TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages

Anonymous ACL submission

Abstract

Being able to thoroughly assess massive multi-task language understanding (MMLU) capabilities is essential for advancing the applicability of multilingual language models. However, preparing such benchmarks in high quality native language is often costly and therefore limits the representativeness of evaluation datasets. While recent efforts focused on building more inclusive MMLU benchmarks, these are conventionally built using machine translation from high-resource languages, which may introduce errors and fail to account for the linguistic and cultural intricacies of the target languages. In this paper, we address the lack of native language MMLU benchmark especially in the under-represented Turkic language family with distinct morphosyntactic and cultural characteristics. We propose two benchmarks for Turkic language MMLU: TUMLU is a comprehensive, multilingual, and natively developed language understanding benchmark specifically designed for Turkic languages. It consists of middle- and high-school level questions spanning 11 academic subjects in Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Tatar, Turkish, Uyghur, and Uzbek. We also present TUMLU-mini, a more concise, balanced, and manually verified subset of the dataset. Using this dataset, we systematically evaluate a diverse range of open and proprietary multilingual large language models (LLMs), including Claude, Gemini, GPT, and LLaMA, offering an in-depth analysis of their performance across different languages, subjects, and alphabets. To promote further research and development in multilingual language understanding, our data set and all corresponding evaluation scripts will be publicly available upon publication.

1 Introduction

Language understanding encompasses a system's ability to interpret and derive meaning from human language, incorporating syntax, semantics, and context. Evaluating language models hinges on this capability, as it ensures coherence, contextual relevance, and accuracy. Benchmarking is integral to assessing these models, particularly with the rapid advancements in Large Language Models (LLMs), which now support multiple languages (Yang et al., 2025; Gemma Team, 2024; Grattafiori et al., 2024) and excel in complex reasoning tasks such as mathematical, scientific, and coding-related inquiries (Hurst et al., 2024; Anthropic, 2024; Gemini Team, 2024; Grattafiori et al., 2024). However, the scarcity of robust natural language understanding (NLU) benchmarks capturing diverse linguistic and cultural contexts remains a challenge. Notably, LLM performance declines in low-resource languages, which are often underrepresented in training data, highlighting the need for more inclusive evaluation frameworks.

The majority of benchmarks included in top leaderboards where cutting-edge LLMs are evaluated are majorly prepared in English (Hendrycks et al., 2021a; Suzgun et al., 2022; Wang et al., 2024b, 2019). In order to extend the applicability of LLM evaluation in more languages, recent efforts were undertaken to build more multilingual NLU benchmarks (Lai et al., 2023), however, most of these either cover a limited set of high-resourced languages, or the multilingual examples are generated by translating original examples from Western-centric languages, thus failing to capture cultural nuances inherent in different languages. Due to the multi-dimensional nature of the reasoning task, language-specific benchmarks especially when translated into other languages also fail to represent the actual usage as well as demonstrating reasoning in the native language., and may further introduce issues such as translationese (Vanmassenhove et al., 2021) and cultural misalignment (Romanou et al., 2025). On the other end of the spectrum, there are efforts to bridge that gap for a particular language, for example, African languages (Bayes et al., 2024), Arabic (Koto et al.,

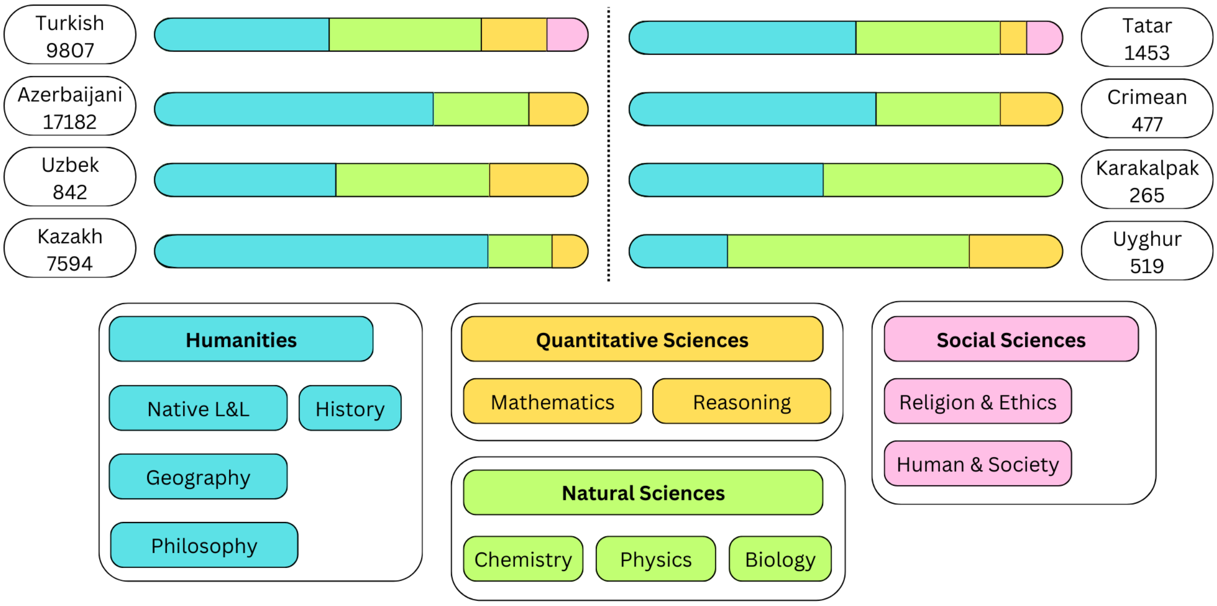

Figure 1. Distribution of subjects across languages in TUMLU. Numbers next to language names indicate the total question count. Left: middle- and high-resource languages; Right: low-resource languages.

2024), Chinese (Li et al., 2024), and Turkish (Yüksel et al., 2024).

