2–5 sentence passage followed by a question and four answer options, with exactly one correct answer (Listing 2).

The dataset was constructed in collaboration with professional linguists and Kyrgyz language educators to ensure linguistic accuracy and cultural relevance. Each passage was written natively in Kyrgyz, preserving natural flow and incorporating idiomatic expressions and culturally specific references. The questions target a range of reading-comprehension skills: (1) *factual understanding*—identifying explicitly stated information; (2) *inference*—drawing logical conclusions from the text; (3) *vocabulary in context*—interpreting word meaning within context; (4) *reasoning across sentences*—connecting ideas and resolving ambiguities.

KyrgyzRC adopts a multiple-choice format to enable automatic evaluation and consistent scoring, mirroring standardized assessments used in Kyrgyz education. The balance across encyclopedic, journalistic, literary, and mathematical genres supports evaluation across varied linguistic registers and domains. Each entry includes metadata for source type and question type.

KyrgyzRC is, to our knowledge, the first publicly available reading-comprehension benchmark designed specifically for Kyrgyz. It addresses a key gap in evaluating context-sensitive understanding for a less-resourced language. *KyrgyzRC* is publicly released and can serve both as a benchmark and as a resource for developing and testing Kyrgyz language models.

3.3 Translated benchmarks

To complement the natively authored datasets, four widely used English benchmarks were translated into Kyrgyz. Specifically: (1) commonsense reasoning—*HellaSwag* (Zellers et al., 2019), which tests plausible sentence continuation, and *WinoGrande* (Sakaguchi et al., 2020), which tests pronoun resolution in context; (2) reading comprehension—*BoolQ* (Clark et al., 2019), which requires answering natural-language questions given a short context; (3) robustness and factuality—*TruthfulQA* (Lin et al., 2022), which probes a model's tendency to produce truthful answers rather than repeat common misconceptions. *GSM8K* (Cobbe et al., 2021) was also translated into Kyrgyz; however, it was excluded from this study for the reasons described in Appendix B.

Collectively, this set provides a compact, high-signal evaluation of core LLM understanding, reasoning, and robustness. All tasks also appear in the

School subject	#Q
Mathematics	1,169
Biology (Bio)	1,550
Physics (Phys)	1,228
Chemistry (Chem)	1,205
Kyrgyz Literature (Lit)	1,169
Geography (Geog)	640
Kyrgyz History (Hist)	440
Kyrgyz Language (Lang)	360
Medicine (Med)	216

Table 1: School subjects in *KyrgyzMMLU*.

Subject	#Q
Math	100
Kyrgyz Wikipedia	100
Kyrgyz News	100
Kyrgyz Literature	100

Table 2: Subjects in KyrgyzRC.

Question: Эгерде почта аркылуу акча жиберүүнүн кызматы үчүн жиберилүүчү сумманын 10% ын төлөө керек болсо, анда Аскар 600 сомду жиберүү үчүн канча сом төлөйт?

Options: (А) 6 (Б) 10 (В) 60 (Г) 100 (Д) 160

Listing 1: Sample test task: "If the post office takes 10% of the total amount for a money transfer, then how many extra soms will Askar pay to send 600 soms?" The correct answer is (B) 60.

Text: Улуу Кыргыз кагандыгы - 9-кылымда Енисей Кыргыз мамлекетинин күчөп турган мезгилиндеги расмий аталышы. 840-жылы Уйгур кагандыгын талкалап, Улуу Кыргыз дөөлөтү Орхондон Чыгыш Түркстанга, Саян-Алтайдан Сыр-Дарыяга чейинки аймактарды ээлеген. Бул доор кыргыз тарыхында Кыргыз улуу державасы деп аталган. Кагандык 924-жылга чейин жашаган.
Question: Улуу Кыргыз кагандыгы кайсы кылымда күчөп турган?
Options: (А) 9-кылымда. (Б) 8-кылымда. (В) 10-кылымда. (Г) 7-кылымда.

Listing 2: A reading comprehension task: "The Great Kyrgyz Khaganate is the name of the Yenisei Kyrgyz state during its height of power in the IX century. In 840, after defeating the Uyghur Khaganate, the Great Kyrgyz state occupied territories from the Orkhon to East Turkestan, and from the Sayan-Altai to the Syr Darya. This era in Kyrgyz history is called the Kyrgyz Great Power. The Khaganate existed until 924. Q: In which century was the Great Kyrgyz Khaganate at its peak?" ((A) 9-кылымда, in the IX century).

Original English:
He never comes to my home, but I always go to his house because the _ is smaller.
Options: (1) home (2) house

Translated Kyrgyz:
Ал менин үйүмө эч качан келбейт, бирок мен ар дайым анын турак жайына барам, анткени _ кичинекей.
Options: (1) үй (2) турак жай

Listing 3: A translated WinoGrande data point.

Lighteval tool and fall into its higher-level benchmark categories.[1] Listing 3 shows an example from the translated *WinoGrande* benchmark.

For each dataset, the following quality-control procedures were applied: (1) automatic translation: source examples were first translated into Kyrgyz by *Claude 4 Sonnet* and independently by *Gemini2.5 Flash*; (2) ensemble validation: the two outputs were compared to identify lexical or semantic divergences; (3) manual post-editing: Kyrgyz linguists and domain experts reviewed all examples to resolve ambiguities, preserve idiomatic usage, and ensure cultural appropriateness; (4) quality assurance: back-translation checks and spot-checks of 10% of entries confirmed fidelity to original meaning.

The translation and validation were conducted in an academic setting as part of a supervised university course at the Department of Computational Linguistics. A total of 19 students and 4 supervisors/curators, all native Kyrgyz speakers, participated in the process. The workflow followed a peer-review structure: each instance was edited by one student and independently verified by another, with oversight from course instructors.

