KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan

Anonymous ACL submission

Abstract

Despite having a population of twenty million, Kazakhstan's culture and language remain underrepresented in the field of natural language processing. Although large language models (LLMs) continue to advance worldwide, progress in Kazakh language has been limited, as seen in the scarcity of dedicated models and benchmark evaluations. To address this gap, we introduce KazMMLU, the first MMLU-style dataset specifically designed for Kazakh language. KazMMLU comprises 23,000 questions that cover various educational levels, including STEM, humanities, and social sciences, sourced from authentic educational materials and manually validated by native speakers and educators. The dataset includes 10,969 Kazakh questions and 12,031 Russian questions, reflecting Kazakhstan's bilingual education system and rich local context. Our evaluation of several state-of-the-art multilingual models (Llama-3.1, Qwen-2.5, GPT-4, and DeepSeek V3) demonstrates substantial room for improvement, as even the best-performing models struggle to achieve competitive performance in Kazakh and Russian. These findings underscore significant performance gaps compared to high-resource languages. We hope that our dataset will enable further research and development of Kazakh-centric LLMs. [1]

1 Introduction

With a population exceeding twenty million, the Republic of Kazakhstan in Central Asia remains underrepresented in the field of natural language processing (NLP) (Joshi et al., 2020). This gap is highlighted by the limited progress in developing large language models (LLMs) and evaluation benchmarks specifically tailored to the languages and the cultural context of Kazakhstan. Kazakh, a Turkic language spoken by more than fourteen million people (around 70% of the population), holds substantial cultural and geopolitical significance in Central Asia. Russian, used by approximately 15% of the population, serves as the country's second primary language.[2]

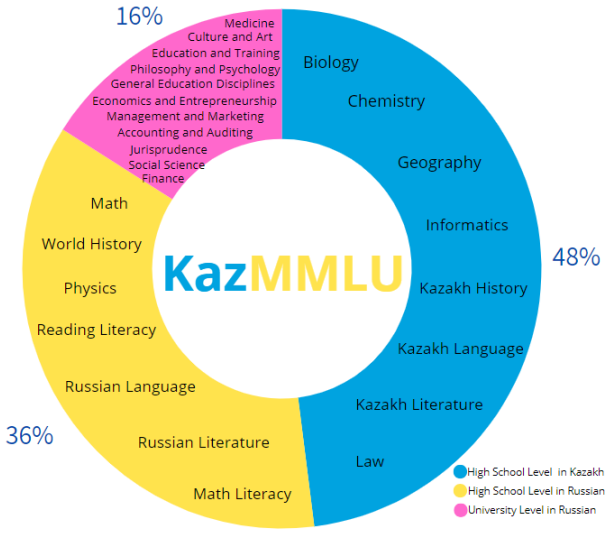

Figure 1: Overview of the dataset. This diagram illustrates the distribution of questions by educational level (High School and University) and language (Kazakh and Russian), along with the variety of subjects covered.

Although Kazakh appears in certain multilingual datasets (Yeshpanov et al., 2024, 2022; Yeshpanov and Varol, 2024), most of these resources rely heavily on translations from English, lacking the cultural richness essential for inclusive LLM development. Previous work has primarily addressed classic NLP tasks such as named entity recognition (Yeshpanov et al., 2022) and sentiment analysis (Yeshpanov and Varol, 2024). Meanwhile, recent developments in LLM research have shifted toward more reasoning-focused evaluation (Meta AI, 2024; OpenAI, 2024), highlighting a clear research gap for inclusive NLP in the Kazakh context.

Here, we aim to bridge this gap. In partic-

[1] Data and code will be made available upon acceptance.

[2] https://glottolog.org/

ular, we introduce KazMMLU, a curated dataset of school- and university-level questions from Kazakhstan, available in both Kazakh and Russian. KazMMLU follows the framework of the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021; Koto et al., 2024; Li et al., 2024; Koto et al., 2023), which features multiple-choice questions across various subjects and education levels. MMLU has become a standard benchmark for evaluating LLMs' reasoning and knowledge capabilities (Meta AI, 2024; Team et al., 2024; Qwen et al., 2025). Unlike general MMLU, KazMMLU incorporates Kazakhstan-specific content, including topics on Kazakh history, traditions, and linguistics, while also reflecting the country's multilingual landscape by providing questions in both Kazakh and Russian.

As shown in Figure 1, the dataset is divided into two categories: High School and University. KazMMLU consists of approximately 48% of the questions in Kazakh and 52% in Russian. The High School section includes questions in both Kazakh and Russian, covering subjects such as Mathematics, Physics, and Kazakh Literature. The University section only features questions in Russian, focusing on professional disciplines such as Law, Economics, and Medicine. This structure aligns with Kazakhstan's bilingual education system and provides a more representative benchmark for evaluating LLMs in the region. KazMMLU is sourced from authentic educational materials, including national exams, textbooks, and professional certification repositories. Each question is accompanied by metadata, including the subject, level, source, and correct answer key, ensuring transparency and usability for downstream evaluations.

