School of Information Technology and
Engineering at the ADA University

School of Engineering and Applied Science
at the George Washington University

EVALUATING LANGUAGE UNDERSTANDING AND WORLD KNOWLEDGE OF LARGE LANGUAGE MODELS IN TURKIC LANGUAGES

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Jafar Isbarov

April, 2024

This Thesis by: Jafar Isbarov
Entitled: *Evaluating Language Understanding and World Knowledge of Large Language Models in Turkic Languages*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Samir Rustamov	Rustamov S.	30.04.2025
(Advisor)		(Date)
Samir Rustamov	Rustamov S.	30.04.2025
(Program Director)		(Date)
Abzatdin Adamov		
(Dean)		(Date)

The most damning revelation you can make about yourself is that you do not know what is interesting and what is not.

— Kurt Vonnegut

ABSTRACT

MMLU is one of the most widely used and referred-to benchmarks for LLM performance [19]. It consists of multiple-choice questions on elementary, high school, college, and professional topics. MMLU benchmark was designed with English in mind, but we have seen a rise of multilingual benchmarks alongside multilingual models in recent years. One example is the MMMLU, which has multiple alternatives now. Such benchmarks allow researchers to compare the performance of LLMs across languages, speeding up the development in this area.

While multilingual benchmarks exist, such works tend to have a global scope. As a result, low-resource languages are left out of the spotlight yet again. This applies to Turkic languages as well. With the exception of Turkish, Turkic languages have been left of most such benchmarks. In this work, we have concentrated on evaluation of Turkic languages, including extremely low-resource languages such as Tatar and Crimean Tatar. We have prepared a benchmark consisting of high-school-level multiple-choice questions. This benchmark has allowed us to evaluate the NLU capabilities of SOTA LLMs in Turkic languages.

Turkish is the only Turkic language with a native MMLU benchmark. TurkishMMLU was released a few months ago [65], and we have collaborated with their team on this project. There have been attempts to create MMLU benchmarks for Turkic (and other) languages by synthetic generation or machine translation. However, these approaches are known to produce potentially erroneous text and do not consider the linguistic and cultural nuances of the target language. Since MMLU was designed for the English language, its contents are biased towards an Anglocentric worldview. Therefore, any successful benchmark should be created from native sources instead of direct or indirect translations. Since TurkishMMLU already exists, our project expanded this to seven more Turkic languages – Azerbaijani, Kazakh, Uzbek, Tatar, Crimean Tatar, Uyghur, and Karakalpak.

In this project, we have created a unified and native language understanding benchmark for the aforementioned languages. We have later used this benchmark to evaluate state-of-the-art LLMs with multilingual capabilities. We have also evaluated the effect of the writing system on the performance of the model. Finally, we have analysed the output of LLMs in low-resource languages and identified that LLMs can respond to questions in low-resource languages using a high-resource language without being prompted to do so.

CONTENTS

1 INTRODUCTION **1**
- 1.1 Turkic Unified Multilingual Language Understanding Benchmark . . . 1
- 1.2 Limitations . . . 2
- 1.3 Thesis outline . . . 2

2 LITERATURE REVIEW **3**
- 2.1 Language modeling . . . 3
- 2.2 Multilingual and low-source language modeling . . . 4
- 2.3 Modeling Turkic languages . . . 5
- 2.4 Evaluation of Large Language Models . . . 6

3 Background **7**
- 3.1 Language modeling . . . 7
- 3.2 Large Language Models . . . 7
- 3.3 Prompting LLMs . . . 10
- 3.4 Evaluating LLMs . . . 12
- 3.5 Multilingual benchmarks . . . 14
- 3.6 Turkic Languages . . . 14
- 3.7 Problem Statement . . . 16

4 Dataset **18**
- 4.1 Dataset Creation . . . 19
- 4.2 Dataset Composition . . . 21
- 4.3 Considerations for Use . . . 22
- 4.4 TUMLU-mini . . . 22
- 4.5 Evaluation Approach . . . 24

5 EXPERIMENTS **25**
- 5.1 Experimental set-up . . . 25
- 5.2 Main findings . . . 30
  - 5.2.1 5-shot results . . . 30
  - 5.2.2 5-shot Chain-of-Thought results . . . 31
  - 5.2.3 Generated language vs. performance . . . 33
  - 5.2.4 Comparing performance on same questions written in different alphabets . 34
- 5.3 Model-specific results . . . 36

5.4 Released Resources 37

6 CONCLUSION AND FUTURE WORK **37**

A Model results **47**

1 INTRODUCTION

1.1 Turkic Unified Multilingual Language Understanding Benchmark

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 [63, 16, 17] and excel in complex reasoning tasks such as mathematical, scientific, and coding-related inquiries [25, 2, 15, 17]. 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 [19, 52, 58, 57]. In order to extend the applicability of LLM evaluation in more languages, recent efforts were undertaken to build more multilingual NLU benchmarks [32], 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 [55] and cultural misalignment [48]. On the other end of the spectrum, there are efforts to bridge that gap for a particular language, for example, African languages [4], Arabic [29], Chinese [33], and Turkish [65]. These benchmarks are more effective in evaluating the performance of LLMs in a particular language, but their mono-lingual nature makes it impossible to perform multi-lingual analysis. We believe it is necessary to develop a benchmark that is multilingual but also capable of capturing the linguistic and cultural nuances of the target languages.

In this work, 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. The 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 6 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 [65]. 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. Figure 7 contains a sample from the Uzbek subset. In the case of Uzbek, Crimean Tatar, and Kazakh, all questions are available both in Cyrillic and Latin scripts, and we report the performance of LLMs in both scripts in the last section. Uyghur is available in Latin and Arabic 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 reasoning [59]. Our initial results show that proprietary models remain the best option for Turkic languages.

1.2 Limitations

The main goal of this project is to create a multilingual and multi-subject language understanding benchmark for Turkic languages. We also aim to use this benchmark to evaluate state-of-the-art LLMs on Turkic languages. Finally, we compare their performance in the same language with different scripts. That being said, we also consider it important to state our non-goals.

1. TUMLU benchmark is not intended to compare performance across languages. While this can be done in a limited sense, these results cannot be considered rigorous since there is a considerable difference in the difficulty level of the same subject in different languages.

2. TUMLU benchmark does not contain open-ended questions. All questions are multiple-choice, with 4 or fewer choices.

3. TUMLU benchmark does not contain multimodal questions. All questions are text only. Mathematical and scientific notations are expressed in LaTeXor Markdown, depending on the level of complexity.

1.3 Thesis outline

The work is structured as follows:

- Section 1 (this) introduces the problem.

- Section 2 outlines the relevant literature on the topic. It covers language modeling, multilingual and low-resource LLMs, Turkic languages, and the evaluation of LLMs.

- Section 3 provides the necessary background information on language modeling, LLM benchmarks, and Turkic languages. It introduces principles of language modeling, discusses LLMs and common prompting techniques, and finally explains the current state of language modeling in Turkic languages.
- Section 4 contains details on the methodology of data collection and contents of the final dataset. It introduces TUMLU-mini, a smaller and manually verified subset of TUMLU that was used to evaluate SOTA LLMs in the next section.
- Section 5 contains the details of the experimental setup and experimental results. It contains general findings and also dives into model-specific results.
- We conclude the thesis with a statement on the limitations and ethical considerations of our work.
- All other results not included in the 5th section can be found in the Appendix A and Appendix A.

2 LITERATURE REVIEW

This section includes review of relevant literature on language modeling, multilingual and low-resource LLMs, and evaluation of LLMs.

2.1 Language modeling

Human languages are known to be redundant and full of patterns. Researchers have attempted to model these patterns in various ways. Earliest methods relied on symbolic AI. An early example of this approach was ELIZA [60]. ELIZA was used to simulate a psychotherapist. Another famous example was SHRDLU [61]. This program, too, relied on rules written by the experts to understand and interact with human languages. Since these methods could not use existing text corpora to "learn", they failed to scale up. As a result, they gradually gave way to implicit language modeling. However, rule-based systems remain an integral part of even the most modern NLP systems. For example, Rasa [6], one of the most sophisticated and modern chatbot development systems, still allows use of pattern matching in its pipelines.

The earliest implicit language learning was purely statistical. The most famous example would be N-gram models [51]. N-gram models rely on the assumption that the probability of the next word relies on a fixed number of previous words. For example, bigrams use the previous two words to predict the next word, trigrams use three words, and so on. Bag-of-words was the simplest

statistical approach to representing a text. A relatively more advanced method was tf-idf [21]. These models were easy to implement, but they lacked advanced representations of the language.

By the 21st century, these models were surpassed by neural language models [5]. Unlike statistical models, neural models were capable of modeling more complex relationships between words. This gives neural models more contextual awareness, making their outputs closer to the actual probability distributions. The main challenge in modeling language was learning the relationship between words. Use of recurrent neural networks (RNNs) made this possible. RNNs were introduced by [49] in 1985, but they were not widely adopted due to computational limitations. Once they were adopted, several issues emerged. Due to gradient vanishing problem, RNNs were less effective when used over long context. Long-short term memory model was developed to solve this problem [22]. RNNs and their derivatives remained the dominant architecture for NLP problems until introduction of attention mechanism [3] and transformer models [56] that relied on it.

Traditionally, language models were trained on a task-specific dataset, with that particular task in mind. Advances in available text data, computational power, and introduction of attention mechanism which was easier to parallelize opened the way for the first foundation models – models that had general-purpose understanding of the language, and could be fine-tuned or prompted to solve various tasks. The first such model was BERT by Google [13]. This was followed by larger and more capable language models in the following years. BERT understood English only, but its multilingual successors were quickly developed. BERT was not intended for language generation, but GPT-2 [44] and GPT-3 models showed that scaling language modeling enabled human-level language generation as well.

