

LLM-as-a-Judge for Multilingual Question Answering Evaluation: An Experimental Study on an English–Turkish Bilingual SQuAD Dataset

Çok Dilli Soru-Cevap Değerlendirmesinde Hakem Olarak LLM: İngilizce–Türkçe İki Dilli SQuAD Veri Kümesi Üzerine Deneysel Bir İnceleme

Ilkay Yelmen

 0000-0002-1684-9717

Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istinye University, Istanbul, Türkiye

ilkay.yelmen@istinye.edu.tr

Özet

Bu çalışma, büyük dil modellerinin İngilizce ve Türkçe soru-cevap değerlendirmesinde otomatik değerlendirici olarak kullanılabilirliğini karşılaştırmalı olarak incelemektedir. Bu kapsamda, beş farklı büyük dil modelinin kısmi insan anotasyonu ve kural tabanlı vekil skorlardan oluşan karma referans puanlama yapısıyla uyumu, sınıflandırma başarımı, diller arası tutarlılığı ve prompt duyarlılığı analiz edilmiştir. Paralel İngilizce–Türkçe SQuAD veri kümesinden seçilen 100 örnek için doğru, kısmen doğru ve yanlış olmak üzere üç farklı aday yanıt oluşturulmuş, bu yanıtlar basit ve rubrik tabanlı iki prompt stratejisiyle değerlendirilmiştir. Model çıktıları Spearman sıra korelasyonu, doğruluk, makro-F1, Cohen'in kappa katsayısı ve Wilcoxon işaretli sıralar testi kullanılarak incelenmiştir. Bulgular, tüm modellerin karma referans puanlarıyla yüksek düzeyde sıra korelasyonu gösterdiğini ortaya koymuştur. GPT-4o-mini İngilizcede en yüksek korelasyonu üretirken, Türkçe değerlendirmede Claude Sonnet 4.6 ve GPT-5.4-mini en yüksek korelasyonu paylaşmış; birleşik İngilizce–Türkçe sonuçlarda ise Claude Sonnet 4.6 en yüksek korelasyonu göstermiştir. Claude Sonnet 4.6, sınıflandırma başarımı ve prompt kararlılığı açısından en dengeli model olarak öne çıkmıştır. Sonuçlar, LLM-as-a-Judge yaklaşımının Türkçe soru-cevap değerlendirmesi için umut verici olduğunu ancak model seçimi, dil etkisi, prompt tasarımı ve puan kalibrasyonunun birlikte ele alınması gerektiğini göstermektedir.

Anahtar kelimeler: Büyük dil modeli, değerlendirici olarak büyük dil modelleri, soru-cevap değerlendirmesi, çok dilli değerlendirme, doğal dil işleme

Abstract

This study conducts a comparative analysis of the feasibility of using large language models as automatic evaluators in English and Turkish question-answering assessments. In this context, the study analyzes the alignment of five different large language models with a hybrid reference scoring system consisting of partial human annotation and rule-based proxy scores, as well as their classification performance, cross-lingual consistency, and prompt sensitivity. For 100 examples selected from the parallel English–Turkish SQuAD dataset, three different candidate responses, namely correct, partially correct, and incorrect responses, were generated, and these responses were evaluated using two prompt strategies: a simple prompt and a rubric-based prompt. The model outputs

were analyzed using Spearman's rank correlation, accuracy, macro-F1, Cohen's kappa coefficient, and the Wilcoxon signed-rank test. The findings revealed that all models exhibited a high level of rank correlation with the hybrid reference scores. While GPT-4o-mini produced the highest correlation in English, Claude Sonnet 4.6 and GPT-5.4-mini shared the highest correlation in Turkish; in the combined English–Turkish results, Claude Sonnet 4.6 achieved the highest correlation. Claude Sonnet 4.6 emerged as the most balanced model in terms of classification performance and prompt stability. The results indicate that the LLM-as-a-Judge approach shows promise for evaluating Turkish question-answering systems, but that model selection, language effects, prompt design, and score calibration must be addressed collectively.

Keywords: Large language model, LLM-as-a-Judge, question-answering evaluation, multilingual evaluation, natural language processing

1. Introduction

The widespread adoption of large language models in question-answering systems has made the question of how to reliably evaluate the responses generated by these systems even more critical. The generated responses often appear not merely as correct or incorrect, but in context-dependent, incomplete, partially correct, or semantically plausible yet erroneous forms. Therefore, the evaluation of question-answering systems requires more flexible evaluation approaches that go beyond simple superficial similarity metrics and take into account the response's contextual alignment, semantic accuracy, and scope.

Traditional evaluation metrics are primarily based on the similarity of words or phrases between the candidate response and the reference response. While metrics such as BLEU and ROUGE are useful for certain tasks, they do not always adequately capture aspects of question-answering evaluation such as the response's contextual appropriateness, semantic accuracy, and completeness. This issue becomes particularly evident when evaluating responses that are partially correct, incomplete, or contextually limited.