Our contributions can be summarized as follows:

- We present the first **MMLU-style dataset** specifically tailored to the Kazakhstan context, covering diverse subject areas across educational and professional levels. The dataset is made available in both Kazakh and Russian.
- We evaluate various **multilingual LLMs**, including Llama-3.1 (Meta AI, 2024), Qwen (Qwen et al., 2025), GPT-4o (OpenAI, 2024), BLOOMZ (Muennighoff et al., 2023), mT0 (Muennighoff et al., 2023), and DeepSeek V3 (DeepSeek-AI, 2024), across different model sizes.
- We conduct a thorough analysis of the top-performing open-source models across various dimensions, encompassing (1) individual **subject areas, educational levels, and Kazakhstan-specific topics**, (2) **few-shot inference performance**, (3) **model confidence**, and (4) the **influence of negation** on model performance. This comprehensive evaluation framework allows us to identify the performance gaps and opportunities for improvement in multilingual LLMs when applied to Kazakh and Russian contexts.

2 Related Work

Language Models in Kazakh and Russian Prominent models such as OpenAI's ChatGPT, Anthropic's Claude, and Yandex's Yandex-GPT are designed to handle multiple languages, including Russian and Kazakh, enabling a wide range of applications from translation to content generation (OpenAI, 2024; Anthropic, 2023; Yandex, 2023). Additionally, open-source models like Meta's Llama series provide multilingual support (Meta AI, 2024). While these models can produce text in Kazakh, they were not specifically trained or fine-tuned for it. In contrast, the Aya model, an open-access multilingual LLM, supports 101 languages, including Kazakh (Üstün et al., 2024).

Kazakh-specific language models have been scarce, with most multilingual models offering only limited support. To bridge this gap, ISSAI (2024) introduced KazakhLLM, a model fine-tuned on Kazakh data from Llama, though its evaluation has primarily relied on machine-translated datasets.

NLP Benchmark for Kazakhstan Context Evaluating LLMs across diverse linguistic and cultural contexts is increasingly critical; however, existing benchmarks overlook Kazakhstan. While benchmarks such as XCOPA (Hu et al., 2021), XFGLUE (Liang et al., 2020), and XTREME (Hu et al., 2020) assess cross-lingual performance, they exclude Kazakh, and GlobalMMLU (Singh et al., 2024) lacks Kazakhstan-specific content. In contrast, several Kazakh-specific datasets exist, including KazNERD (Yeshpanov et al., 2022) for named entity recognition, KazQAD (Yeshpanov et al., 2024) for question answering, and KazSANDRA (Yeshpanov and Varol, 2024) for sentiment analysis. However, these datasets focus on nar-

row tasks and do not assess reasoning, factual recall, or domain-specific knowledge. To address these limitations, KazMMLU presents a large-scale, Kazakhstan-specific benchmark covering STEM, humanities, and social sciences. Unlike previous datasets, KazMMLU supports a holistic evaluation of reasoning and domain-specific knowledge, offering a more accurate assessment of multilingual LLM capabilities and advancing AI for low-resource languages.

To further illustrate the differences between KazMMLU and previous benchmarks, we compare it with two existing datasets, SIGTURK (Maxutov et al., 2024) and INCLUDE (Romanou et al., 2024), in Table 1. As shown, KazMMLU is the only dataset that incorporates **real-world educational materials, professional subjects**, and **domain-specific reasoning** in both Kazakh and Russian, offering a more **localized** and **comprehensive** evaluation of LLMs. This structured assessment underscores the **importance of country-specific benchmarks** in multilingual NLP research and contributes to bridging the gap in Kazakh language understanding.

3 KazMMLU

In Kazakhstan, the K-12 education system operates in a multilingual setting, with Kazakh as the primary language of instruction in most schools. However, Russian and other languages, such as Uzbek, Uyghur, and Tajik, are also used in specific regions. The curriculum includes core subjects such as mathematics, science, and history, with students required to study both Kazakh and Russian as part of their language education. At the university and professional levels, Russian remains the dominant language of instruction, especially in fields such as law, medicine, economics, and engineering.

KazMMLU covers a wide range of subjects spanning multiple disciplines, including STEM, humanities, social sciences, and professional studies. Figure 2 illustrates the distribution of questions across different subjects. The dataset exhibits a strong representation in STEM fields, with subjects like Biology, Mathematics and Physics accounting for a significant portion of the total questions. Humanities and social sciences are also well represented, particularly Kazakh History, World History, and Law, reflecting the importance of these disciplines in the Kazakh educational system.

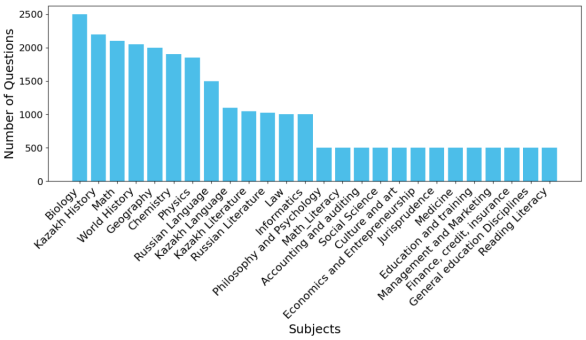

Figure 2: Subject-wise distribution of questions in KazMMLU.

