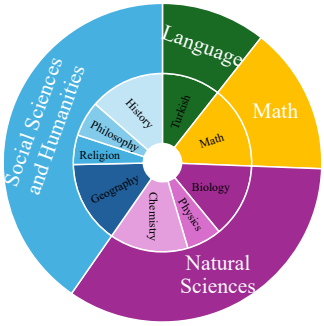
TurkishMMLU: Measuring Massive Multitask Language Understanding in Turkish

Arda Yüksel[1], **Abdullatif Köksal**[2,3,4], **Lütfi Kerem Senel**[2,3], **Anna Korhonen**[4], **Hinrich Schütze**[2,3]

[1]Technical University of Munich
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning
[4]Language Technology Lab, University of Cambridge
arda.yueksel@tum.de, akoksal@cis.lmu.de

Abstract

Multiple choice question answering tasks evaluate the reasoning, comprehension, and mathematical abilities of Large Language Models (LLMs). While existing benchmarks employ automatic translation for multilingual evaluation, this approach is error-prone and potentially introduces culturally biased questions, especially in social sciences. We introduce the first multitask, multiple-choice Turkish QA benchmark, TurkishMMLU, to evaluate LLMs' understanding of the Turkish language. TurkishMMLU includes over 10,000 questions, covering 9 different subjects from Turkish high-school education curricula. These questions are written by curriculum experts, suitable for the high-school curricula in Turkey, covering subjects ranging from natural sciences and math questions to more culturally representative topics such as Turkish Literature and the history of the Turkish Republic. We evaluate over 20 LLMs, including multilingual open-source (e.g., Gemma, Llama, MT5), closed-source (GPT 4o, Claude, Gemini), and Turkish-adapted (e.g., Trendyol) models. We provide an extensive evaluation, including zero-shot and few-shot evaluation of LLMs, chain-of-thought reasoning, and question difficulty analysis along with model performance. We provide an in-depth analysis of the Turkish capabilities and limitations of current LLMs to provide insights for future LLMs for the Turkish language. We publicly release our code for the dataset and evaluation: https://github.com/ArdaYueksel/TurkishMMLU.



Subject: History

Question: Aşağıdakilerden hangisi 1923'ten sonra Türk dış politikasında ele alınıp çözümlenmeye çalışılmış konulardan biri değildir? (Which of the following is not one of the issues that were tried to be addressed and resolved in Turkish foreign policy after 1923?)

A. Boğazlar Sorunu (Turkish straits crisis)
B. Kapitülasyonlar (Capitulations)
C. Dış Borçlar (Foreign Debts)
D. Irak Sınırı (Iraq Border)
E. Nüfus Mübadelesi (Population Exchange)

Correctness Ratio: 40% (Difficulty: Medium)

LLM Evaluations:
Llama-3 70B-IT: E
GPT 4o: B
Gemini 1.5-pro: D
Claude-3 Opus: E

Figure 1: The chart displays the subject distribution of TurkishMMLU. An example from our dataset shows recent multilingual LLMs struggling with a question about Turkish history.

1 Introduction

Benchmarking plays an important role in understanding and measuring the capabilities of language models. Recent multitask multiple-choice question answering (QA) benchmarks like MMLU (Hendrycks et al., 2021) cover a wide range of use cases for language models, making them highly popular as one of the main evaluation benchmarks in recent LLMs such as GPT 4 (OpenAI et al., 2024) and Gemini (Team et al., 2024a). For the multilingual adaptation of the MMLU benchmark, recent works (Lai et al., 2023) have focused on automatic translations. However, automatic translations are often prone to errors and may fail to capture the linguistic and cultural nuances of the target language. Consequently, there have been manual efforts to create multitask multiple-choice bench-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7035–7055
November 12-16, 2024

marks in various languages, including Arabic (ArabicMMLU, Koto et al., 2024), Korean (KMMLU, Son et al., 2024), and Chinese (CMMLU, Li et al., 2023).

In our work, we introduce TurkishMMLU, the first multitask multiple-choice QA benchmark specifically designed for the Turkish language. Our dataset includes 10,032 multiple-choice questions, each with five options, spanning nine subjects categorized into four groups: *Natural Sciences*, *Mathematics*, *Turkish Language and Literature*, and *Social Sciences and Humanities*. These questions are sourced from a high-quality online learning platform created by the Turkish Ministry of Education, which aims to support high school students in preparing for the university entrance exam. A unique feature of TurkishMMLU is the correctness ratio, which reflects the actual performance of students on these questions, offering a more accurate measure of question difficulty. We illustrate the distribution of subjects and an example from TurkishMMLU in Figure 1.

After introducing this dataset for benchmarking in Turkish, we evaluate a wide range of current language models, more than 40, including multilingual autoregressive LLMs, both open models like Gemma (Team et al., 2024b), Llama-3 and Aya-23 (Aryabumi et al., 2024) and closed-source models such as GPT 4o, Claude and Gemini. In addition, we also cover multilingual encoder-decoder models such as MT5, MT0, Aya and Turkish-adapted LLMs such as Trendyol-LLM, a LoRA adaptation of multilingual LLMs. We also cover many different setups including zero-shot, few-shot, and chain-of-thought (Wei et al., 2022). We further provide analysis of LLMs based on subjects and difficulty. Our additional analysis provides insights for the design of future LLMs for Turkish and beyond. We publicly release our code for the dataset and evaluation: `https://github.com/ArdaYueksel/TurkishMMLU`.

Our contributions are as follows:

1. We introduce the first large-scale multitask multiple-choice benchmark for Turkish, consisting of 10,032 questions across nine subjects.

2. We evaluate a wide range of LLMs, varying in size from 60M to 141B, including both open and closed-source models, and provide a comprehensive leaderboard featuring over 40 models.

3. We conduct an in-depth analysis of LLM performance in chain-of-thought setups and based on question difficulty.

2 Related Work

LLM Benchmarking: Benchmarks are crucial for understanding the capabilities of NLP models, identifying their weaknesses and facilitating the development of more capable models. Historically, most NLP benchmarks focused on linguistic tasks (Wang et al., 2018, 2019; Rajpurkar et al., 2016) and followed the paradigm of supervised fine-tuning of a model on a training set and evaluation on an unseen test set. However, with the advent of powerful LLMs, this type of evaluation became obsolete as these models showed impressive zero-shot and few-shot learning skills, even for higher level tasks closer to real world applications. To evaluate the emerging capabilities of the LLMs, new benchmarks are proposed that focus on more advanced capabilities such as common sense reasoning (Levesque et al., 2012), multi-hop reasoning (Yang et al., 2018), programming (Chen et al., 2021) and multi-turn conversations. Additionally, some studies aimed at evaluating these capabilities through extensive datasets that cover a broad range of knowledge-based topics (Srivastava et al., 2023). One prominent example is MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021); it covers 57 diverse fields from basic arithmetic to intricate areas like legal studies and computer science. Although many of these benchmarks have focused on English, there have been significant efforts to adapt and develop similar benchmarks for other languages (Son et al., 2024; Koto et al., 2024; Li et al., 2023; Senel et al., 2024; Conneau et al., 2018; Ponti et al., 2020).