In this paper, we focus on building a truly representative and inclusive single language family benchmark to address previous problems and provide a challenging setting for LLM evaluation. TUMLU (Turkic Unified Multilingual Language Understanding) benchmark covers the following languages: Azerbaijani, Crimean Tatar, Turkish, Uyghur, Uzbek, Karakalpak, Kazakh, and Tatar. The dataset consists of 4-choice questions at middle- and high-school levels. It consists of 38139 questions across 8 languages and 11 subjects (see Figure 1 for a higher-level breakdown across languages). It is the first such benchmark to include Uyghur, Karakalpak, Tatar, or Crimean Tatar. It is also a significant improvement over existing benchmarks for Azerbaijani, Uzbek, and Kazakh. Turkish dataset is TurkishMMLU, which was a separate project (Yüksel et al., 2024). The benchmark is also representative in terms of different scripts by including questions and answers in chosen languages in Latin, Cyrillic, and Arabic scripts. These datasets are transliterated such that it could be possible to get a dual dataset with the same content in two scripts for further comparative studies. We use these dual datasets to compare the performance of LLMs across different scripts.

We also release a more balanced and manually verified version of the dataset called TUMLU-mini, which contains 100 questions per subject (unless there are less than 100 for the said subject in a particular language). We use this version to test SOTA open-source and proprietary models of various sizes. We evaluated them in two settings: few-shot and chain-of-thought (CoT) reasoning (Wei et al., 2024). Our initial results show that proprietary models remain the best option for Turkic languages.

2 Related Work

Language understanding benchmarks Multi-task language understanding evaluation benchmarks play an important role in the evaluation of LLMs. Early benchmarks concentrated on general natural language understanding. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) were two such benchmarks that were widely adopted by the research community. These benchmarks were saturated quickly, due to the development of better LLMs. However, LLMs struggled more against benchmarks that required knowledge and reasoning. MMLU (Hendrycks et al., 2021a) and MMLU-Pro (Wang et al., 2024b) were more challenging since they required not only language understanding but also world knowledge. These general-purpose benchmarks gradually gave way to higher-level and more specialized benchmarks such as MATH (Hendrycks et al., 2021b), GPQA (Rein et al., 2024), and MUSR (Sprague et al., 2024).

Multilingual benchmarks The development of multilingual LLMs also necessitated challenging multilingual benchmarks. Most of these benchmarks were developed through machine translation (Conneau et al., 2018; Singh et al., 2024). However, such datasets have been shown to contain cultural biases and translation artifacts (Vanmassenhove et al., 2021). Global MMLU relied on machine and professional translation to (Singh et al., 2024). INCLUDE consists of native data (Romanou et al., 2025), but it is imbalanced, with different subject distributions in different languages. There is also a significant difference in required knowledge levels between languages, making a direct comparison impossible.

Benchmarks for Turkic languages SeaEval was one of the first LLM benchmarks to include Turkish (Wang et al., 2024a). Global MMLU contains Kyrgyz and Turkish subsets. INCLUDE contains Azerbaijani and Kazakh. MRL 2024 Shared Task on Multi-lingual Multi-task Information Retrieval (Tinner et al., 2024) contains an Azerbaijani dataset, but it contains general language understanding tasks rather than world knowledge. Kardeş-NLU has introduced a multilingual language understanding benchmark (Senel et al., 2024). But again, this benchmark contains general language understanding tasks that require no world knowledge. There are also monolingual benchmarks. Mukayese was one of the earliest general language understanding benchmarks in Turkish (Safaya et al., 2022). TurkishMMLU and TR-MMLU (Bayram et al., 2025) were the first native MMLU alternatives for the Turkish language. Another pilot study was performed to evaluate LLMs in Kazakh language (Maxutov et al., 2024). While there are no peer-reviewed monolingual MMLU alternatives for Azerbaijani, there is a general language understanding benchmark (Isbarov et al., 2024).

3 TUMLU

TUMLU is a multilingual and multitask dataset containing 38139 multiple-choice questions across 8 languages and 11 subjects. All questions are at middle or high school level. The majority are sample or official questions for university entrance exams of respective countries.

Data collection Data was collected from publicly available books and websites. In original form, questions had 2 to 5 choices. In cases where more than 4 choices were available, we removed an incorrect choice. If less than 4 choices were available, we left the question as-is. Except for Language and Literature questions in Crimean Tatar, all questions have 4 choices in the final version.

Figure 2. A sample question from the parallel Uzbek dataset, available in both Cyrillic and Latin alphabets. This enables comparison of LLM performance across different scripts. English translation is provided for clarity.

After collecting the data, native speakers of each language manually verified the quality of a random sample from each subject. In languages such as Azerbaijani where questions were developed by the community, around 10 % of the questions were either invalid or had incorrect answers.

We also created 5 CoT prompts per subject in Azerbaijani, Kazakh, Turkish, and Uzbek. Azerbaijani, Kazakh, and Uzbek prompts were created manually by native speakers. Turkish prompts were adapted from the TurkishMMLU project by changing the number of choices from 5 to 4. These prompts allowed us to compare the no-CoT and CoT performance of models. We did not create CoT prompts for other languages, because we did not have native speakers to validate their quality. CoT samples in Azerbaijani are available in appendix A.

Data composition TUMLU contains eleven subjects: Maths, Physics, Chemistry, Biology, Geography, Native Language & Literature (NL&L), History, Logic, Human & Society, Philosophy, Re-

Language	Question	Answer
Azerbaijani	63.1	28.0
Crimean Tatar	113.5	67.3
Karakalpak	112.3	65.3
Kazakh	96.8	19.7
Tatar	154.2	47.8
Turkish	204.6	69.6
Uyghur	180.1	51.1
Uzbek	161.4	16.2

Table 1. Average length of questions and answers across languages. An answer here refers to all choices, not only the correct ones.

ligion & Ethics. Among these, Logic, Human & Society, Philosophy and Religion & Ethics subjects are available only in one or two languages. Therefore, they have not been included in experiments.

We report the number of characters per question and per choice in Table 1. High variance in question and answer length indicates variable question types and levels across languages.