Kazakh History, High School	
1930 жылдары өзге республикалардың жоғарғы оқу орындары мен техникумдарында білім алған қазақстандың жастар саны:	The number of young people from Kazakhstan who studied at universities and technical schools in other republics in the 1930s:
A. 25 мыңдай	A. About 25 thousand
B. 20 мыңдай	**B. About 20 thousand**
C. 15 мыңдай	C. About 15 thousand
D. 30 мыңдай	D. About 30 thousand
E. 40 мыңдай	E. About 40 thousand
Geography, High School	
Трактор, ауыл шаруашылық машиналар шығаратын қалалары:	Cities producing tractors and agricultural machinery:
A. Теміртау, Павлодар	A. Temirtau, Pavlodar
B. Астана, Теміртау	B. Astana, Temirtau
C. Арқалық, Павлодар	C. Arkalyk, Pavlodar
D. Балқаш, Ақтау	D. Balkash, Aktau
E. Астана, Павлодар	**E. Astana, Pavlodar**

Figure 3: Examples of a Kazakh history question and a geography test question from KazMMLU. The **left** side presents the original text, while the **right** side provides the English translation for reference. The bolded options indicate the correct answers.

Additionally, KazMMLU includes questions from professional domains such as Economics, Finance, Jurisprudence, and Medicine, enabling the evaluation of language models in specialized areas. The inclusion of both Kazakh and Russian language subjects further ensures a balanced linguistic representation. This diverse subject distribution provides a robust benchmark for evaluating multilingual language models across various domains.

To illustrate the question format in KazMMLU, Figure 3 presents a sample multiple-choice question in Kazakh. Answering this question requires an understanding of Kazakhstan's local context, as it covers topics such as history and geography. This example highlights the dataset's diversity across subjects and difficulty levels.

3.1 Data Construction

To construct KazMMLU, we adopted a systematic approach inspired by MMLU datasets (Koto et al., 2023, 2024; Li et al., 2024). The dataset com-

Feature	KazMMLU	SIGTURK	INCLUDE
Public Dataset Size	23,000 questions in Kazakh and Russian	3,000 questions exclusively in Kazakh	23,741 total questions, including 500 in Kazakh
Languages Covered	Kazakh, Russian	Kazakh	44 (including Kazakh)
Kazakh-Specific Content	**Yes**, sourced from local curriculum, national exams	Limited (Kazakh NLP tasks)	Limited
Education Levels	High School, University	Not explicitly structured	General education
Subjects Covered	STEM, Humanities, Social Sciences, Law, Medicine	QA, MT, causal reasoning	Broad general knowledge
Task Type	Bilingual MCQs reflecting real-world knowledge	QA, classification, generative tasks	General MCQs across languages
Model Benchmarking	41 LLMs (GPT-4o, Llama-3.1, DeepSeek V3, etc.)	7 models on Kazakh NLP tasks	Multiple LLMs across 44 languages

Table 1: Comparison of **KazMMLU** with SIGTURK and INCLUDE.

prises questions sourced from national exams, textbooks, and professional certification materials such as iTest.kz, ymnik.kz, oltest.kz and Book - Shyn Kitap. To ensure diversity, we employed three data collection strategies: (1) automated online crawling, (2) manual transcription from scanned books, and (3) manual extraction from online sources. A detailed breakdown of dataset sources is provided in Appendix 6.

For automatic online crawling, we collect question texts, multiple-choice options, correct answer keys, and metadata. For books, authors manually scan materials and apply document processing for machine-readable conversion. Two expert workers fluent in Kazakh and Russian manually extract questions from scanned books and online sources, recording metadata such as source, country, subject, level, and answer key. We compiled 23,000 questions, with (85%) from automated crawling and the rest from manual extraction.

Only questions with valid answer keys were included, while multimodal ones requiring images or videos were excluded. For context-dependent questions, annotators ensured necessary context was included. A training workshop clarified guidelines, and weekly check-ins ensured consistency. Annotators were competitively compensated to maintain quality.

3.2 Quality Control

Our quality control process primarily targets the automatically-crawled data, as the other two data collection strategies involve direct human involvement. To ensure accuracy, we recruited two professional annotators, each holding at least a bachelor's degree and fluent in both Kazakh and Russian. They manually reviewed all questions to verify correctness and completeness. Any question containing errors or missing components (e.g., incomplete contexts or broken answer options) was discarded. Through this extensive human verification, every question included in KazMMLU undergoes manual validation, ensuring a high-quality dataset.

Additionally, we developed scripts to detect duplicates, verify metadata, and eliminate errors like duplicate questions, incorrect answer keys, and formatting issues, enhancing dataset reliability

3.3 Data Statistics

KazMMLU comprises 23,000 multiple-choice questions spanning two educational levels: high school and university. As shown in Figure 1, 48% of the dataset consists of high school-level questions in Kazakh, 36% in Russian, and 16% university-level questions in Russian. Table 2 outlines the subject distribution, covering STEM, Humanities, Social Sciences, and Languages. Notably, the Humanities, Social Sciences, and Language sections contain extensive Kazakhstan-specific knowledge. The dataset maintains a balanced distribution between Kazakh and Russian, reflecting the bilingual nature of Kazakhstan's education system.

Table 3 presents the average question and answer lengths across educational levels and subject areas. While the overall question length remains relatively consistent between high school and university levels, answer lengths (in characters) tend to be longer at the university level. Additionally, questions in Humanities and STEM subjects are generally longer compared to those in Social Sciences and Languages.

Group	Subjects
Humanities	Culture and Art (U), Kazakh History (H), Kazakh Literature (H), Philosophy and Psychology (U), Russian Literature (H), World History (H)
Language	Kazakh Language (H), Reading Literacy (H), Russian Language (H)
Other	General Education Disciplines (U)
STEM	Biology (H), Chemistry (H), Informatics (H), Math (H), Math Literacy (H), Medicine (U), Physics (H)
Social Science	Accounting and Auditing (U), Economics and Entrepreneurship (U), Education and Training (U), Finance, Credit, Insurance (U), General Education Disciplines (U), Geography (H), Jurisprudence (U), State and Law (U), Management and Marketing (U), Social Science (U)

Table 2: Subject groups covered by KazMMLU. "H" indicates high school subjects, and "U" indicates university subjects.