Due to these limitations, the “LLM-as-a-Judge” approach which utilizes large language models as automatic evaluators has gained prominence in recent years. In this approach, large language models can score, rank, or classify candidate responses by evaluating them in conjunction with the question, context, and reference answer. The LLM-as-a-Judge

method is more scalable than human evaluation; it also has the potential to perform more contextual and semantic evaluations compared to traditional metrics [1]. However, previous studies have shown that LLM judges are not equally reliable across all tasks and contexts, and that their alignment with human judgments can vary, particularly in tasks requiring expertise or involving more complex evaluation dimensions such as user satisfaction [2], [3].

However, there is a significant research gap in the context of multilingual evaluation. The majority of LLM-as-a-Judge studies have been conducted primarily in English; the reliability of this approach has been less thoroughly examined in languages such as Turkish, which are morphologically rich and have more limited evaluation resources. Due to its agglutinative structure and distinct syntactic features, Turkish may present different challenges for automatic evaluation compared to English. Therefore, language-specific and systematic analyses are needed to evaluate the outputs of large language models in Turkish. While studies such as TurkBench highlight the importance of specialized benchmarks for Turkish LLM evaluation [4], the alignment of the LLM-as-a-Judge approach with hybrid reference scores, its classification performance, cross-lingual consistency, and prompt sensitivity in Turkish question-answering evaluation have not yet been sufficiently and comprehensively investigated.

This study examines the feasibility and consistency of the LLM-as-a-Judge approach in evaluating English-Turkish question-answering tasks. To this end, 100 examples were selected from the parallel English-Turkish SQuAD dataset; for each example, three different types of candidate answers were generated: correct, partially correct, and incorrect. Five large language models scored these candidate responses on a scale of 1–5 under two different prompt strategies: a simple prompt and a rubric-based prompt. The model outputs were analyzed in terms of agreement with hybrid reference scores, classification accuracy, macro-F1, Cohen’s kappa, cross-language scoring differences, and prompt sensitivity.

The main contributions of the study can be summarized as follows:

- This study provides preliminary empirical evidence on the use of large language models as automatic evaluators in a language like Turkish, which is morphologically complex and underrepresented in the LLM-as-a-Judge literature.
- By evaluating the English and Turkish equivalents of the same question, context, and response levels, this approach directly compares not only the models’ overall performance but also the consistency of their evaluations across languages.
- By considering the “correct,” “partially correct,” and “incorrect” response categories together, it highlights the models’ ability to distinguish “partially correct” responses, which constitute a particularly critical and challenging category for evaluation.
- The study demonstrates that high agreement with reference scores does not always translate to reliable classification performance, highlighting the need to measure score calibration and class discrimination performance separately in LLM-as-a-Judge evaluations.
- By comparing simple and rubric-based prompt strategies, it demonstrates that evaluation outcomes can vary depending on prompt design and language; thus, it positions prompt robustness as a

fundamental evaluation dimension for multilingual LLM-as-a-Judge systems.

- The findings provide a practical evaluation framework for model selection, prompt design, and score calibration in Turkish and multilingual question-answering systems.

The remaining sections of the paper are structured as follows. Section 2 summarizes the literature on the LLM-as-a-Judge approach and multilingual evaluation studies. Section 3 describes the dataset, candidate response generation, judge models, the prompting process, and evaluation metrics. Section 4 presents the experimental setup and reproducibility settings. Section 5 presents results on model performance, agreement with hybrid reference scores, language differences, and prompt sensitivity. Section 6 discusses the findings, while Section 7 presents conclusions and recommendations for future work.

2. Related Work

The LLM-as-a-Judge approach has emerged in recent years as a scalable alternative to human evaluation in automated assessment processes. Li et al. have presented a comprehensive literature review examining this approach across various scenarios, such as evaluation, scoring, ranking, and selection. The study highlights that traditional matching-based or small-model-based evaluation methods have limitations when applied to open-ended and dynamic tasks; the LLM-as-a-Judge approach has been classified based on the dimensions of “what is evaluated,” “how it is evaluated,” and “which benchmarks are used for measurement” [1]. This framework demonstrates that LLM-based evaluators are not merely automated scoring tools, but systems that must be carefully examined in terms of reliability, bias, and assessment design.

Current studies show that LLM judges can produce promising results on various tasks, but that this performance varies depending on the task type and evaluation context. Yu et al. treated the “LLM-as-a-Judge” capability as a general skill of large language models and proposed a two-stage training approach consisting of supervised fine-tuning and direct preference optimization. The study demonstrates that judgment performance can be improved not only through prompt design but also through specialized training and preference optimization processes [5]. In contrast, Huang et al. noted that while judge models fine-tuned based on open-source models may demonstrate high performance on their own test datasets, they still lag behind GPT-4 in terms of generalizability, fairness, and adaptability [6]. These findings indicate that high performance alone is not sufficient in LLM-as-a-Judge systems; reliability analysis is necessary across different data domains, tasks, and usage conditions.