All participants were informed in advance about the purpose of data preparation, the intended research use of the datasets, and the goals of the study. Participation was voluntary and took place as part of the students' regular practical training. The tasks were aligned with the course learning objectives and the participants' primary field of study. In consultation with university representatives, it was verified that course credit and practical experience constituted adequate compensation for the time and effort required. No sensitive personal data were collected.

4 Experimental Setup

We conducted two experimental setups. The first evaluates 14 open-source models on the full benchmark, while the second uses a condensed subset (*KyrgyzLLM Tiny Bench*) to evaluate 12 proprietary models. We additionally evaluated the original English benchmarks as a cross-lingual reference, selecting 18 MMLU subjects analogous to those in KyrgyzMMLU. [2] All evaluations were conducted

[1] See *Lighteval*'s README: https://github.com/huggingface/lighteval/blob/main/README.md

[2] Selected: college_biology, ..._chemistry, ..._mathematics, ..._medicine, ..._physics, ..._mathematics,

Model	MMLU	WinoGrande	BoolQ	HellaSwag	TruthfulQA	Avg
Zero-shot evaluation						
Qwen2.5-0.5B-Instruct	42.7	53.4	59.9	40.6	34.4	46.2
Qwen2.5-1.5B-Instruct	58.6	58.9	75.8	50.8	38.9	56.6
Qwen2.5-3B-Instruct	64.4	66.4	68.6	56.4	50.3	61.2
Qwen2.5-7B-Instruct	71.5	67.1	82.5	**62.0**	**56.2**	**67.9**
Qwen3-0.6B	38.3	52.5	45.5	37.6	34.8	41.7
Qwen3-1.7B	56.0	58.3	30.3	46.1	37.6	45.7
Qwen3-4B	71.6	61.8	79.6	52.3	45.8	62.2
Qwen3-8B	**75.9**	64.8	**85.1**	57.1	45.4	65.7
Gemma-3-270m	20.9	49.6	0.0	25.0	24.1	23.9
Gemma-3-1b-it	33.8	55.6	71.4	43.4	31.5	47.1
Gemma-3-4b-it	56.5	60.7	76.6	55.9	43.3	58.6
Llama-3.2-1B-Instruct	45.6	58.4	69.4	45.7	35.3	50.9
Llama-3.2-3B-Instruct	58.2	64.3	72.6	53.3	42.6	58.2
Llama-3.1-8B-Instruct	65.0	**71.3**	82.5	59.8	46.2	65.0
Few-shot evaluation (5-shot; 10-shot for HellaSwag)						
Qwen2.5-0.5B-Instruct	43.8	54.5	59.7	39.9	36.7	46.9
Qwen2.5-1.5B-Instruct	59.2	60.7	78.9	50.3	42.6	58.3
Qwen2.5-3B-Instruct	65.3	65.3	79.9	56.3	52.1	63.8
Qwen2.5-7B-Instruct	74.2	70.4	85.4	**63.2**	**57.2**	**70.1**
Qwen3-0.6B	47.7	53.8	53.1	37.8	39.6	46.4
Qwen3-1.7B	61.8	56.7	80.3	46.2	42.4	57.5
Qwen3-4B	72.7	61.7	86.0	55.6	47.4	64.7
Qwen3-8B	**77.6**	66.9	**87.6**	58.1	48.3	67.7
Gemma-3-270m	20.9	49.6	0.0	25.0	28.3	24.8
Gemma-3-1b-it	36.9	53.1	70.9	42.3	37.4	48.1
Gemma-3-4b-it	56.2	63.1	0.0	24.6	48.8	38.5
Llama-3.2-1B-Instruct	43.8	56.7	70.8	45.2	40.0	51.3
Llama-3.2-3B-Instruct	58.6	64.2	78.3	53.2	47.0	60.3
Llama-3.1-8B-Instruct	65.5	**74.0**	86.7	62.0	52.4	68.1

Table 3: Zero-shot and few-shot evaluation results on English benchmarks (accuracy %). Colors show gains/losses compared to the zero-shot case under this revised averaging.

Model	$\text{Kyrgyz}_{\text{MMLU}}$	$\text{Kyrgyz}_{\text{RC}}$	WinoGrande	BoolQ	HellaSwag	TruthfulQA	Avg
Zero-shot evaluation							
Qwen2.5-0.5B-Instruct	27.4	53.2	**51.5**	37.9	14.6	33.5	36.4
Qwen2.5-1.5B-Instruct	27.9	60.5	50.1	38.6	22.9	32.5	38.8
Qwen2.5-3B-Instruct	28.6	66.0	50.5	**59.4**	22.0	34.2	43.4
Qwen2.5-7B-Instruct	31.5	70.0	48.7	56.3	10.0	34.1	41.8
Qwen3-0.6B	26.0	61.8	49.8	38.0	11.1	29.9	36.1
Qwen3-1.7B	27.9	61.8	48.9	40.4	24.6	29.6	38.9
Qwen3-4B	30.3	68.2	49.0	38.3	24.5	32.9	40.5
Qwen3-8B	**32.1**	71.8	51.0	39.2	24.6	**34.7**	42.2
Gemma-3-270m	27.5	56.8	48.3	37.9	17.4	**34.7**	37.1
Gemma-3-1b-it	26.7	58.2	50.0	37.9	24.4	34.0	38.5
Gemma-3-4b-it	30.3	70.2	50.6	58.3	24.6	**34.7**	**44.8**
Llama-3.1-8B-Instruct	31.0	**75.2**	50.6	50.3	**26.6**	33.7	44.6
Llama-3.2-1B-Instruct	26.3	58.2	49.4	38.3	0.2	30.1	33.7
Llama-3.2-3B-Instruct	27.8	64.2	49.1	43.1	24.5	31.5	40.0
Few-shot evaluation (5-shot; 10-shot for HellaSwag)							
Qwen2.5-0.5B-Instruct	25.4	54.0	49.7	61.0	25.9	33.4	41.6
Qwen2.5-1.5B-Instruct	28.7	67.5	50.1	58.0	26.5	32.9	43.9
Qwen2.5-3B-Instruct	34.0	73.2	51.3	57.4	23.7	34.4	45.7
Qwen2.5-7B-Instruct	38.5	74.8	50.4	64.6	17.8	36.2	47.1
Qwen3-0.6B	26.8	59.5	50.1	60.1	26.4	30.0	42.2
Qwen3-1.7B	30.8	71.2	48.6	62.0	25.2	30.3	44.7
Qwen3-4B	38.5	77.2	48.1	74.0	24.7	32.5	49.2
Qwen3-8B	**44.5**	**81.8**	50.6	**76.9**	26.4	35.8	**52.7**
Gemma-3-270m	27.0	53.2	48.7	61.5	**27.6**	36.6	42.4
Gemma-3-1b-it	26.5	38.0	48.9	62.8	23.5	31.3	38.5
Gemma-3-4b-it	29.5	25.0	49.6	62.1	24.6	**50.0**	40.1
Llama-3.1-8B-Instruct	38.1	80.5	**51.6**	75.5	21.9	34.4	50.3
Llama-3.2-1B-Instruct	26.1	45.8	49.7	62.0	25.8	30.3	40.0
Llama-3.2-3B-Instruct	29.4	64.8	48.9	62.3	25.3	32.9	43.9