Group	Question	Answer
High School	78.3	16.6
University	84.4	29.6
Humanities	81.3	19.1
Language	49.3	20.3
Others	82.0	37.1
STEM	83.8	15.1
Social Science	52.9	15.4

Table 3: Average question and answer length (in characters) for each educational group and subject area.

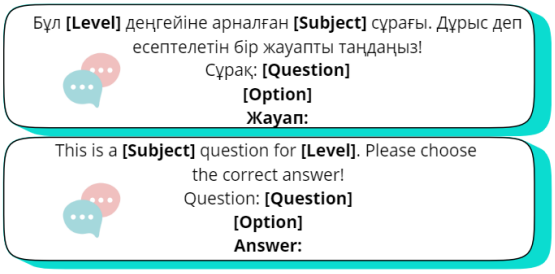

Figure 4: Prompt templates in Kazakh and English.

4 Experiment

4.1 Setup

We evaluate 26 multilingual LLMs of various sizes in both zero-shot and few-shot settings. Our selection includes models from diverse architectures, such as BLOOM (Scao et al., 2023), BLOOMZ, and mT0 (Muennighoff et al., 2023), Falcon (Almazrouei et al., 2023), Llama-3.1 (Touvron et al., 2023a,b), GPT-4o (OpenAI, 2024), mT5 (Xue et al., 2020), Vikhr (Nikolich et al., 2024), and DeepSeek V3 (DeepSeek-AI, 2024).

For the evaluation purpose, we use two distinct prompt configurations to examine the effect of prompt language: (1) a Kazakh prompt with English (Latin script) alphabetic output and (2) an English prompt with English alphabetic output, as illustrated in Figure 4. For placeholders such as `[Subject]` and `[Level]`, we use Kazakh in the Kazakh prompt and translate them into English for the English prompt. However, the question and answer choices remain in their original language (Kazakh or Russian). A complete example prompt is provided in Appendix B (Figure 10).

Following prior studies (Koto et al., 2023; Li et al., 2024), we adopt different answer selection methods based on model accessibility. For open-weight models, we apply the *next-token prediction* approach, computing probabilities for each multiple-choice option (A, B, C, D, or E) and selecting the one with the highest probability. This method is well-suited for autoregressive models that perform token-wise scoring. For closed-weight models (e.g., GPT–4o, Yandex-GPT, and DeepSeek V3), we use a *free-text generation* strategy, prompting the model to generate a textual response, from which the predicted answer is extracted via string matching. This approach is necessary due to the lack of direct token probability outputs in closed-source APIs. For evaluation, we use accuracy as the primary metric, following prior studies (Koto et al., 2023; Li et al., 2024).

4.2 Results and Analysis

First, we observe that LLMs achieve higher accuracy when prompted in English, as shown in Table 4 and Table 7. To provide clearer insights into model performance, we focus on English-prompted results in the main body of the paper.

Results across all models Table 4 presents the average accuracy for each subject area across 26 models on the `KazMMLU` using English prompts. The performance analysis reveals several notable patterns. GPT-4o and DeepSeek V3 emerge as the top performers, achieving remarkably similar average scores of 75.7% and 75.6% respectively,

Model	STEM	Social Science	Humanities	Language	Other	Average
Mistral-7B-Instruct-v0.3	36.5	44.7	41.8	30.0	40.6	38.2
Mistral-7B-v0.3	31.5	37.1	37.9	27.3	34.6	33.3
Vikhr-Nemo-12B-Instruct-R-21-09-24	39.5	49.5	48.1	32.7	44.0	42.2
aya-23-35B	34.5	38.5	38.5	29.4	33.2	35.2
aya-23-8B	29.4	31.9	32.7	25.8	26.5	29.9
bloom-1b1	24.1	20.4	21.0	21.6	20.8	22.2
bloomz-1b7	24.2	22.7	22.7	23.0	25.8	23.4
bloomz-3b	24.2	23.5	23.5	22.1	25.2	23.6
bloomz-7b	24.0	24.0	24.2	23.9	22.5	24.0
Gemma-2-27b	54.8	59.4	55.2	37.3	47.3	**52.7**
Gemma-2-27b-IT	57.0	60.3	54.6	39.1	48.3	**54.0**
Gemma-2-9b	50.8	57.4	53.8	36.2	42.6	50.2
Gemma-2-9b-IT	50.6	52.8	49.1	35.8	44.3	48.1
issai-8b	39.1	45.9	43.7	31.2	38.6	40.1
Llama3.1-70b	57.0	60.0	58.7	41.7	49.3	**55.2**
Llama3.1-70b-instruct	50.6	47.5	47.9	33.9	45.6	**46.5**
Llama3.1-8b	36.6	43.8	42.4	28.9	38.3	38.0
Llama3.1-8b-instruct	41.6	47.4	45.4	30.4	39.3	41.6
mt0-large	24.6	23.9	23.3	22.9	23.5	23.9
mt0-xl	27.9	35.5	28.9	26.5	31.9	29.5
mt0-xxl	30.0	38.9	32.0	29.5	37.9	32.3
qwen-2.5-7b	45.5	42.0	41.2	29.3	39.9	41.0
qwen-2.5-7b-instruct	47.7	49.6	46.4	33.9	42.3	45.4
GPT-4o	69.7	**83.1**	**81.7**	73.4	62.1	**75.7**
DeepSeek V3	78.7	82.3	77.0	61.2	65.1	75.6
YandexGPT	53.7	69.8	52.2	42.6	57.0	54.9