Turkish Benchmarks: One of the initial efforts in Turkish benchmarking was THQUAD (Soygazi et al., 2021), a variant of the SQuAD question-answering benchmark (Rajpurkar et al., 2016) that focuses on extracting information from historical passages and answering questions about Ottoman and Islamic history in an open-book format. MUKAYESE (Safaya et al., 2022), another Turkish benchmark, was created by combining multiple existing datasets for various tasks. However, most of the tasks that are included in MUKAYESE, such as NER (named entity recognition), sentence segmentation and spellchecking, do not effectively capture the knowledge and the language understand-

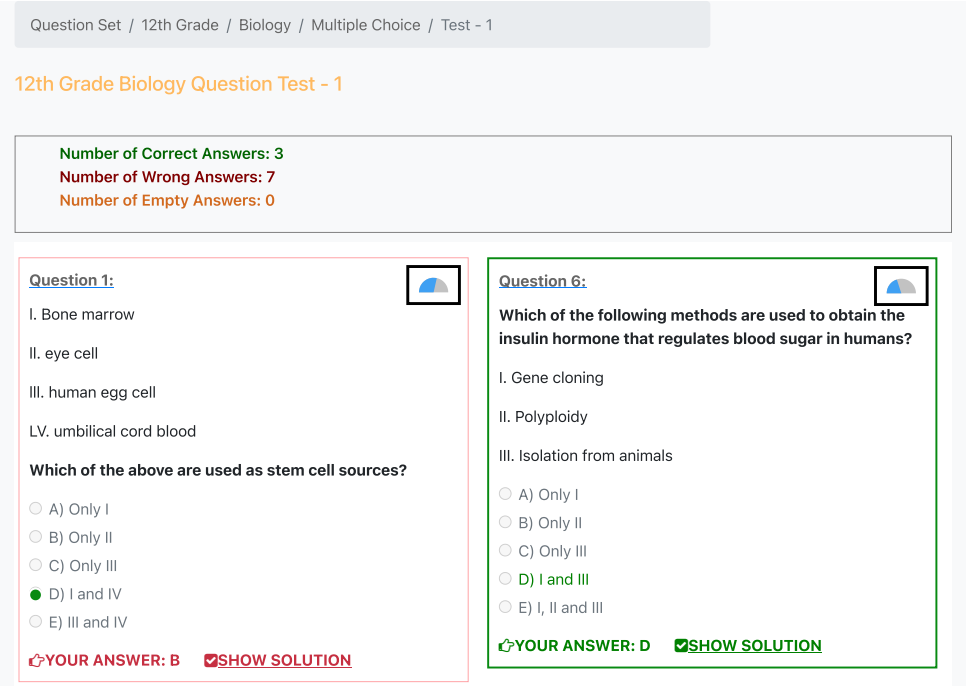


Figure 2: Sample biology test from the EBA Platform (translated to English, see Figure 5 in Appendix for the original Turkish test). Black boxes indicate the correctness ratio (difficulty level). Green borders appear when the user's choice matches the ground-truth answer, while red borders indicate incorrect choices.

ing capabilities of LLMs due to their low level nature. Several other studies that created multilingual benchmarks for specific tasks such as XCOPA (Cross-lingual Choice of Plausible Alternatives) (Ponti et al., 2020) and XNLI (Cross-lingual Natural Language Inference) (Conneau et al., 2018) also include Turkish among several other languages. A recent study that focuses on Turkish LLMs (Acikgoz et al., 2024) created the Turkish versions of the TruthfulQA Multiple Choice (MC) (Lin et al., 2022) and ARC (AI2 Reasoning Challenge) (Clark et al., 2018) datasets to evaluate Turkish LLMs. These benchmarks are constructed by machine translating the English versions of the corresponding datasets, which is usually followed by manual verification and editing to ensure good quality. Overall, despite some efforts to evaluate the capabilities of LLMs for Turkish, Turkish still lacks a high quality and comprehensive evaluation resource that covers multiple domains. In this study, we address this by introducing Turkish MMLU.

3 Dataset

TurkishMMLU is curated using resources from online learning materials for Turkish high school education. In the Turkish educational system, high school education spans four years, and students take the National University Entrance Exams after completing their studies. This exam contains multiple-choice questions covering various subjects from the curricula. To assist students in preparing for these exams, official and commercial exam preparation booklets, video guides, and online practice tests in multiple-choice question-answering format are available. The Turkish Ministry of Education (MEB) has developed an online platform called the Education Information Network (EBA), which aims to provide electronic resources such as lecture notes, videos, tests and solutions, and interactive books to facilitate the learning process for students. This platform[1] contains multiple-choice questions and their solutions that form the basis of our study. The questions are driectly sourced from the platform which ensures that each question is crafted and verified by domain experts, guaranteeing correctness and high quality.

We conducted repeated manual reviews, randomly selecting 30 questions from the complete dataset during each review session to assess their formatting and accuracy. Whenever we detected formatting issues—typically due to irregular presentation on the source website—we revised our parsing code to effectively address these edge cases. This process was continued until we ensured con-

[1] https://ogmmateryal.eba.gov.tr/panel/MSoruDers.aspx

sistent formatting accuracy across multiple consecutive rounds of the sampled questions.

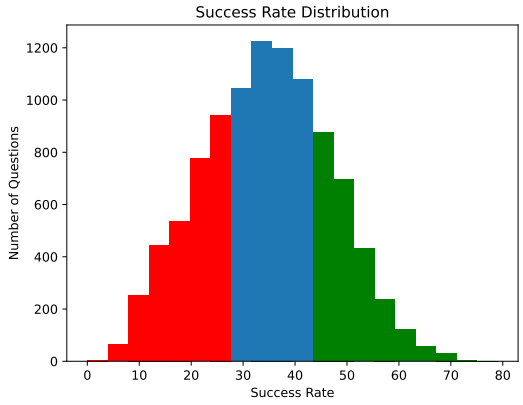


Figure 3: Distribution of correctness ratios. Questions are categorized as Easy (green, top 30%), Medium (blue, middle 40%), or Hard (red, bottom 30%) based on the 30th and 70th percentiles.