Table 4: Combined zero- and few-shot evaluation on Kyrgyz benchmarks (accuracy %). Cell colors indicate few-shot gains or losses relative to zero-shot.

using *Lighteval* (Habib et al., 2023).

We use temperature $= 0.6$ and top-$p = 0.9$, following baseline generation settings commonly recommended in established LLM inference frameworks (NVIDIA, 2025). These values reflect a commonly used moderately stochastic decoding regime that balances stability and diversity (Gao et al., 2023; Hugging Face, 2024).

Open-source models. We evaluated 14 open-source models from Qwen (Bai et al., 2023), Gemma (GemmaTeam et al., 2024), and LLaMA (Grattafiori et al., 2024) families in zero-shot and few-shot settings. Inference was performed on rented NVIDIA RTX6000 Ada and NVIDIA L40S GPUs.

high_school_biology, ..._chemistry, ..._computer_science, ..._european_history, ..._geography, ..._mathematics, ..._physics, ..._statistics, ..._us_history, ..._world_history, prehistory, and professional_medicine.

Few-shot evaluation used 5 examples for most tasks. For *HellaSwag*, we use 10-shot prompting, following common practice in widely adopted benchmarking frameworks and leaderboards.

Responses were parsed using a standard *Lighteval* regex to extract answers, and accuracy was reported as the percentage of correct responses. For open models, we additionally report parallel English baselines in Table 3 as a cross-lingual reference. A detailed analysis is provided in Sections 5–6.

Proprietary models. To assess state-of-the-art closed-source model performance on Kyrgyz in a cost-effective yet representative manner, we conducted an evaluation using a condensed benchmark, *KyrgyzLLM Tiny Bench*. It consists of 100 randomly selected questions per subject from the original suite, sampled with a fixed random seed for reproducibility. Evaluated models include GPT-

5.1 and GPT-5 Mini (OpenAI); Claude 4.5 Sonnet and Haiku (Anthropic); Gemini 2.5 Flash (Google); Grok 4.1 Fast (xAI); Mistral Medium 3.1 and Large (Mistral); DeepSeek V3.2 Exp; Qwen3-Max; Kimi K2 Turbo; and GigaChat-2 Max. Results for Gemini 2.5 Flash (marked *) were affected by safety-related refusals; the reported scores should therefore be treated with caution. The in-context supervision strategy matched that used for open-source models.

5 Evaluation Results

We report accuracy (%) and macro-averages (arithmetic mean). Because tasks probe different LLM capabilities, averages over all tasks should be interpreted as coarse summaries rather than definitive diagnostic scores.

Open-source models. Table 4 shows consistent within-family scaling, with larger instruction-tuned models performing best. In few-shot Kyrgyz evaluation, Qwen3-8B achieves the highest average accuracy (52.7The largest few-shot gains are observed on *BoolQ* (Qwen3-8B: 39.2→76.9). English baselines (Table 3) show Qwen2.5-7B-Instruct leading both zero-shot (67.9%) and few-shot (70.1%).

Proprietary models. Table 5 shows *Claude 4.5 Sonnet* achieving the highest average accuracy in both zero-shot (74.82%) and few-shot (77.06%). Performance is near ceiling on reading comprehension, particularly RC_{Wiki} (98–100% for several models), whereas science-heavy *KyrgyzMMLU* subjects exhibit substantially higher variance, suggesting that out-of-context factual and numerical reasoning in Kyrgyz remains more challenging than extracting answers from a given passage. Few-shot prompting often improves results, but gains are model- and task-dependent.

Scaling and few-shot effects. Few-shot effects differ between open-source and proprietary models. For open-source models, in-context demonstrations yield clear gains on *KyrgyzRC* and *BoolQ*, more moderate improvements on *KyrgyzMMLU*, and limited gains on *HellaSwag* (Table 4). In contrast, proprietary models exhibit less consistent few-shot behavior: *BoolQ* accuracy often stagnates or decreases relative to zero-shot performance, while gains on *KyrgyzRC* are typically small due to already high zero-shot accuracy (Table 5).

Cross-lingual consistency. Model rankings in English and Kyrgyz are broadly preserved on *WinoGrande/BoolQ*, and (to a lesser extent) on *MMLU*, indicating partial transferability of core reasoning and comprehension capabilities across languages. In contrast, *HellaSwag* exhibits the largest performance gap between English and Kyrgyz. This pattern is consistent with the plausibility-shift hypothesis, which posits that translation-induced changes in discourse flow and event continuity disproportionately affect event-completion tasks.