Difficulty levels While TUMLU can be used as a monolingual benchmark for any of the languages included, we are also interested in comparing performance across languages. This raises an important question: how comparable are questions of the same subject across different languages? While we can easily compare the model performance within each language, comparing it across languages proves more challenging. Different language subsets have different levels of difficulty. Uzbek and Turkish datasets are particularly difficult because those questions are designed specifically to imitate university entrance examinations in respective countries. Azerbaijani and Kazakh questions were developed by a community of students and teachers. While they all refer to middle- and high-school topics, there has been no oversight regarding their difficulty levels. For example, we know that maths questions in Kazakh are easier than the ones in other languages because they cover middle-school topics only. While these are explicit discrepancies that can be fixed in the future, there are certainly less obvious differences that are harder to identify and even harder to fix. That being said, even though comparison across languages is challenging, these datasets are more useful for comparison across models in monolingual settings.

TUMLU-mini To make our experiments more balanced and less costly, we developed TUMLU-mini, which consists of 100 randomly selected and manually verified questions per subject. In cases where we had less than 100 questions, we used the entire set. You can find the number of questions per language and subject in Appendix B. If a question had more than 5 answer choices, one was dropped. All choices have been shuffled to make the dataset more robust to simple memorization since it is possible that these questions were a part of the pre-training corpus for these LLMs. We also removed subjects that were available in less than 3 languages. All experiments were run on this subset.

4 Experimental set-up

Data Previous work (Yüksel et al., 2024) has shown that 100 questions per subject are enough to estimate the performance of a larger dataset. Therefore, we run all experiments on TUMLU-mini. While we have performed the experiments and publicly released the data, results on the following subjects are not reported in the paper: Logic, Philosophy, Religion & Ethics, and Human & Society. These subjects are available in one or two languages only, which makes any generalization impossible.

Model choice We have used TUMLU to evaluate both open-source models, such as Llama 3.1 (Grattafiori et al., 2024), Gemma 2 (Gemma Team, 2024), Qwen2.5 (Yang et al., 2025) and proprietary models, such as Gemini 1.5 (Gemini Team, 2024), Claude 3.5 (Anthropic, 2024), GPT-4o (Hurst et al., 2024). The size of selected open-source models varies between 7B and 70B. We do not have this information on proprietary models. This list includes models from the same series, such as Qwen2.5 7B instruct and Qwen2.5 70B instruct (Yang et al., 2025), which allows us to observe the effect of scaling (Hestness et al., 2017; Kaplan et al., 2020) on multilingual performance. All open-source models are instruct-tuned versions. We have omitted this information in the tables to preserve space. Wherever applicable, we have included the performance of Claude 3.5 Sonnet in the paper, since it consistently outperforms all other models. The performance of the remaining models can be found in the appendices C and D.

Prompting We have run experiments in two settings: 5-shot, where we provide 5 example questions and answers on the same subject before asking a question (Brown et al., 2020), and 5-shot CoT, where we provide 5 example questions and explanations of their answers before asking the question (Wei et al., 2024). Few-shot and CoT prompt samples are available in Appendix A. Previous work has demonstrated that (Romanou et al., 2025) providing the prompt in English does not result in performance gains. Due to this, we provide all prompts in respective native languages.

Technical details We run our experiments through OpenAI API, Anthropic API, Google Cloud Gemini API, Together AI API, and Deep Infra API. No model was run on a local machine. We used the following hyperparameters with all APIs: TEMPERATURE = 0.0, MAX_TOKENS = 1024, TOP_P = 1.0.

5 Results

In this section, we present the few-shot and CoT performance of selected models on the TUMLU-mini dataset. We also present an analysis of output language. Lastly, we explore how well LLMs perform on the same questions written in different (Latin, Cyrillic, or Arabic) scripts.

5-shot results We present the average performance of all models in each language in Table 2. Claude 3.5 Sonnet outperforms other models in all languages. The top 5 spots belong to proprietary models, although it has to be noted that there are larger open-source models that have not been included in this benchmark. Among the available open-source models, Qwen2.5 72B Instruct has the best performance. Results also confirm the scaling hypothesis: Llama 3.1 70B significantly outperforms Llama 3.1 8B. The same applies to Qwen2.5 7B/72B and Gemma 2 9B/27B. We can also observe a significant improvement from Llama 3.1 70B to Llama 3.3 70B. While it is not possible to directly compare results across languages, we can observe that low-resource languages, such as Crimean Tatar, Karakalpak, and Uyghur have comparable performance to middle- and high-resource languages. Notably, this trend holds even with the lowest-performing models.

We present the 5-shot evaluation of Claude 3.5 Sonnet in more detail in Table 3. In most languages, Native Language & Literature is the most challenging subject for Claude 3.5 Sonnet. This holds for other models, as well (See Appendix C).

5-shot CoT results We present the average results of the 5-shot CoT evaluation in Table 4. CoT prompts have an overall positive effect on performance. Sporadic negative effects can be explained by incorrect output format, rather than incorrect answers. We avoided manual validation of the output and instead relied on generalizable pattern-matching methods.

Table 5 shows the performance of Claude 3.5 Sonnet on each subject and language. On average, CoT prompts have a net positive effect in each subject and each language.

Generated language vs. performance Benchmark results demonstrate that LLMs can have significant language understanding capabilities even in previously unseen languages, such as Crimean Tatar. This can be explained by linguistic proximity to languages better represented in the training data. Even though LLMs perform surprisingly well in these languages with simple 5-shot prompts. The results are less impressive when we analyze the generated text quality. While quality *per se* is hard to quantify, we can detect the language of generated content. We used Google Cloud Translate API to detect output language. This API supports all languages in our benchmark, except for Karakalpak. We present results for Crimean Tatar in Figure 3. As you can see, although these models have answered the majority of the questions in Crimean Tatar correctly, only a small portion of the generated text is classified as Crimean Tatar. Almost all of the answers are a synthesis between Turkish and Crimean Tatar. A similar issue appears in Kazakh when we switch from Cyrillic to Latin script. Although this has a small negative effect on the performance, the nature of the generated content changes dramatically. While the output of Cyrillic questions is easily detected as Kazakh, the output of Latin questions is easily confused with Tatar language.