The reliability of the LLM-as-a-Judge approach is closely related to its alignment with human judgments and the transparency of its evaluation criteria. Yamasaki et al. proposed the Function-Based Labels approach to more clearly define the ambiguous relationships between product pairs in complementary product recommendations. In this study, a dataset containing 2,759 human-annotated product pairs was created based on the functional features of the products, independent of user behavior history. ModernBERT demonstrated the best machine learning performance with a Macro-F1 score of 0.911; GPT-4o-mini, on the other hand, achieved a consistency score of 0.989 and a classification accuracy of 0.849 when a detailed FBL definition was used [7]. These results are significant in that they demonstrate that

clear and detailed evaluation definitions can improve alignment with human judgments in LLM-based annotations.

Domain-specific studies highlight both the potential and limitations of LLM judges in different usage contexts. Türkmen and Keleş examined the “LLM-as-a-Judge” approach to evaluating RAG systems using 150,000 real e-commerce interactions on the Trendyol QA Assistant. In the study, retrieval and generation-related errors were distinguished by combining user satisfaction labels, LLM evaluator scores, and factor-based discriminant analyses. The findings have shown that LLM-based judge models can generally reflect trends in user satisfaction, but may overlook certain details of dissatisfaction [2]. Szymanski et al., on the other hand, found that LLM judges demonstrated only limited agreement with human experts in specialized fields such as dietetics and mental health. In the study, LLM judges agreed with experts 68% of the time in the field of dietetics and 64% of the time in the field of mental health [3]. These studies show that LLM evaluators should be viewed as complementary assessment tools used in conjunction with human feedback, rather than as a complete replacement for human evaluation.

Research in various fields—such as healthcare, creative writing, and software engineering—also demonstrates that the “LLM-as-a-Judge” approach must be evaluated in a task-sensitive manner. Khan et al. proposed a framework called “AI on Trial” that uses the “LLM-as-a-Judge” approach in clinical decision-making processes and ensures that responses are evaluated based on criteria of quality, completeness, safety, and clarity [8]. Bellard et al. integrated an LLM-as-a-Judge component into an agent-based RAG system for the management of heart failure and demonstrated that this architecture can improve the robustness of responses [9]. Mahadik et al. used multidimensional evaluation criteria such as factual consistency, emotional tone, emotional expression, readability, and overall response quality in medical dialogue systems [10]. Abir et al. focused on evaluating creative and semantically complex texts by using the ROUGE, BERTScore, and LLM-as-a-Judge metrics together in a poetry summarization task [11]. Crupi et al., on the other hand, examined the “LLM-as-a-Judge” approach in code generation and code summarization tasks; they demonstrated that, although GPT-4-turbo was the most successful judge model, it could occasionally misjudge code accuracy and summary quality [12]. These findings indicate that a single, general performance metric is insufficient in LLM-based evaluation systems; instead, a multidimensional evaluation tailored to the specific task, domain, and output type is necessary.

Another key trend emerging in LLM-as-a-Judge research is the expansion of evaluation processes to include multimodal scenarios and those sensitive to different linguistic contexts. The MLLM-as-a-Judge benchmark, developed in this vein, examines the evaluation capabilities of vision-language models by adapting the LLM-as-a-Judge approach to multimodal large language models. Within the scope of this benchmark, the models were evaluated using rating, pairwise comparison, and aggregate ranking tasks. The findings show that the models are capable of making human-like decisions in pairwise comparisons, but may exhibit significant differences from human preferences in rating and aggregate ranking tasks [13].

In general, the literature indicates that the LLM-as-a-Judge approach holds strong potential for scalable and context-aware evaluation; however, for reliable use, factors such as human collaboration, task type, language effects, prompt design, and evaluation metrics must be considered together. In this

context, the present study contributes to addressing the multilingual reliability gap in the literature by jointly examining the agreement of LLM evaluators with hybrid reference scores, their classification performance, cross-lingual consistency, and prompt sensitivity in English-Turkish question-answering evaluations.

3. Methodology

The methodological design of this study was structured to examine the extent to which large language models can reliably score English and Turkish question-answer pairs using an automatic evaluator. In the experimental process, parallel English-Turkish SQuAD examples with the same content structure were used, and candidate answers representing different levels of accuracy were generated for each example. These candidate responses were categorized into three groups: correct, partially correct, and incorrect. This allowed for an analysis not only of the models’ ability to distinguish clearly correct or incorrect answers but also of their capacity to correctly classify partially correct answers, which are more challenging to evaluate. The study design was structured to evaluate model performance across complementary dimensions such as agreement with hybrid reference scores, classification performance, cross-lingual consistency, and prompt sensitivity rather than reducing it to a single metric.