6 Discussion

Scaling trends. Across both Kyrgyz and English evaluations, performance consistently improves as model capacity increases within the same architectural family. Higher-capacity and more extensively instruction-tuned variants outperform smaller counterparts across most subject areas. These trends align with established scaling behaviors in multilingual benchmarks, indicating that model capacity and training quality remain key drivers even for less-resourced languages such as Kyrgyz.

Effect of in-context learning (ICL). Few-shot prompting often improves results on *KyrgyzRC* and *BoolQ* for certain model families, where contextual examples aid comprehension and answer selection. Observed gains of approximately $+5$–$+10$ points suggest that these datasets are suitable for assessing in-context adaptation. In contrast, *KyrgyzMMLU* shows more modest benefits. Chain-of-thought prompting may further improve results but remains outside the scope of this study.

However, our results indicate that the effects of few-shot prompting are not uniform across tasks or models. Few-shot benefits are strongly model- and task-dependent. Using the *Lighteval* framework with Kyrgyz-translated prompts, few-shot prompting occasionally reduces accuracy, suggesting that ICL effects in translated-prompt scenarios are more variable than often expected. In particular, open-source models show substantial improvements on *KyrgyzRC* and *BoolQ* under in-context learning, whereas proprietary models—already strong in zero-shot settings—exhibit limited or even negative gains, especially on *BoolQ*. For reading comprehension, performance saturation limits observable few-shot improvements among closed models.

The pronounced degradation on *HellaSwag* provides empirical support for the plausibility-shift hypothesis and underscores the limitations of directly translating event-continuation benchmarks for less-resourced languages.

The evaluation pipeline, scoring rules, and an-

Model	MMLU										RC									Total
	Bio	Chem	Geog	Hist	Lang	Lit	Math	Med	Phys	**Avg**	Lit	Math	News	Wiki	**Avg**	**BoolQ**	**Hella**	**TQA**	**Wino**	**avg**
Zero-shot evaluation																				
GigaChat-2-Max	64.0	60.0	60.0	70.0	55.0	43.0	49.0	64.0	51.0	57.33	89.0	66.0	87.0	99.0	85.25	83.0	48.0	44.0	48.0	63.76
Claude 4.5 Haiku	53.0	61.0	52.0	69.0	71.0	36.0	54.0	60.0	76.0	59.11	92.0	88.0	86.0	98.0	91.00	87.0	0.0	61.0	48.0	64.06
Claude 4.5 Sonnet	61.0	69.0	57.0	81.0	82.0	49.0	53.0	72.0	77.0	**66.78**	93.0	98.0	89.0	99.0	**94.75**	95.0	77.0	69.0	51.0	**74.82**
DeepSeek V3.2 Exp	60.0	65.0	63.0	83.0	71.0	43.0	53.0	61.0	77.0	64.00	94.0	84.0	84.0	98.0	90.00	84.0	56.0	66.0	56.0	70.47
Gemini 2.5 Flash *	43.0	42.0	45.0	57.0	85.0	31.0	50.0	43.0	54.0	50.00	96.0	81.0	82.0	95.0	88.50	79.0	34.0	4.0	49.0	57.06
GPT-5 Mini	50.0	39.0	61.0	70.0	55.0	38.0	43.0	53.0	42.0	50.11	88.0	66.0	84.0	95.0	83.25	86.0	59.0	62.0	49.0	61.35
GPT-5.1	62.0	54.0	67.0	82.0	77.0	46.0	37.0	63.0	48.0	59.56	94.0	85.0	89.0	98.0	91.50	88.0	79.0	68.0	55.0	70.12
Grok 4.1 Fast	43.0	50.0	64.0	67.0	44.0	41.0	32.0	56.0	40.0	48.56	92.0	64.0	87.0	100	85.75	86.0	59.0	48.0	56.0	60.53
Kimi k2 Turbo	58.0	54.0	63.0	71.0	70.0	40.0	42.0	58.0	55.0	56.78	94.0	70.0	89.0	99.0	88.00	86.0	59.0	50.0	58.0	65.70
Mistral Large	60.0	65.0	63.0	78.0	63.0	38.0	50.0	62.0	74.0	61.44	94.0	81.0	89.0	98.0	90.50	86.0	64.0	56.0	52.0	68.94
Mistral Medium 3.1	46.0	54.0	64.0	67.0	47.0	45.0	41.0	63.0	55.0	53.56	94.0	68.0	88.0	98.0	87.00	76.0	56.0	50.0	50.0	62.47
Qwen3-Max	57.0	63.0	61.0	72.0	72.0	46.0	49.0	59.0	57.0	59.56	92.0	85.0	88.0	99.0	91.00	90.0	66.0	65.0	53.0	69.06
Few-shot evaluation																				
GigaChat-2-Max	65.0	57.0	65.0	77.0	72.0	44.0	46.0	64.0	47.0	59.67	89.0	65.0	89.0	94.0	84.25	69.0	45.0	36.0	50.0	63.82
Claude 4.5 Haiku	65.0	62.0	63.0	75.0	78.0	37.0	53.0	55.0	74.0	62.44	67.0	86.0	72.0	57.0	70.50	14.0	62.0	74.0	50.0	62.59
Claude 4.5 Sonnet	63.0	68.0	68.0	84.0	84.0	51.0	53.0	70.0	76.0	**68.56**	92.0	97.0	91.0	99.0	**94.75**	79.0	84.0	96.0	55.0	**77.06**
DeepSeek V3.2 Exp	63.0	62.0	64.0	82.0	61.0	47.0	49.0	54.0	55.0	59.67	94.0	86.0	92.0	100.0	93.00	70.0	35.0	59.0	48.0	61.01
Gemini 2.5 Flash *	57.0	61.0	74.0	79.0	89.0	52.0	48.0	63.0	69.0	65.78	97.0	88.0	78.0	95.0	89.50	47.0	38.0	21.0	47.0	59.88
GPT-5 Mini	58.0	66.0	73.0	82.0	81.0	36.0	52.0	58.0	77.0	64.78	91.0	93.0	89.0	98.0	92.75	41.0	45.0	26.0	48.0	59.74
GPT-5.1	65.0	62.0	68.0	83.0	85.0	44.0	41.0	64.0	51.0	62.56	93.0	84.0	90.0	98.0	91.25	13.0	79.0	41.0	58.0	68.84
Grok 4.1 Fast	51.0	52.0	67.0	71.0	56.0	41.0	36.0	58.0	50.0	53.56	93.0	58.0	90.0	100.0	85.25	16.0	50.0	67.0	53.0	59.35
Kimi k2 Turbo	58.0	57.0	63.0	72.0	72.0	47.0	42.0	60.0	48.0	57.67	96.0	68.0	92.0	100.0	89.00	79.0	70.0	79.0	51.0	67.88
Mistral Large	61.0	56.0	69.0	84.0	75.0	41.0	41.0	68.0	56.0	61.22	93.0	78.0	91.0	99.0	90.25	63.0	35.0	53.0	57.0	61.65
Mistral Medium 3.1	57.0	52.0	68.0	76.0	58.0	48.0	42.0	64.0	67.0	59.11	95.0	79.0	90.0	100.0	91.00	17.0	46.0	49.0	53.0	62.29
Qwen3-Max	65.0	64.0	64.0	79.0	76.0	42.0	46.0	61.0	57.0	61.56	91.0	87.0	88.0	100.0	91.50	73.0	63.0	98.0	50.0	70.82