Table 4: Performance of different models on the Kazakh MMLU benchmark across different subject categories using **English** prompt.

outperforming other models. Among the open-source models, the Llama3.1 and Gemma families demonstrated strong performance. Llama3.1-70b achieved the highest average score (55.2%), followed by the Gemma-2-27b instruction-tuned model (54.0%) and its base variant (52.7%). Interestingly, the impact of instruction tuning varies across model families - while Gemma-2-27b-it showed a slight improvement over its base model (+1.3%), Llama3.1-70b-instruct performed worse than its base variant (-8.7%). Consistently across all models, the Language category seems to be the most challenging, with scores lower than other categories.

Few-Shot Performance As shown in Figure 5, our few-shot results show a consistent improvement across all models as the number of shots increases, with Qwen-2.5-7B and Mistral-7B-v0.3 benefiting the most. English prompts consistently outperform Kazakh prompts in 1, 2, and 3-shot settings, though this trend does not hold in 0-shot, where Kazakh sometimes performs better. Instruction-tuned models also improve, though the gains are smaller, with Qwen-2.5-7B-Instruct (English prompt) increasing from 47.8% (0-shot) to 58.9% (3-shot). The largest accuracy jumps occur between 0-shot and 1-shot, indicating that even a single in-context example significantly enhances model understanding. Overall, these results highlight the robustness of few-shot learning across diverse model architectures and prompt settings.

Kazakh vs Russian Performance In Figure 6, we compared model performance across two languages: Kazakh and Russian. To ensure a fair comparison, we only considered the High School level subjects and excluded the Professional & University level tasks because they are not available in Kazakh. The results indicate that GPT-4o achieves the highest accuracy in Kazakh, scoring 76.90%, while DeekSeek performs best in Russian with an accuracy of 81.8%. Llama 3.1-70B and Gemma 2-27B show lower but comparable results, with a slight advantage in Russian over Kazakh. Overall, models tend to perform slightly better in Russian than in Kazakh, which could be due to differences in training data availability, language complexity, or tokenization differences.

Results Across Education Level Results in Figure 7 indicate that GPT-4o performs approximately the same across both education levels. Similarly, DeepSeek V3 maintains a balanced performance, but with a slight preference towards High School. In contrast, the open-source models Llama 3.1-70B

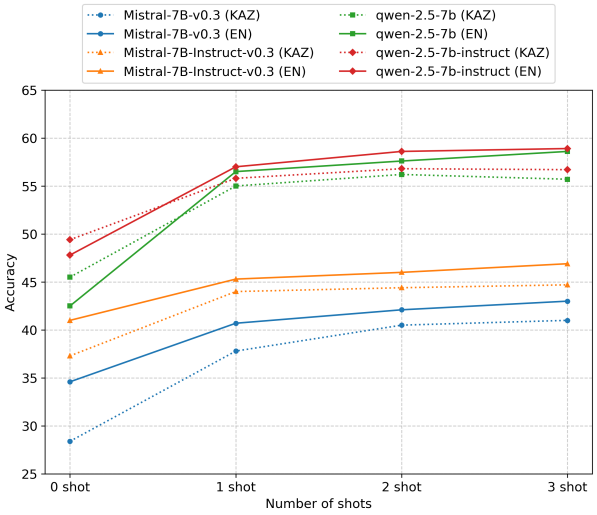

Figure 5: The few-shot accuracy (%) of LLMs on KazMMLU, averaged across all tasks, comparing base models and instruction-tuned models using Kazakh (dotted lines) and English (solid lines) prompts.

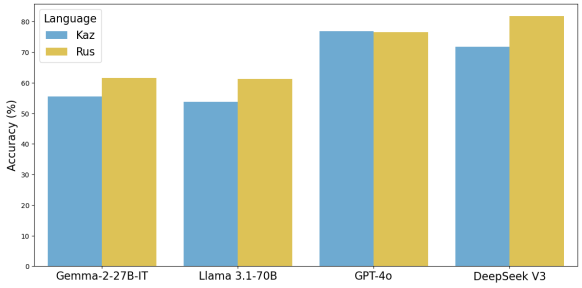

Figure 6: LLM Performance across different languages(only at the high-school level).

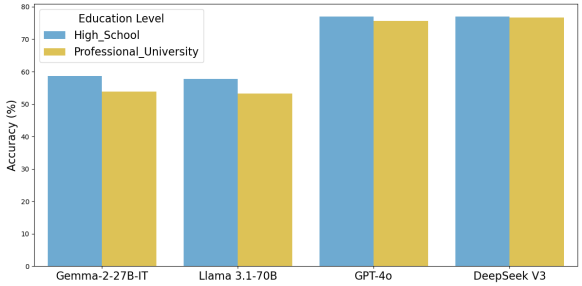

Figure 7: LLM Performance across different education levels

and Gemma 2-27B show considerably lower accuracy and a pronounced gap between education levels. Llama 3.1-70B achieves 57.7% in High School but drops to 53.2% in Professional & University. These results suggest that while proprietary models generalize well across different subject complexities, open-source models struggle more with specialized university-level knowledge.