2.2 Multilingual and low-source language modeling

First, NLP solutions were developed for English, similar demand existed in other languages, too. Multilingual NLP became an active area of research after modern tokenizers enabled the development of such models. XLM-RoBERTa was one of the earliest attempts at building a multilingual foundation model [10]. mT5 [62] was an early multilingual generative language model. In the last years, almost all state-of-the-art models have some level of multilingual capabilities. Even models that were trained on monolingual data have been shown to be useful foundation models for other languages [27].

Multilinguality can be achieved in two general ways: (1) pretraining on a multilingual dataset or (2) fine-tuning a pre-trained model on a multilingual labeled dataset. The latter can be achieved in multiple ways, including LORA [24], QLORA [12], RLHF [39], or DPO [45]. The last two are considered reinforcement learning methods. LORA (Low-Rank Adaptation) is an efficient alternative to traditional fine-tuning methods. Instead of training the existing weights of a Transformer block, this method creates an additional rank decomposition matrix for each block. These new

smaller matrices are trained and used during inference alongside the original weight matrices. This method decreases the number of fine-tuned parameters while preserving the performance of full fine-tuning. QLORA (Quantized Low-Rank Adaptation) is a more efficient alternative to LORA. It performs backpropogation on 4-bit weights of the original frozen model. This greatly reduces the memory requirements for fine-tuning an LLM.

Apart from multilingual models, some low-resource language have their dedicated models. BERT architecture has been replicated for various languages, including Chinese and Azerbaijani [27]. Another approach to developing models for low-resource languages is to fine-tune an existing multilingual model or continue the pre-training stage.

The lack of proper benchmarks is a common barrier to the development of LLMs for low-resource languages. This issue is gradually being resolved as new benchmarks are introduced. In the case of Turkic languages, the similarity of these languages has inspired some researchers to create common benchmarks. Turkic Interlingua is an example of this for the translation task [36]. TurkishMMLU was introduced recently as the first MMLU benchmark for any Turkic language.

2.3 Modeling Turkic languages

Among Turkic languages, **Turkish** has the most established NLP traditions. This is reflected in the fact that Turkish has various treebanks, large-scale text corpora, morphological parsers, and other fundamental NLP tools. While not all of these are prerequisites for developing an LLM or evaluating it, they show the level of maturity of the Turkish ecosystem. There have been attempts to train Turkish LLMs from scratch on Turkish-only text corpus [37], or adapt a large multilingual model to Turkish [1]. An MMLU alternative for Turkish was announced recently [65]. **Azerbaijani** has an active NLP community as well. In recent years, there have been attempts to build large-scale text corpus and a standalone BERT model [27]. This is also the first work to attempt to benchmark NLU capabilities of language models for Azerbaijani. Another work has built a task-specific LLM [68]. However, there are no established benchmarks for Azerbaijani.

A pilot evaluation of several LLMs on the **Kazakh** language was performed [34]. A BERT clone was trained for **Uzbek** [30]. However, just like Azerbaijani, these languages lack standardized benchmarks like TurkishMMLU. There is also a lack of standardization between Turkic languages, which may be desirable due to the linguistic and cultural proximity of the communities. We are not aware of any foundation models trained for Kyrgyz, but there are some works that fine-tune existing foundation models on a Kyrgyz dataset.

While Turkish, Azerbaijani, Uzbek, Kazakh, and Kyrgyz have reached a certain level, other Turkic languages are lagging. Turkmen, Uyghur, Tatar and other languages Turkic languages have significant native speakers, but they have very limited NLP resources.

2.4 Evaluation of Large Language Models

Language modeling was pushed forward to solve certain NLP tasks. These tasks can be classified into the following categories [9]:

- Natural language understanding: sentiment analysis, text classification, natural language inference (NLI), etc. This category contains traditional NLP benchmarks like SentiBench [47], IMDB reviews dataset, or Stanford Natural Language Inference Corpus [7].
- Reasoning. Reasoning capabilities of large language models remain an object of active debates. [43] demonstrates ChatGPT's reasoning abilities on various NLP tasks.
- Natural language generation tasks like summarization [66] and translation [67].
- Multilinguality [23].
- Factuality [38].

Our work concentrates on an NLU task – multiple-choice question-answering. There are several such benchmarks, including MedMCQA [40] for medical question-answering, BIG-bench [14] and MMLU [19] for multitask evaluation. Since MMLU is partially saturated, its more robust alternatives have been developed [58].

Just like we need multilingual models, we also need multilingual benchmarks. While it is easy to build similar benchmarks in high-resource languages like Chinese and French. Mid- and low-resource languages are lagging in this area, too. Since most languages do not have the resources required to build a native benchmark, alternative approaches have been developed. Creating multilingual benchmarks by machine translating English benchmarks is a common approach, but it has been shown to be erroneous [55]. There are more advanced works that try to solve this issue. For example, MURI [31] generates high-quality instruction-tuning datasets that preserve cultural context.

Recently, there have been attempts to create massively multilingual MMLU datasets. Global MMLU [50] has translated an existing dataset into 42 languages. They have achieved this with a combination of professional translation, community translation, and in cases where neither was available, machine translation. Using the same dataset has its advantages – it is easier to compare the performance of LLMs across languages. When using different dataset for each languages, it is harder to make a comparison because difficulty level of the questions may be different. On the other hand, a dataset created in English carries with itself certain cultural and linguistic biases. It would be easier for a Chinese language model to understand them. This issue has been explored in depth in the INCLUDE paper [48]. Instead of translating a single dataset, they have compiled a multilingual dataset with community support across 44 languages.

3 BACKGROUND

3.1 Language modeling

The term "language modeling" refers to an arbitrary function that predicts the next element (token) in text based on previous elements. Formally speaking, a language model is represented as a conditional probability function

$$P(x_i|x_1, x_2, ..., x_{i-1}) \tag{1}$$

where $x_{i-1}, x_{i-2}, ..., x_1$ is the input sequence of tokens and x_i is the next predicted token. We would need to predict more than a single token to produce text, in which case it would look like

$$P(x_1|x_0) * P(x_2|x_0, x_1) * ... * P(x_i|x_0, x_1, ..., x_{i-1}) \tag{2}$$

where x_0 represents the initial input sequence.

Theoretically speaking, a token can be anything from a bit to a sentence. In reality, however, it tends to be between a character and a word, although this is not strictly true.

Early language models were purely statistical. The simplest such model is called n-gram. This model uses n previous tokens to predict the next one. Most commonly used versions were bigrams and trigrams. The larger the n, the larger corpus is required for proper training.

Here's a concise mathematical explanation in LaTeX of how an n-gram model is trained:

Given a sequence of tokens $t_1, t_2, \ldots, t_m$

$$P(t_i \mid t_{i-(n-1)}, \ldots, t_{i-1}) = \frac{C(t_{i-(n-1)}, \ldots, t_{i-1}, t_i)}{C(t_{i-(n-1)}, \ldots, t_{i-1})}$$

In the case of bigrams, this formula would be:

$$P(t_i \mid t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})}$$

Probability function P is later used as a language model. As you can see, this is a simple statistical model that has very fast training and inference steps. This makes n-grams useful even now, despite all the advances in neural language models. N-grams are also used in combination with more advanced algorithms to solve problems such as autocomplete and spelling correction.

3.2 Large Language Models

In the late 2010s, LLMs emerged and quickly dominated the NLP field. Three factors made this possible. First was the introduction of transformer architecture. Unlike its predecessor LSTM,

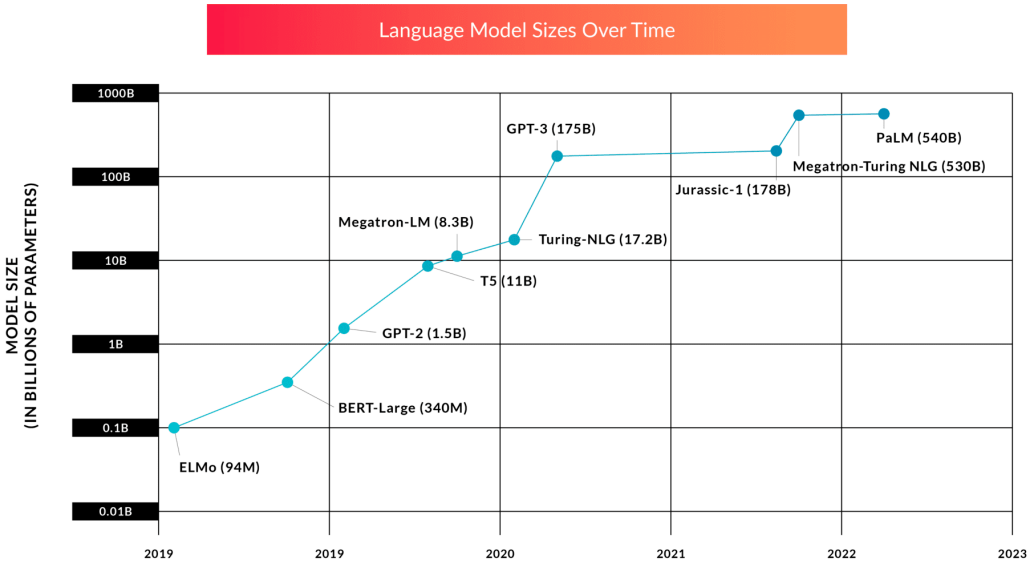

Figure 1: Increase in the number of parameters across years. Image credit: Two Sigma Ventures

transformer models do not treat the input sequence, well, sequentially. This enabled the parallelization of the training process. Second was the abundance of text data available on the web. This was especially the case for English. Third was the continued improvements we achieved in parallel computing. While GPUs were traditionally used for the rendering of 3D graphics, it was also adopted in other industries, including machine learning research. As a result of these three conditions, the NLP community was able to move from task-specific models to **foundation models** – models that were trained in a self-supservised manner. That is, instead of using predefined labels for training, the data itself was used to generate labels. Also, this is not the only model type that was trained, but it is the most successful type and it is the model that we will concentrate on throughout this work. These models were large – hundreds of millions of parameters at first and hundreds of billions of parameters now. Although there was a trend of exponential increase in the number of LLM parameters (See Figure 1, these large and dense models are gradually giving way to two different lines of new models. On the one hand, we are seeing more compact, dense models; on the other hand, we are seeing sparse models with more and more parameters, only a fraction of which are being used during the inference process. These models are called mixture-of-experts.