In the experimental workflow, 100 parallel English-Turkish question-answer pairs were first selected, and then three different candidate responses were generated for each pair using the GPT-4o-mini model. Subsequently, five different large language models were used as evaluators, and each was asked to rate each candidate response on a scale of 1 to 5. The evaluation process was conducted under two different prompt strategies: simple prompt and rubric-based prompt. This structure allowed for an examination of both the models’ overall evaluation performance and whether there were changes in scoring behavior depending on the prompt type. The obtained scores were analyzed using metrics such as Spearman correlation, accuracy, Macro-F1, and Cohen’s kappa; additionally, language differences and prompt sensitivity were evaluated using statistical methods such as the Wilcoxon signed-rank test and the Friedman test. The general experimental process followed in this context is summarized in Figure 1.

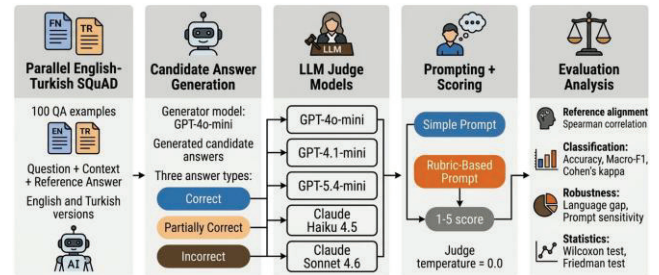


Figure 1. Experimental pipeline for multilingual LLM-as-a-Judge evaluation

3.1. Dataset and Candidate Answer Generation

In this study, English-Turkish parallel examples from the validation split of the Stanford Question Answering Dataset (SQuAD) were used. The experimental dataset consists of 100 question-answer pairs with successfully established English-Turkish alignment. The 100 examples were selected deterministically from the SQuAD v1.1 validation split and its Turkish counterpart, SQuAD-TR. English and Turkish records

were matched using their common SQuAD identifiers, and only examples with non-empty reference answers in both languages were considered eligible. Among the 2,346 aligned English–Turkish examples satisfying these criteria, the first 100 pairs encountered in the original dataset order were included in the study. No random sampling, stratification, or filtering based on question type, answer type, answer length, or context length was applied. When multiple reference answers were available, the first reference answer provided for each example was used consistently in both the candidate-response generation and evaluation procedures. Each example consists of three basic components: a question, a context paragraph, and a reference answer. Using the English and Turkish versions of these components enabled the models to be compared under the same evaluation framework in both languages. Thus, model performance was assessed using parallel English and Turkish versions of the same underlying question-answer content.

Candidate responses were generated separately for each example and each language. At this stage, the GPT-4o-mini model was used as the generative model, and the temperature value was set to 0.7. The model was asked to generate candidate responses at three different levels of accuracy, taking into account the given context and the reference answer. These accuracy levels were defined as “correct,” “partially correct,” and “incorrect.” “Correct” responses represent context-based answers that are fully consistent with the reference answer; “partially correct” responses contain accurate information but are incomplete or limited; and “incorrect” responses are clearly erroneous but may appear superficially plausible.

To construct the reference scoring scheme used in the evaluation process, a hybrid scoring approach was adopted. For 25 of the 100 parallel examples, corresponding to 150 candidate responses in total, the generated answers were manually scored by the author on a 1–5 scale. These manually assigned scores were used as human reference scores for this subset. For the remaining 75 examples, rule-based proxy scores were assigned according to the predefined accuracy levels. Accordingly, correct responses were represented by 5 points, partially correct responses by 3 points, and incorrect responses by 1 point. This proxy scoring scheme was selected to maintain consistency with the score-to-label conversion rule used in the classification analysis, where scores ≤ 2 indicate incorrect responses, a score of 3 indicates partially correct responses, and scores ≥ 4 indicate correct responses. This framework made it possible to examine not only the ability of large language models to distinguish between clearly correct and incorrect responses but also their capacity to identify partially correct responses, which are more ambiguous in terms of evaluation. However, because the human-scored subset was annotated by a single human evaluator and covered only part of the dataset, the resulting reference structure should be interpreted as a partially human-annotated and partially proxy-based scoring framework rather than as a fully human-annotated gold standard. The prompt template used to generate candidate responses was designed to include explicit constraints to ensure that each type of response was formulated in accordance with the specified quality level.

3.2. Judge Models and Prompting Procedure

In this study, five large language models from two different providers were used as evaluator models for the automatic evaluation of candidate responses, as shown in Table 1. The GPT-4o-mini, GPT-4.1-mini, and GPT-5.4-mini

models offered by OpenAI, as well as the Claude Haiku 4.5 and Claude Sonnet 4.6 models offered by Anthropic, were included in the evaluation process. All judge models were run with a temperature value of 0.0 to minimize sampling variability and improve reproducibility. Each model was presented with the question in the relevant language, the reference answer, and the candidate answer; the models were asked to rate the candidate answer on a scale of 1 to 5. The original SQuAD context paragraph was not included in the judge prompt.