Negation Sensitivity Analysis Table 5 presents the accuracy of LLaMA3-70B, Gemma-2-27B-IT, and DeepSeek V3 on *negation-sensitive subjects*, comparing performance with and without negation. The results indicate that DeepSeek V3 consistently outperforms both LLaMA3-70B and Gemma-2-27B-IT, demonstrating greater resilience to negation-based reasoning challenges.

To systematically analyze the impact of negation, we employed a negation phrase filtering method inspired by ArabicMMLU (Koto et al., 2024). Specifically, we identified questions containing common negation phrases in Kazakh: жоқ (no), емес (is not), болмайды (not allowed), жарамайды (prohibited), невозможно (impossible), не (not), нельзя (forbidden). . After applying this filtering, we obtained a total of 2,554 negation-related questions. To validate our filtering accuracy, we randomly sampled 100 questions and manually inspected them. The detection accuracy exceeded 92%, confirming the reliability of our filtering process.

For LLaMA3-70B and Gemma-2-27B-IT, accuracy generally decreases on questions containing negation, suggesting a negative impact on reasoning capabilities. Notably, LLaMA3-70B exhibits a larger drop in accuracy, particularly in *Reading Literacy*, where performance declines from 57.1% to 50.0%. Meanwhile, Gemma-2-27B-IT demonstrates greater robustness in certain cases, showing less fluctuation in performance across negation and non-negation settings.

Model Confidence We analyze whether the evaluated models, including Gemma-2.9B-IT, Qwen-2.5-7B-Instruct, and Llama3.1-70B, are well-calibrated when answering KazMMLU questions by comparing the probability of the correct answers with the actual accuracy for each subject and level combination. The answer probability is obtained through softmax normalization across the available candidate answers. In Figure 8, we observe that the evaluated models exhibit a strong correlation between confidence and accuracy, with correlation scores exceeding 0.9. This indicates that models with higher confidence generally produce more accurate predictions.

Additionally, we investigate the correlation between model confidence and question length in Figure 12. The results show that question length has minimal impact on model confidence, as evidenced by the weak correlation scores across all evaluated

Subject	Model	W/o Negation	W/ Negation
Jurisprudence (University)	Llama3-70B	**56.2**	55.2
	Gemma-2-27B-IT	55.2	**56.5**
	DeepSeek V3	**78.1**	76.4
Law (High School)	Llama3-70B	**60.8**	59.0
	Gemma-2-27B-IT	59.0	**58.1**
	DeepSeek V3	**79.5**	78.1
Reading Literacy (High School)	Llama3-70B	57.1	**50.0**
	Gemma-2-27B-IT	**100.0**	87.5
	DeepSeek V3	**85.7**	83.5
Philosophy and Psychology (University)	Llama3-70B	55.9	**56.6**
	Gemma-2-27B-IT	**56.6**	55.9
	DeepSeek V3	**83.2**	81.9

Table 5: Model accuracy on negation-sensitive questions across various subjects. Bold values indicate higher accuracy in each category.

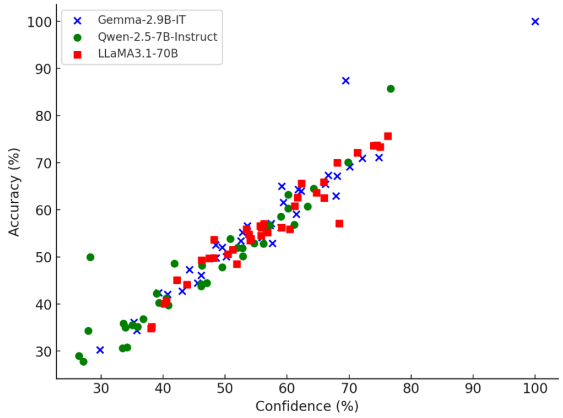

Figure 8: Confidence vs. Accuracy for different models in a zero-shot setting. **Confidence (%)** denotes the average probability scores in percentage.

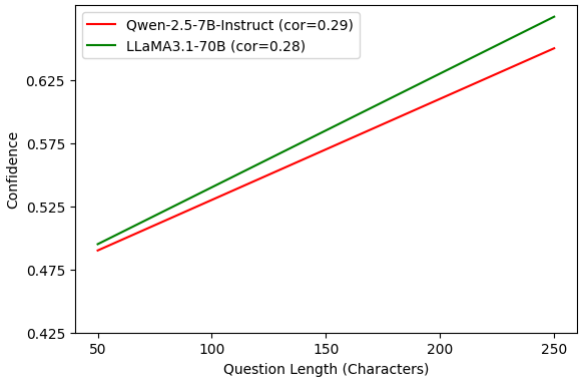

Figure 9: Correlation between model confidence and question length across different models.

models. For Qwen-2.5-7B-Instruct ($r = 0.29$) and Llama3.1-70B ($r = 0.28$) exhibit a mild positive correlation, suggesting that **longer questions slightly increase model confidence**. However, the effect remains weak overall, implying that confidence calibration remains relatively stable across different question lengths.

5 Conclusion and Future Work

We introduced KazMMLU, the first large-scale multitask language understanding dataset designed to evaluate real-world knowledge in Kazakhstan's bilingual setting. Through experiments with over 23K multiple-choice questions spanning various subjects and education levels, we observed that models perform much better in Russian than in Kazakh, with proprietary models such as GPT-4o and DeepSeek V3 achieving the highest accuracy.

