CETVEL: A Unified Benchmark for Evaluating Language Understanding, Generation and Cultural Capacity of LLMs for Turkish

Abrek Er[1*] **Ilker Kesen**[3*] **Gözde Gül Şahin**[1,2] **Aykut Erdem**[1,2]
[1] KUIS AI Center [2] Department of Computer Engineering, Koç University
[3]Department of Computer Science, University of Copenhagen
[*]Equal Contribution aber@ku.edu.tr

Abstract

We introduce CETVEL, a comprehensive benchmark designed to evaluate large language models (LLMs) in Turkish. Existing Turkish benchmarks often lack either task diversity or culturally relevant content, or both. CETVEL addresses these gaps by combining a broad range of both discriminative and generative tasks ensuring content that reflects the linguistic and cultural richness of Turkish language. CETVEL covers 23 tasks grouped into seven categories, including tasks such as grammatical error correction, machine translation, and question answering rooted in Turkish history and idiomatic language. We evaluate 33 open-weight LLMs (up to 70B parameters) covering different model families and instruction paradigms. Our experiments reveal that Turkish-centric instruction-tuned models generally underperform relative to multilingual or general-purpose models (e.g. Llama 3 and Mistral), despite being tailored for the language. Moreover, we show that tasks such as grammatical error correction and extractive question answering are particularly discriminative in differentiating model capabilities. CETVEL offers a comprehensive and culturally grounded evaluation suite for advancing the development and assessment of LLMs in Turkish.

1 Introduction

Large language models (LLMs) have recently achieved remarkable performance on widely used English-centric benchmarks such as (Super)GLUE (Wang et al., 2018, 2019) and MMLU (Hendrycks et al., 2021). Their success across a broad spectrum of tasks and domains (Jiang et al., 2023; Touvron et al., 2023; Yang et al., 2024) has spurred the development of evaluation suites in languages beyond English (Park et al., 2021; Elmadany et al., 2023; Nielsen, 2023). In this work, we extend these efforts to Turkish by introducing CETVEL[1] , a comprehensive benchmark designed to evaluate LLMs across a diverse set of natural language processing (NLP) tasks, with a particular emphasis on cultural and linguistic relevance to Turkish.

Existing Turkish NLP benchmarks typically suffer from one or both of the following limitations: insufficient task diversity and a lack of culturally relevant content. CETVEL addresses both shortcomings. First, it provides broad task coverage, extending well beyond the multiple-choice question answering (MCQA) format predominant in recent Turkish benchmarks (Yüksel et al., 2024; Bayram et al., 2024; Alhajar, 2024). Specifically, CETVEL includes 23 tasks grouped into seven categories: Text Classification (TC), Multiple Choice Question Answering (MCQA), Extractive Question Answering (QA), Grammatical Error Correction (GEC), Machine Translation (MT), Summarization (SUM), and Natural Language Inference (NLI).

Second, CETVEL prioritizes content deeply rooted in Turkish language and culture, an aspect often missing in multilingual or machine-translated benchmarks, which tend to reflect Western cultural biases (Singh et al., 2024; Acikgoz et al., 2024). To counter this, CETVEL includes tasks based on grammatical error correction, figurative language processing, and extractive QA centered on Turkish and Islamic history. We also introduce a novel circumflex-based word sense disambiguation task[2], further enriching the benchmark's linguistic specificity.

We evaluate 33 open-weight LLMs on CETVEL, spanning a broad range of parameter scales (1B to 70B) and model families (e.g., Llama 3, Qwen2.5), including both general-purpose/multilingual and Turkish-specific models. Among all models, Llama 3 variants consistently deliver the strongest overall performance within their respective size categories. However, more importantly, our results

[1] CETVEL means *ruler* in Turkish, i.e. a rectangular shaped object used for measuring the distance between two points.

[2] In Turkish, *hala* means aunt, whereas *hâlâ* means still.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1052–1085

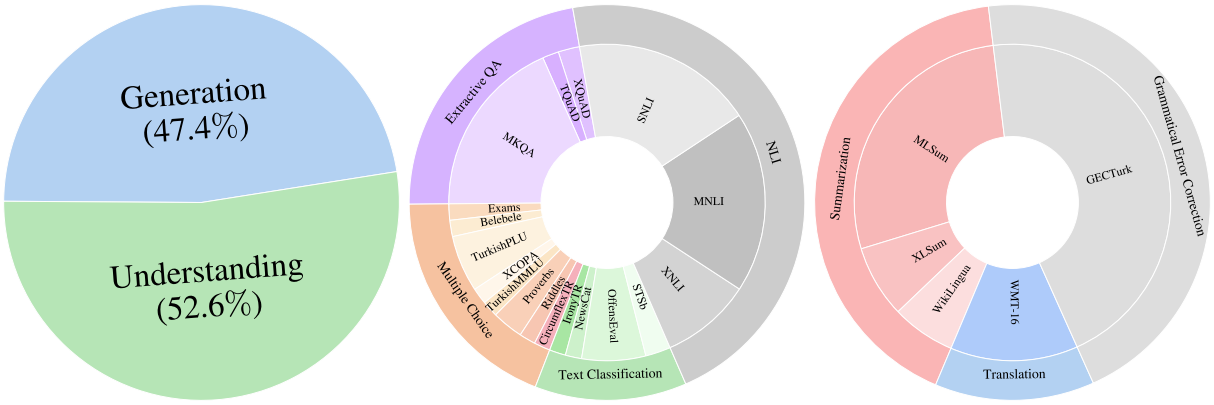

Figure 1: **Task taxonomy in the CETVEL benchmark.** The leftmost pie chart illustrates the overall distribution of tasks across two primary categories: **language understanding** and **language generation**. The middle chart details the subtypes within the understanding category, including extractive question answering, multiple-choice QA, text classification, and natural language inference, along with their associated datasets. The rightmost chart breaks down the generation tasks into three subtypes: summarization, machine translation, and grammatical error correction.

show that most instruction-tuned LLMs specifically developed for Turkish do not outperform general-purpose models such as Llama 3 variants and Mistral. Our findings suggest that Turkish-centric LLMs can benefit from improved instruction-tuning, continued pretraining, and more rigorous validation strategies. Nonetheless, we find that there exists some exceptions: Cere-Llama-3-8B achieves the best performance on grammatical error correction and extractive question answering about Turkish and Islamic history, even outmatching the 70 billion parameter model Llama-3.3-70B-Instruct.

Additionally, to better understand task-level variability, we assess the informativeness of each task using Gini coefficient-based analysis. Our findings indicate that grammatical error correction and extractive QA are particularly effective in differentiating model capabilities, positioning these tasks as highly valuable resources for benchmarking LLMs in Turkish.

Our contributions are as follows:

- We present CETVEL[3], a new Turkish LLM benchmark that combines broad task diversity with high linguistic and cultural relevance.
- We evaluate 33 open-weight LLMs spanning multiple families, language specializations, and parameter scales (up to 70B).
- Our results reveal that most Turkish-centric models do not outperform general-purpose LLMs such as Llama 3 and Mistral at comparable scales.
- We exceptionally find that Cere-Llama-3-8B excels among all other Turkish-centric LLMs, even surpassing a 70 billion parameter model on grammatical error correction and extractive QA about Turkish and Islam history.
- We identify grammatical error correction and extractive QA as the most informative tasks for evaluating Turkish LLMs.

2 Related Work

2.1 LLM Benchmarks

Early benchmark suites such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) have been pivotal in evaluating English-centric language understanding for smaller-scale models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). As LLMs have evolved (LLaMA-Team, 2024; Yang et al., 2024), more challenging benchmarks have emerged, targeting skills such as commonsense reasoning (Bisk et al., 2020), mathematical problem-solving (Cobbe et al., 2021), coding proficiency (Liu et al., 2023, 2024b), and domain-specific knowledge (e.g., scientific QA) (Hendrycks et al., 2020; Rein et al., 2024). Instruction-tuned models are further assessed through alignment and safety benchmarks (Bai et al., 2024). Inspired by GLUE and SuperGLUE, CETVEL brings a similar unification of tasks and datasets but with a specific focus on evaluating LLMs in Turkish.

[3]Code and data: KUIS-AI/cetvel

Benchmark	NLG	Native	Culture	# Cat.	# Tasks
CETVEL	✓	✓	✓	7	23
Mukayese	✓	✓	×	7	11
TurkishMMLU	×	✓	✓	1	9
TUMLU	×	✓	✓	1	9
XTREME	×	✓	×	4	9
TR-MMLU	×	✓	✓	1	62
OpenLLM Turkish	✓	×	×	2	6

Table 1: An overview of existing Turkish NLP benchmarks. **NLG** indicates the presence of generation tasks, while **Native** and **Culture** indicate whether tasks are native-authored and culturally grounded. **# Cat.** and **# Tasks** report the number of task categories and tasks. In CETVEL, TurkishMMLU$_{sub}$ is included as an aggregated subset counted as one task. Overall, CETVEL uniquely combines linguistic and cultural grounding with broad task diversity.

2.2 Multilingual Benchmarks

Multilingual benchmarks such as XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), and XGLUE (Liang et al., 2020) include Turkish among other languages, covering a variety of tasks like question answering (QA), natural language inference (NLI), machine translation (MT), and named entity recognition (NER). However, these benchmarks typically provide only one or two datasets per task, limiting their comprehensiveness. The MEGA benchmark (Ahuja et al., 2023) extends multilingual evaluation by focusing on generative LLMs, featuring 16 datasets, including XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019), XLSum (Hasan et al., 2021a), and WikiANN (Rahimi et al., 2019), across 70 languages and multiple evaluation settings (monolingual, translated, zero-shot cross-lingual). Despite this breadth, MEGA reports significant performance gaps between Turkish and English, as well as among other non-English languages. Recent efforts such as M2QA (Engländer et al., 2024) and SeaEval (Wang et al., 2023) show that LLM performance can vary significantly by domain and language. Additionally, large-scale multilingual classification benchmarks have been developed (Ma et al., 2025; Adelani et al., 2024). Notably, TUMLU (Isbarov et al., 2025) evaluates LLMs across eight Turkic languages and 11 school-subject domains, offering a culturally aware assessment framework for Turkic language models.

2.3 Benchmarks tailored for Turkish

Several benchmarks have been designed specifically for Turkish NLP. Mukayese (Safaya et al., 2022) includes seven tasks, primarily targeting the evaluation of pre-LLM-era multilingual models using fine-tuning. Benchmarks such as Acikgoz et al. (2024) and Alhajar (2024) are built from machine-translated datasets, including ARC (Clark et al., 2018), TruthfulQA (Lin et al., 2022), and GSM8K (Cobbe et al., 2021). TurkishMMLU (Yüksel et al., 2024), a localized version of MMLU, provides 10K high-school-level questions spanning nine subjects with zero-shot and few-shot evaluations. TR-MMLU (Bayram et al., 2024) expands this further with 6,200 questions across 62 categories, including law and healthcare.

Table 1 compares CETVEL against prior benchmarks, where CETVEL advances beyond these benchmarks in three key ways:

(i) it covers a broader set of 23 tasks that include both discriminative and generative settings, unlike MCQA-heavy benchmarks such as TurkishMMLU, TR-MMLU, and TUMLU;

(ii) it includes tasks explicitly designed around Turkish linguistic and cultural content, a critical shortcoming of machine-translated benchmarks;

(iii) it is more comprehensive and up-to-date than Mukayese, evaluating a much wider range of open-weight LLMs (up to 70B parameters) and replacing earlier tasks such as language modeling, sentence segmentation, and spell checking with instruction-oriented tasks like summarization, grammatical error correction, and culturally grounded question answering.

3 Tasks and Datasets

CETVEL includes a diverse set of tasks designed to comprehensively evaluate large language models in Turkish. These tasks are grouped into two high-level categories: natural language understanding (NLU) and natural language generation (NLG). In total, CETVEL spans 23 tasks drawn from publicly available benchmarks and curated datasets, with particular emphasis on linguistic and cultural relevance. Figure 1 provides a visual breakdown of task categories and subtypes.

3.1 Language Understanding Tasks

NLU tasks are organized into four subcategories: (i) extractive question answering (QA), (ii) multiple-choice question answering (MCQA), (iii) text classification (TC), and (iv) natural language inference (NLI).

Extractive Question Answering

In extractive QA, the model is presented with a question and a contextual passage that contains the answer. The objective is to extract the correct answer span from the context. CETVEL includes the following resources: **XQuAD** (Artetxe et al., 2020) extends the English SQuAD dataset (Rajpurkar et al., 2016) with crowd-sourced translations into 11 languages, including Turkish. **MKQA** (Longpre et al., 2021) offers 10K aligned question–answer pairs across 26 languages. It lacks contextual passages, we retain this original format. **TQuAD** contains context-based questions on Turkish and Islamic history, making it uniquely suited for culturally grounded QA in Turkish[4].