Figure 2 illustrates the EBA platform interface. Users generate tests by specifying grade level and subject, upon which the platform provides multiple 10-question tests. After test completion, users can review ground-truth answers and video solutions. Each question's difficulty is denoted by a Correctness Ratio (black boxes in Figure 2), calculated as the percentage of correct user responses. For each test, we extract question text, multiple-choice options, correct answer, topic, subject, grade, and difficulty level.

Table 1 details the distribution of test questions by grade and subject in TurkishMMLU. The dataset includes nine high school subjects across four domains: Math (Mathematics); Natural Sciences (Biology, Chemistry, Physics); Language (Turkish Language and Literature); and Humanities and Social Sciences (History, Geography, Philosophy, Religion and Ethics). The test set comprises 9,807 multiple-choice questions, with an additional 225 (25 per subject) in the development set. While Philosophy is limited to grades 10 and 11, other subjects span all four grades. Many questions include mathematical formulas/notations (in LaTeX or text) and images, however, we exclude image-based questions to focus on evaluating text models.

Figure 3 displays the distribution of Correctness Ratios. Questions are categorized as Easy (top 30%), Medium (middle 40%), or Hard (bottom 30%), with percentile thresholds at 41 and 28, respectively.

We manually selected 25 questions per subject for the development set, maintaining subject-grade distributions and mirroring the overall difficulty distribution. For few-shot examples, we focus on 5-shot experiments with 5 questions per subject due to context window constraints and compute budget limitations, each with different correct answers to avoid selection bias. For Chain-of-Thought (COT) prompting, we manually provide step-by-step solutions for these 5 questions per subject.

Subject	**Grade**				**Total**
	9	10	11	12	
Turkish L & L	251	336	208	246	1041
Mathematics	565	470	64	379	1478
Physics	194	93	78	246	611
Chemistry	283	474	340	309	1406
Biology	273	328	401	323	1325
History	342	398	281	316	1337
Geography	331	364	494	290	1479
Religion and Ethics	120	229	122	42	513
Philosophy	0	332	285	0	617

Table 1: Distribution of test questions of TurkishMMLU by subject and grade.

The large scale of our test dataset, including 9,807 questions, raises significant challenges. Experiments with state-of-the-art proprietary models like GPT 4 and Claude-Opus face budget constraints, while using Chain-of-thought (COT) prompting with open-source models generates excessively long responses, resulting in long inference times. To address these issues while maintaining comprehensive evaluations, we create a smaller version of TurkishMMLU, called TurkishMMLU$_{\text{sub}}$ with 100 randomly selected questions per subject, totaling 900. We uniformly sampled 25 questions per grade for each subject, except for Philosophy, which has 50 questions evenly distributed between grades 10 and 11. This sample is representative of grades and subjects, enabling in-depth model evaluation, but can be easily used in resource-constrained scenarios. We measure the correlation between TurkishMMLU$_{\text{sub}}$ and TurkishMMLU in §4.5, finding a strong correlation across 32 models.

4 Evaluation Results

After finalizing TurkishMMLU, we now evaluate various multilingual and Turkish-adapted open- and closed-source LLMs. We cover a wide range of models, from 60M to 141B parameters, and various experimental setups.

Experimental Setup Our main evaluation setup is **5-shot in-context learning evaluation**, following the prior evaluation setups in recent LLMs (Team et al., 2024b; OpenAI et al., 2024) on English MMLU (Hendrycks et al., 2021). From the development set proposed in §3, we select a fixed set of questions for each subject and include 5 of them in our few-shot prompt, with the question, multiple-choice options, and the answer. We carefully design these prompts to ensure that each question has a different option (in our dataset, the five options are always A, B, C, D, E) as the answer. For evaluation, we report accuracy by using the lm-evaluation-harness framework from EleutherAI (Gao et al., 2023). For open-source models, we perform log-prob based evaluation; for closed-source models we perform greedy decoding and then parse the prediction. In order to ensure validity of parsing, we utilize regex patterns based on the few-shot example structure which can be formulated as `Question: ...; Choices: A. ... E. ... ; Answer: [A-E] .`

Our second evaluation is a **zero-shot evaluation** to compare few-shot and zero-shot performance of the models. Additionally, we evaluate LLMs with a **5-shot chain-of-thought (CoT) evaluation**. Especially for questions requiring further reasoning and elaboration, such as mathematics, directly giving answers may be a limitation in our main evaluation. Therefore, we evaluate a wide range of models, including closed-source models, with CoT reasoning (Wei et al., 2022). In this setup, we provide CoT solutions for each question in our few-shots for each subject and perform greedy decoding. We put the final answer option at the end of the solution in the prompts, and then parse the predicted answer in the generated solution. We formatted the CoT few-shot examples as they would end with the Turkish phrase of `The correct choice is [A-E]`. This way we can parse the predicted answer easily from the generated solution.

Since TurkishMMLU includes real-world data for difficulty, we also conduct a **difficulty analysis** to evaluate models. This expands our evaluation setup from comparing models on different subjects to varying difficulty levels. In all of our evaluations, we use a small subset of TurkishMMLU, TurkishMMLU$_{\text{sub}}$, because the closed-source experiments are quite expensive.[2] With public models, we calculate performance on both TurkishMMLU and TurkishMMLU$_{\text{sub}}$ to test our assumption that they would yield similar results.

Language Models: We evaluate a diverse range of models, including Turkish-adapted, multilingual open-source and closed-source LLMs.

For Turkish-adapted models, we use Trendyol-LLM 7B, a Llama-2 model further pretrained on Turkish[3], available in base, chat, and chat-dpo forms on HuggingFace. We also include Kanarya (Safaya et al., 2022), a pretrained autoregressive 2B Turkish model.

In the multilingual open-source category, we evaluate models with encoder-decoder architectures such as mT5 (Xue et al., 2021) (from small to xxl), mT0 (Muennighoff et al., 2023) (with the same sizes as mT5), and Cohere's Aya-101 (Üstün et al., 2024). For autoregressive models, we include Meta's Llama-2 (Touvron et al., 2023) (7B, 7B-Chat, 13B, 13B-Chat) and Llama-3 (8B, 8B-Instruct, 70B, and 70B-Instruct). From MistralAI, we evaluate Mistral 7B variants (Jiang et al., 2023), Mixtral 8x22B, and 8x7B (Jiang et al., 2024). We also include Cohere4AI's Command-R and Aya-23 models (Aryabumi et al., 2024), Google's Gemma (Team et al., 2024b) (7B and 2B with their instruction versions), and Microsoft's Phi-3-Mini (Abdin et al., 2024).