Comparing performance on same questions written in different alphabets Some Turkic languages, such as Crimean Tatar, Kazakh, and Uzbek have both Cyrillic and Latin alphabets that are actively used. As a result, the text corpora that are used to train LLMs contain both versions. Also, transliteration between these scripts can be done automatically with a negligible error rate. Using these facts, we developed dual datasets for the languages above (see Figure 2). We evaluated models in both versions and compared their performance.

Model	Mean	aze	crh	kaa	kaz	tat	tur	uig	uzb
Claude 3.5 Sonnet	**79.0**	**84.4**	**81.0**	**75.3**	**83.0**	**84.0**	**84.9**	**70.9**	**68.6**
GPT-4o	75.1	82.4	70.5	70.8	81.0	80.4	83.7	66.5	65.4
Gemini 1.5 Pro	66.7	74.7	59.7	67.4	78.3	77.3	59.1	55.8	61.1
Gemini 1.5 Flash	64.6	72.1	67.5	61.2	68.4	66.7	73.1	57.4	50.0
Claude 3.5 Haiku	63.3	70.6	61.7	54.9	69.9	67.3	77.6	54.4	50.0
Qwen2.5 72B	61.5	70.1	61.8	54.6	62.4	62.5	73.9	56.0	50.4
Llama 3.3 70B	57.6	66.0	58.7	49.2	59.7	69.5	68.1	45.6	44.1
Gemma 2 27B	51.5	58.1	49.8	47.6	58.4	54.7	64.3	42.2	36.9
Llama 3.1 70B	50.9	67.6	56.1	49.0	47.7	54.6	64.7	28.5	39.0
Gemma 2 9B	46.8	53.7	46.8	40.8	49.1	51.8	60.1	35.8	36.1
Qwen2.5 7B	42.1	48.0	42.6	37.2	45.0	40.5	55.6	33.4	34.6
Llama 3.1 8B	39.7	48.4	35.7	33.4	46.4	44.1	44.4	35.0	29.9

Table 2. Average 5-shot performance of models on Azerbaijani (aze), Crimean Tatar (crh), Kara-Kalpak (kaa), Kazakh (kaz), Tatar (tat), Turkish (tur), Uyghur (uig), and Uzbek (uzb) datasets.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	89.0	89.0	91.0	71.0	85.0	73.0	93.0
Crimean Tatar	81.6	75.0	87.0	88.4	70.4	75.0*	89.7
Karakalpak	78.0	85.7	75.0	-	-	42.2	95.6
Kazakh	92.0	73.0	78.0	78.0	96.0	76.0	88.0
Tatar	94.0	84.0	83.0	91.0	86.3	69.0	81.0
Turkish	84.0	87.0	94.0	91.0	77.0	76.0	85.0
Uyghur	73.0	66.0	-	-	75.8	66.0	73.5
Uzbek	69.0	73.0	64.0	68.0	70.0	55.0	81.0

Table 3. Subject-wise 5-shot performance of Claude 3.5 Sonnet across Turkic languages. Missing values indicate the absence of data for that language in the given subject. NL&L refers to Native Language and Literature. *This subset contains questions with 2 or 3 choices.

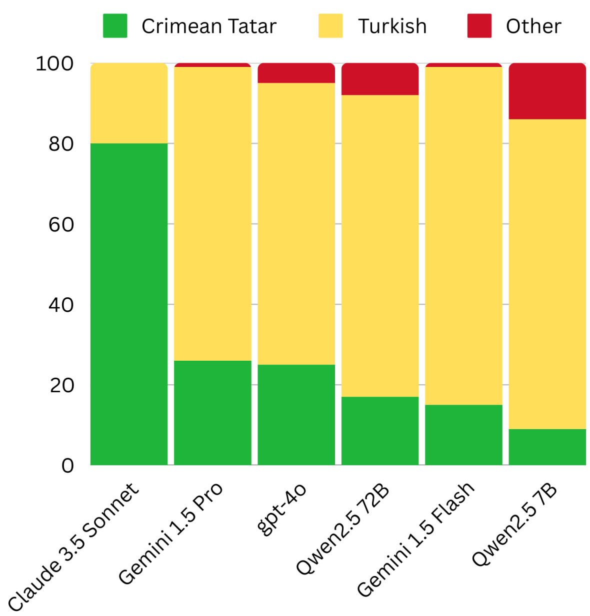

Figure 3. Language distribution of model responses to Crimean Tatar queries, as detected by Google Cloud Translation API.

We present some of the results in Table 6. While the results initially seem irregular, they follow a simple pattern:

1. In Crimean Tatar questions, all three models perform better in the Latin script. FineWeb 2 (Penedo et al., 2024), one of the largest multilingual text corpora, contains 21,365,608 Latin and 1,934,168 Cyrillic words in Crimean Tatar.

2. In Kazakh questions, all three models perform better in the Cyrillic script. This aligns with the fact that most of the Kazakh text data on the web is written in the Cyrillic script. For example, the FineWeb 2 corpus contains 1,837,049,585 Cyrillic and 0 Latin words in Kazakh.

3. In Uyghur questions, all three models perform better in the Arabic script. While Uyghur is not represented in Fineweb 2 corpus, virtually all Uyghur text is written in Arabic script.