KazMMLU provides a **bilingual** (Kazakh and Russian) evaluation framework tailored to Kazakhstan's educational and professional landscape, distinguishing itself from previous multilingual benchmarks. Unlike SIGTURK, which focused on Kazakh NLP tasks, and INCLUDE, which lacks Kazakhstan-specific content, KazMMLU enables a localized and comprehensive assessment of LLMs in Kazakhstan.

Future research directions include extending KazMMLU to multimodal evaluations, improving reasoning-based question assessments, and mitigating biases in data sources. We hope that this benchmark will encourage further development of high-performance LLMs for Kazakh and other low-resource languages.

Limitations

While we are confident that our benchmark will significantly advance the development of Kazakh LLMs, it is important to acknowledge certain limitations that need to be addressed in future research.

While we are confident that our benchmark will significantly advance the development of Kazakh LLMs, it is important to acknowledge certain limitations that need to be addressed in future research. We outline these limitations as follows:

Limited Modality KazMMLU is focused solely on text-based assessment, and the exploration of multimodal questions (including those involving images, audio, or other media types) has been excluded. Future work could explore the integration of multimodal content to better reflect real-world applications, such as vision-language tasks, speech recognition, and interactive assessments.

Lack of Explicit Reasoning Evaluation While KazMMLU provides a broad and representative set of multiple-choice questions, it does not explicitly evaluate reasoning processes beyond answer selection. Investigating how models approach complex reasoning, justification, and open-ended question answering would be a valuable direction for further improvement.

Static Evaluation Limitation KazMMLU primarily evaluates static model performance on predefined questions, which may not fully capture how models generalize to dynamic, real-world language use. Exploring benchmarks that assess interactive and adaptive reasoning, as well as domain adaptation, could enhance our understanding of model capabilities in evolving contexts.

By addressing these limitations, future research can further refine the evaluation of Kazakh LLMs, ensuring more robust, fair, and practically useful language models for Kazakhstan and beyond.

Ethics and Broader Impact

We adhered to the internal policies of web resources while scraping data and included only publicly available information verified by authorities.

All human subjects in our study provided informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time. They were also appropriately compensated as part of their job.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. Technical report, Technology Innovation Institute.

Anthropic. 2023. Claude. Accessed: 2025-02-10.

DeepSeek-AI. 2024. Deepseek llm: Advancing open large language models with high-quality pretraining and instruction tuning. Available at: https://huggingface.co/deepseek-ai.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Junjie Hu et al. 2020. Xtreme: A massively multilingual benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*.

Junjie Hu et al. 2021. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP 2021*.

ISSAI. 2024. Llama-3.1-kazllm-1.0-8b. Accessed: 2025-02-10.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Hongyu Liang et al. 2020. Xfglue: A benchmark dataset for cross-lingual pretraining, understanding and generation. In *Proceedings of NeurIPS*.

Maxutov et al. 2024. Sigturk 2024: Do llms speak kazakh? a pilot evaluation of seven models. In *Proceedings of SIGTURK 2024*.

Meta AI. 2024. Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date. Accessed: 2024-11-29.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for russian. *Preprint*, arXiv:2405.13929.

OpenAI. 2024. GPT-4o. Accessed: 2024-11-29.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Na-

joung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hass-

abis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yandex. 2023. Yandexgpt. Accessed: 2025-02-10.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. 2024. Kazqad: Kazakh open-domain question answering dataset. *Preprint*, arXiv:2404.04487.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. Kazsandra: Kazakh sentiment analysis dataset of reviews and attitudes. *Preprint*, arXiv:2403.19335.

A Appendix

A.1 Additional Prompt Examples

Figure 10 illustrates the Kazakh and English prompts used in our evaluation. The formatting emphasizes key aspects such as the subject (e.g., **Law**) and the educational level (e.g., **High School**). The structure maintains consistency in question presentation, ensuring uniformity across both languages.

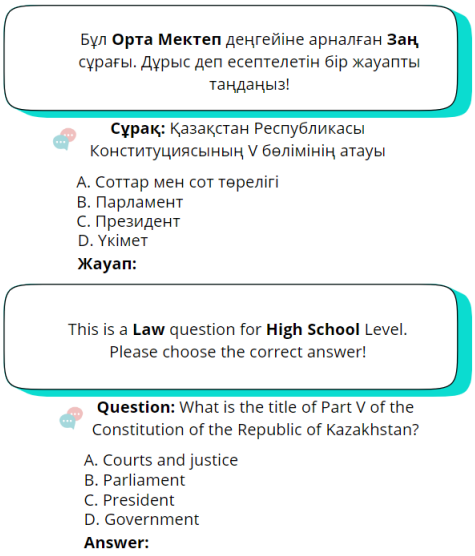

Figure 10: Examples of Kazakh and English prompts. The placeholders are dynamically replaced based on the question context.

B Appendix B:

B.1 Additional Multiple-Choice Question Example

In addition to the Kazakh-language question provided in the main document, we present an example of a Russian-language multiple-choice question (Figure 11). This figure highlights a range of subjects from social sciences to STEM disciplines, demonstrating the dataset's diversity.

B.2 Breakdown of Dataset Sources

Table 6 provides a categorized breakdown of dataset sources used in KazMMLU, covering national exams, professional certification tests, and textbooks.