Multiple Choice Question Answering

The multiple-choice question answering (MCQA) is well-suited for zero-shot evaluation, and hence it serves as the main format in many recent NLP benchmarks. In CETVEL, we include datasets spanning three subdomains:

(i) **Exam-style Assessments: Exams** (Hardalov et al., 2020) features 393 questions drawn from Turkish high-school subjects such as mathematics and religion. **TurkishMMLU** (Yüksel et al., 2024) is an adaptation of MMLU offering 10k questions across a wide range of academic domains; we use their provided subset of 900 questions. **Belebele** (Bandarkar et al., 2024) includes 900 reading comprehension questions, translated by professionals into 122 languages, including Turkish.

(ii) **Procedural and Commonsense Reasoning: Turkish-PLU** (Uzunoglu and Şahin, 2023) contains four tasks adapted from WikiHow, including goal inference, next-event prediction, step inference, and step ordering. **XCOPA** (Ponti et al., 2020) is a multilingual benchmark requiring causal reasoning, in which models must infer either a cause or an effect given a premise.

(iii) **Specific to Turkish: Turkish Proverbs**[5] comprises 1,730 Turkish proverbs paired with definitions from official linguistic resources. Distractors are generated using Llama3.3-70B embeddings. **BilmeceBench**[6] has 442 riddles converted into MCQA format with randomized distractors. **CircumflexTR** is curated specifically for CETVEL, and targets minimal pairs distinguished by the circumflex diacritic (e.g., kar "snow" vs. kâr "interest").

Text Classification

We frame text classification tasks as MCQA by presenting labels as choices. The task involves selecting the most appropriate label for a given input. We include the following four datasets: **OffensEval** (Çöltekin, 2020) for hate speech detection, performed on user-generated social media data. **IronyTR** (Ozturk et al., 2021) for detecting irony within sentences. **STSb-TR**, a machine-translated variant of the STS benchmark, for predicting semantic similarity between two given languages. Lastly, we include NewsCat (Amasyali and Yildirim, 2004) for news article categorization.

Natural Language Inference

Natural language inference (NLI) involves predicting the logical relationship (entailment, contradiction, or neutrality) between a premise and a hypothesis. Similar to the text classification tasks, we treat the NLI task as a multiple-choice question answering task where we use Turkish versions of the widely adopted **XNLI** (Conneau et al., 2018), **SNLI** (Bowman et al., 2015; Budur et al., 2020), and **MNLI** (Williams et al., 2018; Budur et al., 2020) datasets.

3.2 Language Generation Tasks

We include three NLG tasks: summarization, machine translation, and grammatical error correction.

Summarization

The goal is to generate a concise summary from a paragraph-length input. We evaluate models on the Turkish portions of **MLSum** (Scialom et al., 2020) (news summaries), **XLSum** (Hasan et al., 2021b) (single-sentence summaries from BBC articles), and **WikiLingua** (Ladhak et al., 2020) (step-by-step instructional summaries from WikiHow).

Machine Translation

We also evaluate English-to-Turkish and Turkish-to-English translation tasks using **WMT-16** (Bojar et al., 2016). These tasks measure cross-lingual language processing and Turkish generation quality.

Grammatical Error Correction

This task involves correcting grammatical mistakes in Turkish sentences. We use **GECTurk** (Kara

[4] TQuad/turkish-nlp-qa-dataset

[5] datasets/furkanunluturk/turkce-atasozleri

[6] datasets/selimc/bilmecebench

	QA	MC	TC	NLI	SUM	MT	GEC	Avg.
■ Llama-3.3-70B-Instruct	16.1	**60.1**	**58.1**	32.4	79.5	84.3	44.1	**53.5**
■ Aya-Expanse-32B	26.2	55.6	55.3	**43.3**	83.3	**87.6**	4.5	50.8
■ Cere-Llama-3-8B	24.2	44.8	43.7	34.0	79.9	72.7	**46.0**	49.3
■ Aya-23-35B	23.7	48.8	38.0	37.6	78.1	86.9	30.8	49.1
● Llama-3.1-8B	19.3	45.8	44.8	32.2	80.9	82.6	35.3	48.7
● Llama-3-8B	20.9	43.0	40.6	33.9	81.1	81.6	34.1	47.9
■ Llama-3.1-8B-Instruct	18.0	50.1	40.1	36.0	81.6	75.4	31.5	47.5
■ Ministral-2410-8B-Instruct	14.2	42.8	38.0	34.0	81.2	78.2	39.1	46.8
● Qwen2.5-14B	**26.7**	52.6	37.7	34.0	81.2	70.7	18.9	46.0
■ Aya-101-13B	5.2	40.6	42.3	27.8	**85.2**	84.7	32.4	45.5

Table 2: Performances of the top-10 models. Bold-face indicates best performances. Shapes next to model ids denote the model type: Base pretrained models are represented by circles and instruction-tuned LLMs are denoted by squares. Colors indicate the language focus: blue for English-centric, yellow for multilingually-pretrained, and red for Turkish-centric LLMs. **QA** denotes extractive QA, **MC** denotes multiple-choice QA, **TC** denotes text classification, **NLI** denotes natural language inference, **SUM** denotes summarization, **MT** denotes machine translation, and **GEC** denotes grammatical error correction. Llama-3.3-70B-Instruct achieves the best overall performance, followed by Aya-Expanse-32B. Among Turkish-centric LLMs, Cere-Llama-3-8B achieves an exceptional performance, even surpassing Llama-3.3-70B-Instruct on GEC and extractive QA about Turkish history.

et al., 2023), which contains 22k sentence pairs synthetically generated using 25 expert-defined grammar rules.

4 Experimental Setup

This section outlines the models evaluated in CETVEL, the metrics used for assessment, and implementation details of our experimental pipeline.

4.1 Models

We evaluate 33 open-weight models collected from the Huggingface Transformers package (Wolf et al., 2020). Models are grouped into three main categories based on language coverage and pretraining objectives:

General-purpose LLMs

These models are primarily pretrained on English but might include additional language data during pretraining. For this category of models, we cover **Mistral** (Jiang et al., 2023), **Mixtral** (Jiang et al., 2024) and **Llama 3** (LLaMA-Team, 2024).

Multilingual LLMs

These models are pretrained to support a wide range of languages. We include **Aya-101** (Üstün et al., 2024), **Aya-23** (Aryabumi et al., 2024), **Aya-Expanse** (Dang et al., 2024), **Llama 3.1**, **Llama 3.2**, **Llama 3.3** (LLaMA-Team, 2024), and **Qwen2.5** (Yang et al., 2024) models.

Turkish-centric LLMs

These models are either pretrained exclusively on Turkish or further finetuned on Turkish data. We include **Kanarya** (Safaya et al., 2022), **Turna** (Uludoğan et al., 2024), **Commencis-LLM-7B** (Commencis, 2024), **Trendyol-LLM-7B** (Trendyol, 2024), and **Cere-Llama-3-8B** (CerebrumTech, 2024). Kanarya and Turna models are pretrained on solely Turkish. The remaining three models, Commencis-LLM-7B, Trendyol-LLM-7B and Cere-Llama-3-8B are finetuned on Turkish instruction following data by Turkish tech companies. Specifically, Commencis-LLM-7B and Trendyol-LLM-7B use Mistral-7B as base model, and Cere-Llama-3-8B is built upon Llama3-8B model.

Proprietary LLMs

We also include several strong proprietary models, which are GPT-4.1 (OpenAI, 2024a), GPT-4o (OpenAI, 2024b), GPT-5 (OpenAI, 2025) and Claude Haiku 3.5 (Anthropic, 2024).

We further categorize models by their architecture (decoder-only vs. encoder-decoder), training paradigm (pretraining-only vs. instruction-tuned), and parameter count. Within CETVEL, all of the evaluated LLMs have fewer than 70B parameters and are open-weights, publicly available models. Exceptionally, Turna and Aya-101 models employ an encoder-decoder architecture built upon T5 (Raffel et al., 2020).

	QA	MC	TC	NLI	SUM	MT	GEC	Avg.
Llama-3.3-70B-Instruct	32.0	78.9	73.5	52.7	**58.8**	84.0	56.0	**62.3**
GPT-4o-mini	32.3	75.0	68.0	52.3	55.3	86.6	58.0	61.1
GPT-4.1-mini	29.7	67.3	69.2	55.0	53.7	87.5	**63.0**	60.8
GPT-4o	**36.3**	80.8	72.8	51.3	55.3	**88.6**	27.0	58.9
GPT-4.1-nano	33.7	68.5	72.0	44.0	53.2	87.6	49.0	58.3
GPT-4.1	16.3	**82.8**	**76.2**	56.0	41.7	80.7	50.0	57.7
GPT-5	35.7	81.9	73.2	**59.7**	25.6	87.1	29.0	56.0
GPT-5-mini	34.0	75.8	69.8	55.7	18.6	87.1	29.0	52.8
Claude-Haiku-3.5	30.7	49.7	55.2	39.0	54.5	71.3	54.0	50.6
GPT-5-nano	12.3	63.6	64.2	45.7	15.9	83.3	35.0	45.7

Table 3: Performances of the known proprietary models are evaluated on 100 randomly selected samples per task. Llama-3.3-70B-Instruct model is added for comparison with the main results. Bold-face indicates best performances. Llama-3.3-70B-Instruct model has the highest average score.

4.2 Evaluation Metrics

We use standard automatic metrics tailored to each task type. **Language understanding tasks** are evaluated using **accuracy**. For MCQA tasks, candidate answers are scored based on per-token perplexity, and the option with the lowest perplexity is selected. **Extractive QA** is evaluated using **Exact Match (EM)** (Rajpurkar et al., 2016). **Summarization** is evaluated using **BERTScore** (Zhang* et al., 2020) computed with TabiBERT (Türker et al., 2026), **Machine translation** using **COMET-22** (Rei et al., 2022), and **Grammatical Error Correction** using **macro-F1**. We do not employ LLM-as-a-judge metrics (Zheng et al., 2023), as they have been shown to be unreliable in multilingual settings (Fu and Liu, 2025).

4.3 Implementation Details

All experiments are conducted using the **LM Evaluation Harness** (Gao et al., 2024), a framework that supports evaluation of Huggingface-compatible models and integrates with the **vLLM** inference backend (Kwon et al., 2023) for efficient model serving. For NLU tasks, we use a batch size of 4. For generation tasks, we process one instance at a time and limit outputs to a maximum of 64 tokens, following the protocol used in Mukayese (Safaya et al., 2022). We use **beam search** decoding with a beam width of 5 across all generative tasks, ensuring deterministic evaluation. We run experiments for each single model on eight NVIDIA A40 GPUs. Experiment duration depends on the model size, for instance, the entire set of experiments for an 8 billion parameter model completes less than two days. We conduct each experiment with exactly one single forward run per model. Due to the computational constraints, we evaluate proprietary models on a subset of CETVEL, where each task includes 100 randomly sampled examples. As opposed to open-weight models, we force proprietary models to generate open-ended text for language understanding tasks and then parse the generated text to map it to an option or a label. This is because the APIs of proprietary models only generate text and do not provide token-level probability information.

5 Results

We present our evaluation results, focusing on model performance with respect to parameter size, multilingual coverage, and training paradigm. Table 2 shows the top-10 models[7] across all task categories, and Figure 2 visualizes average performance grouped by model architecture and size.

5.1 Overall Results

Our results indicate that LLaMA 3 models consistently outperform alternatives within comparable parameter ranges. The best-performing model overall is **Llama-3.3-70B-Instruct**, which exceeds the second-best model, **Aya-Expanse-32B**, by 2.7 points in average score. Notably, **Llama-3.1-8B** performs comparably to larger models such as **Aya-23-35B** and **Aya-Expanse-32B**, indicating strong performance scaling efficiency. Unexpectedly, base pretrained **Qwen2.5** models outperform their instruction-tuned variants across all parameter sizes, except for the smallest 0.5B model.

[7] See the appendix for full results.