Since the beginning of this process, LLMs had both a general understanding of the language and world knowledge. In most cases, fine-tuning or prompting these models to perform specific tasks proved more successful than training task-specific models.

An LLM is trained on an unlabeled text corpus. This results in a language model that has learned the language implicitly. Given an input text, it will automatically complete it in the most statistically likely manner. However, such a model has limited practical use. In order to make these base models, as they are called, into chatbots, we need to fine-tune them. Base LLMs fine-tuned to follow instructions are called instruct models. For example, google/gemma-2-9b[1] is a base model, and google/gemma-2-9b-it [2] is an instruction model fine-tuned from google/gemma-2-9b.

Initial works were performed in English using datasets like BookCorpus and WebText (or OpenWebText, its open-source alternative). These works outperformed all existing models that were trained directly on a labeled dataset. The release of GPT and BERT models with permissive licenses promoted the scientific work in this area, resulting in more effective and more efficient alternatives. These advances were made by (1) training on larger and better datasets, (2) an improved training process, (3) larger models, and (4) improved model architecture. The international research community was quick to replicate these results in other languages, too. Both monolingual and multilingual alternatives emerged with varying levels of success. The level of success depended on several factors, including the amount of available text data and existing NLP infrastructure. Both of these varied widely across the world.

[1] `https://huggingface.co/google/gemma-2-9b`
[2] `https://huggingface.co/google/gemma-2-9b-it`

3.3 Prompting LLMs

LLMs are autoregressive models, i.e., they use tokens $x_1, x_2, ..., x_n$ to predict x_{n+1}. Giving a certain input $x_1...x_n$ to get a certain output $x_{n+1}...xn + m$ is called prompting.

In the following example, Baku is predicted by the LLM based on the provided input because it has the lowest perplexity among options (i.e., vocabulary of the tokenizer):

> The capital of Azerbaijan is Baku

The desired output is not necessarily a single sequence. For example, if the prompt is a question, LLM can answer it correctly in multiple ways. For example, both of these are correct:

> What is the capital of Azerbaijan? The capital of Azerbaijan is Baku.

or

> What is the capital of Azerbaijan? Baku.

These are simple questions, but the latest LLMs are capable of more than answering geography trivia. In our benchmark, we expect them to answer questions on high-school mathematics and physics. In such complex cases, it is important to keep the output predictable so that we can automatically calculate the performance of LLM on multiple-choice questions. Our goal is to extract the choice made by LLM. This is not an easy task. For example, an LLM could answer a maths question in the following ways.

> ..., therefore the answer is A.

> ..., the answer B is not correct ... the correct answer is A.

> ...Answer: A.

> ...answer is the fourth option.

...answer is 3/5.

As you can see, LLM does not even necessarily output option A, B, C, or D. It can answer the question but not mention which option it was. In order to solve this problem, we employ a method called few-shot learning [8]. Here, we provide multiple samples with the same structure, followed by the actual question. This ensures that LLM is going to answer in a predictable format, making automatic extraction possible.

Question: <Sample question 1>
Answer: A
Question: <Sample question 2>
Answer: B
Question: <Sample question 3>
Answer: B
Question: <Sample question 4>
Answer: A
Question: <Sample question 5>
Answer: C
Question: <Question>
Answer: B.

The actual 5-shot prompt used in our experiments can be found in section 5.1.

Chain-of-thought is a more advanced prompting technique to elicit reasoning in LLMs [59]. First introduced in 2022 by Google Research Brain Team, Chain-of-Thought forces LLMs to start not with the answer but with the explanation, improving the performance. Zero-shot Chain-of-Thought is achieved by simply appending "Let's think step-by-step" to the question.

I am 25 years old. My brother is twice as old as me. How old is he going to be when I am 50? Let's think step-by-step. Your brother is twice as old as you, so he is 25 * 2 = 50 years old. The difference in your ages now is 50 - 25 = 25 years. Your brother will still be 25 years older than you, so when you are 50, your brother will be 50 + 25 = 75 years old.

Few-shot Chain-of-Thought is a combination of few-shot and Chain-of-Thought prompts.

Question: <Sample question 1>
Explanation: <Sample explanation>
Question: <Sample question 2>
Explanation: <Sample explanation>
Question: <Sample question 3>
Explanation: <Sample explanation>
Question: <Sample question 4>
Explanation: <Sample explanation>
Question: <Sample question 5>
Explanation: <Sample explanation>
Question: <Question>
Explanation: Let's think step-by-step. Your brother...

5-shot Chain-of-Thought is a powerful method, but it creates additional problems. Traditional 5-shot prompts create predictable outputs, but 5-shot Chain-of-Thought prompts are less manageable.

We have used 5-shot and 5-shot Chain-of-Thought prompting methods in our experiments. Actual prompts can be found in section 5.1.

3.4 Evaluating LLMs

LLMs are known to be capable in both NLU and NLG tasks – people use them daily as summarizers, information retrievers, email writers, etc. But long before these became possible, researchers wanted to quantify the capabilities of language models and created challenges. Most of the widely used benchmarks are still classification-based. This is because evaluating text generation is more challenging. As a result, generative problems are formulated as classification problems and solved as such. A famous example would be an extractive question-answering dataset called SQuAD [46], where question-answering – a traditionally generative task – was converted into a classification problem. But even if we convert the problem, state of the art LLMs remain generative models. This means that we also need to map the output of the LLM to a predefined class, which is error-prone. This step can be represented as follows:

$$f_{\text{extract}} : y_{\text{gen}} \mapsto c_{\text{pred}}, \quad c_{\text{pred}} \in C \tag{3}$$

Where y_{gen} is the generated output and c_{pred} is the final prediction extracted from this text. Using this formula, we can represent the evaluation of LLMs on a classification task as follows:

$$\text{score} = s\big(f_{\text{extract}}(P(x, p_{\text{sys}}; \theta, \phi)), c_{\text{true}}\big) \tag{4}$$

where c_{true} is the true label, x is the input question, p_{sys} is the system prompt, θ is the model weights, and ϕ is the inference hyperparameters, such as temperature and top K.

In fact, LLMs emerged as a result of attempts to beat these challenges. Earlier challenges were mostly classification-based, and they have already been saturated. LLMs made most of these benchmarks obsolete, so more advanced benchmarks like MMLU were introduced to replace them. In the last 4 years, MMLU has also been mostly solved (see Figure 2), and researchers are now working to beat even more challenging benchmarks like MMLU-Pro. The problem is that while English models have surpassed the levels at which MMLU was a valuable benchmark, we cannot say the same about most other languages. Especially so for the Turkic language family.

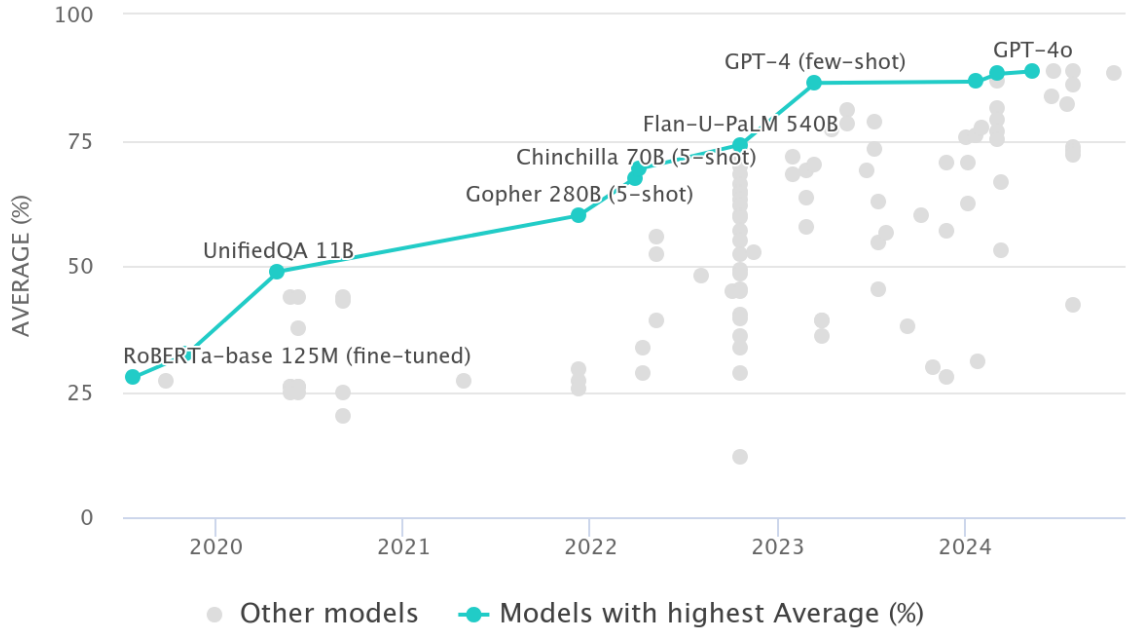

Figure 2: MMLU Leaderboard, 2019-2024.

LLM benchmarks have been more and more challenging. This has happened in two directions. The first direction is specialization. We have more specialized and challenging benchmarks like AIME for Mathematics [41], LEGALBENCH for Law [18], and MedQA for Medicine.

The second direction is multilingualism. While SOTA LLMs have solved most of the benchmarks in English, similarly challenging benchmarks in other languages remain unsolved or partially solved. Multilingual LLM research has skyrocketed in recent years, increasing the number of multilingual LLMs and multilingual benchmarks.

3.5 Multilingual benchmarks

Evaluating LLMs showed strong performance in English. This led researchers to assess models in other languages. As LLMs became more multilingual, the need for evaluation tools in various languages grew [32, 50].