Table 1. Judge models used in the evaluation

Model	Provider	Model Family
GPT-4o-mini	OpenAI	GPT-4o
GPT-4.1-mini	OpenAI	GPT-4.1
GPT-5.4-mini	OpenAI	GPT-5
Claude Haiku 4.5	Anthropic	Claude Haiku
Claude Sonnet 4.6	Anthropic	Claude Sonnet

Two different guidance strategies, simple prompting and rubric-based prompting, were used in the prompting process. While the simple prompt required the candidate’s response to be scored based on the criteria of correctness and clarity, the rubric-based prompt explicitly listed correctness, clarity, and completeness as evaluation criteria, providing a more structured evaluation instruction than the simple prompt. The same evaluation workflow was applied to both English and Turkish examples, thereby enabling a comparison of the models’ cross-lingual consistency and sensitivity to different prompt types. To improve reproducibility, representative examples of the prompt templates are provided here. In the simple-prompt condition, the evaluator was instructed: “Score the candidate answer from 1 to 5 based on correctness and clarity.” In the rubric-based condition, the evaluator was instructed: “Evaluate the candidate answer based on correctness, clarity, and completeness.” Both prompt types then presented the question, reference answer, and candidate answer, and required the model to return a JSON-formatted score and short justification. Equivalent Turkish-language templates were used for the Turkish evaluation condition. As part of the experimental design, a total of 6,000 evaluations were conducted using a combination of 5 models, 100 examples, 2 languages, 3 response labels, and 2 prompt strategies.

3.3. Scoring Protocol and Evaluation Metrics

In this study, the judge models were asked to score each candidate response on a scale of 1 to 5. During the scoring process, the candidate response was evaluated alongside the relevant question and reference answer. The original SQuAD context paragraph was not provided to the judge models. A hybrid reference scoring scheme was used. For 25 of the 100 parallel examples, corresponding to 150 candidate responses, scores were manually assigned by the author on a 1–5 scale. For the remaining 75 examples, rule-based proxy scores were used: correct answers were assigned 5 points, partially correct answers 3 points, and incorrect answers 1 point. In this way, the scores generated by the model for each candidate response were made comparable to partially human-annotated and partially proxy-based reference scores. This framework was used to examine whether the models ranked the responses in a manner consistent with the reference scoring scheme.

A predefined threshold rule was applied when converting the model scores into a three-class evaluation framework.

Accordingly, responses with a score ≤ 2 were classified as “incorrect,” those with a score = 3 as “partially correct,” and those with a score ≥ 4 as “correct.” Thanks to this transformation, the models were evaluated not only on their ability to assign scores between 1 and 5, but also on their ability to distinguish between correct, partially correct, and incorrect answers. In particular, maintaining the “partially correct” class as a separate category made it possible to examine the extent to which the models could distinguish answers that were incomplete or only partially correct.

Multiple complementary metrics were used in the evaluation. The agreement between model scores and reference scores was measured using Spearman’s rank correlation. Pearson correlation was additionally reported as a complementary measure to assess the linear association between model scores and reference scores and to compare it with the rank-based agreement captured by Spearman’s correlation. Classification performance was calculated using accuracy and macro-F1 scores following score-to-label conversion. While accuracy provides the ratio of examples correctly classified by the model to the total number of examples, macro-F1 was used to evaluate performance across the three classes in a balanced manner. The agreement between the labels generated by the model and the reference labels was measured using Cohen’s kappa coefficient.

Prompt sensitivity was defined as the absolute difference between the mean scores obtained for the same example under the simple and rubric-based prompt conditions, calculated as $|\text{mean_score_simple} - \text{mean_score_rubric}|$. The absolute difference was used to quantify the magnitude of prompt-induced score variation irrespective of direction; therefore, this metric does not indicate whether the rubric-based prompt produced systematically higher or lower scores. Language difference, on the other hand, was assessed based on the difference between the scores for the same example in English and Turkish. The 95% confidence intervals for the metrics were obtained using the percentile bootstrap method. In the bootstrap process, the number of resamples was set to 1,000, and the random seed was fixed at 42 (BOOTSTRAP_RANDOM_SEED) to ensure reproducibility. In each bootstrap iteration, sampling was performed with replacement to preserve the original sample size. The sampling unit was defined at the response-pair level, depending on the metric in question; for classification metrics, the reference–prediction pairs were used, while for correlation metrics, the model-score–reference-score pairs were used. Grouped bootstrap was not applied; correct, partially correct, and incorrect candidate responses to the same question were treated as independent response pairs. The original sample size of $n = 600$ was maintained for each cell; confidence intervals were reported using the 2.5th and 97.5th percentiles of the bootstrap distribution. The statistical significance of language and prompt effects was assessed using the Wilcoxon signed-rank test; for the prompt-sensitivity analyses, p-values were additionally adjusted using the Benjamini–Hochberg procedure to account for multiple comparisons. Differences between models were analyzed using the Friedman test. Because the human-scored subset was annotated by a single evaluator, the findings should be interpreted as agreement with a hybrid reference structure rather than validation against a fully human-annotated gold standard. Although the experimental design yielded 6,000 model judgments, these judgments were derived from 100 underlying parallel question–answer pairs. Therefore, the 6,000 judgments should be interpreted as repeated evaluation

observations generated from 100 fundamental examples rather than as 6,000 statistically independent samples.