For multilingual closed-source models, we evaluate OpenAI's GPT models (3.5, 4-Turbo, and 4o), Anthropic's Claude-3 models (Haiku, Sonnet, and Opus versions), and Google's Gemini models (pro versions 1.0 and 1.5).

4.1 Few-shot Evaluation

We present the 5-shot evaluation of models in Table 2. We show scores in four categories: Natural Sciences, Math, Turkish Language & Literature, and Social Sciences and Humanities, as well as the macro-averaged scores over nine subjects. The best-performing model is a closed-source model, GPT 4o, with **83.1%** accuracy. It outperforms all other models in each category as well. The best-performing open-source model is Llama-3 70B-IT (Instruction-Tuned) with **67.3%** accuracy. While it is better than many closed-source models such as Claude-3 Sonnet and Gemini 1.0-pro, it is still 15.8% worse than GPT 4o. Another interesting point is that the best encoder-decoder model,

[2] For example, a 5-shot CoT evaluation with Claude-3 Opus on the entire dataset would cost more than $750.

[3] https://huggingface.co/Trendyol/Trendyol-LLM-7b-base-v0.1

Model	Source	Average	Natural Sciences	Math	Turkish L & L	Social Sciences and Humanities
GPT 4o*	Closed	**83.1**	**75.3**	**59.0**	**82.0**	**95.3**
Claude-3 Opus*	Closed	79.1	71.7	**59.0**	77.0	90.3
GPT 4-turbo*	Closed	75.7	70.3	57.0	67.0	86.5
Llama-3 70B-IT	Open	67.3	56.7	42.0	57.0	84.3
Claude-3 Sonnet*	Closed	67.3	67.3	44.0	58.0	75.5
Llama-3 70B	Open	66.1	56.0	37.0	57.0	83.3
Claude-3 Haiku*	Closed	65.4	57.0	40.0	61.0	79.3
Gemini 1.0-pro	Closed	63.2	52.7	29.0	63.0	79.8
C4AI Command-r+	Open	60.6	50.0	26.0	57.0	78.0
Aya-23 35B	Open	55.6	43.3	31.0	49.0	72.5
C4AI Command-r	Open	54.9	44.7	29.0	49.0	70.5
Mixtral 8x22B	Open	54.8	45.3	27.0	49.0	70.3
GPT 3.5-turbo*	Closed	51.0	42.7	39.0	45.0	61.8
Llama-3 8B-IT	Open	46.4	36.7	29.0	39.0	60.0
Llama-3 8B	Open	46.2	37.3	30.0	33.0	60.3
Mixtral 8x7B-IT	Open	45.2	41.3	28.0	39.0	54.0
Aya-23 8B	Open	45.0	39.0	23.0	31.0	58.5
Gemma 7B	Open	43.6	34.3	22.0	47.0	55.0
Aya-101	Open	40.7	31.3	14.0	38.0	55.0
Trendyol-LLM 7B-C-D	Open	34.1	30.3	22.0	28.0	41.5
mT0-xxl	Open	33.9	29.3	28.0	21.0	42.0
Mistral 7B-IT	Open	32.0	34.3	26.0	38.0	30.3
Llama-2 7B	Open	22.3	25.3	26.0	20.0	19.8
mT5-xxl	Open	18.1	19.3	24.0	14.0	16.8

Table 2: 5-shot experiments on TurkishMMLU$_{sub}$. Many closed models shift to chain-of-thought-like detailed explanations, we indicate this with the * symbol. Natural Sciences consists of Biology, Chemistry, and Physics. Turkish L&L is the Turkish Language and Literature subject. Social Sciences and Humanities consists of History, Geography, Philosophy, and Religion and Ethics.

Aya-101, performs much worse than autoregressive models, achieving only 40.7% accuracy.

The results suggest that mathematics is the most difficult subject for almost all models, as it is usually challenging to answer these questions correctly in a single token, given that they require multi-hop reasoning. The easiest category in TurkishMMLU$_{sub}$ is Social Sciences and Humanities. For STEM courses, models perform poorly compared to other subjects. We also observe that many closed-source models switch to COT-like problem-solving rather than providing the answer directly, even though we provided single-answer style few-shots. We parse the predicted option in those answers with manually-designed patterns and indicate these "CoT" models with the * symbol in Table 2.

Among 7B-8B models, Llama-3 8B-IT exhibits the best performance, but Aya-23 and Gemma show comparable results. Mistral 7B-IT and Llama-2 7B lag more than 10% behind these three models. Among mT5-xxl (13B) based models, Aya-101 achieves the best performance, however, encoder-decoder based models perform worse than autoregressive models of similar sizes.

We note that recent open-source models such as Llama-3, Command-R, Aya-23, and Mixtral 8x22B (all released after April 2024) outperform older closed-source models like GPT 3.5 (released in March 2022), signaling promise for open-source models. However, Turkish-adapted models like Trendyol-LLM, despite outperforming their base model (Llama-2 7B), are significantly behind newer variants of similar size (Llama-3 8B).

We provide the results for all nine subjects and all models in the Appendix in Table 6.

4.2 Zero-Shot Evaluation

To assess the performance gain from few-shots, we also compare models in zero-shot settings. Table 4 summarizes the results for selected open-source models. We observe the most significant perfor-