However, creating good multilingual benchmarks presented challenges. A common method involves translating existing English benchmarks, like MMLU, into other languages [50]. This approach seems quick but has significant drawbacks. Translated text often suffers from "translationese" [55]. This means the language might be grammatically correct but sound unnatural or awkward to native speakers. It carries traces of the original language's structure and style.

Translations can also miss important cultural details and language nuances [48]. Benchmarks created in English reflect the cultural context of English-speaking regions. Directly translating these benchmarks imposes this viewpoint onto other languages and cultures. This creates a biased test that may not accurately measure an LLM's understanding of a different language's native context.

Another approach uses synthetic generation, where models create new benchmark questions. This method can also produce incorrect or unnatural text. Some advanced methods have been suggested to incorporate translation and synthetic generation as steps of a more advanced method. MURI is a prime example of this approach [31]. Nevertheless, generating high-quality, culturally relevant questions automatically remains difficult.

These problems highlight the need for benchmarks built from original sources in the target language. Native benchmarks avoid translation issues and better reflect the actual language use and cultural context. Creating these native benchmarks is more difficult but provides a more accurate and fair evaluation of LLM abilities in diverse languages. This need is clear for language groups like the Turkic languages, which have unique linguistic features and varied resource availability.

3.6 Turkic Languages

The Turkic language family includes languages spoken across a wide area of Eurasia (see Figure 4). The development and use of LLMs for these languages show significant variation. Some languages benefit from established research communities and available resources, while others have fewer tools and datasets. Figure 3 shows the amount of available text data in major Turkic languages and

Figure 3: Amount of available text data in each language, according to FineWeb 2 [42]

.

compares it against other world languages. While not a comprehensive comparison, it highlights the resource gap between Turkic languages and high-resource languages. While the lack of readily available natural text cannot be solved in a short period of time, other problematic situations, such as the lack of benchmarks, can be improved upon. This work focuses on improving evaluation resources for several Turkic languages.

Turkish: Turkish has the most developed NLP resources among Turkic languages. It possesses various text corpora, treebanks [53], and morphological analyzers [64]. Researchers have trained LLMs specifically for Turkish [54] and adapted existing multilingual models [1]. A dedicated MMLU-style benchmark, TurkishMMLU [65], exists to evaluate LLM performance in Turkish. The TUMLU benchmark incorporates data from this existing Turkish resource [26].

Azerbaijani: The Azerbaijani NLP community is active. Efforts include building text corpora and training specific models like BERT [27]. Some work exists on task-specific LLMs [68] and initial NLU benchmarking [27]. However, before TUMLU, there were no established, broad-coverage benchmarks comparable to MMLU for Azerbaijani.

Kazakh: Research includes pilot evaluations of existing LLMs on Kazakh tasks [35]. Like Azerbaijani, standardized, comprehensive benchmarks were lacking before this project. According to FineWeb 2, one of the largest multilingual text corpora, most online Kazakh text uses the Cyrillic script [42]. TUMLU provides the first large-scale NLU benchmark for Kazakh.

Uzbek: A BERT language model specifically for Uzbek exists [30]. Resources and standardized benchmarks remain limited compared to higher-resource languages. Both Latin and Cyrillic scripts are in common use [42]. TUMLU introduces a significant benchmark dataset for Uzbek.

Kyrgyz: Kyrgyz datasets are available as a part of larger multilingual benchmarks, such as Global MMLU [50], but this dataset is translated and we are not aware of any native MMLU-style benchmarks in Kyrgyz.

Tatar, Crimean Tatar, Uyghur, and Karakalpak: These languages generally have fewer available NLP resources and less LLM development compared to the languages above. Tatar primarily uses the Cyrillic script. Crimean Tatar and Karakalpak use both Latin and Cyrillic scripts. Uyghur primarily uses an Arabic-derived script. For these languages, TUMLU represents the first major effort to create a substantial NLU benchmark dataset, enabling evaluation of current LLMs and supporting future model development.

The differing levels of resources and benchmark availability highlight the need for unified evaluation tools like TUMLU to better understand LLM capabilities across the Turkic language family.

3.7 Problem Statement

Our main goal is to create a comprehensive benchmark for understanding Turkic languages. This benchmark, called TUMLU, draws inspiration from existing benchmarks like MMLU [19] and Turk-

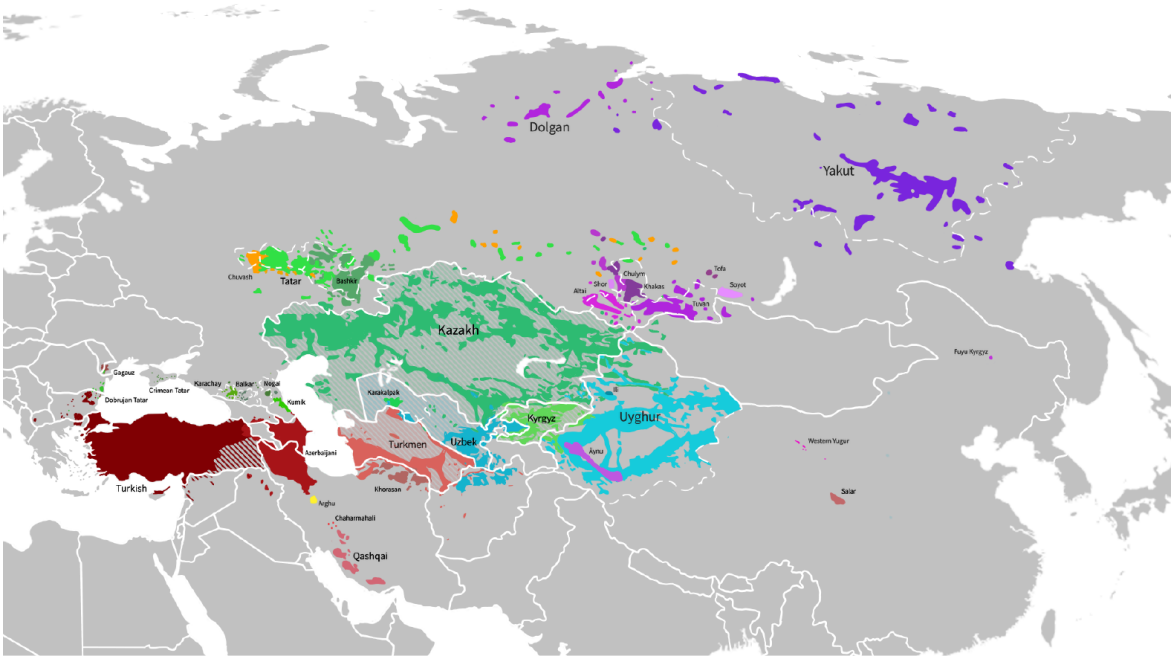

Figure 4: Distribution of Turkic languages.

ishMMLU [65]. However, TUMLU focuses specifically on the Turkic language family. We aim to provide a unified tool for evaluating large language models (LLMs) in these languages.

The benchmark consists of multiple-choice questions covering subjects typically taught in middle and high school. It includes eight Turkic languages: Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Tatar, Turkish, Uyghur, and Uzbek. For Turkish, we incorporate data from the existing TurkishMMLU project [65]. For the other languages, we gathered questions from native sources. We focused on subjects like History, Geography, Native Language & Literature, Biology, Physics, Chemistry, and Mathematics. We tried to include at least 100 questions per subject for each language, though this was not always possible. We also added other subjects where sufficient data existed.

A key part of this work involves creating high-quality evaluation data. We decided against using machine translation or synthetic text generation. Instead, we collected questions directly from native language materials. This approach requires more effort but ensures the benchmark accurately reflects real language use and cultural context. It helps avoid the pitfalls of translationese and cultural bias often found in translated datasets [65, 48]. Native speakers verified samples of the

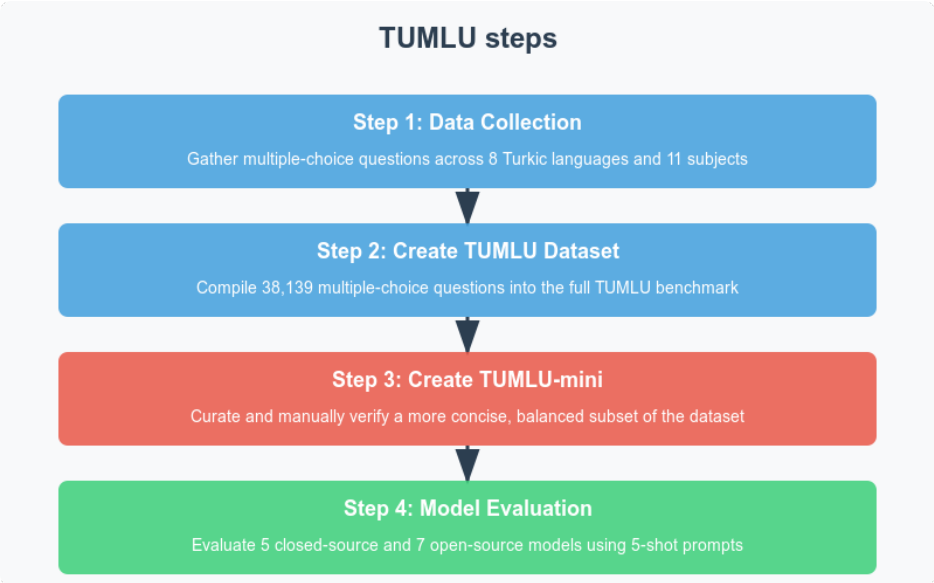

Figure 5: Main steps of the project.

collected data. We also developed native-language chain-of-thought prompts for several languages to test different evaluation methods [59]. We intend to use TUMLU to assess the performance of current leading LLMs, both open-source and proprietary models. The results will help establish a leaderboard, offering guidance for projects involving Turkic languages. To support further research, we will publicly release the TUMLU dataset and the code used for evaluation. This work aims to fill a gap in evaluation resources and promote better language model development for the Turkic language family.