4. Experimental Setup

4.1. Model Configuration

In the experimental evaluation, five large language models from two different API providers were used as evaluation models. The GPT-4o-mini, GPT-4.1-mini, and GPT-5.4-mini models from OpenAI, and the Claude Haiku 4.5 and Claude Sonnet 4.6 models from Anthropic were included in the evaluation process. The evaluator models were used to evaluate candidate responses generated from English and Turkish question-answering examples. All evaluator models were run with a temperature parameter of 0.0 to minimize sampling variability and improve reproducibility. Candidate response generation was carried out as a separate step; in this step, GPT-4o-mini was used as the generative model, and the temperature value was set to 0.7.

The experimental setup is based on 100 parallel English/Turkish examples selected from the SQuAD validation set. For each example, fields for the English and Turkish questions, the context paragraph, and the reference answer were used. For each example and language, three different candidate response labels—correct, partially correct, and incorrect—were evaluated. Additionally, each evaluator model scored the candidate responses using two different prompt strategies: simple and rubric-based. The experiment was designed with 5 models, 100 examples, 2 languages, 3 response labels, 2 prompt strategies, and 1 repetition. As a result of this structure, a total of 6,000 model judgments were obtained. These judgments were generated from 100 underlying parallel question–answer pairs and therefore do not represent 6,000 independent experimental units.

4.2. Evaluation Procedure

The evaluation process was conducted on 100 parallel English/Turkish examples selected from the SQuAD validation set. For each example, the question, reference answer, and candidate answer were provided to the relevant evaluator model. The original context paragraph was not included in the judge-model input. The candidate responses were created at three different levels of accuracy: correct, partially correct, and incorrect; each candidate response was evaluated in both the English and Turkish versions. The evaluator models were instructed to generate a score between 1 and 5 by comparing the given candidate response with the reference response.

Each model was run through the same evaluation workflow. Each of the five evaluators scored candidate responses for 100 examples across two languages, three candidate response labels, and two prompt strategies. Two different types of prompts were used in the prompting process: simple prompts and rubric-based prompts. In the simple prompt, evaluation was based on the criteria of correctness and clarity; in the rubric-based prompt, however, evaluation was structured around the dimensions of correctness, clarity, and completeness. The experimental design was implemented as $5 \text{ models} \times 100 \text{ examples} \times 2 \text{ languages} \times 3 \text{ labels} \times 2 \text{ prompts} \times 1 \text{ repetition}$, resulting in a total of 6,000 judgments.

The model scores obtained were converted to class labels using a predefined score-to-label conversion rule. Accordingly, responses with a score ≤ 2 were considered

incorrect, responses with a score = 3 were considered partially correct, and responses with a score ≥ 4 were considered correct. Following this process, the model scores were compared with the hybrid reference scores, and the transformed model labels were compared with the reference labels. The evaluation results were analyzed using Spearman’s rank correlation, accuracy, macro-F1, and Cohen’s kappa coefficient. Additionally, prompt sensitivity and English–Turkish score differences were calculated in a paired manner using the same examples.

4.3. Reproducibility Settings

To ensure the reproducibility of the experiments, a fixed and predefined experimental framework was followed during the evaluation process. The dataset was limited to 100 parallel English/Turkish examples selected from the SQuAD validation set; for each example, the question, context paragraph, and reference answer fields were used during dataset preparation and candidate-response generation. The candidate responses were generated under three categories correct, partially correct, and incorrect and all judge models evaluated them using the same candidate response structure. Thus, model comparisons were conducted based on the same data, the same language conditions, and the same response categories.

The model settings were also kept constant to ensure comparability. While the GPT-4o-mini model was used with a temperature of 0.7 to generate candidate responses, all evaluator models were run with a temperature of 0.0 during the evaluation phase. Each model was evaluated on the same 100 examples in two languages—English and Turkish—using three candidate response labels and two prompt strategies. The experimental design was implemented as 5 models $\times$ 100 examples $\times$ 2 languages $\times$ 3 labels $\times$ 2 prompts $\times$ 1 repetition, resulting in a total of 6,000 judgments.