Model	Source	Average	Natural Sciences	Math	Turkish L & L	SocSci/ Humanities
GPT 4o	Closed	**88.2** (+5.1)	**86.3** (+11.0)	**84.0** (+25.0)	**81.0** (–1.0)	**92.5** (–2.8)
Claude-3 Opus	Closed	81.8 (+2.7)	77.0 (+5.3)	74.0 (+15.0)	76.0 (–1.0)	88.8 (–1.5)
GPT 4-turbo	Closed	79.2 (+3.5)	75.3 (+5.0)	75.0 (+18.0)	69.0 (+2.0)	85.8 (–0.8)
Gemini 1.5-pro*	Closed	70.1 (+45.1)	65.0 (+43.7)	51.0 (+27.0)	54.0 (+7.0)	82.7 (+60.2)
Llama-3 70B-IT	Open	68.1 (+0.8)	62.0 (+5.3)	57.0 (+15.0)	53.0 (–4.0)	79.2 (–5.0)
Claude-3 Haiku	Closed	66.1 (+0.7)	56.7 (–0.3)	45.0 (+5.0)	64.0 (+3.0)	79.0 (–0.3)
Llama-3 70B	Open	63.3 (–2.8)	57.3 (+1.3)	34.0 (–3.0)	54.0 (–3.0)	77.5 (–5.8)
Claude-3 Sonnet	Closed	60.7 (–6.6)	58.7 (–8.6)	38.0 (–6.0)	62.0 (+4.0)	67.5 (–8.0)
GPT 3.5-turbo	Closed	58.2 (+7.2)	52.3 (+9.6)	42.0 (+3.0)	51.0 (+6.0)	68.5 (+6.7)
Gemini 1.0-pro	Closed	54.1 (–9.1)	42.7 (–10.0)	39.0 (+10.0)	48.0 (–15.0)	68.0 (–11.8)
C4AI command-r	Open	49.6 (–5.3)	40.0 (–4.7)	28.0 (–1.0)	41.0 (–8.0)	64.2 (–6.2)
Llama-3 8B-IT	Open	40.6 (–5.8)	35.0 (–1.7)	20.0 (–9.0)	29.0 (–10.0)	52.8 (–7.2)
Mixtral 8x7B-IT	Open	40.1 (–5.1)	33.0 (–8.3)	33.0 (+5.0)	39.0 (+0.0)	47.5 (–6.5)
Gemma 7B	Open	34.0 (–9.6)	26.3 (–8.0)	17.0 (–5.0)	27.0 (–20.0)	45.8 (–9.2)
Llama-3 8B	Open	28.2 (–18.0)	24.3 (–13.0)	7.0 (–23.0)	27.0 (–6.0)	36.8 (–23.5)
Trendyol-LLM 7B-C	Open	27.7 (–10.3)	24.0 (–6.3)	6.0 (–12.0)	26.0 (–9.0)	36.2 (–13.2)

Table 3: 5-shot chain-of-thought (CoT) evaluation results on TurkishMMLU$_{sub}$. The table presents accuracy for four subject categories and the macro-average, with performance changes from non-CoT experiments in parentheses. * Gemini 1.5-pro's large improvement (+45.1) is due to a model behavior that causes mispredictions in non-CoT, rather than true CoT gains.

Model	Zero-Shot	5-Shot
Llama-3 70B-IT	**64.6**	**67.3** (+2.7)
C4AI Command-r+	50.6	60.6 (+10.0)
Mixtral 8x22B	46.8	54.8 (+8.0)
Aya-23 35B	45.3	55.6 (+10.3)
mT0-xxl	44.8	33.9 (–10.9)
C4AI Command-r	42.4	54.9 (+12.5)
Llama-3 8B-IT	38.3	46.4 (+8.1)
Aya-101	37.4	40.7 (+3.3)
Mixtral 8x7B-IT	35.8	45.2 (+9.4)
Trendyol-LLM 7B-C-D	33.3	34.1 (+0.8)
Mistral 7B-IT	24.6	32.0 (+7.4)
Gemma 7B	23.1	43.6 (+20.5)

Table 4: 5-shot and zero-shot accuracy on TurkishMMLU$_{sub}$ for open-source language models.

mance improvement via few-shot in the Gemma 7B model. Llama-3 70B-IT, the best-performing model in the few-shot setting, also leads in the zero-shot setting among public models with a minimal performance drop of just 2.7%.

Interestingly, mT0-xxl performs considerably better in the zero-shot setting than in the few-shot setting, contrary to the trends in the other models. We attribute this to mT0's (Muennighoff et al., 2023) primary focus on zero-shot adaptation. This finding suggests that mT0's zero-shot performance even surpasses Aya's few-shot performance.

4.3 Chain-of-Thought Evaluation

We evaluate 5-shot chain-of-thought (CoT) in Table 3, showing the performance difference between non-CoT and CoT few-shot experiments. We include CoT evaluations for three reasons: (i) to evaluate reasoning capabilities of recent LLMs, which show promising results (Team et al., 2024a), (ii) some subjects like mathematics require multi-hop reasoning, and (iii) CoT also indicates NLG performance of models in Turkish, complementing our NLU evaluation.

All models performing below 60% accuracy in the non-CoT few-shot scenario, except GPT 3.5-turbo, show worse performance with CoT reasoning. This suggests these models may have limited generation and reasoning capabilities in Turkish. Across all subjects, the most significant improvement is observed in mathematics, with +25.0% for the best-performing model, GPT 4o. With this approach, GPT 4o sets the best performance on TurkishMMLU$_{sub}$ at 88.2% accuracy across all settings. We also observe improvements in Natural Sciences, though not as substantial as in Mathematics. However, for Turkish Language & Literature and Social Sciences and Humanities, we observe no consistent improvements and even performance drops across models, including strong ones.

One exception to our findings is Gemini 1.5-pro.

Models	r_{pb}	Accuracy (%)		
		Easy	Medium	Hard
GPT 4o	0.211***	96.1	88.0	80.1
Claude-3 Opus	0.175***	89.4	81.7	73.7
GPT 4-turbo	0.143***	86.6	79.1	71.4
Gemini 1.5-pro	0.228***	80.3	73.7	54.5
Llama-3 70B-IT	0.193***	79.2	68.3	56.0
Claude-3 Haiku	0.265***	80.6	66.0	50.8
Llama-3 70B	0.287***	76.1	68.3	43.2
Claude-3 Sonnet	0.193***	68.7	64.3	47.4
GPT 3.5-turbo	0.220***	71.1	57.1	45.9
Gemini 1.0-pro	0.175***	65.8	52.6	43.6
C4AI Command-r	0.199***	60.9	50.9	36.5
Llama-3 8B-IT	0.197***	48.9	44.0	27.1
Mixtral 8x7B-IT	0.164***	48.2	42.0	29.3
Gemma 7B	0.130***	40.8	34.6	25.9
Llama-3 8B	0.193***	36.6	28.0	19.5
Trendyol-LLM 7B-C	0.152***	36.6	26.6	19.5

Table 5: Chain-of-thought results in $\text{TurkishMMLU}_{\text{sub}}$ for selected models with respect to question difficulty. The 'r_{pb}' column shows the point-biserial correlation coefficient, indicating the strength and direction of the relationship between model performance and question difficulty. All models show a significant positive correlation ($p < 0.001$), confirming that model performance decreases as question difficulty increases. Easy, Medium, and Hard labels are based on the 30th and 70th percentiles of the correctness ratio distribution (28% and 41%, respectively).