Model	Mean	aze	kaz	tur	uzb
Claude 3.5 Sonnet	**82.2**	**87.1 (+2.7)**	**84.1 (+1.1)**	87.4 **(+1.5)**	70.1 (+1.5)
GPT-4o	78.4	82.9 (-0.5)	80.7 (-0.3)	84.0 (+0.3)	66.0 (+0.6)
Gemini 1.5 Pro	69.4	74.6 (-0.1)	74.7 (-3.6)	75.0 (+15.9)	53.4 (-7.7)
Claude 3.5 Haiku	67.0	77.0 (+6.4)	74.0 (+4.1)	70.7 (-6.9)	46.3 (-3.7)
Qwen2.5 72B	66.1	72.1 (+2.0)	63.9 (+1.5)	78.4 (+4.5)	49.9 (-0.5)
Gemini 1.5 Flash	65.1	73.6 (+1.5)	67.6 (-0.8)	69.4 (-3.7)	50.0 (0)
Llama 3.3 70B	59.5	69.9 (+3.9)	69.3 (+9.6)	75.9 (+7.8)	23.0 (-21.1)
Gemma 2 27B	58.0	62.9 (+4.8)	61.6 (+3.2)	66.4 (+2.1)	41.1 (+4.2)
Llama 3.1 70B	52.8	59.0 (-8.6)	61.7 (+14.0)	73.3 (+8.6)	17.3 (-21.7)
Gemma 2 9B	51.4	57.1 (+3.4)	52.6 (+3.5)	62.3 (+2.2)	33.7 (-2.4)
Qwen2.5 7B	47.1	48.0 (0.0)	46.4 (+1.4)	56.3 (+0.7)	37.9 (+2.3)
Llama 3.1 8B	36.4	40.7 (-7.7)	38.9 (-7.5)	45.0 (+0.6)	20.9 (-9.0)

Table 4. Average 5-shot performance of models on Turkic languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	89.0 (0.0)	96.0 (+7.0)	91.0 (0.0)	78.0 (+7.0)	83.0 (-2.0)	76.0 (+3.0)	97.0 (+4.0)
Kazakh	96.0 (+4.0)	74.0 (+1.0)	78.0 (0.0)	80.0 (+2.0)	95.0 (-1.0)	79.0 (+3.0)	87.0 (-1.0)
Turkish	86.0 (+2.0)	85.0 (-2.0)	95.0 (+1.0)	93.0 (+2.0)	77.0 (0.0)	87.0 (+11.0)	89.0 (+4.0)
Uzbek	77.0 (+8.0)	73.0 (0.0)	65.0 (+1.0)	65.0 (-3.0)	79.0 (+9.0)	45.0 (-10.0)	87.0 (+6.0)

Table 5. Subject-wise 5-shot CoT performance of Claude 3.5 Sonnet across Turkic languages.

4. In Uzbek questions, results are less predictable. This can be explained by the fact that Cyrillic and Latin are more evenly distributed in Uzbek text. FineWeb 2 corpus contains 616,563,348 Latin and 492,264,125 Cyrillic words in Uzbek.[1]

While these patterns hold across multiple models, there are exceptions. For example, on Uyghur questions, GPT-4o performs similarly with Arabic and Latin scripts. Llama 3.1 70B has an average accuracy of 28.48 on Uyghur questions with Arabic script and 41.30 with Latin script.

6 Conclusion

We introduce TUMLU, a unified and native language understanding benchmark for Turkic languages. It contains 38139 multiple-choice questions in 8 languages and 11 subjects. Latin, Cyrillic, and Arabic scripts are represented in the benchmark. Uzbek, Crimean Tatar, and Kazakh are available in both Cyrillic and Latin. Uyghur is available both in Arabic and Latin. We also release TUMLU-mini, a smaller, more balanced and manually verified version that is more suitable for large-scale experiments. We use TUMLU-mini to benchmark 5 proprietary and 7 open-source LLMs. Results show that LLMs have a reasonably good understanding of all 8 languages, including ones that are not explicitly included in the training data of LLMs. However, LLMs are less capable of generating text in these languages, usually answering multiple-choice questions correctly, but in another, similar high-resource language.

7 Limitations

TUMLU benchmark has two main limitations.

Mismatched difficulty levels Native language & literature subset contained both literature and language questions in some languages, while it contained only language questions in others. Similarly, the history subset contained both world and national history questions in some languages, while it contained only national questions in others. Maths questions in Kazakh are at middle-school level, which results in very high scores.

Missing major languages TUMLU covers 8 Turkic languages with more than 180 million native speakers. However, some major Turkic languages, such as Turkmen, Kyrgyz and Bashkir are not included. We are hoping to extend our benchmark with more languages in further editions.

[1]In this work, Uzbek refers to Northern Uzbek.

Language	Claude 3.5 Sonnet			Qwen2.5 72B			Gemma 2 27B		
	Cyrillic	Latin	Arabic	Cyrillic	Latin	Arabic	Cyrillic	Latin	Arabic
Crimean Tatar	66.1	**80.0**	—	47.6	**61.8**	—	43.5	**49.8**	—
Kazakh	**82.7**	78.0	—	**64.3**	54.1	—	**58.5**	46.3	—
Uyghur	—	64.5	**70.8**	—	53.4	**56.1**	—	36.0	**42.2**
Uzbek	67.9	**68.6**	—	**51.1**	50.4	—	**39.4**	36.9	—

Table 6. Performance comparison (%) of three LLMs on Turkic languages with their native writing systems: Arabic and Latin for Uyghur, Cyrillic and Latin for Kazakh, Crimean Tatar, and Uzbek. Bold numbers indicate the best script performance per language-model pair. Dashes (—) denote script combinations not used in practice.

References

Anthropic. 2024. Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet.

Edward Bayes, Israel Abebe Azime, Jesujoba O Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A Etori, Shamsuddeen Hassan Muhammad, et al. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948*.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Banu Diri, Savaş Yıldırım, and Öner Aytaş. 2025. Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation. *Preprint*, arXiv:2501.00593.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep Learning Scaling is Predictable, Empirically. *Preprint*, arXiv:1712.00409.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jafar Isbarov, Kavsar Huseynova, Elvin Mammadov, Mammad Hajili, and Duygu Ataman. 2024. Open foundation models for Azerbaijani language. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 18–28, Bangkok, Thailand and Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *Preprint*, arXiv:2001.08361.