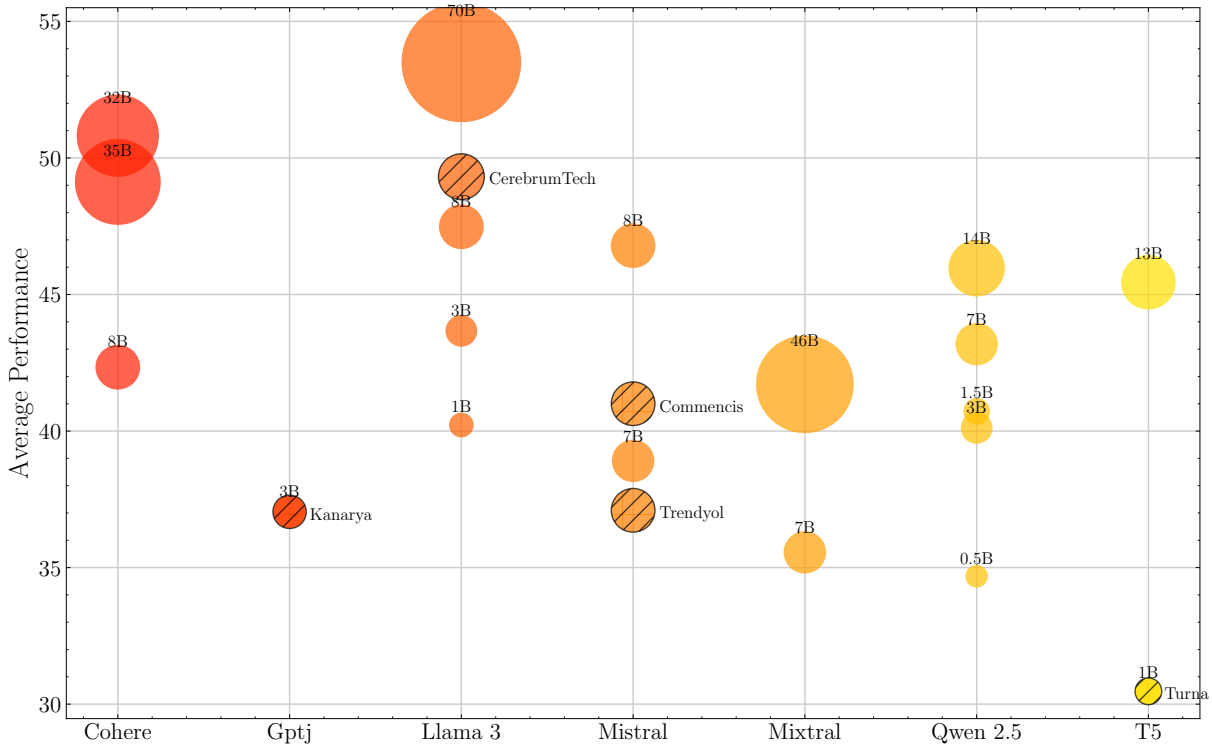

Figure 2: Overall performances on CETVEL grouped by model family. Model size is indicated by the size of the corresponding sphere. A striped sphere indicates that the corresponding model is Turkish-centric LLM. Our experiments reveal that Llama-3.3-70B-Instruct achieves the best overall performance

Turkish-centric instruction-tuned models generally lag behind multilingual and English-centric models. In particular, **Trendyol-LLM-7B** underperforms relative to its base model **Mistral-7B**. Models pretrained from scratch in Turkish also show weak results: **Turna-1B** ranks last overall, and **Kanarya-2B** achieves only an average score of 37.0. One exception among Turkish-centric models is **Cere-Llama-3-8B**, which excels in **Grammatical Error Correction** and **Extractive QA** on culturally specific datasets (e.g., TQuAD), outperforming even Llama-3.3-70B-Instruct on these tasks. However, Cere-Llama-3-8B underperforms in **knowledge-intensive tasks**, likely due to the lack of English exposure and general-domain fine-tuning. This highlights the importance of including cross-lingual and domain-diverse data during instruction tuning for low-resource language models.

5.2 Language Understanding Tasks

For non-generative tasks, **Llama-3.3-70B-Instruct** again leads overall, particularly on knowledge-intensive benchmarks such as Turkish proverbs and riddles. Larger models tend to outperform smaller ones on exam-style tasks (TurkishMMLU, Belebele, Exams), likely due to increased memorization and reasoning capacity. However, performance on commonsense reasoning (e.g., XCOPA) is less sensitive to model size. For instance, **Qwen2.5-0.5B** achieves 53.6% accuracy on XCOPA, just 11.4 points below the strongest model, **Qwen2.5-14B**. On the extractive QA task, **Qwen2.5** models outperform LLaMA models on XQuAD, with **Qwen2.5-14B** ranking highest.

For TQuAD, however, **Cere-Llama-3-8B** achieves the highest score, outperforming all other models despite its smaller size. This result highlights the advantages of task-specific tuning for culturally grounded datasets.

5.3 Language Generation Tasks

On generative tasks, **Aya** models take the lead in summarization and machine translation, likely due to their strong multilingual pretraining. These models particularly benefit from overlapping multilingual content in training corpora (e.g., WikiLingua), which may enhance memorization and transfer. In contrast, **Cere-Llama-3-8B** achieves the best results on grammatical error correction even surpassing 70b parameters model Llama-3.3-70B-Instruct, but performs poorly in summarization and translation tasks-again pointing to the impor-

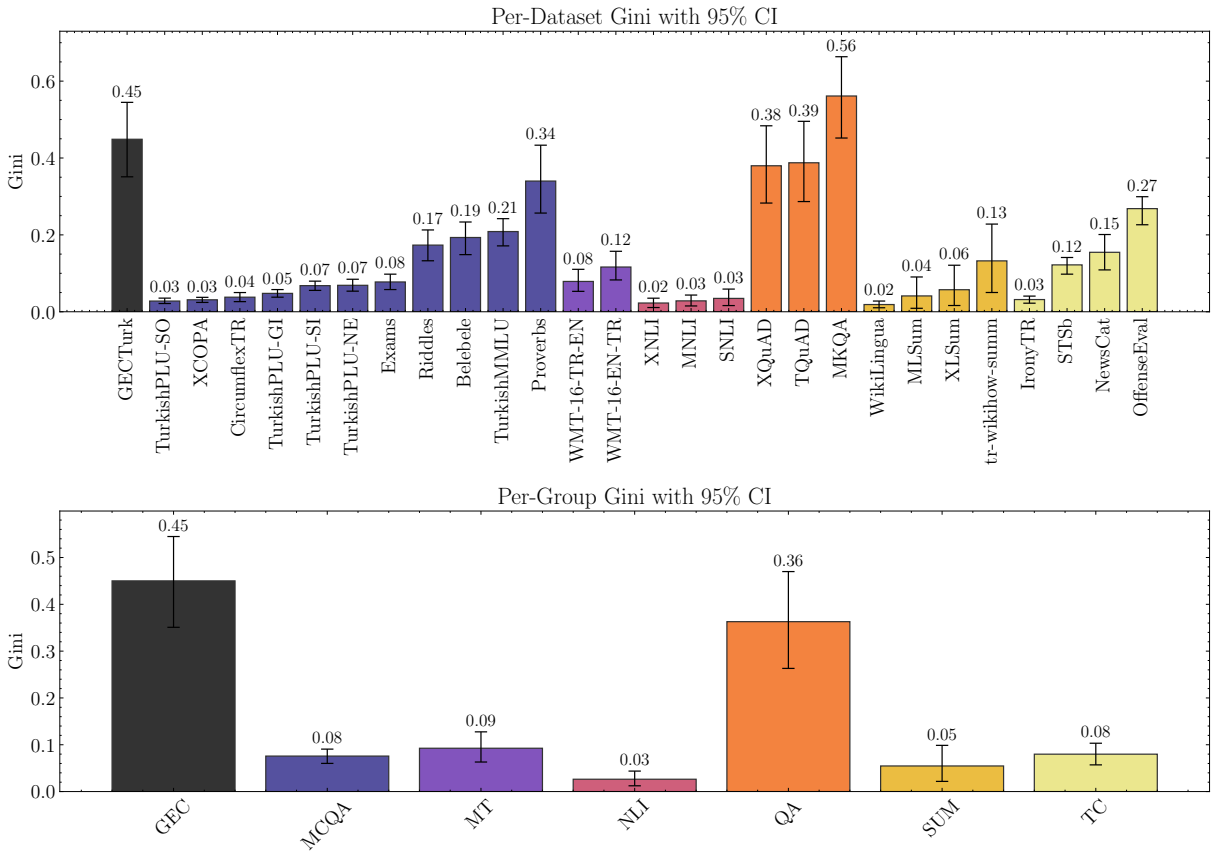

Figure 3: Bar plots with 95% CI of the task-specific and category-wise bootstrapped Gini coefficients. We find that grammatical error correction and extractive question answering tasks are the strongest indicators of differentiating model performances. Remaining tasks contribute less to differentiating model performances.

tance of balanced cross-lingual training. All remaining Turkish-centric models perform extremely poorly on machine translation due to exclusion of English during instruction-tuning phase, where Commencis-7B is the highest performing model, attaining a COMET-22 score of 75.6, followed by Cere-Llama-3-8b with a score of 72.7.

5.4 Turkish-Centric LLMs

Overall, Turkish-centric models fall behind English-centric and multilingual LLMs. This is the case for both LLMs pretrained from scratch on Turkish or LLMs instruction-tuned on Turkish after pretraining. Turkish LLMs pretrained from scratch, Turna-1B and Kanarya-2B, rank in the lower places. LLMs finetuned on Turkish instructions perform better, yet they underperform against their base models. Trendyol-LLM-7B achieves overall performances below its base LLM, Mistral-7B. As we mentioned earlier, they perform poorly on machine translation, due to catastrophic forgetting English (Liu et al., 2024a). These models attain mediocre overall performances, highlighting that there is a large room for improvement in developing LLMs that can effectively process Turkish. As we mentioned earlier, Cere-Llama-3-8B achieves the highest score on TQuAD and GECturk datasets, even outperforming the largest model Llama-3.3-70B-Instruct. Nonetheless, Cere-Llama-3-8B also suffers from catastrophic forgetting like other LLMs instruction-tuned on Turkish.

5.5 Proprietary Models

Table 3 presents the results of the evaluated proprietary models. As the evaluation protocol differs from that used for open-weight models, we additionally include the Llama-3.3-70B-Instruct model to enable a more meaningful comparison. Overall, larger GPT variants achieve over 80% accuracy on several MCQA benchmarks, including those that require external knowledge; however, they lag behind smaller GPT models on the generative tasks of GEC and MT. Nonetheless, the open-weights Llama-3.3-70B-Instruct model outperforms all of the tested proprietary models on CETVEL under the described settings.

5.6 Task Discrimination Analysis

To assess which tasks most effectively differentiate model capabilities, we compute bootstrapped Gini coefficients (Dorfman, 1979) across task categories and within the tasks using the following formula,

$$G = \frac{1}{2\,n^2\,\bar{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|, \quad \bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k. \tag{1}$$

As shown in Figure 3, Grammatical Error Correction (**GEC**) and Question Answering (**QA**) exhibit the highest discrimination power, with coefficients of 0.45 and 0.362, respectively. These tasks consistently produce wide performance gaps across models, making them particularly informative for benchmarking. Conversely, Natural Language Inference (**NLI**), Summarization (**SUM**), Text Classification (**TC**), Machine Translation (**MT**), and Multiple-Choice Question Answering (**MCQA**) have much lower Gini coefficients (0.039, 0.05, 0.08, 0.08, and 0.09, respectively), suggesting less utility in differentiating LLMs in the Turkish setting.

Additionally, some task categories contain tasks with substantially varying discriminative power. For example, the **XCOPA** and **Proverbs** tasks, both from the **MCQA** category, have Gini coefficients of 0.03 and 0.34, respectively, indicating room for improvement in this category. Despite this variability, tasks from the **QA** and **NLI** categories consistently show high and low Gini coefficients, respectively, which may indicate overall category-level discrimination power rather than dataset selection.

6 Conclusion

In this work, we introduced CETVEL, a task-diverse NLP benchmark designed to evaluate large language models in Turkish, with particular attention to linguistic and cultural specificity. CETVEL addresses key limitations of previous efforts, which often lacked task diversity or overlooked culturally grounded content.

By incorporating underexplored phenomena such as proverbs and riddles, CETVEL broadens the scope of evaluation beyond standard NLP tasks and provides a more comprehensive testbed for both multilingual and Turkish-specific models.

Our extensive experiments reveal that instruction-tuned Turkish LLMs consistently underperform compared to general-purpose models that have been pretrained on multilingual corpora including Turkish. These results point to the need for more effective instruction-tuning strategies tailored to Turkish, including higher-quality prompts, culturally relevant tasks, and improved validation pipelines.

Among the evaluated models, Llama 3 variants deliver the strongest overall performance across tasks and parameter ranges. Furthermore, our task discrimination analysis shows that Grammatical Error Correction and Extractive Question Answering are particularly effective in distinguishing model capabilities, while NLI and Text Classification tasks contribute less to differentiation.

We hope CETVEL serves as a valuable resource for advancing Turkish NLP and guiding the development of more robust and culturally aware LLMs.

Limitations

In this version of CETVEL, proprietary model evaluations are constrained by data size, and experiments are conducted on a subset of CETVEL with 100 examples per task. Our analysis is also restricted to the zero-shot setting. While this provides a controlled and reproducible baseline, incorporating one-shot and few-shot evaluations remains an important direction for future iterations of the benchmark. Finally, we note that CETVEL might include user-generated web data, which can be noisy as shown by Cengiz et al. (2025). Nonetheless, we retain these data resources because the underlying information for solving the tasks remains accurate.