4 DATASET

This section outlines our approach to data collection and describes the resulting TUMLU benchmark designed for LLM evaluation in Turkic languages. The TUMLU (Turkic Unified Multilingual Language Understanding) benchmark is a multilingual and multitask dataset. It contains 38,139 multiple-choice questions covering 11 subjects across 8 languages. The questions target middle or high school level knowledge. Many questions are sample or official items from university entrance exams in the respective countries. A few sample questions from Tatar subset of the dataset can be found in Figure 8

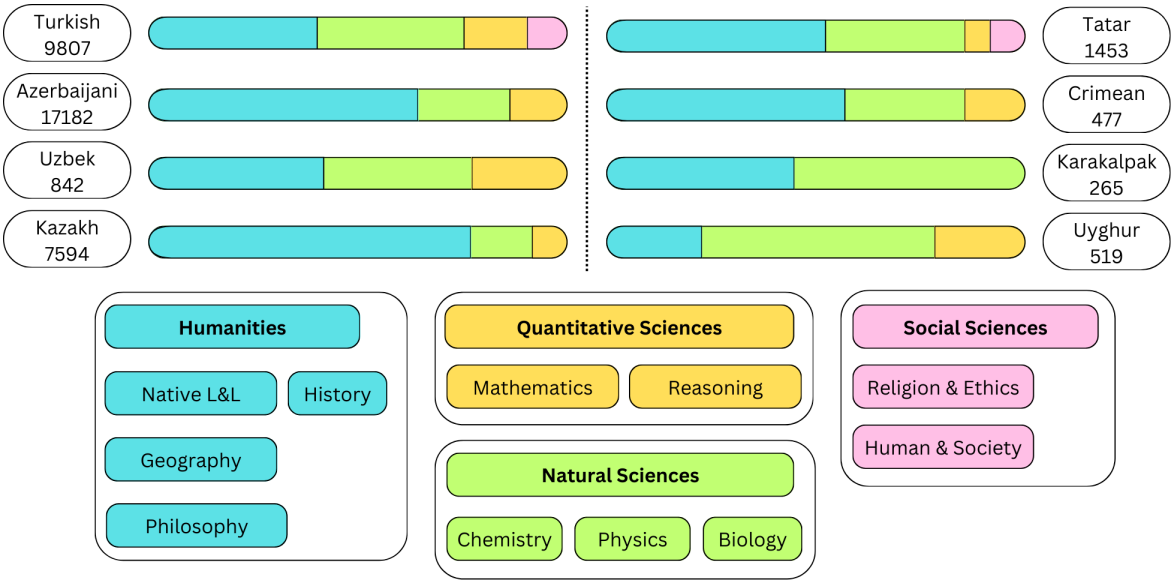

Figure 6: Distribution of subjects across languages in TUMLU. Numbers next to language names indicate the total question count. Left: middle- and high-resource languages; Right: low-resource languages.

4.1 Dataset Creation

Our goal was to create MMLU-like datasets [19] for several Turkic languages using native sources. We aimed to avoid issues like translationese [55] and cultural misalignment [48] common in translated benchmarks. The process varied for each language due to unique circumstances and resource availability.

Source Collection: We collected data from publicly available books and websites. For Turkish, we incorporated the existing TurkishMMLU dataset [65]. For Azerbaijani, we initially used a dataset collected by scraping several public websites. Closer inspection revealed issues, such as incorrect answer sheets in biology questions. We curated this data to form the final Azerbaijani subset. For Kazakh and Uzbek, we collaborated with native speakers to identify useful websites containing public multiple-choice questions, mostly samples for the Unified National Test (Kazakh) or Entrance Exam (Uzbek). Kazakh math questions came from 7th-grade materials. Data for Tatar, Crimean Tatar, Uyghur, and Karakalpak were gathered from similar native sources.

Question Formatting: Original questions had between 2 and 5 answer choices. If a question had more than 4 choices, we removed one incorrect option. If it had fewer than 4, we kept the question as

Figure 7: A sample question from the parallel Uzbek dataset, available in both Cyrillic and Latin alphabets. This enables a comparison of LLM performance across different scripts. English translation is provided for clarity.

is. Most questions in the final dataset have 4 choices, except for Language and Literature questions in Crimean Tatar.

Data Verification: Native speakers for each language manually verified the quality and correctness of a random sample from each subject. In languages like Azerbaijani, where initial data came from community collections, this verification step was crucial. We found that around 10% of the initially collected Azerbaijani questions were invalid or had incorrect answers, which we then corrected or removed. Native speakers confirmed the integrity of the final questions used in the benchmark.

question string · lengths	answer string · classes	subject string · classes	choices sequence · lengths
10⌓84 57.6%	4 values	8 values	4 4
Түбәндә китерелгән кайсы үлчәм көч берәмлегенә туры килә?	D	Physics	["", "$\frac{кг * м}{с^3}$", "$\frac{кг}{с}$", "$\frac{кг * м}{с^2}$"]
Электромагнитик индукция күренешен кем ачкан?	C	Physics	["Эрстед", "Ленц", "Фарадей", "Герц"]
4 Н көч пружинаны 0,02 м га озынайта. Пружинаның катылык коэффициентын табарга.	D	Physics	["2 $\frac{Н}{м}$", "0,5 $\frac{Н}{м}$", "0,02 $\frac{Н}{м}$", "200 $\frac{Н}{м}$"]
l = 40 м озынлыктагы математик маятникның периодын табарга.	A	Physics	["12,5 с", "$\frac{1}{12}$ с", "2 с", "$\frac{1}{2}$ с"]

Figure 8: A sample of TUMLU dataset. Language: Tatar. Subject: Physics.

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.

4.2 Dataset Composition

The final TUMLU benchmark includes Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Tatar, Turkish, Uyghur, and Uzbek.

Content: It covers eleven subjects: Maths, Physics, Chemistry, Biology, Geography, Native Language & Literature (NL&L), History, Logic, Human & Society, Philosophy, and Religion & Ethics. Figure 1 shows the distribution of questions across languages and subjects. Table 1 shows the average length (in characters) of questions and answer choices per language, indicating variability.

Scripts: The benchmark represents linguistic diversity by including questions in Latin, Cyrillic, and Arabic scripts, reflecting the writing systems used for these languages (See Figure 7 for an Uzbek example). We also created transliterated versions to enable comparisons of LLM performance across different scripts using the same content.

Subjects Excluded from Experiments: Some subjects (Logic, Human & Society, Philosophy, Religion & Ethics) were available only in one or two languages. While included in the full dataset release, they were not used in the experiments reported later in this paper due to limited comparability.

4.3 Considerations for Use

While TUMLU supports monolingual evaluation within each language, comparing model performance across languages requires care.

Difficulty Levels: The difficulty of questions within the same subject can vary between languages. Uzbek and Turkish questions, often based on specific university entrance exams, may be harder. Azerbaijani and Kazakh questions came from community efforts without strict difficulty oversight. For instance, Kazakh math questions cover only middle-school topics, making them easier than those in other languages. It is practically impossible to ensure same difficulty levels, because different countries have different high-school cuuricula. This would only be possible via translation, but translated datasets are known to contain cultural and regional biases [48], as well as a phenomenon called translationese [55]. These differences in difficulty levels make direct cross-lingual performance comparison challenging, though the benchmark remains valuable for comparing different models within a single language.

Value of Native Data: Creating benchmarks from native sources, despite being more effort, provides a more accurate measure of LLM understanding in a specific cultural and linguistic context compared to direct translation or synthetic generation [65, 48]. It is possible that such native benchmarks will reveal insights regarding model performance that would otherwise be missed by synthetic or translated benchmarks.

4.4 TUMLU-mini

To facilitate more balanced and efficient experiments, we created TUMLU-mini.

Creation: This subset consists of 100 randomly selected and manually verified questions per subject per language. If a subject had fewer than 100 questions in a language, we included all available questions. Answer choices were shuffled to reduce simple memorization effects. We removed subjects available in fewer than 3 languages. Table 2 provides the exact composition.

Answers in TUMLU-mini were not distributed equally. We would expect 25 % of each option, but as you can see in Figure 9, this is not the case. Therefore, we shuffled the answers before using them in the benchmark. The publicly available dataset is not shuffled.

All experiments reported in Section 5 were run using the TUMLU-mini subset.

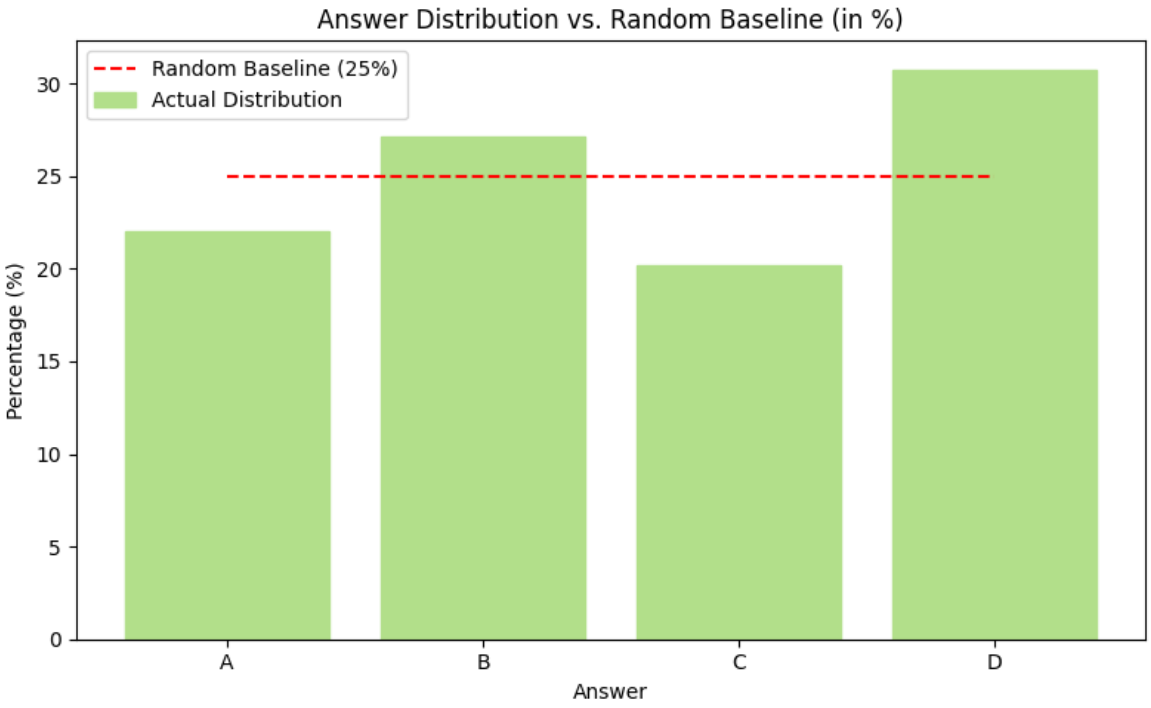

Figure 9: Distribution of correct answers before shuffling.