During the analysis phase, predefined rules were used for scoring and labeling. A hybrid reference scoring scheme was adopted: 25 examples, corresponding to 150 candidate responses, were manually scored by the author on a 1–5 scale, while the remaining 75 examples were scored using rule-based proxy scores. For the proxy-scored examples, correct, partially correct, and incorrect answers were assigned scores of 5, 3, and 1, respectively. When converting model scores to labels, the following rule was applied: a score of ≤ 2 was considered incorrect, a score of 3 was considered partially correct, and a score of ≥ 4 was considered correct. Confidence intervals were calculated at the 95% level using 1,000 bootstrap resamples; the significance of language and prompt effects was assessed using the Wilcoxon signed-rank test, with Benjamini–Hochberg adjustment applied to the prompt-sensitivity analyses, while differences between models were evaluated using the Friedman test. The human-scored subset was included to support the reference scoring framework, but because it was annotated by a single evaluator, it was not used to estimate inter-annotator agreement.

5. Results

5.1. Model Performance Results

The performance of the evaluator models was analyzed from two complementary perspectives: their agreement with the hybrid reference scores and their ability to classify candidate responses into the correct, partially correct, and

incorrect categories. This distinction is important because a model may preserve the overall ranking of response quality while still showing limited performance in discrete class assignment. Accordingly, Table 2 reports both rank-based agreement and classification-based performance metrics for each model under the English and Turkish evaluation settings.

Table 2. Agreement and classification performance metrics by model and language.

Model	Language	Spearman ρ	Accuracy	Macro-F1	Cohen’s κ
GPT-4o-mini	English	0.925	0.747	0.695	0.620
GPT-4o-mini	Turkish	0.867	0.728	0.687	0.593
Claude Sonnet 4.6	English	0.908	0.772	0.735	0.658
Claude Sonnet 4.6	Turkish	0.901	0.753	0.717	0.630
GPT-5.4-mini	English	0.903	0.670	0.552	0.505
GPT-5.4-mini	Turkish	0.901	0.692	0.611	0.538
Claude Haiku 4.5	English	0.890	0.678	0.577	0.518
Claude Haiku 4.5	Turkish	0.870	0.705	0.642	0.558
GPT-4.1-mini	English	0.882	0.702	0.639	0.553
GPT-4.1-mini	Turkish	0.849	0.705	0.656	0.558

As shown in Table 2, all models produced high Spearman correlation coefficients in both English and Turkish evaluations, indicating that their score rankings were generally consistent with the expected ordering of response quality. In the English evaluation, the highest Spearman correlation was obtained by GPT-4o-mini ($\rho = 0.925$), followed by Claude Sonnet 4.6 ($\rho = 0.908$) and GPT-5.4-mini ($\rho = 0.903$). In the Turkish evaluation, the highest Spearman correlation was observed for Claude Sonnet 4.6 and GPT-5.4-mini, both with $\rho = 0.901$, followed by Claude Haiku 4.5 ($\rho = 0.870$) and GPT-4o-mini ($\rho = 0.867$). However, the classification metrics reveal a different pattern. Claude Sonnet 4.6 achieved the highest accuracy, macro-F1, and Cohen’s κ values in both languages, with an accuracy of 0.772, a macro-F1 score of 0.735, and $\kappa = 0.658$ in English, and an accuracy of 0.753, a macro-F1 score of 0.717, and $\kappa = 0.630$ in Turkish. Although GPT-5.4-mini produced one of the highest Spearman correlation values in both languages, its classification performance was lower, particularly in English, where it achieved an accuracy of 0.670, a macro-F1 score of 0.552, and $\kappa = 0.505$. This finding indicates that high rank correlation does not necessarily imply equally strong classification performance across the three response categories, namely correct, partially correct, and incorrect.

These results show that evaluating differences between models based solely on correlation coefficients is not sufficient. In particular, differences in classification performance indicate that, although the models exhibit similar ranking behavior, they may apply decision boundaries differently. Therefore, a visual comparison of model accuracy helps to highlight these performance differences more clearly. This distinction is particularly important when assessing the practical reliability of judge models, since similar correlation levels may conceal meaningful differences in categorical decision performance. To provide a more direct comparison of classification performance across models and languages, Figure 2 visualizes the accuracy scores reported in Table 2.

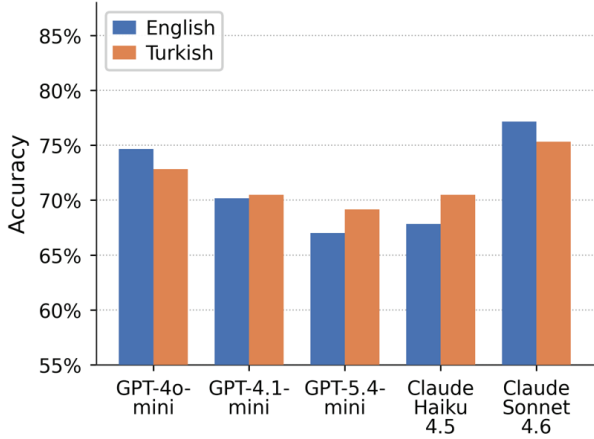


Figure 2. Classification accuracy of LLM-as-a-Judge models in English and Turkish.