Acknowledgments

We thank Mustafa Cemil Güney and Demir Ekin Arıkan for their contributions on the early stages of the development. Ilker Kesen was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101135671 (TrustLLM). This work has been supported by the Scientific and Technological Research Council of Türkiye (TÜBITAK) as part of the project "Automatic Learning of Procedural Language from Natural Language Instructions for Intelligent Assistance" with the number 121C132. We gratefully acknowledge Google for providing Gemini credits that supported this research. We also gratefully acknowledge KUIS AI Center for providing computational support.

References

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. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 242–268, Miami, Florida, USA. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mohamad Alhajar. 2024. Open llm turkish leaderboard. `https://huggingface.co/spaces/malhajar/OpenLLMTurkishLeaderboard`.

MF Amasyali and T Yildirim. 2004. Automatic text categorization of news articles. In *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004.*, pages 224–226. IEEE.

Anthropic. 2024. Claude haiku 3.5.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7421–7454. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

M Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Banu Diri, Savaş Yıldırım, and Öner Aytaş. 2024. Setting standards in turkish nlp: Tr-mmlu for large language model evaluation. *arXiv preprint arXiv:2501.00593*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Ayşe Aysu Cengiz, Ahmet Kaan Sever, Elif Ecem Ümütlü, Naime Şeyma Erdem, Burak Aytan, Büşra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. Evaluating the quality of benchmark datasets for low-resource languages: A case study on turkish. *Preprint*, arXiv:2504.09714.

CerebrumTech. 2024. CerebrumTech/cere-llama-3-8b-tr · Hugging Face.

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. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Commencis. 2024. Commencis/Commencis-LLM · Hugging face.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2qa: Multi-domain multilingual question answering. *arXiv preprint arXiv:2407.01091*.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual llm-as-a-judge? *Preprint*, arXiv:2505.12201.

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. 2024. A framework for few-shot language model evaluation.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nvakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. *arXiv preprint arXiv:2011.03080*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021a. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021b. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Ilshat Saetov, Rinat Kharisov, Saule Belginova, et al. 2025. Tumlu: A unified and native language understanding benchmark for turkic languages. *arXiv preprint arXiv:2502.11020*.

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. *Preprint*, arXiv:2310.06825.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Atakan Kara, Farrin Marouf Sofian, Andrew Bond, and Gözde Şahin. 2023. GECTurk: Grammatical error correction and detection dataset for Turkish. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 278–290, Nusa Dua, Bali. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. 2024a. More than catastrophic forgetting: Integrating general capabilities for domain-specific LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7531–7548, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jiawei Liu, Songrun Xie, Junhao Wang, Yuxiang Wei, Yifeng Ding, and Lingming Zhang. 2024b. Evaluating language models for efficient code generation. In *First Conference on Language Modeling*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

LLaMA-Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. Taxi1500: A dataset for multilingual text classification in 1500 languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Asli Umay Ozturk, Yesim Cemek, and Pinar Karagoz. 2021. Ironytr: Irony detection in turkish informal texts. *International Journal of Intelligent Information Technologies (IJIIT)*, 17(4):1–18.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. 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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

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.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trendyol. 2024. Trendyol/Trendyol-LLM-7b-base-V1.0 · Hugging face.

Melikşah Türker, A. Ebrar Kızıloğlu, Onur Güngör, and Susan Üsküdarlı. 2026. Tabibert: A large-scale modernbert foundation model and a unified benchmark for turkish. *Preprint*, arXiv:2512.23065.

Gökçe Uludoğan, Zeynep Balal, Furkan Akkurt, Meliksah Turker, Onur Gungor, and Susan Üsküdarlı. 2024. TURNA: A Turkish encoder-decoder language model for enhanced understanding and generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10103–10117, Bangkok, Thailand. Association for Computational Linguistics.

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

Arda Uzunoglu and Gözde Şahin. 2023. Benchmarking procedural language understanding for low-resource languages: A case study on Turkish. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 804–819, Nusa Dua, Bali. Association for Computational Linguistics.

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 R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Şenel, Anna Korhonen, and Hinrich Schütze. 2024. Turkishmmlu: Measuring massive multitask language understanding in turkish. *arXiv preprint arXiv:2407.12402*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A AI Assistant Usage

Within this work, we only used AI Assistants for writing purposes. We mainly used chatbots for refining our initial writing, i.e., proof-reading, improving clarity and coherence. We did not use them for generating textual content based on instructions.

B Complete Results

This appendix subsection includes overall and task-specific results for all models tested within CETVEL.

B.1 Overall Results

	QA	MC	TC	NLI	SUM	MT	GEC	Avg.
■ Llama-3.3-70B-Instruct	16.1	**60.1**	**58.1**	32.4	79.5	84.3	44.1	**53.5**
■ Aya-Expanse-32B	26.2	55.6	55.3	**43.3**	83.3	**87.6**	4.5	50.8
■ Cere-Llama-3-8B	24.2	44.8	43.7	34.0	79.9	72.7	**46.0**	49.3
■ Aya-23-35B	23.7	48.8	38.0	37.6	78.1	86.9	30.8	49.1
● Llama-3.1-8B	19.3	45.8	44.8	32.2	80.9	82.6	35.3	48.7
● Llama-3-8B	20.9	43.0	40.6	33.9	81.1	81.6	34.1	47.9
■ Llama-3.1-8B-Instruct	18.0	50.1	40.1	36.0	81.6	75.4	31.5	47.5
■ Ministral-2410-8B-Instruct	14.2	42.8	38.0	34.0	81.2	78.2	39.1	46.8
● Qwen2.5-14B	**26.7**	52.6	37.7	34.0	81.2	70.7	18.9	46.0
■ Aya-101-13B	5.2	40.6	42.3	27.8	**85.2**	84.7	32.4	45.5
● Qwen2.5-7B	20.5	50.6	51.6	34.0	81.1	56.9	22.3	45.3
● Llama-3.2-3B	14.2	40.6	38.7	32.5	80.9	78.5	26.8	44.6
■ Qwen2.5-14B-Instruct	0.4	56.0	43.3	35.2	80.9	82.0	10.9	44.1
■ Llama-3-8B-Instruct	9.0	49.3	40.6	33.8	81.9	83.3	10.0	44.0
■ Llama-3.2-3B-Instruct	15.5	39.1	36.9	34.0	80.0	76.9	16.7	42.7
■ Aya-23-8B	18.5	45.7	45.3	33.3	64.0	85.5	4.3	42.4
■ Aya-Expanse-8B	15.1	52.2	43.6	34.7	58.2	87.0	5.5	42.3
● Qwen2.5-1.5B	22.1	39.8	35.3	33.9	80.9	68.6	14.2	42.1
■ Mixtral-v0.1-7B-Instruct	11.0	45.1	45.6	33.7	73.1	79.9	3.6	41.7
● Qwen2.5-3B	19.7	41.5	40.2	34.0	80.9	60.3	14.7	41.6
● Commencis-7B	5.1	38.1	39.4	32.5	72.6	75.6	23.7	41.0
● Llama-3.2-1B	4.0	37.7	43.7	34.1	80.6	63.5	18.1	40.2
■ Qwen2.5-7B-Instruct	0.5	50.9	48.4	35.0	75.1	64.6	6.6	40.2
■ Qwen2.5-1.5B-Instruct	10.4	42.8	33.0	33.9	80.8	70.7	3.6	39.3
● Mistral-v0.3-7B	16.8	39.0	37.7	34.0	52.6	70.0	23.5	39.1
■ Qwen2.5-3B-Instruct	3.1	45.2	35.7	34.0	79.0	69.2	4.1	38.6
■ Mistral-v0.3-7B-Instruct	8.7	38.6	42.5	31.7	71.9	74.8	1.0	38.4
● Trendyol-v1.0-7B-Base	0.3	38.4	41.8	34.3	80.5	64.3	0.1	37.1
■ Kanarya-2B	1.0	41.0	47.8	33.1	79.9	49.8	6.6	37.0
● Mistral-v0.1-7B	16.6	39.2	37.3	33.9	30.2	70.8	20.8	35.6
■ Qwen2.5-0.5B-Instruct	10.1	36.9	27.8	33.9	80.4	55.7	1.8	35.2
● Qwen2.5-0.5B	3.8	37.8	30.7	33.9	80.6	51.3	0.9	34.1
● Turna-1B	0.0	35.9	36.6	34.1	77.2	29.5	0.0	30.5

Table 4: Overall results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.2 Grammatical Error Correction Results

	GECTurk
■ Cere-Llama-3-8B	**46.0**
■ Llama-3.3-70B-Instruct	44.1
■ Ministral-2410-8B-Instruct	39.1
● Llama-3.1-8B	35.3
● Llama-3-8B	34.1
■ Aya-101-13B	32.4
■ Llama-3.1-8B-Instruct	31.5
■ Aya-23-35B	30.8
● Llama-3.2-3B	26.8
● Commencis-7B	23.7
● Mistral-v0.3-7B	23.5
● Qwen2.5-7B	22.3
● Mistral-v0.1-7B	20.8
● Qwen2.5-14B	18.9
● Llama-3.2-1B	18.1
■ Llama-3.2-3B-Instruct	16.7
● Qwen2.5-3B	14.7
● Qwen2.5-1.5B	14.2
■ Qwen2.5-14B-Instruct	10.9
■ Llama-3-8B-Instruct	10.0
■ Kanarya-2B	6.6
■ Qwen2.5-7B-Instruct	6.6
■ Aya-Expanse-8B	5.5
■ Aya-Expanse-32B	4.5
■ Aya-23-8B	4.3
■ Qwen2.5-3B-Instruct	4.1
■ Mixtral-v0.1-7B-Instruct	3.6
■ Qwen2.5-1.5B-Instruct	3.6
■ Qwen2.5-0.5B-Instruct	1.8
■ Mistral-v0.3-7B-Instruct	1.0
● Qwen2.5-0.5B	0.9
● Trendyol-v1.0-7B-Base	0.1
● Turna-1B	0.0

Table 5: Grammatical error correction results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.3 Multiple Choice Question Answering Results

	XCOPA	PLU	PLU_GI	PLU_NEP	PLU_SI	PLU_SO	Exams	Belebele	Proverbs	TurkishMMLU	BilmeceBench	CircumflexTR	Avg.
■ Llama-3.3-70B-Instruct	65.0	**54.2**	**49.1**	**58.0**	39.1	65.1	**39.2**	**86.8**	**92.5**	**64.6**	**72.6**	**67.1**	**60.1**
□ Qwen2.5-14B-Instruct	**66.6**	48.5	40.6	49.8	35.1	62.2	29.8	84.7	78.3	59.4	57.0	58.6	56.0
□ Aya-Expanse-32B	59.2	51.8	44.8	55.1	39.2	63.0	36.9	83.4	82.4	56.9	41.2	57.1	55.6
○ Qwen2.5-14B	64.6	48.7	41.3	48.7	35.3	62.9	33.1	81.2	75.4	56.2	47.5	58.6	52.6
□ Aya-Expanse-8B	57.8	50.2	43.0	51.1	**40.4**	61.4	31.6	73.6	72.3	46.6	48.9	54.3	52.2
□ Qwen2.5-7B-Instruct	61.8	47.1	41.1	46.7	32.2	61.3	30.3	73.4	71.2	47.6	52.0	54.3	50.9
○ Qwen2.5-7B	59.8	48.3	42.5	47.2	32.2	63.4	29.5	73.9	73.5	49.3	48.4	57.1	50.6
■ Llama-3.1-8B-Instruct	60.8	48.5	40.9	44.6	34.3	**65.7**	32.3	70.8	75.5	38.1	41.6	64.3	50.1
■ Llama-3-8B-Instruct	58.6	47.1	37.6	46.6	33.5	63.5	27.0	66.3	69.5	38.1	38.5	61.4	49.3
□ Aya-23-35B	60.4	48.8	43.0	52.1	35.1	59.7	29.8	72.9	56.9	45.3	34.8	60.0	48.8
● Llama-3.1-8B	62.6	47.6	38.8	46.0	35.1	63.2	31.3	61.4	54.1	30.6	32.1	58.6	45.8
□ Aya-23-8B	59.6	49.3	42.1	48.9	37.3	62.7	27.0	60.7	45.0	33.0	34.4	58.6	45.7
□ Qwen2.5-3B-Instruct	56.2	44.8	38.9	42.0	32.2	59.1	27.5	67.4	60.1	37.8	33.0	54.3	45.2
■ Mixtral-v0.1-7B-Instruct	56.4	47.1	44.6	46.1	31.7	59.1	29.3	58.6	51.5	35.8	34.2	57.1	45.1
■ Cere-Llama-3-8B	60.2	48.7	41.8	46.9	35.9	63.1	28.0	51.4	48.1	25.6	33.9	58.6	44.8
● Llama-3-8B	61.8	46.5	36.9	46.1	32.8	62.8	30.3	51.4	44.0	25.4	29.6	54.3	43.0
■ Ministral-2410-8B-Instruct	57.4	45.3	37.5	44.1	31.9	60.6	31.3	60.9	40.5	26.4	24.9	58.6	42.8
□ Qwen2.5-1.5B-Instruct	54.6	42.5	35.7	40.3	28.1	58.2	22.9	53.4	34.7	28.9	29.2	48.6	42.8
○ Qwen2.5-3B	55.2	43.9	38.0	40.6	29.7	59.5	26.5	61.9	43.5	22.6	24.4	55.7	41.5
■ Kanarya-2B	64.2	49.3	45.9	45.6	35.8	62.5	30.0	28.1	0.0	18.0	27.1	54.3	41.0
□ Aya-101-13B	59.6	41.3	37.4	35.0	27.3	57.1	22.9	22.9	1.0	37.4	47.1	57.1	40.6
● Llama-3.2-3B	57.0	45.4	40.0	43.2	31.5	59.5	29.5	47.3	19.9	29.0	29.0	57.1	40.6
○ Qwen2.5-1.5B	54.2	42.1	35.8	39.7	27.3	57.6	21.6	46.7	23.0	23.0	29.9	50.0	39.8
● Mistral-v0.1-7B	56.6	45.2	42.8	39.5	29.2	60.2	24.2	37.4	30.8	20.3	25.3	57.1	39.2
■ Llama-3.2-3B-Instruct	54.6	44.0	35.5	39.4	27.8	63.7	26.2	55.8	1.1	34.4	31.0	54.3	39.1
● Mistral-v0.3-7B	58.4	43.4	41.0	38.6	26.6	58.4	24.2	41.1	27.6	26.9	23.5	57.1	39.0
■ Mistral-v0.3-7B-Instruct	57.2	43.0	41.2	40.9	27.0	55.3	25.4	46.1	30.0	19.6	21.5	50.0	38.6
● Trendyol-v1.0-7B-Base	61.0	46.9	46.4	43.2	32.4	58.6	28.5	36.2	0.0	24.8	23.1	57.1	38.4
● Commencis-7B	58.0	41.3	34.8	38.6	27.6	56.5	24.7	32.3	22.7	24.7	24.2	58.6	38.1
○ Qwen2.5-0.5B	54.8	40.8	36.2	35.7	26.5	56.5	21.1	29.9	20.3	17.9	25.1	47.1	37.8
● Llama-3.2-1B	55.6	42.1	36.2	37.3	29.2	57.7	28.5	29.6	21.7	18.9	22.4	52.9	37.7
□ Qwen2.5-0.5B-Instruct	53.6	41.6	36.8	35.6	28.1	57.4	23.7	30.0	28.3	21.1	24.2	54.3	36.9
● Turna-1B	55.8	40.3	38.0	38.3	27.3	51.2	23.7	22.6	19.2	19.3	24.2	51.4	35.9