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 2: 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 Chain-of-Thought prompts. Language codes are from ISO 639-3.

4.5 Evaluation Approach

We used the TUMLU-mini dataset to assess the language understanding capabilities of current LLMs.

Models: We evaluated both closed-source/proprietary models (like GPT-4o, Gemini 1.5 Pro, Gemini 1.5 Flash, Claude 3.5 Sonnet, and Claude 3.5 Haiku) and open-source models (like Qwen2.5 72B, Qwen2.5 7B, Llama- 3.1 405B, Llama 3.3 70B, Llama 3.1 70B, Llama 3.1 8B, Gemma 2 27B, and Gemma 2 9B). Section 5 provides details on the specific models.

Methods: We ran experiments in two main settings:

5-shot: Providing 5 example question-answer pairs from the same subject before the target question [8]. This method pushes the language model to follow a certain pattern, making automatic evaluation easier.

5-shot Chain-of-Thought: Providing 5 examples with step-by-step reasoning leading to the answer before the target question [59]. While this method improves the performance, it also complicates the answer detection process because the final answer can be formatted in different ways. While we try to force a certain template by providing 5 examples, chain-of-thought reasoning usually ends with more varying outputs than 5-shot.

Chain-of-Thought Prompts: We created 5 native-language Chain-of-Thought prompts per subject for Azerbaijani, Kazakh, Turkish, and Uzbek. Native speakers created the prompts for Azerbaijani, Kazakh, and Uzbek. Turkish prompts were adapted from the TurkishMMLU project [65] (changing from 5 choices to 4). We did not create Chain-of-Thought prompts for other languages due to a lack of native speakers available for validation. All prompts were provided in the respec-

tive native languages, as prior work suggests English prompts do not improve performance [48]. Sample prompts are in the next section.

5 EXPERIMENTS

5.1 Experimental set-up

Data Previous work [65] has shown that 100 questions per subject are enough to estimate the performance of a larger dataset. Therefore, we ran all experiments on TUMLU-mini, a subset of the TUMLU dataset containing no more than 100 questions per subject in each language. While we have performed the experiments and publicly released the data, results on the following subjects are not reported in this report: 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 [17], Gemma 2 [16], Qwen2.5 [63] and proprietary models, such as Gemini 1.5 [15], Claude 3.5 [2], GPT-4o [25]. 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 [63], which allows us to observe the effect of scaling [20, 28] 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 A and A.

These experiments started in December 2024, and models were selected at this time. There are several notable omissions that we believe should be included here:

- We did not include reasoning models, such as DeepSeek-R1 [11] by DeepSeek or o1 model by OpenAI. However, it is possible that some of the closed-source models included in our benchmark also use similar reasoning techniques without mentioning it.
- Our benchmark does not include some important open-source models, such as Mistral-Large-Instruct-2411 by Mistral or Llama-3.1-Tulu-3-70B by Allen AI.
- We also did not include some major proprietary models, such as Grok models.

The reason for these omissions varies from case to case, but it was either due to resource limitations or limited multilingual support offered by these models.

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 [8], and 5-shot Chain-of-Thought, where we provide 5 example questions and explanations of their answers before asking the question [59]. Below is a fewshot sample for Biology questions in Azerbaijani:

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:

And here is the Chain-of-Thought sample for the same subset:

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

Həll: Addım-addım düşünək.Difteriyaya tutulan 10 nəfərdən 3-4 nəfər sağ qalır. Deməli 10 nəfərdən 6-7 nəfər ölər. Faizlə ifadə etsək, xəstələrin 60-70 faizi ölər. Deməli, 50 nəfərdən 30-35 nəfər ölər. Düzgün cavab D variantıdır.

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

Həll: Addım-addım düşünək. Mitoz bölünmə nəticəsində bir hüceyrədən 2 hüceyrə yaranır. Bu hüceyrələrin hər birinin xromosom sayı ilk hüceyrə ilə eyni olur. İlk hücyerədə 24 xromosom var. Deməli, əmələ gələn hüceyrələrin hər birində 24 xromosom olacaq. Düzgün cavab B 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:

Previous work has demonstrated that [48] providing the prompt in English does not result in performance gains. Due to this, we provide all prompts in their respective native languages. These prompts were created manually without the use of LLMs. All of them were created by native speakers and validated by at least one additional person.

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. It is important to note that, in the case of proprietary models like GPT-4o, even setting the temperature parameter to 0 does not ensure deterministic output. Nevertheless, we attempted to make our results as replicable as possible.

Specifically, we used the following APIs:

- OpenAI API: GPT-4o
- Antrhopic API: Claude 3.5 Sonnet, Claude 3.5 Haiku
- Google Cloud Gemini API: Gemini 1.5 Pro, Gemini 1.5 Flash
- Together AI API: Gemma 2 27B, Gemma 2 9B

Model	Mean	aze	crh	kaa	kaz	tat	tur	uig	uzb
Claude 3.5 Sonnet	**79.3**	**84.4**	**81.2**	**75.3**	**83.0**	**84.0**	**85.7**	**71.3**	**69.1**
GPT-4o	75.1	82.4	70.5	70.8	81.0	80.5	83.7	66.5	65.4
Gemini 1.5 Pro	74.0	78.6	70.3	68.2	78.4	80.5	80.0	71.0	65.1
Gemini 1.5 Flash	65.6	72.4	68.0	61.2	68.6	68.3	76.6	57.8	52.1
Claude 3.5 Haiku	63.9	70.6	62.9	55.2	69.9	67.5	78.0	56.6	50.3
Llama- 3.1 405B	62.9	65.9	69.5	60.0	69.0	70.4	59.7	58.2	50.4
Qwen2.5 72B	61.5	70.1	61.8	54.6	62.6	62.5	73.9	56.0	50.4
Llama 3.3 70B	58.4	66.0	58.7	49.2	60.0	69.5	68.4	51.6	44.1
Llama 3.1 70B	57.6	68.1	57.3	49.9	56.4	66.2	64.9	52.4	45.3
Gemma 2 27b	51.6	58.1	49.8	47.6	58.4	54.9	64.3	42.2	37.6
Gemma 2 9b	46.8	53.7	46.8	40.8	49.1	51.8	60.4	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	40.1	48.4	35.7	33.4	46.4	44.1	47.7	35.0	29.9

Table 3: 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.

- Deep Infra API: Qwen2.5 72B, Qwen2.5 7B, Llama- 3.1 405B, Llama 3.3 70B, Llama 3.1 70B, Llama 3.1 8B

We had to use two different APIs for open source models due to technical reasons, but we performed small-scale experiments to ensure that their outputs match each other.

We ran all experiments in Python 3.10.15. We used beautifulsoup4 4.13.3 to extract structured text data from HTML files. google-generativeai, openai, anthropic, and togetherai clients were used for sending LLM requests.

5.2 Main findings

In this section, we present the few-shot and Chain-of-Thought 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.

5.2.1 5-shot results

We present the average performance of all models in each language in Table 3. Claude 3.5 Sonnet outperforms other models in all languages. The top 5 spots belong to proprietary models, although

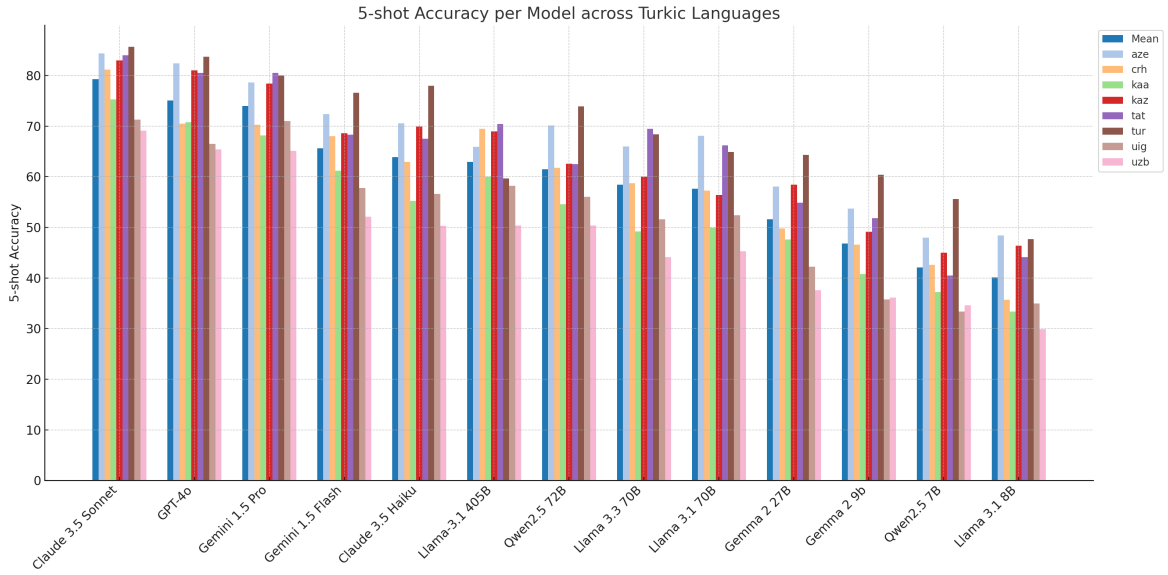

Figure 10: 5-shot performance on TUMLU-mini benchmark.