In our 5-shot non-CoT experiments, we found that Gemini 1.5-pro generates solutions for all questions in the few-shot, even when provided with gold answers. This prevents us from getting predictions for test questions since it exceeds our maximum generation length (it attempts to generate solutions for 5 few-shot questions + 1 test question). This causes mispredictions in many 5-shot non-CoT cases for Gemini 1.5-pro. Therefore, the apparent large improvement (+45.1) between non-CoT and CoT settings for Gemini 1.5-pro is misleading. In the CoT setting, we see that Gemini is the fourth-best model overall, placing it in a competitive position.

4.4 Difficulty Analysis

We analyze model performance across question difficulty levels using the correctness ratio in $\text{TurkishMMLU}_{\text{sub}}$, categorizing questions as Easy, Medium, or Hard based on the 30th and 70th percentiles. Table 5 presents these results along with point-biserial correlation coefficients (r_{pb}), which all show statistically significant positive correlations ($p < 0.001$), confirming that model performance decreases as question difficulty increases. This pattern holds across all models, from smaller ones like Trendyol-LLM 7B-C ($r_{pb} = 0.152$) to state-of-the-art models like GPT 4o ($r_{pb} = 0.211$), *validating the difficulty categorization in TurkishMMLU*. On the other hand, when we apply point-biserial correlation coefficients to the grade instead of the question difficulty, we do not observe any significant correlation ($p > 0.1$) for any of the models. Surprisingly, difficult questions at the lower grades seem to be as hard for models as difficult questions at the higher grades. Models generally perform well on easy questions (up to 96.1% accuracy) but struggle with hard ones (19.5% to 80.1%). We also observe that for some models, the largest differences come from the hard questions. For example, Gemini 1.5-pro is only 6% lower than GPT 4-turbo in easy and medium questions, however the gap is 17% in hard questions.

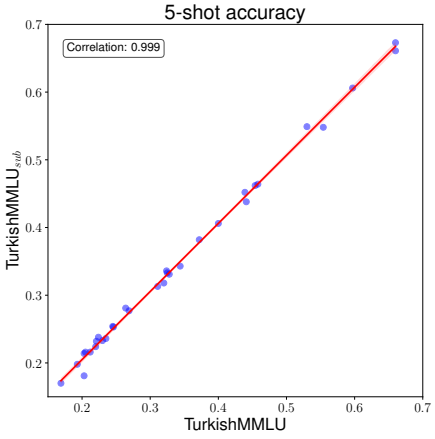


Figure 4: 5-shot accuracy comparison of 32 open-source models for $\text{TurkishMMLU}_{\text{sub}}$ and TurkishMMLU (each point corresponds to an LLM). Pearson's r correlation between them is 0.999.

4.5 Small Set - All Set Correlation

To reduce the inference time and cost of the experiments, many analyses in this paper are conducted on $\text{TurkishMMLU}_{\text{sub}}$. In this section, we computed 5-shot average scores for the open-source models in both the small and full sets. The correlation plot is shown in Figure 4. Pearson's r correlation between the two sets is 0.999, confirming that findings based on $\text{TurkishMMLU}_{\text{sub}}$ are likely to hold as well for TurkishMMLU.

5 Conclusion

In this study, we introduced TurkishMMLU, the first Turkish multitask Question Answering benchmark designed for evaluating LLMs. Our dataset

consists of 10,032 multiple-choice questions covering nine subjects from the Turkish high school curriculum and university entrance exams, complete with correctness ratios to indicate question difficulty. We evaluated a wide range of LLMs, including Turkish-adapted and multilingual models, in various setups such as zero-shot, few-shot, and chain-of-thought reasoning. Our results highlighted the superior performance of closed-source models like GPT 4o and Claude-3 Opus and the notable improvements in newer open-source autoregressive models like Llama-3 70B-IT. The benchmark demonstrates significant performance variation by subject and question difficulty, emphasizing the strengths and limitations of current LLMs in understanding and reasoning in Turkish. Furthermore, as LLMs mature, it will become increasingly crucial to shift the focus of the field from English to broader coverage of the languages of the world. We see TurkishMMLU as a promising contribution towards ensuring that all language communities will be equally served by NLP in the future.

6 Limitations

While we believe TurkishMMLU will significantly contribute to Turkish NLP and the design of next multilingual LLMs, it does have some limitations. First, TurkishMMLU is focused solely on text-based assessment. Exploring multimodal questions that involve images or audio is left for future work. Second, the dataset covers high school curriculum and university entrance exam questions in a multiple-choice format. However, future efforts should aim to expand Turkish benchmarking datasets to include assessments of generative abilities and more open-ended questions. One other limitation of our study is the potential risk of knowledge leakage, as some large language models (LLMs) may have been pre-trained on datasets that overlap with or are sourced from similar data used in our benchmarks, which could artificially inflate their performance. Although the platform does not display gold answers in its interface, models that encountered these questions during pretraining may still perform better than others. As noted by Pezeshkpour and Hruschka (2024), the order of choices may cause variations in model performance; however, we maintained the original order of the choices in our dataset.

Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (project SCHU 2246/14-1).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. 2024. Bridging the bosphorus: Advancing turkish large language models through strategies for low-resource language adaptation and benchmarking.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

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

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

"Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin". 2024. "arabicmmlu: Assessing massive multitask language understanding in arabic".

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

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal

Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

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

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean.

Fatih Soygazi, Okan Çiftçi, Uğurcan Kök, and Soner Cengiz. 2021. Thquad: Turkish historic question answering dataset for reading comprehension. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 215–220.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo,

Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese,

Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jin-

dal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez

Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi

M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024a. Gemini: A family of highly capable multimodal models.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024b. Gemma: Open models based on gemini research and technology.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for

diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

A Question Examples

B Leaderboard

For a comprehensive overview of model performance across all nine subjects, we provide detailed leaderboard in this sectoon. Table 6 presents the 5-shot evaluation scores for 43 models, covering a wide range of LLMs. This detailed breakdown allows for a deeper analysis of model performance variations across different subjects, providing valuable insights into the strengths and weaknesses of each model.