B.3 Additional Evaluation Details

For further insights into our evaluation framework, we provide additional examples and implementation details. Table 4 summarizes our experimental settings.

Culture and Art, University	
Когда в Казахстане появились театры, концертные залы и филармонии: A. Конец 20 века B. 19 век C. 18 век **D. Первая половина 20 века** E. 21 век	When the theater, concert hall and philharmonic appeared in Kazakhstan: A. The end of the 20th century B. 19th century C. 18th century **D. First half of the 20th century** E. 21th century
Social Science, University	
По Конституции РК не должны состоять в партиях A. Члены Парламента B. Работники Аппарата Президента **C. Военнослужащие, работники органов национальный безопасности, правоохранительные органы, судьи** D. Студенты E. Государственные чиновники	According to the Constitution of the Republic of Kazakhstan, there should be parties A. Member of Parliament B. Employees of the Presidential Executive Office **C. Military personnel, employees of national security agencies, law enforcement agencies, judges** D. Students E. Government officials

Figure 11: Example of a Russian-language multiple-choice question from KazMMLU. The **left** side shows the original text, while the **right** side provides the English translation for illustrative purposes. The bold options represent the correct answer keys.

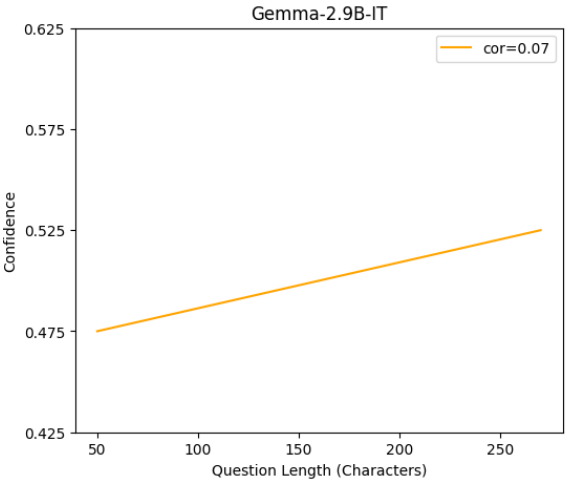

Figure 12: Correlation between model confidence and question length

Main Source	Language	Level	Subjects
itest.kz	Kazakh	High School	Biology, Chemistry, Geography, Informatics, Kazakh History, Kazakh Literature, Law, Math, Math Literacy, Physics, Reading Literacy, Russian Language, Russian Literature, World History
oltest.kz	Russian	University and Professional	Accounting and Auditing, Biology, Culture and Art, Economics and Entrepreneurship, Education and Training, Finance, General Education Disciplines, Jurisprudence, Management and Marketing, Medicine, Philosophy and Psychology, Social Science
ymnik.kz	Russian	High School	Biology, Geography, Kazakh History, Kazakh Language, World History
Book - Shyn Kitap	Kazakh	High School	Biology, Geography, Kazakh History, Kazakh Language, World History

Table 6: Breakdown of dataset sources in KazMMLU, categorized by domain and subject area. The alternating row colors improve readability.

Model	STEM	Social Science	Humanities	Language	Other	Average
Mistral-7B-Instruct-v0.3	33.3	41.2	37.6	28.0	31.9	34.9
Mistral-7B-v0.3	28.7	29.2	31.2	24.6	25.2	28.6
Vikhr-Nemo-12B-Instruct-R-21-09-24	38.3	52.0	48.5	32.8	41.9	42.3
aya-23-35B	29.8	32.6	34.4	25.8	26.5	30.6
aya-23-8B	26.8	26.9	27.3	23.2	22.8	26.2
bloom-1b1	24.3	20.0	20.7	22.9	19.1	22.4
bloomz-1b7	23.3	21.1	21.1	23.3	20.1	22.3
bloomz-3b	23.7	21.4	20.9	21.9	17.8	22.3
bloomz-7b	22.7	21.0	22.5	21.1	21.5	22.0
gemma-2-27b	57.2	64.4	59.8	41.1	46.6	56.3
gemma-2-27b-it	58.7	63.1	57.1	40.9	46.3	56.1
gemma-2-9b	47.3	54.6	52.4	33.6	39.3	47.4
gemma-2-9b-it	50.5	51.9	48.3	35.2	41.9	47.6
issai-8b	36.0	39.6	39.5	28.4	33.9	36.1
Llama3.1-70b	57.9	65.8	64.0	44.4	50.3	58.4
Llama3.1-70b-instruct	56.3	57.4	55.5	39.5	50.3	53.4
Llama3.1-8b	33.7	38.4	39.0	28.0	31.5	34.7
Llama3.1-8b-instruct	36.9	39.5	39.9	28.0	31.5	36.4
mt0-large	24.5	24.8	23.1	22.9	24.8	24.0
mt0-xl	28.4	37.1	29.8	26.8	34.6	30.3
mt0-xxl	28.8	37.2	31.1	29.3	36.2	31.2
qwen-2.5-7b	46.5	48.2	44.4	30.5	42.3	43.6
qwen-2.5-7b-instruct	47.4	50.6	47.1	33.2	43.3	45.5
GPT4-o	69.7	83.1	81.7	73.4	62.1	75.7
DeepSeek V3	78.5	81.8	77.1	61.4	65.3	72.8
YandexGPT	39.3	50.7	48.1	44.0	34.6	44.4

Table 7: Performance of different models on the Kazakh MMLU benchmark across different subject categories using **Kazakh** prompt.