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 4. In most languages, Native Language & Literature is the most challenging subject for Claude 3.5 Sonnet. This holds for other models, as well. Subject-language performance matrices of other models are available in Appendix A. These tables contain the performance with 5-shot chain-of-thought prompts, but they do not contain performance comparisons with vanilla 5-shot prompts.

5.2.2 5-shot Chain-of-Thought results

We present the average results of the 5-shot Chain-of-Thought evaluation in Table 5. Chain-of-thought prompts have an overall positive effect on performance. Sporadic negative effects can be

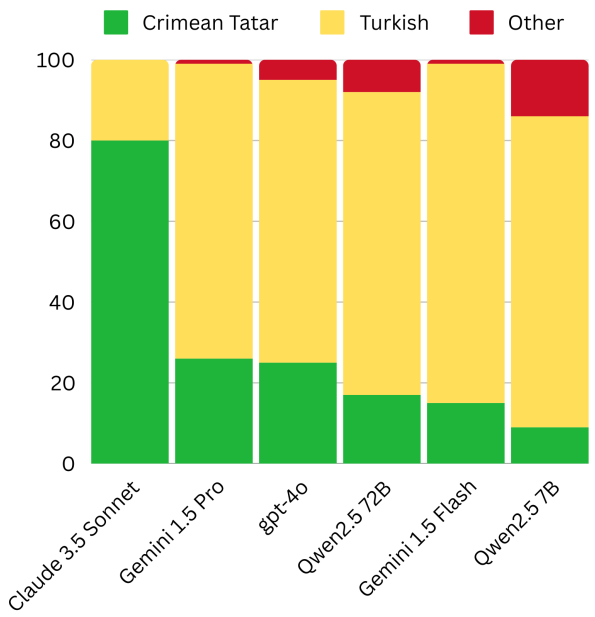

Figure 11: Language distribution of model responses to Crimean Tatar queries, as detected by Google Cloud Translation API.

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	89.9	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	88.0	94.0	92.0	78.0	79.0	85.0
Uyghur	75.0	66.0	-	-	75.8	66.0	73.5
Uzbek	71.0	73.0	64.0	70.0	70.0	55.0	81.0

Table 4: 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.

explained by incorrect output format rather than incorrect answers. We avoided manual validation of the output and instead relied on generalizable pattern-matching methods. Apart from this, chain-of-thought prompts proved to be a better approach than vanilla 5-shot prompts. The only model that experienced deteriorated performance with Llama 3.1 8B.

Table 6 shows the performance of Claude 3.5 Sonnet on each subject and language. On average, Chain-of-Thought prompts have a net positive effect in each subject and each language.

5.2.3 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 the output language. This API supports all languages in our benchmark except for Karakalpak. We present results for Crimean Tatar in Figure 11. 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 the Tatar language.

Model	Mean	aze	kaz	tat	tur	uzb
Claude 3.5 Sonnet	84.0	**87.1 (+2.7)**	**84.1 (+1.1)**	**87.9 (+3.9)**	**87.9 (+2.1)**	**72.9 (+3.7)**
GPT-4o	79.4	82.9 (+0.4)	80.7 (-0.3)	83.0 (+2.5)	84.0 (+0.3)	66.3 (+0.9)
Gemini 1.5 Pro	76.6	80.0 (+1.4)	75.1 (-3.3)	79.9 (-0.6)	81.0 (+1.0)	67.0 (+1.9)
Llama 3.1 405B	68.9	73.4 (+7.6)	68.7 (-0.3)	65.0 (-5.4)	80.7 (+21.0)	56.4 (+6.0)
Claude 3.5 Haiku	70.0	77.0 (+6.4)	74.0 (+4.1)	72.2 (+4.8)	77.6 (-0.4)	49.0 (-1.3)
Gemini 1.5 Flash	68.1	73.9 (+1.4)	69.0 (+0.4)	70.1 (+1.8)	73.6 (-3.0)	54.1 (+2.0)
Qwen2.5 72B	67.1	72.1 (+2.0)	63.9 (+1.3)	67.6 (+5.1)	78.4 (+4.6)	53.6 (+3.1)
Llama 3.3 70B	66.8	70.6 (+4.6)	69.3 (+9.3)	66.3 (-3.2)	77.4 (+9.0)	50.4 (+6.3)
Gemma 2 27B	59.4	63.0 (+4.9)	61.6 (+3.1)	61.0 (+6.1)	66.4 (+2.1)	44.9 (+7.3)
Llama 3.1 70B	56.2	59.4 (-8.7)	61.7 (+5.3)	58.6 (-7.6)	73.3 (+8.4)	27.9 (-17.4)
Gemma 2 9B	52.0	57.3 (+3.6)	52.7 (+3.6)	50.1 (-1.7)	62.3 (+1.9)	37.4 (+1.3)
Qwen2.5 7B	46.4	48.1 (+0.1)	46.4 (+1.4)	43.0 (+2.5)	56.3 (+0.7)	38.0 (+3.4)
Llama 3.1 8B	38.2	40.7 (-7.7)	38.9 (-7.6)	39.7 (-4.4)	45.1 (-2.6)	26.6 (-3.3)

Table 5: 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 6: Subject-wise 5-shot Chain-of-Thought performance of Claude 3.5 Sonnet across Turkic languages.

This experiment was inspired by the author's personal experience with early versions of ChatGPT supported by the GPT-3.5 model. While this model responded to questions in Azerbaijani with reasonable accuracy, the quality of its contents was often questionable, and the generated text was a mixture of Azerbaijani and Turkish. Of course, this is not a singular experience but a well-attested phenomenon. As far as we are concerned, this is the first work that presents an analysis of this kind.

5.2.4 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

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 7: 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.

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 7). We evaluated models in both versions and compared their performance. We present some of the results in Table 7. 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 [42], 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.

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.[3]

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.

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

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 8: Accuracy scores for Qwen2.5 72B 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 9: Accuracy scores for Qwen2.5 7B Instruct model across languages.

5.3 Model-specific results

In this part, we are going to present a breakdown of the performance of some models across languages. Tables 8 and 9 show the performance of Qwen2.5 72B Instruct and Qwen2.5 7B Instruct models, respectively. Comparing these results, we can observe that the scaling hypothesis holds not only in the aggregate but also across all subjects. In other words, larger Qwen2.5 outperforms the smaller model in every single subject.

Among the open-source models, Llama 3.1 405B has the best performance, with an average accuracy of 62.9 across all languages. We evaluate three models with 70 billion parameters – Qwen2.5 72B, Llama 3.3 70B, Llama 3.1 70B. Among these, Qwen2.5 72B has the best performance with an average of 61.5.

Gemma 27B is the only model in its size, and it predictable sits between 8B and 70B models. Among the smallest models (7B to 9B), Gemma 2 9b has the best performance with an average of

46.8.

A similar pattern holds with chain-of-thought prompting as well, but we also observe some changes. Notably, Llama 3.1 405B beats Claude 3.5 Haiku and Gemini 1.5 Flash, some of the most commonly used proprietary models. Gemma 2 27B beats Llama 3.1 70B despite its size. These results indicate that not all models react similarly to chain-of-thought prompting. More comprehensive experiments may be necessary to identify such patterns in full.

One of the most surprising results of chain-of-thought prompts is observed on Llama 3.1 8B. It loses accuracy in all subjects and is the only model that sees decreases in average accuracy.

5.4 Released Resources

As a result of this work, we have released a database and a code repository. We believe that these resources will be useful for future research projects.

The main resource is TUMLU-mini, which is hosted on the Hugging Face platform [4]. You need to send a request to access the database, but these requests are approved automatically. TUMLU-mini contains a manually verified subset of the TUMLU benchmark. It was used in our experiments. The TUMLU dataset is not released publicly, but we intend to use it to create alternative benchmarks in the future.

We also release the codebase associated with this work. This repository is hosted on GitHub [5]. It includes the Python scripts used in experiments. It also includes the experimental results themselves, which can be used to regenerate the tables and graphs available in this work.

Both the dataset and the codebase are released under the Creative Commons Attribution 4.0 license. This license allows commercial use, distribution, modification, or private use of the resources with proper attribution.

6 CONCLUSION AND FUTURE WORK

Multitask benchmarks are important resources for evaluating the performance of LLMs. Such benchmarks are available in high-resource languages, but this is not the case for most of the Turkic languages. While certain Turkic languages have been studied in this regard, these works have either concentrated on an individual language or languages across the globe. No work has treated Turkic languages as a subject of its own. In this work, we do exactly that – we introduce a unified and native language understanding benchmark for major Turkic languages and use this benchmark to evaluate SOTA LLMs.

[4] https://huggingface.co/datasets/jafarisbarov/TUMLU-mini
[5] https://github.com/ceferisbarov/TUMLU

We also evaluate SOTA LLMs using this benchmark. Comparing the performance of these LLMs against their performance on translated benchmarks reveals interesting patterns. It turns out that LLMs struggle more with native benchmarks. We believe this is due to the fact that native benchmarks contain more local information that has not been available in the pretraining data of these LLMs.

Our work 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 future editions.

Apart from these, we also believe that our work can be expanded in the following directions:

- Multi-modal datasets containing questions with images in Physics, Biology, and other subjects, questions with audio in the Native Language subject. The current dataset only contains text data, which is a very limiting setting to evaluate world knowledge and reasoning capabilities.

- Creating an open-ended dataset: TUMLU is a multiple-choice question dataset. It only evaluates whether or not the model chose the correct option out of 4. A generative benchmark would provide additional insights into how LLMs reason and answer these questions and whether or not these steps are correct.

- Higher-level questions. TUMLU contains questions at middle and high-school level. Creating an alternative dataset at the undergraduate level would challenge the LLMs further and also offer an opportunity to create a more standardized test, enabling comparison across languages.