Table 5: Zero-show and few-shot accuracy (%) on Kyrgyz-LLM Tiny Bench (100 sample subset). * Gemini 2.5 Flash scores were impacted by safety filter refusals.

Model	Medicine		History		Literature		Lang		Biology		Chemistry		Math		Physics		Geography		Average	
	ZS	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ
Qwen2.5-0.5B-Instruct	31.5	-2.3	33.9	-2.3	26.3	-2.4	27.2	-8.3	30.8	-2.8	20.7	+2.9	22.8	+1.4	26.1	-0.4	26.9	-3.8	27.4	-2.0
Qwen2.5-1.5B-Instruct	29.6	+3.7	35.9	-4.1	27.7	-0.1	29.4	-6.6	31.8	-1.4	20.5	+5.6	25.1	+4.8	26.6	+2.1	24.8	+3.2	27.9	+0.8
Qwen2.5-3B-Instruct	32.9	+6.0	36.8	+3.4	27.7	+3.2	26.7	+0.8	32.1	+4.2	23.0	+7.4	26.9	+5.3	26.5	+4.6	25.0	+10.5	28.6	+5.1
Qwen2.5-7B-Instruct	**35.2**	+4.6	40.7	+4.1	29.1	+1.7	25.3	+10.8	**34.1**	+5.6	26.0	+11.4	31.8	+6.4	**30.3**	+9.2	30.9	+9.4	31.5	+7.0
Qwen3-0.6B	28.2	+1.9	33.4	+0.0	27.1	+1.6	26.1	-3.3	30.4	-2.8	19.5	+5.3	21.4	+5.0	24.9	-1.2	23.1	+0.8	26.0	+0.8
Qwen3-1.7B	29.6	+1.0	36.4	-1.2	28.3	+3.3	25.3	-6.7	30.7	+5.6	22.1	+10.1	26.9	+4.0	25.7	+2.3	26.1	+7.5	27.9	+2.9
Qwen3-4B	28.7	+9.3	36.1	+4.8	28.0	+3.7	27.5	+0.3	32.8	+8.3	**28.0**	+17.2	32.1	+4.8	29.9	+13.4	29.5	+12.5	30.3	+8.2
Qwen3-8B	32.4	+10.2	42.5	+10.7	29.8	+3.2	28.9	+6.1	33.6	+15.3	27.4	+21.2	**34.0**	+6.8	28.8	+18.0	31.7	+19.7	**32.1**	+12.4
Gemma-3-270m	32.4	-7.4	31.6	-5.7	27.9	+1.2	**34.7**	+6.7	30.4	-1.0	19.3	+2.4	21.8	-0.2	24.8	-2.6	24.4	+2.0	27.5	-0.5
Gemma-3-1b-it	27.8	+0.9	34.3	-4.5	27.0	-1.1	30.8	+0.3	29.4	+0.7	21.7	+0.4	22.0	-2.1	24.1	+0.8	23.6	+2.3	26.7	-0.2
Gemma-3-4b-it	33.3	+5.6	39.8	+3.6	**29.9**	-1.2	29.4	+11.4	32.6	-5.4	23.3	-2.8	28.1	-9.1	28.3	-5.2	28.3	-4.4	30.3	-0.8
Llama-3.1-8B-Instruct	**35.2**	+1.8	**43.9**	+4.5	29.4	+0.9	25.0	+13.1	32.8	+10.5	24.9	+10.5	27.7	+4.7	27.2	+5.9	**32.5**	+12.7	31.0	+7.1
Llama-3.2-1B-Instruct	26.9	+0.9	29.8	-4.1	27.8	+1.0	28.9	+0.0	30.8	-4.3	21.8	+1.4	22.0	+2.1	25.2	-1.6	23.4	+3.2	26.3	-0.2
Llama-3.2-3B-Instruct	27.3	+2.3	34.8	-2.1	27.8	-1.2	25.3	+2.8	29.6	+2.9	23.7	+2.8	26.7	+1.6	25.2	+1.3	29.5	+4.6	27.8	+1.6

Table 6: Zero-shot accuracy on *KyrgyzMMLU* (ZS, %); few-shot change (Δ = few-shot − zero-shot, % points).