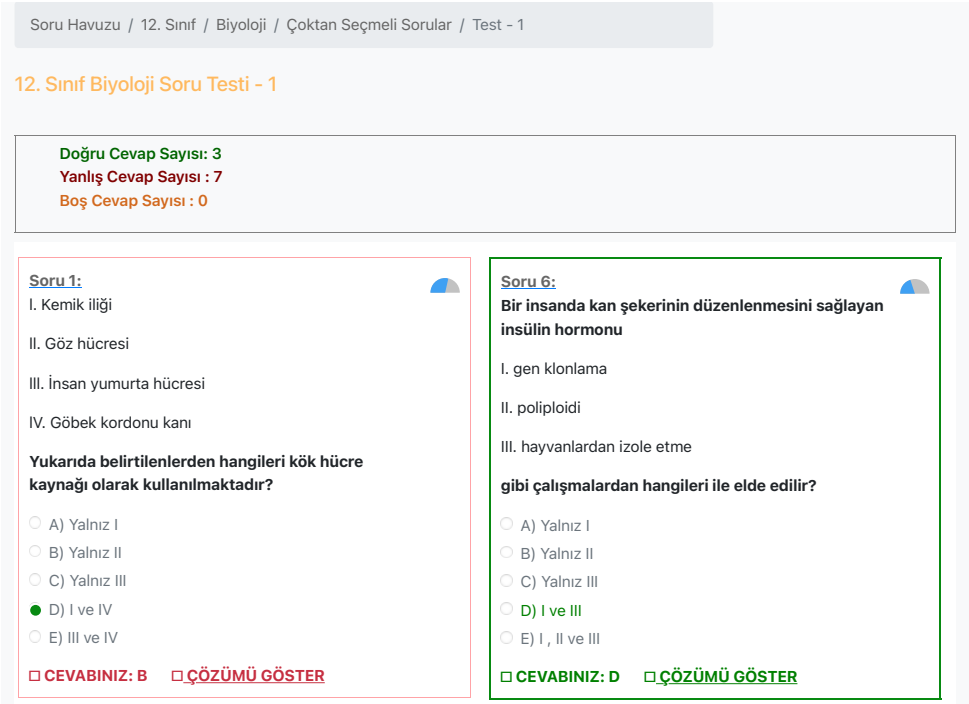


Figure 5: Sample biology test from the EBA Platform, the original version of Figure 2 in Turkish.

Model	Source	All	Biology	Physics	Chemistry	Math	Turkish	History	Geography	Philosophy	R&E
GPT 4o	Closed	**83.1**	78.0	**77.0**	**71.0**	**59.0**	**82.0**	**96.0**	**95.0**	**98.0**	92.0
Claude-3 Opus	Closed	79.1	**82.0**	76.0	57.0	59.0	77.0	87.0	87.0	91.0	**96.0**
GPT 4-Turbo	Closed	75.7	73.0	76.0	62.0	57.0	67.0	83.0	88.0	89.0	86.0
Llama-3 70B-IT	Open	67.3	59.0	59.0	52.0	42.0	57.0	86.0	85.0	85.0	81.0
Claude-3 Sonnet	Closed	67.3	76.0	64.0	62.0	44.0	58.0	75.0	77.0	86.0	64.0
Llama-3 70B	Open	66.1	66.0	51.0	51.0	37.0	57.0	81.0	83.0	89.0	80.0
Claude-3 Haiku	Closed	65.4	61.0	61.0	49.0	40.0	61.0	71.0	80.0	85.0	81.0
Gemini 1.0-pro	Closed	63.2	63.0	53.0	42.0	29.0	63.0	76.0	75.0	86.0	82.0
C4AI Command-r+	Open	60.6	57.0	50.0	43.0	26.0	57.0	75.0	69.0	85.0	83.0
Aya-23 35B	Open	55.6	42.0	45.0	43.0	31.0	49.0	61.0	73.0	78.0	78.0
C4AI command-r	Open	54.9	52.0	44.0	38.0	29.0	49.0	65.0	67.0	78.0	72.0
Mixtral 8x22B	Open	54.8	44.0	41.0	51.0	27.0	49.0	63.0	72.0	75.0	71.0
GPT 3.5-turbo	Closed	51.0	47.0	43.0	38.0	39.0	45.0	58.0	57.0	72.0	60.0
Llama-3 8B-IT	Open	46.4	38.0	41.0	31.0	29.0	39.0	51.0	51.0	65.0	73.0
Llama-3 8B	Open	46.2	37.0	38.0	37.0	30.0	33.0	51.0	53.0	71.0	66.0
Mixtral 8x7B-IT	Open	45.2	43.0	46.0	35.0	28.0	39.0	47.0	48.0	60.0	61.0
Aya-23 8B	Open	45.0	40.0	42.0	35.0	23.0	31.0	53.0	52.0	69.0	60.0
Gemma 7B	Open	43.6	33.0	41.0	29.0	22.0	47.0	47.0	55.0	63.0	55.0
Aya-101	Open	40.7	30.0	32.0	32.0	14.0	38.0	42.0	38.0	74.0	66.0
Trendyol-LLM 7B-C	Open	38.0	28.0	31.0	32.0	18.0	35.0	47.0	51.0	55.0	45.0
Trendyol-LLM 7B-C-D	Open	34.1	29.0	33.0	29.0	22.0	28.0	41.0	50.0	39.0	36.0
mT0-xxl	Open	33.9	34.0	29.0	25.0	28.0	21.0	27.0	40.0	43.0	58.0
Mistral 7B-v0.2	Open	33.1	32.0	39.0	30.0	27.0	34.0	31.0	35.0	38.0	32.0
Mistral 7B-v0.1	Open	32.9	31.0	39.0	26.0	28.0	31.0	29.0	35.0	43.0	34.0
Mistral 7B-IT	Open	32.0	32.0	39.0	32.0	26.0	38.0	20.0	35.0	40.0	26.0
Trendyol-LLM 7B	Open	31.7	24.0	29.0	31.0	19.0	31.0	33.0	31.0	46.0	41.0
mT0-xl	Open	28.1	26.0	28.0	24.0	21.0	25.0	30.0	20.0	41.0	38.0
Gemma 7B-IT	Open	27.3	28.0	26.0	26.0	25.0	25.0	25.0	30.0	31.0	30.0
Phi-3-mini-4k-instruct	Open	26.1	28.0	30.0	24.0	27.0	27.0	26.0	30.0	25.0	18.0
Llama-2 13B-C	Open	25.8	27.0	33.0	23.0	27.0	23.0	23.0	19.0	33.0	24.0
Llama-2 13B	Open	25.6	28.0	28.0	24.0	31.0	22.0	23.0	25.0	25.0	24.0
Gemini 1.5-pro	Closed	25.0	23.0	22.0	19.0	24.0	47.0	14.0	29.0	25.0	22.0
mT5-base	Open	23.8	26.0	25.0	19.0	19.0	21.0	30.0	28.0	23.0	23.0
Gemma 2B	Open	23.4	28.0	29.0	19.0	16.0	22.0	21.0	25.0	28.0	23.0
Gemma 2B-IT	Open	23.2	33.0	22.0	25.0	19.0	22.0	17.0	26.0	28.0	17.0
Llama-2 7B-C	Open	23.2	19.0	25.0	19.0	26.0	19.0	23.0	23.0	27.0	28.0
Llama-2 7B	Open	22.3	25.0	30.0	21.0	26.0	20.0	16.0	21.0	25.0	17.0
mT5-xl	Open	21.6	25.0	23.0	26.0	15.0	22.0	20.0	18.0	19.0	26.0
mT0-large	Open	21.6	16.0	16.0	27.0	23.0	21.0	19.0	19.0	26.0	27.0
mT0-base	Open	21.4	21.0	19.0	21.0	25.0	19.0	22.0	18.0	22.0	26.0
Kanarya 2B	Open	19.8	23.0	17.0	18.0	18.0	18.0	21.0	21.0	17.0	25.0
mT5-xxl	Open	18.1	19.0	20.0	19.0	24.0	14.0	19.0	19.0	17.0	12.0
mT5-large	Open	17.0	14.0	15.0	18.0	17.0	27.0	12.0	19.0	19.0	12.0