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.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Akylbek Maxutov, Ayan Myrzakhmet, and Pavel Braslavski. 2024. Do LLMs speak Kazakh? a pilot evaluation of seven models. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 81–91, Bangkok, Thailand and Online. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Francesco Tinner, Raghav Mantri, Mammad Hajili, Chiamaka Chukwuneke, Dylan Massey, Benjamin A. Ajibade, Bilge Deniz Kocak, Abolade Dawud, Jonathan Atala, Hale Sirin, Kayode Olaleye, Anar Rzayev, Jafar Isbarov, Dursun Dashdamirov, David Adelani, and Duygu Ataman. 2024. Findings of the 2nd shared task on multi-lingual multi-task information retrieval at MRL 2024. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 365–376, Miami, Florida, USA. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew 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, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing* *and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Preprint*, arXiv:2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt samples

Sual: 1 saatda 1 bakteriyadan neçə nəsil törəyər?
A) 1 nəsil
B) 8 nəsil
C) 3 nəsil
D) 9 nəsil

Cavab: C

Sual: Difteriyaya tutulan 50 nəfərdən heç biri müalicə serumu almazsa onda onların neçə nəfəri ölər? (10 nəfərdən 3-4 sağ qalır)
A) 12-14
B) 3-4
C) 6-7
D) 30-35

Cavab: D

Sual: Mitoz bölünmədən əvvəl bitki hüceyrəsində 24 xromosom olarsa əmələ gələn cavan hüceyrələrin hər birində neçə xromosom olar?
A) 24
B) 12
C) 6
D) 48

Cavab: D

Sual: İkiqat mayalanmaya hazırlıq mərhələsində əmələ gələn tozcuqların nüvəsi bölündükdən sonra cəmi 186 nüvə yaranır. Neçə tozcuq hüceyrəsi bu bölünmədə iştirak etmişdir?
A) 186
B) 185
C) 372
D) 93

Cavab: D

Sual: İnsan bütün ömrü boyu maksimum neçə yeni diş çıxarır?
A) 28
B) 12
C) 52
D) 32
Cavab: C

Sual: Selülozanın inşaat funksiyasına nə aiddir?
A) Qan damarlarının divarı
B) İnşaat funksiyası yoxdur
C) Bitki hüceyrələrinin qılafı
D) Plazmatik membran

Cavab:

Figure 4. 5-shot prompt sample for Biology questions in Azerbaijani.

Sual: 1 saatda 1 bakteriyadan neçə nəsil törəyər?
A) 3 nəsil
B) 9 nəsil
C) 8 nəsil
D) 1 nəsil

Həll: Addım-addım düşünək. Bakteriyalar hər 20 dəqiqədə bir dəfə çoxalır. 1 saat 60 dəqiqədir. 1 bakteriyalar 1 saatda 3 nəsil törəyər. Düzgün cavab A variantıdır.

Sual: İkiqat mayalanmaya hazırlıq mərhələsində əmələ gələn tozcuqların nüvəsi bölündükdən sonra cəmi 186 nüvə yaranır. Neçə tozcuq hüceyrəsi bu bölünmədə iştirak etmişdir?
A) 185
B) 186
C) 93
D) 372

Həll: Addım-addım düşünək. İkiqat mayalanmayanın hazırlıq mərhələsində hər tozcuq mitoz bölünmə keçirir. Bilirik ki, mitoz bölünmə nəticəsində hüceyrənin nüvəsi ikiyə bölünür. Bu o deməkdir ki, mitoz bölünmə zamanı 186 nüvə yaranmışdırsa, deməli 93 tozcuq hüceyrəsi bu bölünmədə iştirak etmişdir. Düzgün cavab C variantıdır.

...

Sual: İnsan bütün ömrü boyu maksimum neçə yeni diş çıxarır?
A) 52
B) 32
C) 12
D) 28

Həll: Addım-addım düşünək. İnsanlar iki dəfə diş çıxarır:
1. Süd dişləri (20 ədəd): Körpəlikdə çıxan və daha sonra tökülən dişlərdir.
2. Daimi dişlər (32 ədəd): Süd dişləri töküldükdən sonra çıxan və ömür boyu qalan dişlərdir. Bunlara 4 ədəd ağız boşluğu dişi (ağıllı dişlər) də daxildir.
İnsanın ömrü boyu çıxara biləcəyi maksimum yeni diş sayı 52-dir. Düzgün cavab: A variantıdır.

Sual: Selülozanın inşaat funksiyasına nə aiddir?
A) İnşaat funksiyası yoxdur
B) Plazmatik membran
C) Qan damarlarının divarı
D) Bitki hüceyrələrinin qılafı

Həll:

Figure 5. 5-shot CoT prompt sample for Biology questions in Azerbaijani.

B TUMLU-mini

Language (code)	NL&L	History	Geography	Chemistry	Physics	Biology	Maths
Azerbaijani (aze)	100	100	100	100	100	100	100
Crimean Tatar (crh)	100	69	23	32	39	38	54
Karakalpak (kaa)	64	0	28	28	45	50	0
Kazakh (kaz)	100	100	100	100	100	100	100
Tatar (tat)	100	100	100	100	100	100	100
Turkish (tur)	100	100	100	100	100	100	100
Uyghur (uig)	100	0	0	97	98	100	99
Uzbek (uzb)	100	100	100	100	100	100	100

Table 7. Composition of TUMLU-mini, a more balanced and manually verified subset of TUMLU benchmark. All experiments in this paper have been run on TUMLU-mini. These numbers exclude sample questions used in 5-shot and 5-shot CoT prompts. Language codes are from ISO 639-3.