Model	Lit		Math		News		Wiki		Avg	
	ZS	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ	**ZS**	Δ
Qwen2.5-0.5B-Instruct	67.0	+12.0	32.0	-3.0	44.0	0.0	70.0	-6.0	53.3	+0.7
Qwen2.5-1.5B-Instruct	77.0	+8.0	45.0	-7.0	48.0	+22.0	72.0	+5.0	60.5	+7.0
Qwen2.5-3B-Instruct	79.0	+5.0	50.0	-6.0	60.0	+12.0	75.0	+18.0	66.0	+7.3
Qwen2.5-7B-Instruct	81.0	+5.0	55.0	-14.0	61.0	+15.0	83.0	+13.0	70.0	+4.8
Qwen3-0.6B	74.0	+2.0	50.0	-14.0	52.0	+5.0	71.0	-2.0	61.8	-2.3
Qwen3-1.7B	66.0	+13.0	53.0	-4.0	61.0	+11.0	67.0	+18.0	61.8	+9.5
Qwen3-4B	80.0	0.0	54.0	+2.0	63.0	+16.0	76.0	+18.0	68.3	+9.0
Qwen3-8B	80.0	+8.0	66.0	-3.0	66.0	+16.0	75.0	+19.0	71.8	+10.0
Gemma-3-270m	75.0	-10.0	28.0	-13.0	49.0	-2.0	75.0	+11.0	56.8	-3.5
Gemma-3-1b-it	79.0	-51.0	43.0	-19.0	45.0	-22.0	66.0	+11.0	58.3	-20.3
Gemma-3-4b-it	82.0	-82.0	50.0	-50.0	71.0	-71.0	78.0	+22.0	70.3	-45.3
Llama-3.1-8B-Instruct	82.0	+3.0	65.0	-7.0	75.0	+10.0	79.0	+15.0	75.3	+5.2
Llama-3.2-1B-Instruct	71.0	-25.0	39.0	-15.0	53.0	-17.0	70.0	+7.0	58.3	-12.5
Llama-3.2-3B-Instruct	77.0	-5.0	45.0	0.0	62.0	-9.0	73.0	+16.0	64.3	+0.5

Table 7: Zero-shot accuracy on *KyrgyzRC* (ZS, %) and few-shot change (Δ = few-shot − zero-shot, % points).

swer extraction procedures are identical in zero- and few-shot settings; observed differences therefore arise from model outputs rather than from the evaluation protocol.

Cross-lingual transfer and translation effects. CComparing English and Kyrgyz results reveals broadly preserved family-wise rankings on *MMLU*, *WinoGrande*, and *BoolQ*, suggesting partial transfer of core reasoning and comprehension capabilities across languages. In contrast, *HellaSwag* displays a pronounced performance gap, with consistently low accuracy in Kyrgyz compared to English. This gap is attributable to translation-induced plausibility shifts—changes in colloquial flow, discourse markers, and event continuity that disrupt completion naturalness. This supports the use of natively authored event-continuation benchmarks rather than literal translations.

Dataset quality and cultural alignment. The native components of KyrgyzLLM-Bench, particularly *KyrgyzMMLU* and *KyrgyzRC*, show that LLMs trained predominantly on non-Turkic corpora exhibit partial transfer, but with accuracy sub-

stantially below English counterparts. This highlights both data scarcity and cultural–linguistic mismatches, as idioms, syntax, and referential forms in Kyrgyz differ markedly from Indo-European patterns. Consequently, even instruction-tuned multilingual models may misinterpret pragmatic cues and culturally grounded reasoning. Expanding native Kyrgyz corpora and developing pretraining data with balanced linguistic registers are essential next steps.

Actionable recommendations. Based on our analysis, we suggest the following improvements for future KyrgyzLLM-Bench releases and for practical use of multilingual LLMs on Kyrgyz-language tasks: (1) enforce strict multiple-choice formatting in prompts and robust parsing for answer extraction; (2) audit translated datasets (especially *HellaSwag*) for cultural and plausibility alignment; consider fully native Kyrgyz rewrites; (3) provide subject- and genre-level breakdowns for *KyrgyzMMLU* and *KyrgyzRC* to reveal domain-specific strengths and weaknesses; (4) explore chain-of-thought prompting and rationale-based few-shot examples to test higher-order reasoning; (5) consider logit-based option scoring as a potential alternative to text-based parsing in order to reduce format sensitivity and improve replicability.

Overall, model scaling, instruction tuning, and, in some cases, in-context learning contribute to improved performance on Kyrgyz-language tasks, yet the gap between English and Kyrgyz remains substantial. This disparity reflects imbalances in multilingual pretraining data and underscores the need for more culturally grounded evaluation and training resources.

7 Conclusion

We present a systematic evaluation of large language models for Kyrgyz using *KyrgyzLLM-Bench*, analyzing model performance under zero-shot and few-shot settings and providing a detailed account of benchmark composition, construction, and annotation. It consists of three components: (i) *KyrgyzMMLU*, a large-scale multitask multiple-choice dataset derived from the national curriculum; (ii) *KyrgyzRC*, a native reading comprehension dataset built from authentic Kyrgyz texts across encyclopedic, literary, journalistic, and mathematical domains; (iii) a translated benchmark set encompassing *WinoGrande*, *HellaSwag*, *BoolQ*, and *TruthfulQA*, enabling cross-lingual evaluation of commonsense reasoning, comprehension, and factual robustness. We evaluated 26 multilingual open-source and proprietary LLMs under zero-shot and few-shot conditions, revealing substantial variability across tasks, subjects, and prompting regimes. While modern instruction-tuned models demonstrate notable generalization capabilities, their performance on natively authored Kyrgyz tasks remains substantially below English baselines, highlighting persistent challenges in less-resourced and morphologically rich languages.