REFERENCES

[1] Acikgoz, E. C., Erdogan, M., and Yuret, D. Bridging the bosphorus: Advancing turkish large language models through strategies for low-resource language adaptation and benchmarking, 2024.

[2] Anthropic. Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet, 2024.

[3] Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Bayes, E., Azime, I. A., Alabi, J. O., Kgomo, J., Eloundou, T., Proehl, E., Chen, K., Khadir, I., Etori, N. A., Muhammad, S. H., et al. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948* (2024).

[5] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 1137–1155.

[6] Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. Rasa: Open source language understanding and dialogue management, 2017.

[7] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal, Sept. 2015), L. Màrquez, C. Callison-Burch, and J. Su, Eds., Association for Computational Linguistics, pp. 632–642.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 1877–1901.

[9] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol. 15*, 3 (Mar. 2024).

[10] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, pp. 8440–8451.

[11] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025.

[12] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2023), NIPS '23, Curran Associates Inc.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, pp. 4171–4186.

[14] et al., A. S. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023).

[15] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[16] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size, 2024.

[17] Grattafiori, A., et al. The Llama 3 Herd of Models, 2024.

[18] Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., K, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems* (2023), A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., pp. 44123–44279.

[19] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations* (2021).

[20] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep Learning Scaling is Predictable, Empirically, 2017.

[21] Hiemstra, D. A probabilistic justification for using tf$\times$ idf term weighting in information retrieval. *International Journal on Digital Libraries 3* (2000), 131–139.

[22] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Comput. 9*, 8 (Nov. 1997), 1735–1780.

[23] Holtermann, C., Röttger, P., Dill, T., and Lauscher, A. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics: ACL 2024* (Bangkok, Thailand, Aug. 2024), L.-W. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, pp. 4476–4494.

[24] Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2022).

[25] Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[26] Isbarov, J., Akhundjanova, A., Hajili, M., Huseynova, K., Gaynullin, D., Rzayev, A., Tursun, O., Saetov, I., Kharisov, R., Belginova, S., Kenbayeva, A., Alisheva, A., Turdubaeva, A., Köksal, A., Rustamov, S., and Ataman, D. Tumlu: A unified and native language understanding benchmark for turkic languages, 2025.

[27] Isbarov, J., Huseynova, K., Mammadov, E., Hajili, M., and Ataman, D. Open foundation models for Azerbaijani language. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)* (Bangkok, Thailand and Online, Aug. 2024), D. Ataman, M. O. Derin, S. Ivanova, A. Köksal, J. Sälevä, and D. Zeyrek, Eds., Association for Computational Linguistics, pp. 18–28.

[28] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, 2020.

[29] Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., and Baldwin, T. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024* (Bangkok, Thailand, Aug. 2024), L.-W. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, pp. 5622–5640.

[30] Kuriyozov, E., Vilares, D., and Gómez-Rodríguez, C. BERTbek: A pretrained language model for Uzbek. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (Torino, Italia, May 2024), M. Melero, S. Sakti, and C. Soria, Eds., ELRA and ICCL, pp. 33–44.

[31] Köksal, A., Thaler, M., Imani, A., Üstün, A., Korhonen, A., and Schütze, H. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions, 2024.

[32] Lai, V. D., Van Nguyen, C., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039* (2023).

[33] Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., and Baldwin, T. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024* (Bangkok, Thailand, Aug. 2024), L.-W. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, pp. 11260–11285.

[34] Maxutov, A., Myrzakhmet, A., and Braslavski, P. 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)* (Bangkok, Thailand and Online, Aug. 2024), D. Ataman, M. O. Derin, S. Ivanova, A. Köksal, J. Sälevä, and D. Zeyrek, Eds., Association for Computational Linguistics, pp. 81–91.

[35] Maxutov, A., Myrzakhmet, A., and Braslavski, P. 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)* (Bangkok, Thailand and Online, Aug. 2024), D. Ataman, M. O. Derin, S. Ivanova, A. Köksal, J. Sälevä, and D. Zeyrek, Eds., Association for Computational Linguistics, pp. 81–91.

[36] Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., Hajili, M., Ivanova, S., Khaytbaev, A., Laverghetta Jr., A., Moydinboyev, B., Onal, E., Pulatova, S., Wahab, A., Firat, O., and Chellappan, S. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing (Online and Punta Cana, Dominican Republic, Nov. 2021), M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Association for Computational Linguistics, pp. 5876–5890.

[37] Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., Hajili, M., Ivanova, S., Khaytbaev, A., Laverghetta Jr., A., Moydinboyev, B., Onal, E., Pulatova, S., Wahab, A., Firat, O., and Chellappan, S. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic, Nov. 2021), M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Association for Computational Linguistics, pp. 5876–5890.

[38] Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., and Shoham, Y. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (St. Julian's, Malta, Mar. 2024), Y. Graham and M. Purver, Eds., Association for Computational Linguistics, pp. 49–66.

[39] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2022), NIPS '22, Curran Associates Inc.

[40] Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning* (07–08 Apr 2022), G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., vol. 174 of *Proceedings of Machine Learning Research*, PMLR, pp. 248–260.

[41] Patel, B., Chakraborty, S., Suttle, W. A., Wang, M., Bedi, A. S., and Manocha, D. Aime: Ai system optimization via multiple llm evaluators, 2024.

[42] Penedo, G., Kydlíček, H., Sabolčec, V., Messmer, B., Foroutan, N., Jaggi, M., von Werra, L., and Wolf, T. Fineweb2: A sparkling update with 1000s of languages, Dec. 2024.

[43] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore, Dec. 2023), H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, pp. 1339–1384.

[44] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners.

[45] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

[46] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, Nov. 2016), J. Su, K. Duh, and X. Carreras, Eds., Association for Computational Linguistics, pp. 2383–2392.

[47] Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., and Benevenuto, F. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science 5* (2016), 1–29.

[48] Romanou, A., Foroutan, N., Sotnikova, A., Nelaturu, S. H., Singh, S., Maheshwary, R., Altomare, M., Chen, Z., Haggag, M. A., A, S., Amayuelas, A., Amirudin, A. H., Boiko, D., Chang, M., Chim, J., Cohen, G., Dalmia, A. K., Diress, A., Duwal, S., Dzenhaliou, D., Florez, D. F. E., Farestam, F., Imperial, J. M., Islam, S. B., Isotalo, P., Jabbarishiviari, M., Karlsson, B. F., Khalilov, E., Klamm, C., Koto, F., Krzemiński, D., de Melo, G. A., Montariol, S., Nan, Y., Niklaus, J., Novikova, J., Ceron, J. S. O., Paul, D., Ploeger, E., Purbey, J., Rajwal, S., Ravi, S. S., Rydell, S., Santhosh, R., Sharma, D., Skenduli, M. P., Moakhar, A. S., soltani moakhar, B., Tarun, A. K., Wasi, A. T., Weerasinghe, T. O., Yilmaz, S., Zhang, M., Schlag, I., Fadaee, M., Hooker, S., and Bosselut, A. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations* (2025).

[49] Rumelhart, D. E., and McClelland, J. L. *Learning Internal Representations by Error Propagation.* 1987, pp. 318–362.

[50] Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ko, W.-Y., Smith, M., Bosselut, A., Oh, A., Martins, A. F. T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024.

[51] Suen, C. Y. N-gram statistics for natural language understanding and text processing. *IEEE transactions on pattern analysis and machine intelligence*, 2 (1979), 164–172.

[52] Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261* (2022).

[53] Türk, U., Atmaca, F., Özateş, Ş. B., Köksal, A., Ozturk Basaran, B., Gungor, T., and Özgür, A. Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th Linguistic Annotation Workshop* (Florence, Italy, Aug. 2019), A. Friedrich, D. Zeyrek, and J. Hoek, Eds., Association for Computational Linguistics, pp. 166–177.

[54] Turker, M., Ari, M. E., and Han, A. Vbart: The turkish llm, 2024.

[55] Vanmassenhove, E., Shterionov, D., and Gwilliam, M. 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* (Online, Apr. 2021), P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Association for Computational Linguistics, pp. 2203–2213.

[56] Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[57] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[58] Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark, 2024.

[59] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2024), NIPS '22, Curran Associates Inc.

[60] Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM 9*, 1 (Jan. 1966), 36–45.

[61] Winograd, T. Procedures as a representation for data in a computer program for understanding natural language.

[62] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, June 2021), K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds., Association for Computational Linguistics, pp. 483–498.

[63] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 Technical Report, 2025.

[64] Yildiz, O. T., Avar, B., and Ercan, G. An open, extendible, and fast Turkish morphological analyzer. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (Varna, Bulgaria, Sept. 2019), R. Mitkov and G. Angelova, Eds., INCOMA Ltd., pp. 1364–1372.

[65] Yüksel, A., Köksal, A., Senel, L. K., Korhonen, A., and Schuetze, H. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (Miami, Florida, USA, Nov. 2024), Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Association for Computational Linguistics, pp. 7035–7055.

[66] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics 12* (2024), 39–57.

[67] Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024* (Mexico City, Mexico, June 2024), K. Duh, H. Gomez, and S. Bethard, Eds., Association for Computational Linguistics, pp. 2765–2781.

[68] Ziyaden, A., Yelenov, A., Hajiyev, F., Rustamov, S., and Pak, A. Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages. *PeerJ Computer Science 10* (2024), e1974.

A MODEL RESULTS

This appendix includes results for all models, except for Claude 3.5 Sonnet which is available in Table 4. Tables 1-20 contain 5-shot results and the remaining tables contain 5-shot CoT results.

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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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 19: 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 20: Accuracy scores for QWEN/Qwen2.5 72B-INSTRUCT model across languages.

This appendix includes 5-shot CoT results for all models, except for Claude 3.5 Sonnet which is available in Table 6.

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 21: 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 22: 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: 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 29: 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 30: 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 31: Accuracy scores for QWEN/Qwen2.5 72B-INSTRUCT model across languages.