C 5-shot results

This appendix includes 5-shot results for all models, except for Claude 3.5 Sonnet which is available in Table 3.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	78.00	79.00	72.00	61.00	65.00	58.00	81.00
Crimean Tatar	63.16	53.12	73.91	56.52	57.41	62.00	65.52
Karakalpak	60.00	60.71	53.57	-	-	20.31	80.00
Kazakh	79.00	66.00	72.00	64.00	76.00	58.00	74.00
Tatar	74.00	63.00	73.00	74.00	63.16	59.00	65.00
Turkish	71.00	82.00	84.00	85.00	71.00	70.00	80.00
Uyghur	57.00	44.33	-	-	59.60	52.00	59.18
Uzbek	53.00	50.00	48.00	49.00	54.00	41.00	55.00

Table 8. Accuracy scores for Claude 3.5 Haiku-20241022 model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	80.00	80.00	78.00	55.00	72.00	56.00	84.00
Crimean Tatar	78.95	59.38	73.91	59.42	61.11	67.00	72.41
Karakalpak	70.00	82.14	53.57	-	-	31.25	68.89
Kazakh	88.00	60.00	73.00	57.00	68.00	57.00	76.00
Tatar	86.00	68.00	74.00	68.00	64.21	47.00	60.00
Turkish	75.00	72.00	78.00	76.00	78.00	59.00	74.00
Uyghur	71.00	53.61	-	-	59.60	57.00	45.92
Uzbek	59.00	43.00	49.00	57.00	69.00	22.00	51.00

Table 9. Accuracy scores for GEMINI-1.5-FLASH model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	81.00	86.00	76.00	56.00	78.00	57.00	89.00
Crimean Tatar	71.05	50.00	65.22	56.52	61.11	52.00	62.07
Karakalpak	76.00	71.43	57.14	-	-	43.75	88.89
Kazakh	88.00	68.00	75.00	73.00	94.00	70.00	80.00
Tatar	95.00	78.00	81.00	84.00	80.00	59.00	64.00
Turkish	51.00	61.00	61.00	70.00	64.00	50.00	57.00
Uyghur	70.00	42.27	-	-	49.49	66.00	51.02
Uzbek	59.00	69.00	61.00	54.00	79.00	30.00	76.00

Table 10. Accuracy scores for GEMINI-1.5-PRO model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	61.00	57.00	59.00	47.00	45.00	42.00	65.00
Crimean Tatar	50.00	37.50	56.52	52.17	33.33	53.00	44.83
Karakalpak	48.00	46.43	35.71	-	-	25.00	48.89
Kazakh	67.00	37.00	63.00	41.00	38.00	48.00	50.00
Tatar	69.00	53.00	63.00	54.00	35.79	41.00	47.00
Turkish	65.00	55.00	76.00	75.00	45.00	48.00	57.00
Uyghur	40.00	26.80	-	-	37.37	38.00	36.73
Uzbek	49.00	33.00	39.00	41.00	31.00	26.00	34.00

Table 11. Accuracy scores for GOOGLE/GEMMA-2-9B-IT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	70.00	59.00	62.00	46.00	38.00	55.00	77.00
Crimean Tatar	44.74	46.88	60.87	49.28	35.19	60.00	51.72
Karakalpak	52.00	64.29	42.86	-	-	23.44	55.56
Kazakh	79.00	44.00	65.00	56.00	52.00	51.00	62.00
Tatar	71.00	58.00	71.00	63.00	43.16	32.00	45.00
Turkish	73.00	65.00	81.00	78.00	41.00	55.00	57.00
Uyghur	50.00	35.05	-	-	34.34	51.00	40.82
Uzbek	46.00	27.00	40.00	44.00	35.00	26.00	40.00

Table 12. Accuracy scores for GOOGLE/GEMMA-2-27B-IT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	91.00	93.00	89.00	75.00	70.00	67.00	92.00
Crimean Tatar	60.53	59.38	69.57	86.96	57.41	70.00	89.66
Karakalpak	80.00	82.14	71.43	-	-	35.94	84.44
Kazakh	93.00	71.00	76.00	77.00	85.00	77.00	88.00
Tatar	98.00	78.00	88.00	92.00	69.47	69.00	68.00
Turkish	86.00	79.00	95.00	94.00	63.00	82.00	87.00
Uyghur	84.00	54.64	-	-	62.63	65.00	66.33
Uzbek	70.00	68.00	65.00	69.00	56.00	51.00	79.00

Table 13. Accuracy scores for GPT-4o model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	76.00	72.00	77.00	59.00	45.00	49.00	84.00
Crimean Tatar	68.42	53.12	69.57	72.46	40.74	58.00	48.28
Karakalpak	58.00	46.43	64.29	-	-	21.88	55.56
Kazakh	80.00	42.00	71.00	62.00	52.00	51.00	60.00
Tatar	88.00	67.00	83.00	82.00	67.37	51.00	48.00
Turkish	76.00	58.00	86.00	83.00	41.00	65.00	68.00
Uyghur	37.00	46.39	-	-	48.48	49.00	46.94
Uzbek	52.00	38.00	52.00	50.00	44.00	31.00	42.00

Table 14. Accuracy scores for META-LLAMA/LLAMA-3.3-70B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	53.00	50.00	57.00	45.00	36.00	47.00	51.00
Crimean Tatar	42.11	34.38	30.43	44.93	18.52	45.00	34.48
Karakalpak	38.00	39.29	32.14	-	-	21.88	35.56
Kazakh	60.00	33.00	64.00	54.00	32.00	41.00	41.00
Tatar	55.00	40.00	61.00	55.00	33.68	33.00	31.00
Turkish	51.00	37.00	63.00	50.00	34.00	35.00	41.00
Uyghur	38.00	27.84	-	-	36.36	42.00	30.61
Uzbek	31.00	22.00	38.00	33.00	25.00	29.00	31.00

Table 15. Accuracy scores for META-LLAMA/META-LLAMA-3.1-8B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	78.00	67.00	78.00	60.00	48.00	62.00	80.00
Crimean Tatar	63.16	40.62	65.22	66.67	29.63	62.00	65.52
Karakalpak	58.00	53.57	60.71	-	-	28.12	44.44
Kazakh	57.00	27.00	55.00	57.00	40.00	50.00	48.00
Tatar	72.00	23.00	67.00	72.00	49.47	52.00	47.00
Turkish	70.00	55.00	88.00	70.00	40.00	62.00	68.00
Uyghur	24.00	16.49	-	-	20.20	45.00	36.73
Uzbek	42.00	39.00	53.00	34.00	36.00	27.00	42.00