Our analysis further shows that evaluation methodology strongly influences measured outcomes, particularly for benchmarks that rely on translated prompts or culturally sensitive plausibility judgments. Tasks such as *HellaSwag* illustrate the limitations of automatic multilingual benchmarking and underscore the importance of natively authored or carefully post-edited datasets for reliable and interpretable evaluation.

Across multiple tasks, we observe that few-shot prompting can lead to unpredictable performance changes, including accuracy drops relative to zero-shot settings, especially for translated benchmarks and for models already operating near saturation. We hypothesize that this instability arises from a combination of factors: prompt translation artifacts, increased sensitivity to example ordering and surface form in morphologically rich languages, and interactions between in-context demonstrations and instruction-tuning objectives that were predominantly optimized for English or other languages. As a result, few-shot evaluation in low-resource languages should not be assumed to be uniformly beneficial and must be interpreted with caution, particularly when translated prompts or plausibility-based tasks are involved.

KyrgyzLLM-Bench fills a critical gap in the evaluation of less-resourced languages by providing culturally grounded benchmarks for Kyrgyz, an underrepresented Turkic language, and enabling systematic analysis of model behavior across native and translated tasks. We hope this work will motivate broader inclusion of Kyrgyz in multilingual benchmarks, support more equitable progress in LLM development for Central Asian languages, and improve the accessibility of AI technologies across diverse linguistic communities. We release datasets, evaluation code, and model results to facilitate future research and reproducibility.

We used OpenAI's ChatGPT service to improve the readability of selected parts of this paper.

Limitations

Several limitations of this study should be noted.

First, performance on translated event-continuation benchmarks, most notably Kyrgyz *HellaSwag*, likely underestimates attainable model capability due to translation-induced plausibility shifts. Changes in discourse flow, event sequencing, and colloquial coherence introduced during translation can substantially alter task difficulty and completion naturalness. As a result, low scores on such benchmarks should not be interpreted as definitive evidence of weak commonsense reasoning in Kyrgyz. Future work will address this limitation by developing natively authored event-continuation datasets.

Second, few-shot evaluation with prompts and exemplars translated into Kyrgyz introduces additional sources of sensitivity that complicate direct comparison with zero-shot results. Few-shot performance was observed to be highly variable across tasks and models, and in some cases degraded relative to zero-shot evaluation. This instability likely reflects interactions between translated prompt structure, example ordering, morphological complexity, and instruction-tuning objectives optimized primarily for English. We therefore recommend interpreting few-shot results in translated-prompt settings with caution and considering zero-shot performance as a complementary and more stable reference point.

Third, while multiple-choice formatting enables automatic evaluation and comparability across models, deviations from strict answer formatting can still affect measured accuracy. Enforcing stricter multiple-choice constraints in prompting and adopting alternative scoring strategies, such as logit-based option ranking, may reduce format sensitivity and improve robustness in future evaluations.

Finally, comparisons involving proprietary models are subject to external service constraints beyond the control of this study. In particular, Gemini 2.5 Flash exhibited safety-related refusals for otherwise benign Kyrgyz-language prompts, which affected result completeness and reliability. Consequently, scores reported for such models should be interpreted with caution, as they may reflect service-level filtering behavior rather than intrinsic model capability.

References

Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: challenges, progress, and future. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 3–39. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

B. Bakanovs. 2024. Large language model evaluation and improvements for the latvian language. Master's thesis, University of Latvia.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, MN, USA.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, and *et al.* 2021. Training verifiers to solve math word problems. In *Proc. International Conference on Learning Representations (ICLR)*, Virtual.

Roberts Darģis, Guntis Barzdins, Inguna Skadiņa, and Baiba Saulīte. 2024. Evaluating open-source llms in low-resource languages: Insights from latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Albert DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and 1 others. 2023. Language model evaluation harness. `https://github.com/EleutherAI/lm-evaluation-harness`.

GemmaTeam, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, and *et al.* 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of ACL*.

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 540 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hugging Face. 2024. Open llm leaderboard (archived evaluation protocol). https://huggingface.co/docs/leaderboards/en/open_llm_leaderboard/archive.

Murat Jumashev, Alina Tillabaeva, Aida Kasieva, Turgunbek Omurkanov, Akylai Musaeva, Meerim Emil Kyzy, Gulaiym Chagataeva, and Jonathan Washington. 2025. The kyrgyz seed dataset submission to the wmt25 open language data initiative shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1088–1102.

Anastasia Kan. 2024. Akylai smart speaker: Artificial intelligence speaking kyrgyz language (june 18th, 2024). https://web.archive.org/web/20240619010036/https://24.kg/english/296874_AkylAI_smart_speaker_Artificial_intelligence_speaking_Kyrgyz_language/. Accessed: 2024-09-14.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and *et al.* 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Z. Lin, J. Hilton, and O. Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Behzod Moydinboyev, Sardana Ivanova, and *et al.* 2021. Evaluating multiway multilingual nmt in the turkic languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 518–530.

NVIDIA. 2025. Nvidia nemo microservices: Model parameter tuning guidelines. https://docs.nvidia.com/nemo/microservices/latest/design-synthetic-data-from-scratch-or-seeds/configure-models.html.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proc. Thirty-Fourth AAAI Conf. on Artificial Intelligence (AAAI)*, New York, NY, USA.

R. Salmorbekova, A. Alymbaev, and A. Tukhtamatov. 2023. Kyrgyz and russian languages in the eurasian space. *Bulletin of Science and Practice*, 9(6):722–733. In Russian.