Table 6: Multiple choice question answering results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.4 Machine Translation Results

	WMT16EN-TR	WMT16TR-EN	Avg.
Aya-Expanse-32B	**20.1**	**35.0**	**27.6**
Aya-23-35B	18.5	31.7	25.1
Llama-3.3-70B-Instruct	13.6	34.4	24.0
Aya-Expanse-8B	17.0	30.6	23.8
Aya-23-8B	16.9	28.4	22.6
Llama-3.1-8B	15.6	26.7	21.2
Llama-3-8B-Instruct	13.6	27.7	20.7
Aya-101-13B	17.2	21.5	19.3
Llama-3-8B	11.3	26.9	19.1
Qwen2.5-14B-Instruct	12.4	24.9	18.7
Mixtral-v0.1-7B-Instruct	9.5	24.8	17.2
Llama-3.2-3B-Instruct	8.2	23.7	16.0
Llama-3.1-8B-Instruct	15.6	16.2	15.9
Ministral-2410-8B-Instruct	11.2	20.3	15.8
Llama-3.2-3B	7.9	20.0	13.9
Qwen2.5-14B	8.1	19.3	13.7
Mistral-v0.3-7B-Instruct	5.9	20.9	13.4
Qwen2.5-7B	5.5	18.3	11.9
Mistral-v0.1-7B	5.0	18.9	11.9
Mistral-v0.3-7B	3.8	18.3	11.0
Qwen2.5-1.5B-Instruct	4.7	16.1	10.4
Qwen2.5-7B-Instruct	5.4	15.1	10.2
Qwen2.5-3B-Instruct	6.6	13.3	9.9
Qwen2.5-3B	4.1	15.3	9.7
Qwen2.5-1.5B	4.1	13.9	9.0
Cere-Llama-3-8B	11.9	4.4	8.2
Commencis-7B	4.7	11.6	8.1
Trendyol-v1.0-7B-Base	1.0	10.7	5.9
Llama-3.2-1B	1.2	8.6	4.9
Qwen2.5-0.5B-Instruct	3.1	4.0	3.6
Qwen2.5-0.5B	1.4	3.4	2.4
Kanarya-2B	3.0	1.5	2.3
Turna-1B	0.1	0.3	0.2

Table 7: Machine translation results of the models, sorted by their average BLEU-4 scores (Papineni et al., 2002). Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

	WMT16EN-TR	WMT16TR-EN	Avg.
■ Aya-Expanse-32B	**88.1**	**87.0**	**87.6**
■ Aya-Expanse-8B	87.8	86.2	87.0
■ Aya-23-35B	87.3	86.5	86.9
■ Aya-23-8B	86.7	84.2	85.5
■ Aya-101-13B	87.5	81.9	84.7
■ Llama-3.3-70B-Instruct	82.1	86.4	84.3
■ Llama-3-8B-Instruct	83.4	83.1	83.3
● Llama-3.1-8B	82.0	83.2	82.6
■ Qwen2.5-14B-Instruct	81.3	82.7	82.0
● Llama-3-8B	79.5	83.7	81.6
■ Mixtral-v0.1-7B-Instruct	76.6	83.2	79.9
● Llama-3.2-3B	75.3	81.8	78.5
■ Ministral-2410-8B-Instruct	81.0	75.4	78.2
■ Llama-3.2-3B-Instruct	73.6	80.1	76.9
● Commencis-7B	76.0	75.2	75.6
■ Llama-3.1-8B-Instruct	85.9	64.8	75.4
■ Mistral-v0.3-7B-Instruct	69.3	80.3	74.8
■ Cere-Llama-3-8B	75.5	69.9	72.7
● Mistral-v0.1-7B	63.3	78.2	70.8
■ Qwen2.5-1.5B-Instruct	65.4	76.1	70.7
● Qwen2.5-14B	68.7	72.8	70.7
● Mistral-v0.3-7B	61.8	78.2	70.0
■ Qwen2.5-3B-Instruct	70.8	67.5	69.2
● Qwen2.5-1.5B	62.6	74.6	68.6
■ Qwen2.5-7B-Instruct	56.3	72.9	64.6
● Trendyol-v1.0-7B-Base	62.3	66.3	64.3
● Llama-3.2-1B	61.6	65.3	63.5
● Qwen2.5-3B	53.2	67.4	60.3
● Qwen2.5-7B	50.4	63.5	56.9
■ Qwen2.5-0.5B-Instruct	51.9	59.4	55.7
● Qwen2.5-0.5B	42.8	59.8	51.3
■ Kanarya-2B	49.2	50.3	49.8
● Turna-1B	23.0	36.0	29.5

Table 8: Machine translation results of the models, sorted by their COMET-22 scores (Rei et al., 2022). Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.5 Natural Language Inference Results

	MNLI	SNLI	XNLI	Avg.
Aya-Expanse-32B	**42.5**	**47.1**	**40.5**	**43.3**
Aya-23-35B	37.3	37.3	38.1	37.6
Llama-3.1-8B-Instruct	36.7	35.9	35.4	36.0
Qwen2.5-14B-Instruct	35.3	36.3	34.1	35.2
Qwen2.5-7B-Instruct	35.6	35.5	33.8	35.0
Aya-Expanse-8B	36.1	31.6	36.3	34.7
Trendyol-v1.0-7B-Base	35.2	33.7	34.0	34.3
Llama-3.2-1B	35.0	33.7	33.5	34.1
Turna-1B	34.9	33.8	33.4	34.1
Llama-3.2-3B-Instruct	34.8	33.8	33.4	34.0
Mistral-v0.3-7B	34.9	33.4	33.6	34.0
Qwen2.5-3B-Instruct	34.8	33.7	33.4	34.0
Qwen2.5-14B	34.8	33.7	33.4	34.0
Ministral-2410-8B-Instruct	34.8	33.7	33.4	34.0
Cere-Llama-3-8B	34.9	33.7	33.3	34.0
Qwen2.5-3B	34.8	33.7	33.4	34.0
Qwen2.5-7B	34.8	33.6	33.4	34.0
Qwen2.5-0.5B-Instruct	34.8	33.7	33.3	33.9
Mistral-v0.1-7B	34.8	33.6	33.4	33.9
Qwen2.5-1.5B-Instruct	34.8	33.7	33.3	33.9
Qwen2.5-1.5B	34.8	33.7	33.3	33.9
Llama-3-8B	34.8	33.6	33.4	33.9
Qwen2.5-0.5B	34.7	33.7	33.4	33.9
Llama-3-8B-Instruct	35.1	33.3	33.1	33.8
Mixtral-v0.1-7B-Instruct	34.8	32.2	34.1	33.7
Aya-23-8B	32.7	33.6	33.7	33.3
Kanarya-2B	33.4	31.7	34.2	33.1
Llama-3.2-3B	32.0	32.5	33.0	32.5
Commencis-7B	33.3	31.5	32.7	32.5
Llama-3.3-70B-Instruct	32.1	31.7	33.3	32.4
Llama-3.1-8B	31.9	34.3	30.3	32.2
Mistral-v0.3-7B-Instruct	31.3	31.8	32.0	31.7
Aya-101-13B	27.9	25.6	30.0	27.8

Table 9: Natural language inference results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.6 Open Ended Question Answering Results

	XQUAD	TQUAD	MKQA	Avg.
○ Qwen2.5-14B	**40.3**	34.8	5.0	**26.7**
□ Aya-Expanse-32B	31.9	30.2	16.4	26.2
■ Cere-Llama-3-8B	21.2	**49.2**	2.2	24.2
□ Aya-23-35B	30.9	20.6	**19.4**	23.7
○ Qwen2.5-1.5B	31.8	34.3	0.3	22.1
● Llama-3-8B	20.8	28.5	13.5	20.9
○ Qwen2.5-7B	31.9	28.0	1.4	20.5
○ Qwen2.5-3B	32.3	26.8	0.1	19.7
● Llama-3.1-8B	20.9	27.6	9.2	19.3
□ Aya-23-8B	24.7	20.6	10.0	18.5
■ Llama-3.1-8B-Instruct	21.4	23.3	9.1	18.0
● Mistral-v0.3-7B	17.1	21.9	11.5	16.8
● Mistral-v0.1-7B	16.7	21.0	12.0	16.6
■ Llama-3.3-70B-Instruct	14.5	17.4	16.3	16.1
■ Llama-3.2-3B-Instruct	23.0	18.7	4.7	15.5
□ Aya-Expanse-8B	25.0	13.5	7.0	15.1
● Llama-3.2-3B	15.5	21.2	6.0	14.2
■ Ministral-2410-8B-Instruct	22.9	17.6	2.2	14.2
■ Mixtral-v0.1-7B-Instruct	10.7	9.8	12.5	11.0
□ Qwen2.5-1.5B-Instruct	16.5	14.7	0.2	10.4
□ Qwen2.5-0.5B-Instruct	13.4	16.3	0.7	10.1
■ Llama-3-8B-Instruct	9.7	12.9	4.2	9.0
■ Mistral-v0.3-7B-Instruct	11.4	9.5	5.0	8.7
□ Aya-101-13B	7.6	5.4	2.5	5.2
● Commencis-7B	6.6	5.4	3.2	5.1
● Llama-3.2-1B	4.9	6.3	0.8	4.0
○ Qwen2.5-0.5B	4.0	7.2	0.1	3.8
□ Qwen2.5-3B-Instruct	6.1	3.3	0.1	3.1
■ Kanarya-2B	0.8	1.7	0.6	1.0
□ Qwen2.5-7B-Instruct	0.9	0.6	0.0	0.5
□ Qwen2.5-14B-Instruct	0.9	0.3	0.0	0.4
● Trendyol-v1.0-7B-Base	0.0	0.8	0.1	0.3
● Turna-1B	0.0	0.0	0.1	0.0

Table 10: Open ended question answering results results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.7 Summarization Results