Table 16. Accuracy scores for META-LLAMA/META-LLAMA-3.1-70B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	41.00	58.00	53.00	37.00	59.00	30.00	58.00
Crimean Tatar	36.84	37.50	39.13	39.13	55.56	42.00	48.28
Karakalpak	30.00	42.86	39.29	-	-	25.00	48.89
Kazakh	38.00	44.00	54.00	31.00	64.00	31.00	53.00
Tatar	41.00	38.00	44.00	42.00	56.84	28.00	34.00
Turkish	42.00	59.00	62.00	69.00	58.00	40.00	59.00
Uyghur	34.00	29.90	-	-	38.38	40.00	24.49
Uzbek	35.00	31.00	30.00	34.00	52.00	21.00	39.00

Table 17. Accuracy scores for QWEN/QWEN2.5-7B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	76.00	77.00	74.00	53.00	73.00	54.00	84.00
Crimean Tatar	65.79	53.12	60.87	72.46	55.56	66.00	58.62
Karakalpak	50.00	75.00	50.00	-	-	35.94	62.22
Kazakh	60.00	55.00	64.00	52.00	75.00	52.00	79.00
Tatar	68.00	60.00	65.00	78.00	71.58	41.00	54.00
Turkish	79.00	73.00	84.00	85.00	61.00	56.00	79.00
Uyghur	60.00	51.55	-	-	64.65	52.00	52.04
Uzbek	53.00	49.00	44.00	55.00	63.00	28.00	61.00

Table 18. Accuracy scores for QWEN/Qwen2.5 72B-INSTRUCT model across languages.

D 5-shot CoT results

This appendix includes 5-shot CoT results for all models, except for Claude 3.5 Sonnet which is available in Table 5.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	81.00	83.00	81.00	69.00	76.00	61.00	88.00
Kazakh	82.00	65.00	75.00	61.00	87.00	64.00	84.00
Turkish	74.00	78.00	91.00	85.00	28.00	71.00	68.00
Uzbek	50.00	50.00	37.00	47.00	63.00	24.00	53.00

Table 19. Accuracy scores for Claude 3.5 Haiku-20241022 model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	81.00	83.00	81.00	52.00	71.00	59.00	88.00
Kazakh	79.00	61.00	69.00	51.00	85.00	50.00	78.00
Turkish	71.00	67.00	69.00	69.00	76.00	60.00	74.00
Uzbek	50.00	56.00	41.00	49.00	70.00	17.00	67.00

Table 20. Accuracy scores for GEMINI-1.5-FLASH model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	79.00	92.00	78.00	53.00	78.00	62.00	80.00
Kazakh	83.00	70.00	63.00	65.00	92.00	67.00	83.00
Turkish	73.00	76.00	84.00	79.00	85.00	51.00	77.00
Uzbek	53.00	67.00	43.00	40.00	73.00	17.00	81.00

Table 21. Accuracy scores for GEMINI-1.5-PRO model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	64.00	67.00	65.00	39.00	45.00	47.00	73.00
Kazakh	61.00	42.00	64.00	46.00	61.00	32.00	62.00
Turkish	72.00	60.00	74.00	71.00	45.00	50.00	64.00
Uzbek	41.00	31.00	40.00	41.00	26.00	25.00	32.00

Table 22. Accuracy scores for GOOGLE/GEMMA-2-9B-IT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	72.00	76.00	71.00	41.00	55.00	48.00	77.00
Kazakh	74.00	52.00	63.00	46.00	70.00	53.00	73.00
Turkish	77.00	64.00	80.00	77.00	53.00	49.00	65.00
Uzbek	43.00	43.00	48.00	53.00	42.00	10.00	49.00

Table 23. Accuracy scores for GOOGLE/GEMMA-2-27B-IT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	90.00	91.00	88.00	73.00	75.00	70.00	93.00
Kazakh	89.00	63.00	78.00	82.00	85.00	84.00	84.00
Turkish	86.00	80.00	97.00	92.00	70.00	80.00	83.00
Uzbek	73.00	68.00	70.00	74.00	59.00	39.00	79.00

Table 24. Accuracy scores for GPT-4o model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	81.00	78.00	75.00	54.00	70.00	52.00	79.00
Kazakh	80.00	57.00	70.00	65.00	74.00	59.00	80.00
Turkish	74.00	67.00	84.00	87.00	77.00	67.00	75.00
Uzbek	46.00	15.00	8.00	33.00	35.00	10.00	14.00

Table 25. Accuracy scores for META-LLAMA/LLAMA-3.3-70B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	43.00	44.00	57.00	36.00	20.00	32.00	53.00
Kazakh	52.00	28.00	55.00	46.00	17.00	32.00	42.00
Turkish	50.00	37.00	58.00	61.00	20.00	39.00	50.00
Uzbek	34.00	5.00	37.00	29.00	9.00	23.00	9.00

Table 26. Accuracy scores for META-LLAMA/META-LLAMA-3.1-8B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	52.00	67.00	69.00	49.00	55.00	50.00	71.00
Kazakh	77.00	51.00	65.00	65.00	56.00	53.00	65.00
Turkish	72.00	64.00	83.00	88.00	69.00	60.00	77.00
Uzbek	31.00	24.00	8.00	20.00	12.00	11.00	15.00

Table 27. Accuracy scores for META-LLAMA/META-LLAMA-3.1-70B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	42.00	54.00	54.00	37.00	51.00	35.00	63.00
Kazakh	38.00	41.00	48.00	31.00	66.00	33.00	68.00
Turkish	51.00	59.00	65.00	48.00	70.00	43.00	58.00
Uzbek	40.00	30.00	40.00	42.00	52.00	17.00	44.00

Table 28. Accuracy scores for QWEN/QWEN2.5-7B-INSTRUCT model across languages.

Language	Biology	Chemistry	Geography	History	Maths	NL&L	Physics
Azerbaijani	72.00	88.00	80.00	51.00	80.00	46.00	88.00
Kazakh	64.00	50.00	70.00	50.00	84.00	52.00	77.00
Turkish	78.00	78.00	89.00	84.00	79.00	57.00	84.00
Uzbek	54.00	55.00	46.00	50.00	49.00	27.00	68.00

Table 29. Accuracy scores for QWEN/Qwen2.5 72B-INSTRUCT model across languages.