Aditya Siddhant, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the International Conference on Machine Learning*, volume 2020, pages 4411–4421.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Timur Turatali, Aida Turdubaeva, Islam Zhenishbekov, Zhoomart Suranbaev, Anton Alekseev, and Rustem Izmailov. 2025. Bridging the gap in less-resourced languages: Building a benchmark for kyrgyz language models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 1673–1677.

UNESCO-IITE. 2022. A chatbot for teenagers about puberty, relationships, and health launched in kyrgyzstan (may 24th, 2022). http://web.archive.org/web/20240525072322/https://iite.unesco.org/highlights/oilo-chatbot-sex-ed-kyrgyzstan-en/. Accessed: 2024-09-14.

E. Vanmassenhove, D. Shterionov, and M. Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Yana Veitsman and Mareike Hartmann. 2025. Recent advancements and challenges of turkic central asian language processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 309–324.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, and Felix Hill. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018.

Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 353–355.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China.

A Prompting strategies

For benchmarks other than *KyrgyzMMLU* or *KyrgyzRC*, the prompts we have used are direct translations of original English queries to Kyrgyz; in this section, we provide the prompts for original datasets only in Listings A1, A2, and A3. The translated prompts, as well as those presented here are available in the code of **KyrgyzLLM-Bench**.

```
Сизге бир темага байланыштуу бир нече үзүндү текст берилген. Бардык үзүндүлөрдү кунт коюп окуп, андан кийин төмөндөгү суроолорго жооп бериңиздер.
Суроо менен 2-4 жооп варианты берилет, туура жооптун НОМЕРИН (индексин) гана кайтарышыңыз керек.

Текст: {example_01_text}
Суроо: {example_01_question}
Сунушталган жооптор:
0. {example_01_choices[0]}
1. {example_01_choices[1]}
2. {example_01_choices[2]}
Туура жоопту тандаңыз: {example_01_answer}

Текст: {example_02_text}
Суроо: {example_02_question}
Сунушталган жооптор:
0. {example_02_choices[0]}
1. {example_02_choices[1]}
2. {example_02_choices[2]}
3. {example_02_choices[3]}
Туура жоопту тандаңыз: {example_02_answer}

Текст: {example_03_text}
Суроо: {example_03_question}
Сунушталган жооптор:
0. {example_03_choices[0]}
1. {example_03_choices[1]}
Туура жоопту тандаңыз: {example_03_answer}

Текст: {text}
Суроо: {question}
Сунушталган жооптор:
0. {choices[0]}
1. {choices[1]}
2. {choices[2]}
3. {choices[3]}
Туура жоопту тандаңыз:
```

Listing A1: Prompt for few-shot solution for *KyrgyzRC* (actual prompt is built dynamically in Python code, some details have been removed).

```
Сиз билимиңизге жана жөндөмүңүзгө жараша суроолорго жооп берген AIсыз.
Сизге суроо жана 2-5 жооп варианты берилет, туура жооптун НОМЕРИН (индексин) гана кайтарышыңыз керек.

Суроо: {example_01_question}
Сунушталган жооптор:
0. {example_01_choices[0]}
1. {example_01_choices[1]}
2. {example_01_choices[2]}
3. {example_01_choices[3]}
4. {example_01_choices[4]}
Туура жоопту тандаңыз: {example_01_answer}

Суроо: {example_02_question}
Сунушталган жооптор:
0. {example_02_choices[0]}
1. {example_02_choices[1]}
2. {example_02_choices[2]}
3. {example_02_choices[3]}
Туура жоопту тандаңыз: {example_02_answer}

Суроо: {question}
Сунушталган жооптор:
0. {choices[0]}
1. {choices[1]}
2. {choices[2]}
3. {choices[3]}
4. {choices[4]}
Туура жоопту тандаңыз:
```

Listing A2: Prompt for few-shot solution for *KyrgyzMMLU* (actual prompt is built dynamically in Python code, some details have been removed).

```
Сиз билимиңизге жана жөндөмүңүзгө жараша суроолорго жооп берген AIсыз.
Сизге суроо жана 2-5 жооп варианты берилет, туура жооптун НОМЕРИН (индексин) гана кайтарышыңыз керек.

Текст: {text}
Суроо: {question}
Сунушталган жооптор:
а. {choices[0]}
б. {choices[1]}
в. {choices[2]}
г. {choices[3]}
Туура жоопту тандаңыз:
```

Listing A3: Prompt for zero-shot solution for *KyrgyzMMLU* / *KyrgyzRC* (actual prompt is built dynamically in Python code, some details such as choices' list building have been removed for better readability).

B *GSM8K* Translation

Although the *GSM8K* dataset was translated into Kyrgyz, we do not include its results in the main reported version of the benchmark. This decision was made for several reasons. First, the evaluation protocol of *GSM8K* differs substantially from that of the other benchmarks considered in this work, as it relies on a strict exact-match metric rather than standard accuracy. Second, our preliminary experiments indicate that the standard *Lighteval* evaluation script, which we adopt to ensure comparability with existing language benchmarks, is highly sensitive to output formatting. In the Kyrgyz setting, this sensitivity makes it difficult to disentangle genuine mathematical problem-solving ability from formatting effects, such as mismatches between the expected output patterns and language-specific

conventions for expressing numerical values. As a result, performance on *GSM8K* under this setup may reflect evaluation artifacts rather than model competence, and we therefore exclude it from the primary analysis.

More generally, this design choice reflects our intention to keep the benchmark focused on a coherent and interpretable set of evaluation signals. By restricting the reported results to benchmarks that share a common evaluation paradigm, we aim to ensure that the aggregate metrics convey a clear and atomic message about model performance in the Kyrgyz setting, without mixing heterogeneous sources of error (where possible). At the same time, we recognize that excluding numerically intensive reasoning tasks such as *GSM8K* limits the scope of the current analysis. Developing evaluation protocols that can robustly accommodate language-specific numerical expressions and disentangle reasoning ability from formatting effects remains an important direction for future work.