	XLSum	WikiLingua	WikiHowSumm	MLSum	Avg.
■ Aya-101-13B	**24.6**	**22.2**	**16.9**	**31.8**	**23.9**
■ Aya-Expanse-32B	21.4	21.5	16.6	30.1	22.4
■ Aya-23-35B	13.3	18.4	12.7	25.9	17.6
■ Llama-3.3-70B-Instruct	16.3	12.0	8.5	28.1	16.2
■ Llama-3-8B-Instruct	13.5	8.1	7.9	25.8	13.8
■ Llama-3.1-8B-Instruct	12.4	7.9	7.9	25.8	13.5
● Llama-3.1-8B	12.4	7.9	7.9	25.8	13.5
● Qwen2.5-14B	13.1	6.8	6.6	25.7	13.0
■ Ministral-2410-8B-Instruct	13.8	6.9	6.7	24.0	12.8
■ Qwen2.5-14B-Instruct	15.7	7.1	4.7	23.7	12.8
● Qwen2.5-7B	12.3	6.9	6.6	25.3	12.8
■ Aya-23-8B	14.1	11.4	1.7	24.0	12.8
● Llama-3-8B	11.1	6.6	6.6	25.0	12.3
● Qwen2.5-3B	11.4	6.5	6.5	24.9	12.3
■ Aya-Expanse-8B	13.3	13.4	0.5	21.4	12.2
■ Cere-Llama-3-8B	10.9	6.4	6.6	24.8	12.2
● Llama-3.2-3B	12.0	6.4	6.5	23.3	12.1
■ Qwen2.5-3B-Instruct	12.0	6.6	6.8	22.7	12.0
■ Qwen2.5-1.5B-Instruct	11.0	6.1	6.3	24.6	12.0
● Qwen2.5-1.5B	11.9	5.9	6.3	24.0	12.0
■ Llama-3.2-3B-Instruct	11.1	6.9	6.6	22.2	11.7
● Qwen2.5-0.5B	9.2	5.6	6.1	24.3	11.3
■ Mistral-v0.3-7B-Instruct	10.9	6.2	3.7	24.1	11.2
■ Qwen2.5-7B-Instruct	11.6	6.3	5.8	20.5	11.0
■ Kanarya-2B	9.3	4.5	5.3	24.7	10.9
■ Mixtral-v0.1-7B-Instruct	10.6	6.0	4.9	22.2	10.9
● Llama-3.2-1B	9.7	5.6	5.9	22.2	10.8
■ Qwen2.5-0.5B-Instruct	9.0	5.0	5.8	22.9	10.7
● Commencis-7B	9.5	6.3	7.2	16.1	9.8
■ Llama-3.2-3B-Instruct	6.7	4.1	6.6	14.8	8.1
● Mistral-v0.3-7B	6.3	6.0	1.9	17.3	7.9
● Turna-1B	6.1	5.6	5.7	11.0	7.1
● Trendyol-v1.0-7B-Base	5.0	2.9	4.6	8.8	5.3
● Mistral-v0.1-7B	1.2	6.0	1.1	6.3	3.6

Table 11: Summarization results of the models, sorted by their average ROUGE-2 scores (Lin, 2004). Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

	XLSum	WikiLingua	WikiHowSumm	MLSum	Avg.
Aya-101-13B	**84.7**	**86.0**	**84.7**	**85.3**	**85.2**
Aya-Expanse-32B	80.2	85.6	83.8	83.4	83.3
Llama-3-8B-Instruct	81.8	80.9	81.2	83.9	81.9
Llama-3.1-8B-Instruct	81.2	80.7	81.0	83.6	81.6
Qwen2.5-14B	81.3	79.9	80.0	83.4	81.2
Ministral-2410-8B-Instruct	81.3	79.9	80.4	83.1	81.2
Qwen2.5-7B	81.0	80.1	80.2	83.2	81.1
Llama-3-8B	80.6	80.0	80.5	83.3	81.1
Llama-3.1-8B	80.5	79.3	80.6	83.3	80.9
Qwen2.5-3B	80.6	79.9	80.1	83.1	80.9
Qwen2.5-14B-Instruct	82.3	80.0	77.9	83.4	80.9
Qwen2.5-1.5B	80.7	79.6	80.3	82.9	80.9
Llama-3.2-3B	80.9	79.4	80.4	82.8	80.9
Qwen2.5-1.5B-Instruct	80.5	80.0	79.8	83.0	80.8
Qwen2.5-0.5B	80.1	79.4	80.2	82.9	80.6
Llama-3.2-1B	80.2	79.6	79.9	82.5	80.6
Trendyol-v1.0-7B-Base	81.3	78.5	79.6	82.8	80.5
Qwen2.5-0.5B-Instruct	80.0	79.2	79.9	82.5	80.4
Llama-3.2-3B-Instruct	80.6	80.4	76.5	82.5	80.0
Kanarya-2B	80.5	76.3	79.4	83.5	79.9
Cere-Llama-3-8B	78.8	78.0	80.5	82.3	79.9
Llama-3.3-70B-Instruct	82.4	82.2	69.1	84.3	79.5
Qwen2.5-3B-Instruct	78.4	79.4	76.9	81.4	79.0
Aya-23-35B	69.1	84.8	75.4	83.0	78.1
Turna-1B	77.9	77.6	79.9	73.3	77.2
Qwen2.5-7B-Instruct	70.4	79.6	76.6	73.8	75.1
Mixtral-v0.1-7B-Instruct	78.6	70.7	61.7	81.2	73.1
Commencis-7B	73.5	69.2	69.0	78.8	72.6
Mistral-v0.3-7B-Instruct	80.1	79.1	45.8	82.7	71.9
Aya-23-8B	80.1	81.9	12.5	81.4	64.0
Aya-Expanse-8B	66.2	83.1	3.2	80.4	58.2
Mistral-v0.3-7B	47.4	78.9	22.5	61.5	52.6
Mistral-v0.1-7B	9.1	78.9	12.1	20.9	30.2

Table 12: Summarization results of the models, sorted by their average BERTScore (Zhang* et al., 2020) computed with TabiBERT (Türker et al., 2026). Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

B.8 Text Classification Results

	STSb	OffensEval	NewsCat	IronyTR	Avg.
■ Llama-3.3-70B-Instruct	12.9	**83.1**	78.0	58.2	**58.1**
■ Aya-Expanse-32B	21.5	67.1	**82.8**	50.0	55.3
● Qwen2.5-7B	17.0	77.4	54.8	57.3	51.6
■ Qwen2.5-7B-Instruct	18.3	80.3	40.0	55.0	48.4
■ Kanarya-2B	12.9	61.6	66.8	50.0	47.8
■ Mixtral-v0.1-7B-Instruct	13.0	62.9	54.0	52.5	45.6
■ Aya-23-8B	23.0	34.2	72.4	51.7	45.3
● Llama-3.1-8B	17.0	34.6	69.2	58.5	44.8
■ Cere-Llama-3-8B	22.1	34.0	68.4	50.2	43.7
● Llama-3.2-1B	17.1	46.7	58.0	52.8	43.7
■ Aya-Expanse-8B	21.0	26.8	76.0	50.5	43.6
■ Qwen2.5-14B-Instruct	24.9	54.7	32.4	**61.3**	43.3
■ Mistral-v0.3-7B-Instruct	12.9	45.2	61.2	50.7	42.5
■ Aya-101-13B	17.0	79.9	20.0	52.2	42.3
● Trendyol-v1.0-7B-Base	15.5	20.3	81.2	50.0	41.8
■ Llama-3-8B-Instruct	14.2	30.8	62.8	54.5	40.6
● Llama-3-8B	16.4	21.9	72.4	51.5	40.6
● Qwen2.5-3B	12.9	48.4	44.8	54.7	40.2
■ Llama-3.1-8B-Instruct	19.6	23.6	66.0	51.3	40.1
● Commencis-7B	14.9	24.3	62.4	56.0	39.4
● Llama-3.2-3B	13.2	25.4	66.4	50.0	38.7
■ Aya-23-35B	**25.4**	21.0	55.6	50.2	38.0
■ Ministral-2410-8B-Instruct	21.4	20.3	60.4	50.0	38.0
● Qwen2.5-14B	20.4	22.0	52.4	56.2	37.7
● Mistral-v0.3-7B	14.2	20.7	66.0	49.8	37.7
● Mistral-v0.1-7B	13.6	20.5	65.2	50.2	37.3
■ Llama-3.2-3B-Instruct	12.9	20.6	64.0	50.2	36.9
● Turna-1B	14.2	51.0	32.8	48.3	36.6
■ Qwen2.5-3B-Instruct	16.8	37.6	37.2	51.3	35.7
● Qwen2.5-1.5B	12.9	27.4	48.4	52.3	35.3
■ Qwen2.5-1.5B-Instruct	12.9	20.7	48.8	49.7	33.0
● Qwen2.5-0.5B	12.9	33.7	26.8	49.3	30.7
■ Qwen2.5-0.5B-Instruct	13.1	21.4	29.2	47.3	27.8

Table 13: Text classification results of the models, sorted by their average scores. Base and instruction-tuned variants are represented by circles and squares, respectively. Colors indicate the language focus: blue for English-focused, yellow for multilingual-focused, and red for Turkish-focused models.

C Task Samples

This appendix section includes sample instances for all tasks & datasets included within CETVEL.

Belebele

Tüm notalara doğru şekilde basmaya devam ederken elinizin mümkün olduğu kadar rahat olduğundan emin olun - aynı zamanda parmaklarınızla fazladan hareketler yapmamaya çalışın. Bu şekilde kendinizi olabildiğince az yormuş olacaksınız. Unutmayın ki piyanoda olduğu gibi daha fazla ses için tuşlara çok güçlü vurmanıza gerek yoktur. Akordeon üzerinde, ekstra hacim elde etmek için körüğü daha fazla basınç veya hızda kullanırsınız.
Metne göre, hangisi akordeonu başarılı bir şekilde çalmak için uygun bir tavsiye değildir?

A. Daha fazla ses çıkarmak için tuşlara daha güçlü basın

B. Yorulmamak için gereksiz hareketleri en aza indirin

C. Eliniz rahat pozisyondayken notalara doğru şekilde basın

D. Ekstra ses elde etmek için körüğü daha hızlı kullanın

------- English Translation -------

While continuing to press all the notes correctly, make sure your hand is as relaxed as possible – at the same time, try not to make extra movements with your fingers. This way, you will tire yourself as little as possible. Remember that, just like on the piano, you don't need to hit the keys very hard to produce more sound. On the accordion, to achieve extra volume, you use the bellows with more pressure or speed.
According to the text, which of the following is *not* an appropriate piece of advice for playing the accordion successfully?

A. Press the keys harder to produce more sound

B. Minimize unnecessary movements to avoid fatigue

C. Press the notes correctly while keeping your hand relaxed

D. Use the bellows faster to achieve extra volume

BilmeceBench

Bilmece: Kuyruklu kumbara yemek taşır ambara.
Bilmecenin anlamı aşağıdakilerden hangisidir?

A. BALTA

B. ARMUT

C. KAŞIK

D. AYAKKABI

------- English Translation -------

Riddle: A piggy bank with a tail carries food to the storage.
What is the meaning of the riddle?

A. AXE

B. PEAR

C. SPOON

D. SHOE

Circumflex

Kelime: Hakim
Kelimenin anlamı aşağıdakilerden hangisidir?
Cevap:

A. Sıfat: Egemenliğini yürüten, buyruğunu yürüten, sözünü geçiren

B. Sıfat: Bilge

------- English Translation ------- Word: Hakim What is the meaning of the word? Answer:

A. Adjective: One who exercises authority, enforces command, and has influence

B. Adjective: Wise

Exams

Glikokortikoidler olarak adlandırılan hormonlar nerede sentezlenirler:

A. tiroid bezinde.

B. hipofizde.

C. pankreasta.

D. böbrek üstü bezinin kabuğunda.

- - - - - - - English Translation - - - - - - -

Where are the hormones called glucocorticoids synthesized:

A. in the thyroid gland.

B. in the pituitary gland.

C. in the pancreas.

D. in the cortex of the adrenal gland.

IronyTR

Cümle: *ODTÜden mezun olmadan yapılacak 100 şey* Madde 101: Pikapla kampüsten kaçmak
Soru: Bu cümlede ironi var mı?

A. Hayır

B. Evet

- - - - - - - English Translation - - - - - - - Sentence: *100 things to do before graduating from METU* Item 101: Escape the campus with a pickup truck Question: Is there irony in this sentence?

A. No

B. Yes

Natural Language Inference

Aşağıda iki cümle verilmektedir:
Cümle 1: "Evet, sanırım en sevdiğim restoran her zaman en yakın restorandır. En yakın olanı biliyorsun. En düşük kriterlere uyduğu sürece."
Cümle 2: "En sevdiğim restoranlar her zaman evimden en az yüz mil uzakta."

Bu iki cümle arasındaki ilişki nedir:

A. TUTARLI

B. ALAKASIZ

C. ÇELİŞKİLİ

- - - - - - - English Translation - - - - - - -

Below are two sentences:

Sentence 1: "Yes, I guess my favorite restaurant is always the closest one. You know the closest. As long as it meets the lowest standards."
Sentence 2: "My favorite restaurants are always at least a hundred miles away from my home."

What is the relationship between these two sentences?