Table 6: 5-Shot Experiments for all models on TurkishMMLU$_{sub}$. The Turkish column refers to the subject of the Turkish Language and Literature, while R&E is the Religion and Ethics course.

Model	All (macro)	All (micro)	Biology	Physics	Chemistry	Math	Turkish	History	Geography	Philosophy	R&E
Meta-Llama-3-70B	66.0	63.3	65.4	59.6	53.4	30.6	58.0	77.9	78.7	89.8	81.1
Meta-Llama-3-70B-Instruct	66.0	63.7	62.9	62.8	53.5	35.1	57.9	76.6	80.7	87.2	77.4
c4ai-command-r-plus	59.7	56.7	51.8	48.4	42.2	25.0	56.1	74.5	73.9	87.4	77.8
Mixtral-8x22B-v0.1	55.4	53.1	49.6	48.3	44.2	32.3	46.1	63.9	66.3	79.1	68.4
aya-23-35B	53.9	50.6	45.6	47.3	36.9	24.2	47.6	62.6	66.6	81.7	72.3
c4ai-command-r-v01	53.0	50.2	45.8	43.5	36.1	26.0	48.1	62.4	65.7	80.4	68.8
Meta-Llama-3-8B-Instruct	45.8	43.4	39.5	37.8	34.9	23.9	38.8	52.3	55.9	68.2	60.4
Meta-Llama-3-8B	45.4	43.4	39.1	33.9	36.1	25.1	39.3	51.3	55.7	72.0	56.5
aya-23-8B	44.3	41.8	37.7	38.6	33.4	21.2	36.8	50.8	53.4	71.6	55.2
gemma-7b	44.1	41.9	36.4	37.8	34.4	25.6	36.5	47.9	54.0	69.0	55.2
Mixtral-8x7B-Instruct-v0.1	43.9	41.9	39.7	37.6	33.1	26.9	36.9	47.7	53.4	62.7	56.7
aya-101	40.0	37.2	30.6	30.1	29.3	19.7	37.8	44.8	45.0	68.4	54.2
Trendyol-LLM-7b-chat-v1.0	37.2	35.9	33.4	35.0	30.7	22.5	36.5	41.8	42.5	51.2	41.5
Trendyol-LLM-7b-chat-dpo-v1.0	34.4	33.4	31.2	32.1	29.2	20.5	34.2	39.1	40.3	47.6	35.7
Mistral-7B-v0.1	32.8	32.1	32.5	31.9	31.4	23.7	28.6	34.3	36.8	43.1	32.7
Mistral-7B-v0.2-hf	32.5	31.9	34.3	35.7	30.7	21.2	29.1	33.6	37.6	39.5	31.0
mt0-xxl	32.4	30.2	30.0	34.5	26.8	20.2	19.1	28.6	38.9	42.8	50.7
Mistral-7B-Instruct-v0.2	32.0	31.4	31.1	32.4	30.4	23.5	32.5	29.7	36.8	43.1	28.3
Trendyol-LLM-7b-base-v1.0	31.1	30.0	28.7	26.2	27.2	20.2	28.3	34.7	33.9	45.9	35.3
gemma-7b-it	26.9	26.4	25.4	26.2	25.7	23.3	23.7	28.9	27.2	33.4	27.9
mt0-xl	26.4	24.7	23.8	23.1	22.7	19.2	22.0	26.8	22.7	40.7	36.5
Phi-3-mini-4k-instruct	25.7	25.8	26.9	28.6	27.7	24.7	23.2	25.5	25.4	30.3	18.9
Llama-2-13b-chat-hf	24.6	24.0	23.4	31.3	23.2	20.4	22.5	26.2	22.8	32.6	19.5
Llama-2-13b-hf	24.5	24.1	25.4	25.5	23.0	21.4	22.3	24.7	24.4	29.0	24.8
gemma-2b	23.5	23.2	23.4	30.6	21.9	21.4	20.8	25.1	23.1	25.3	19.7
gemma-2b-it	23.0	23.0	23.2	21.3	23.0	22.3	22.9	23.8	22.6	27.9	20.1
mt5-base	22.4	22.1	24.0	30.1	21.9	18.6	19.1	23.1	23.1	21.1	20.5
Llama-2-7b-chat-hf	22.1	21.9	19.4	24.4	22.9	21.4	21.2	22.5	22.3	24.3	20.3
Llama-2-7b-hf	22.0	22.0	21.3	21.8	23.6	21.4	22.2	21.7	21.4	24.3	20.3
mt5-xl	21.2	21.1	22.0	19.0	23.1	20.1	20.7	20.7	20.2	19.9	25.1
mt0-large	20.4	20.2	17.7	15.9	22.7	19.9	20.6	19.9	19.1	23.0	25.1
mt5-xxl	20.3	20.4	22.6	21.3	21.1	20.8	20.4	19.6	19.0	20.9	17.2
mt0-base	20.3	20.6	18.1	17.0	23.2	21.4	20.7	20.4	21.5	21.1	18.9
kanarya-2b	19.3	19.5	20.5	18.0	17.6	20.8	19.3	19.1	20.5	20.4	17.7
mt5-large	16.9	17.0	13.9	13.6	15.3	19.1	20.0	17.4	18.2	18.0	16.8

Table 7: 5-Shot Experiments for open models on TurkishMMLU. The Turkish column refers to the subject of the Turkish Language and Literature, while R&E is the Religion and Ethics course.