A. ENTAILMENT

B. NEUTRAL

C. CONTRADICTION

NewsCat

Cümle: Hırsız, Hietanen'in başını yaktı DENİZLİSPORLU Hietanen'i hırsız yaktı. Porto maçı için kampta olduğu saatte evine giren hırsız, yatak odasına geçip, içki içti. Eşi Riiena eve döndüğünde yatağı dağınık görünce, "Bana kampta olduğunu söylüyorsun, eve kadın getiriyorsun" diyerek ayrılmak istedi. Rıza Çalımbay'dan izin alan futbolcu, eşini ikna etti.
...
Soru: Bu cümlenin konusu nedir?
Cevap:

A. spor

B. magazin

C. siyaset

D. sağlık

E. ekonomi

- - - - - - - English Translation - - - - - - -

Sentence: The thief got Hietanen in trouble. The thief caused trouble for Denizlispor player Hietanen. While he was at camp for the Porto match, a thief entered his home, went into the bedroom, and drank alcohol. When his wife Riiena returned home and saw the bed messy, she said, "You tell me you're at camp, but you

bring a woman home," and wanted to leave him. The footballer got permission from Rıza Çalımbay and convinced his wife.
...
Question: What is the topic of this sentence?
Answer:

A. sports

B. celebrity news

C. politics

D. health

E. economy

OffenseEval

Cümle: Hala Hogwarts mektubum gelmediğinden oluyor tüm bunlar.
Soru: Bu cümle nefret söylemi içermekte midir?

A. Hayır

B. Evet

- - - - - - - English Translation - - - - - - -

Sentence: All of this is happening because I still haven't received my Hogwarts letter.
Question: Does this sentence contain hate speech?

A. No

B. Yes

STSb

Aşağıda iki cümle verilmektedir:

Cümle 1: "Bir kız saçlarını şekillendirmekte."
Cümle 2: "Bir kız saçını fırçalıyor."

Bu iki cümle arasında ne kadar benzerlik vardır:

A. Benzerlik Yok

B. Düşük Benzerlik

C. Orta Benzerlik

D. Yüksek Benzerlik

E. Çok Yüksek Benzerlik

F. Mantıksal Olarak Aynı

- - - - - - - English Translation - - - - - - -

Below are two sentences:

Sentence 1: "A girl is styling her hair."
Sentence 2: "A girl is brushing her hair."

How similar are these two sentences?

A. No Similarity

B. Low Similarity

C. Moderate Similarity

D. High Similarity

E. Very High Similarity

F. Logically Equivalent

TQUAD

Kaynak: Kemaleddin ibn Yunus ya da Musa ibn Yunus (doğum yılı ve yeri: 1156 Musul - ölüm yılı ve yeri: 1241 Musul).Astronom, matematikçi ve İslam bilgini.Tam adı Musa bin Yunus bin Muhammed bin Men'a'dır, Künyesi ise Ebu'l-Feth'tir, lakabı Kemaleddin olup ayrıca İbn-i Yunus ve Mewsilî diye de bilinir.İlk eğitimini babası Şeyh Yunus Rızauddin'in yanında fıkıh ve hadis ilimleri öğrendi, ardından Bağdat'taki Nizamiye Medreseleri'nde okumaya devam etti. Burada Şerafeddin el-Tusî'den matematik derslerini aldı, ardından Batlamyus'un Almagest adlı eserini de öğrenir. Ardından Musul'a döndü, Emir Zeyneddin Camii'nde dersler verdi. İlim öğretmeye elverişli olarak inşa edilen bu cami Kemaliyye Medresesi olarak anıldı. Kısa zamanda şöhreti etrafa yayılan Musa Kemaleddin ibn Yunus pek çok çevreden gelen talebelere ilim öğretti.

Soru: Kemaleddin ibn Yunus lakabı dışında hangi isimlerle bilinir?

Cevap:
İbn-i Yunus ve Mewsilî

- - - - - - - English Translation - - - - - - -

Source: Kemaleddin ibn Yunus or Musa ibn

Yunus (born 1156 Mosul – died 1241 Mosul). Astronomer, mathematician, and Islamic scholar. His full name is Musa bin Yunus bin Muhammad bin Men'a; his kunyah is Abu'l-Feth, his laqab is Kemaleddin, and he is also known as İbn-i Yunus and Mewsilî. He received his initial education in fiqh and hadith from his father, Sheikh Yunus Rızauddin, then continued at the Nizāmiyya Madrasas in Baghdad, studying mathematics under Sharaf al-Din al-Tusi and learning Ptolemy's Almagest. He returned to Mosul and taught at the Zayn al-Dīn Mosque, also called the Kemaliyye Madrasa, gaining fame and attracting many students.
Question: Aside from the laqab Kemaleddin ibn Yunus, by what other names is he known?
Answer:
İbn-i Yunus ve Mewsilî

Turkce Atasozleri

Atasözü: aba altında er yatar
Yukarıdaki atasözünün tanımı aşağıdakilerden hangisidir?

A. Giyim kuşam kişiliğe ölçü olamaz.

B. Tanrı'dan korkmayan kimse, insana her türlü kötülüğü yapabilir.

C. İnsan kendinde herhangi bir kusur varken başkalarını aynı kusurla suçlamamalıdır.

D. Ortaya çıkan bir yanlışlık çok geç de olsa düzeltilebilir.

- - - - - - - English Translation - - - - - - -

Proverb: Beneath a coarse cloak may lie a noble man What is the meaning of the proverb above?

A. Clothing and appearance are not reliable measures of character.

B. One who does not fear God is capable of doing all kinds of harm to others.

C. A person should not accuse others of a fault they themselves possess.

D. A mistake that has come to light can be corrected, even if belatedly.

TurkishPLU / Goal Inference

Örnek Adım: İşletme adının hemen altındaki ekranın sağ tarafında bulunan "Yer İşareti" düğmesine dokunun. Hedef:

A. Yelp'e İşletme Fotoğrafı Eklemek

B. Audacity'de İz İşaretleri Eklemek

C. Yelp'te Bir İşletmeye Yer İşareti Eklemek

D. Yelp'te Yinelenen İşletme Girişlerini Bildirmek

- - - - - - - English Translation - - - - - - -

Example Step: Tap the "Bookmark" button located on the right side of the screen just below the business name. Goal:

A. Add a Business Photo on Yelp

B. Add Track Markers in Audacity

C. Bookmark a Business on Yelp

D. Report Duplicate Business Listings on Yelp

TurkishPLU / Next Event Prediction

Hedef: Dâhilî Numara Nasıl Aranır? Adım: Arama cevaplandığı anda dâhilî numarayı gireceksen bir "duraklama" ekle. Bir sonraki adım:

A. Eğer dâhilî numara sadece tüm menü oynatıldıktan sonra çevrilebiliyorsa bir "bekleme" ekle.

B. daha önce yapmadıysan, gizli Geliştirici Seçenekleri butonunu görüntülemek için seri numarana 7 kez dokun.

C. Ekran görüntüsünü Command ve V tuşlarını basılı tutarak veya Düzenle menüsünden Yapıştır'ı seçerek bir kelime işleme belgesine, bir e-postaya veya bir görüntü düzenleyiciye yapıştır.

- - - - - - - English Translation - - - - - - -

Goal: How to Dial an Extension Number? Step: If you'll enter the extension as soon as the call is answered, insert a "pause." Next step:

A. If the extension can only be dialed after the full menu has played, insert a "wait."

B. If you haven't done so already, tap your serial number 7 times to reveal the hidden Developer Options button.

C. Paste the screenshot into a word processing document, an email, or an image editor by holding down the Command and V keys or selecting Paste from the Edit menu.

TurkishPLU / Step Inference

Hedef: Obsesif Kompulsif Kişilik Bozukluğu Nasıl Tanınır? Örnek Adım:

A. VPN'in sınırlamalarını bil.

B. Hedef SGPT seviyenin ne olduğunu bil.

C. SNM'nin prensiplerini benimse.

D. OKKB'nin tanı kriterini bil.

------- English Translation -------

Goal: How to Recognize Obsessive-Compulsive Personality Disorder (OCPD)? Example Step:

A. Know the limitations of your VPN.

B. Know what your target SGPT level is.

C. Adopt the principles of SNM.

D. Know the diagnostic criteria for OCPD.

TurkishPLU / Step Ordering

Hedef: Tarayıcınızı Güncellemek

A. Önce: Tarayıcıya uygulanmasını istediğiniz tüm Internet Explorer güncellemelerinin yanına bir onay işareti koyun. Sonra: Internet Explorer için herhangi bir güncelleme olup olmadığını görmek için güncelleme listesini gözden geçirin.

B. Önce: Internet Explorer için herhangi bir güncelleme olup olmadığını görmek için güncelleme listesini gözden geçirin. Sonra: Tarayıcıya uygulanmasını istediğiniz tüm Internet Explorer güncellemelerinin yanına bir onay işareti koyun.

------- English Translation -------

Goal: Updating Your Browser

A. First: Place a check mark next to all Internet Explorer updates you want to apply to the browser. Then: Review the update list to see if there are any updates for Internet Explorer.

B. First: Review the update list to see if there are any updates for Internet Explorer. Then: Place a check mark next to all Internet Explorer updates you want to apply to the browser.

TurkishMMLU

220 V gerilimle çalışan ve direnci 484 Ω olan bir klima günde 5 saat süreyle çalıştırılıyor. Elektrik enerjisinin kWh'i 40 kuruş olduğuna göre klimanın harcadığı 30 günlük enerji bedeli kaç ₺'dir?

A. 4

B. 5

C. 6

D. 10

E. 15

------- English Translation -------

A 220 V air conditioner with a resistance of 484 Ω is operated for 5 hours per day. Given that the cost of 1 kWh of electricity is 40 kuruş, what is the energy cost in ₺ for 30 days of use?

A. 4

B. 5

C. 6

D. 10

E. 15

XCOPA

Ürün balonlu naylonla paketlenmişti bu yüzden

A. kırılgandı.

B. küçüktü.

------- English Translation -------

Sentence: The product was packaged with bubble wrap, so

A. it was fragile.

B. it was small.

XQUAD

Kaynak: Akademi Ödülü kazananı Marlee Matlin Amerikan İşaret Dili(ASL) çevirisini yaparken altı kez Grammy kazanan ve Akademi Ödülü adayı Lady Gaga ulusal marşı söylemiştir.

Soru: Lady Gaga kaç Grammy kazanmıştır?

Cevap: altı

------- English Translation -------

Source: Academy Award winner Marlee Matlin performed the American Sign Language (ASL) interpretation while six-time Grammy winner and Academy Award nominee Lady Gaga sang the national anthem.

Question: How many Grammys has Lady Gaga won?

Answer: six

GECTurk

Verilen cumlenin yazım hatalarını duzeltin.
Hatalı Cümle: Büyük yıldızlar transfer ederek, ya'da büyük hocalar getirerek hokus pokus başarıların gelmediği gerçeğinin farkına vardılar.

Düzeltilmiş hali: Büyük yıldızlar transfer ederek, ya da büyük hocalar getirerek hokus pokus başarıların gelmediği gerçeğinin farkına vardılar.

------- English Explanation -------

In Turkish, the coordinating conjunction "ya da" ("or") is always written as two separate words without an apostrophe.

MLSum

Başlık: İzmir'deki orman yangınının tehdit ettiği oteller tahliye edildi

Metin: İZMİR'in Menderes ilçesindeki tatil beldesi Özdere'de otluk alanda yangın çıktı. Dumanların etkilediği iki otel ise boşaltıldı. İzmir'in tatil beldelerinden Özdere Cumhuriyet Mahallesindeki otluk alanda yangın çıktı. Çıkan yangını söndürmek için 38 arazöz ve 4 dozer aralıksız olarak çalışmalarını sürdürüyor. Öte yandan dumanların etkilediği iki otel boşaltıldı. Helikopter kullanılamıyor Menderes Belediyesi tarafından yapılan açıklamada, "Menderes Belediyesi ekipleri olarak ilk müdahaleyi gerçekleştirdik. İzmir Büyükşehir Belediyemizle de irtibata geçerek İZSU ve İzmir itfaiyemiz hemen yangın alanında müdahaleye başladılar. İzmir Orman Bölge Müdürlüğümüze bağlı ekipler de yangın söndürme çalışmalarını gerçekleştiriyorlar. Havanın karanlık olması nedeniyle uçak ve helikopter ile yangına müdahale gerçekleştirilemiyor. Şu an yol trafiğe kapatılmış durumda. Gümüldür yolu üzerinden ve Ahmetbeyli yolu üzerinden araç trafiği verilmekte. Rüzgarın etkisi çalışmaları zorlaştırsa da ekipler canla, başla yangını kontrol altına almaya çalışıyorlar. Umarız en kısa sürede yangını kontrol altına alabiliriz. Ne yazık ki milli servetimiz, ciğerlerimiz yanıyor. Bir an önce bu felaketin son bulması için canla, başla çalışıyoruz" denildi.

Özet:
İZMİR'in Menderes ilçesi Özdere bölgesindeki ormanlık alanda elektrik tellerinin sürtünmesi sonucu orman yangını çıktı. Alevlerin tehdit ettiği bölgede bulunan otellerde konaklayanlar, ekipler tarafından tahliye edildi.

------- English Translation -------

Title: Hotels threatened by the forest fire in İzmir evacuated

Text: A fire broke out in a grassy area

in Özdere, a holiday resort in the Menderes district of İzmir. Smoke affected two hotels, which were evacuated. To extinguish the fire, 38 fire trucks and 4 bulldozers have been working nonstop. Menderes Municipality teams performed the first intervention, and İzmir Metropolitan Municipality's water and fire departments joined the efforts. Teams from the İzmir Forestry Regional Directorate are also fighting the blaze. Due to darkness, aircraft and helicopters cannot be used. Roads are closed to traffic, with vehicles rerouted via the Gümüldür and Ahmetbeyli roads. Despite challenging winds, teams are striving to bring the fire under control. "Our nation's natural treasure is ablaze." officials said, urging that the disaster end as soon as possible.

Summary:
A fire broke out in a grassy area in Özdere, Menderes district of İzmir. Guests at hotels threatened by the flames were evacuated by response teams.

XLSum

Başlık: 'İklim değişikliği penguenleri tehdit ediyor'

Metin: ABD'li, İngiliz ve Hollandalı araştırmacılar tarafından yürütülen ve iklim değişikliğinin penguenler üzerindeki etkisini konu alan çalışma, "Nature Climate Change" adlı bilimsel dergide yayımlandı. Makalede, "Büyük penguen" olarak da anılan ve Antartika'da yaşayan bu kuş türüne yönelik asıl tehdidin deniz-buz oranındaki değişimden iddia edildi. Buna göre Antartika'daki buz ve su oranı değişirse, penguenlerin çoğalmaları ve beslenmeleri olumsuz etkilenecek. Çalışma, penguen grupları arasında farklı dinamiklerin etkili olacağını ancak yine de tüm gruplarda sayının azalacağını savunuyor. Araştırmacılar, devletlerin penguenleri "nesli tükenmekte olan kuşlar" olarak korumaya alması önerisinde bulundu. Ancak korumaya yönelik tedbirler, turizm ve balıkçılık alanında kısıtlamalara neden olabiliyor. 'Neslinin tükenmesi tehdidi var' Çalışmayı yürüten ekibin başında "Woods Hole Oceanographic" Enstitüsü'nden Stephanie Jenouvrier yer alıyor. Doktor Jenouvrier, tüm penguen nüfusunun yüzde 19 ile 33 arasında bir oranda azalacağını belirtiyor. Jenouvrier penguenlerin "Yakın bir gelecekte önemli oranda nüfusunu kaybedeceğini ve muhtemelen neslinin tükenmesi tehlikesiyle karşı karşıya kalacağını" söyledi. Antarktika'daki Ross denizi çevresinde yaşayan penguen gruplarının iklim değişikliğinden en son etkilenenler olacağını kaydeden Jenouvrier'a göre bunun sebebi, bölgedeki deniz ve buz dağılımının penguenler için hala elverişli olması. Jenouvrier sözlerine şöyle devam etti: "Ross denizinde penguenlerin yaşadığı bölgenin korunması ve geliştirilmesi, nesil tükenmesi tehdidine karşı zaman kazandıracaktır. Böylece sera gazının azaltılması konusunda gerekli müzakereler yapılabilecek, stratejiler belirlenebilecektir." Aylarca yol kat ediyor, yemek arıyorlar Penguenler, yavrularını beslemek için aylarca yuvadan uzak, yemek arıyorlar. Antarktika buzulları boyunca uzun müsefaler kat eden penguenler, denize eriştikleri yerlerden karides gibi yiyecekler topluyorlar. Penguenler yemek ararken yırtıcı hayvanlardan korunmak gibi çeşitli nedenlerle ideal miktarda buzul tabakaya ihtiyaç duyuyor. Buzul ve deniz miktarındaki değişimin penguenlerin beslendiği karides gibi canlıların verimliliğini de etkileyeceği belirtiliyor. Penguenlerin ana besin kaynağı olan karides ve benzeri deniz kabuklularının üremesinin, buzul deniz dağılımından etkilendiği ifade ediliyor. Buzulların artması karides ve diğer kabuklular için olumlu olarak değerlendiriliyor. Ancak bu durum, penguenlerin denize ulaşmak için daha uzun mesafe kat etmesi anlamına geliyor. Uydudan yapılan ölçümlerde, Antartika'da buzul-su seviyesinin daha önce görülmeyen bir seviyeye yükseldiği görülüyor. Ancak iklim modelleme yazılımları, bu durumun ileride tersine döneceğini belirtiyor.

Özet:
İklim değişikliğinin Antarktika'daki penguen nüfusunu olumsuz etkileyebileceği belirtiliyor. Yapılan bir çalışmaya göre, sayıları 600 bini bulan penguenlerin 2100 yılı itibariyle beşte

biri oranında azalabileceği ifade ediliyor.
- - - - - - - English Translation - - - - - - -

Title: 'Climate change threatens penguins'

Text: A study conducted by US, British, and Dutch researchers on the impact of climate change on penguins was published in the scientific journal "Nature Climate Change." The article states that the main threat to this bird species, also known as the "Emperor penguin" living in Antarctica, is claimed to be changes in the sea–ice ratio. According to this, if the ratio of ice to water in Antarctica changes, penguins' breeding and feeding will be negatively affected. The study argues that different dynamics will be at play among penguin colonies, but that numbers will decline in all groups. The researchers recommended that governments list penguins as "endangered birds." However, conservation measures can lead to restrictions in tourism and fishing. Under the leadership of Stephanie Jenouvrier from the Woods Hole Oceanographic Institution, the team determined that the total penguin population will decrease by between 19% and 33%. Jenouvrier said that penguins will lose a significant portion of their population in the near future and may face the risk of extinction. According to Jenouvrier, the colonies living around the Ross Sea in Antarctica will be the last to be affected by climate change, because the distribution of sea and ice in that region remains favorable for penguins. Jenouvrier continued: "Protecting and enhancing the region where penguins live in the Ross Sea will buy time against the threat of extinction. This will allow for necessary negotiations and strategies to reduce greenhouse gases to be developed." Penguins travel for months and search for food to feed their chicks. Penguins traverse long distances along Antarctic ice, collecting krill and similar prey from sea access points. Penguins need an ideal amount of ice shelf for various reasons, such as protection from predators while foraging. Changes in ice and sea levels will also affect the productivity of krill and other crustaceans that penguins eat. An increase in ice benefits krill and other crustaceans, but forces penguins to travel longer distances to reach the sea. Satellite measurements show that the ice–water ratio in Antarctica has reached unprecedented levels. However, climate models predict this trend will reverse in the future.

Summary:
It is stated that climate change could negatively affect the penguin population in Antarctica. According to a study, the population of around 600,000 penguins could decrease by one fifth by 2100.

WikiHowSum

Metin: Yatabileceğin en doğal uyku pozisyonunda uzan. Birşey tutma, bacaklarını yatakta tut, başını kaldırma. Eğer normalda sırt üstü uyuyorsan numaradan uyurken de öyle yap. Böylece seni tanıyan insanlar şüphelenmez. Doğal uykunda çok az hareket edersin. Gerçekten uyuyormuş izlenimi yaratmak için en iyisi hiç hareket etmemek. Biri seni uzun bir süre boyunca izlemediği sürece hareket etmen beklenmez. Göz kapaklarını fazlaca sıkarak kapatmaktan kaçın. En iyi uyuyor izlenimi için, göz kapakların dâhil tüm kaslarının rahat olmalı. Gözlerini kapattıktan sonra göz kapaklarının kırpışmasını engellemek için aşağı doğru bak. Uyurken gözlerin her zaman tam kapalı olmaz. Göz kapaklarının düşerek nazikçe kapanmasına izin ver; hâlâ göz kapaklarının arasından etrafı biraz görebilirsin. Yavaş, hatta derin nefesler al. Nefes almanı rahatlatmalı ve mümkün olduğunca eşit aralıklarla nefes alıp vermelisin. Nefes alırken kafandan sayıp, aynı sürede vermeye çalış. Bunu her nefesinde yap. Eğer yüksek bir ses duyarsan ya da biri sana dokunursa kısa ve ani bir nefes al ve vücudunu hafifçe titret. Uyurken bile, vücutlarımız etrafımızda olan şeylerin farkındadır. Sahte uykunu, odadaki seslere ve hareketlere bilinçsiz görünen tepkiler ekleyerek sat. Rahatsızlığa tepki verdikten sonra, vücudunun gevşemesine ve nefesinin yavaş ve dengeli bir duruma dönmesine izin ver. Sakın gülümseme ve gözlerini açma, yoksa aslında uyanık olduğun hemen anlaşılır.

Özet: Doğal bir uyku pozisyonu seç. Yatakta hareketsiz bir şekilde yat. Gözlerini nazikçe kapat. Ritmik bir şekilde nefes alıp ver. Seslere ve dokunmaya tepki ver.
- - - - - - - - English Translation - - - - - - - -

Text: Lie in the most natural sleep position possible. Don't hold anything, keep your legs on the bed, don't lift your head. If you normally sleep on your back, do so here as well. That way, people who know you won't suspect. You move very little in natural sleep. To create the impression of truly sleeping, it's best not to move at all. As long as no one watches you for a long time, movement isn't expected. Avoid squeezing your eyelids tightly shut. For the best sleeping impression, all your muscles—including your eyelids—should be relaxed. After closing your eyes, look downward to prevent eyelid twitching. Eyes are never fully closed when sleeping. Allow your eyelids to fall gently and close; you can still see a little through them. Breathe slowly, even deeply. Your breathing should be relaxed and at as equal intervals as possible. Count in your head as you inhale, and try to exhale in the same time. Do this with each breath. If you hear a loud noise or someone touches you, take a short, sudden breath and slightly shake your body. Even during sleep, our bodies are aware of things around us. Sell your fake sleep by adding unconscious-seeming reactions to sounds and movements in the room. After reacting, allow your body to relax and your breathing to return to a slow, balanced state. Never smile or open your eyes, or it will immediately reveal you are actually awake.
Summary: Choose a natural sleep position. Lie motionless on the bed. Gently close your eyes. Breathe rhythmically. Respond to sounds and touch.

WMT16$_{\text{EN-TR}}$

Translate English to Turkish.

English: Norway's rakfisk: Is this the world's smelliest fish?
Turkish: Norveç'in rakfisk'i: Dünyanın en kokulu balığı bu mu?

Language Understanding Tasks

Task	Dataset	Source / Domain	License	#examp.	Metric
Extractive Question Answering	XQUAD	Wikipedia	CC BY-SA 4.0	1190	Exact Match
	TQUAD	Wikipedia	MIT	892	
	MKQA	Wikipedia	CC BY-SA 3.0	10000	
Multiple Choice Question Answering	Exams	High School Exams	CC-BY-SA 4.0	393	Accuracy
	Belebele	Online Web Pages	CC-BY-SA 4.0	900	
	TurkishPLU	WikiHow	CC BY-NC-SA 3.0	3124	
	XCOPA	Crowd Sourcing	CC-BY-SA 4.0	600	
	TurkishMMLU	High School Exams	-	900	
	Proverbs	Turkish Dictionaries	GPL-3.0	1730	
	BilmeceBench	Turkish riddles	MIT	442	
	CircumflexTR	Turkish Dictionaries	MIT	72	
Text Classification	IronyTR	Twitter	-	600	Accuracy
	NewsCat	Online Newspapers	-	250	
	OffensEval	Twitter	CC BY-SA 2.0	3528	
	STSb	Multi-Domain	CC BY-SA 4.0	1389	
Natural Language Inference	XNLI	SNLI & MNLI	CC BY 4.0	5010	Accuracy
	MNLI	Crowd Sourcing	CC BY-SA 3.0	10000	
	SNLI	Image Captions	CC BY-SA 4.0	10000	

Language Generation Tasks

Task	Dataset	Source / Domain	License	#examp.	Metric
Summarization	MLSum	Online Newspapers	Other	12775	BERTScore
	XLSum	BBC News	CC BY-NC-SA 4.0	3400	
	WikiLingua	WikiHow	CC BY-NC-SA 3.0	3000	
Machine Translation	$\text{WMT16}_{\text{EN-TR}}$	Online Newspapers	-	3000	COMET-22
Grammatical Correction	GECTurk	Newspapers	Apache-2.0	20800	Exact Match

Table 14: The overview of the tasks & datasets included in the CETVEL benchmark. Most of the listed datasets are the Turkish splits of a larger multilingual dataset with few exceptions which are TurkishPLU, TQUAD, IronyTR, NewsCat and GECTurk.