According to Figure 2, the highest classification accuracy was achieved by the Claude Sonnet 4.6 model. This model achieved an accuracy of approximately 77% for English and approximately 75% for Turkish. Although GPT-4o-mini demonstrated strong performance in English, it produced lower accuracy in Turkish. In contrast, the GPT-5.4-mini and Claude Haiku 4.5 models showed relatively lower accuracy in English. Overall, the results indicate that the models’ classification performance is at a medium-to-high level, but not all models demonstrate the same level of balanced performance.

Figure 3 presents the confusion matrix of Claude Sonnet 4.6 under the Turkish evaluation setting. This model was examined in greater detail because it achieved the most balanced overall classification performance among the evaluated judge models. The confusion matrix provides a more fine-grained view of model behavior by showing how often each true response category was assigned to the correct or incorrect predicted label.

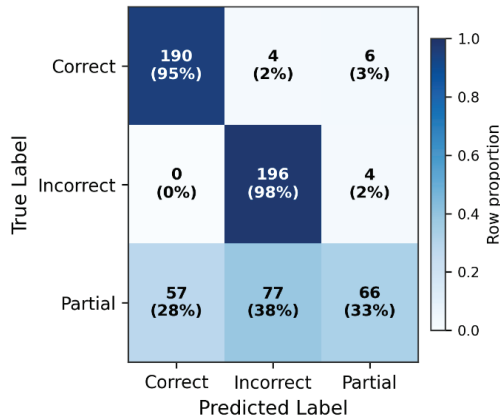


Figure 3. Confusion matrix of Claude Sonnet 4.6 for Turkish evaluation.

As shown in Figure 3, Claude Sonnet 4.6 classified the clearly defined categories with high reliability, correctly identifying 98% of incorrect responses and 95% of correct responses. In contrast, performance on the partially correct category was substantially weaker, with only 33% of partial responses correctly assigned to the partial class. Most errors in this category involved confusion with the incorrect label,

accounting for 38% of partial responses, whereas 28% were misclassified as correct. This pattern indicates that the model was highly effective in separating clearly correct and clearly incorrect answers, but less reliable in identifying responses with intermediate or borderline quality.

Figure 4 presents the confusion matrix of Claude Sonnet 4.6 under the English evaluation setting, allowing a class level comparison of prediction behavior across correct, incorrect, and partially correct responses.

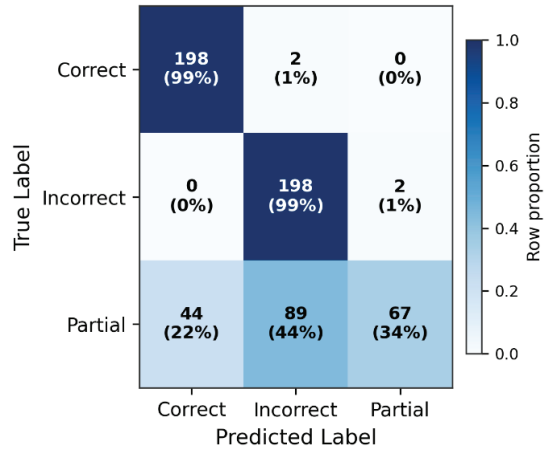


Figure 4. Confusion matrix of Claude Sonnet 4.6 for English evaluation.

As shown in Figure 4, Claude Sonnet 4.6 demonstrated very high classification accuracy for the clearly defined categories in the English evaluation setting, correctly identifying 99% of correct responses and 99% of incorrect responses. In contrast, performance on partially correct responses remained limited, with only 34% assigned to the partial class. Most errors in this category were due to confusion with the incorrect label, which accounted for 44% of partial responses, whereas 22% were misclassified as correct. This pattern confirms that, even under the English evaluation condition, the model was highly reliable for clearly correct and clearly incorrect answers but substantially less effective in identifying responses with intermediate quality.

To examine overall agreement across languages, the English and Turkish evaluation results were combined and analyzed jointly. Table 3 reports the combined correlation results between judge scores and the hybrid reference scores for each model.

Table 3. Combined English-Turkish correlation between judge scores and hybrid reference scores (n = 1,200 per model).

Model	Spearman ρ	Lower CI	Upper CI	Pearson r	p value
GPT-5.4-mini	0.902	0.885	0.916	0.886	< 0.001
Claude Sonnet 4.6	0.905	0.888	0.920	0.900	< 0.001
GPT-4o-mini	0.896	0.881	0.910	0.891	< 0.001
Claude Haiku 4.5	0.880	0.864	0.896	0.873	< 0.001
GPT-4.1-mini	0.865	0.843	0.885	0.857	< 0.001

As shown in Table 3, all models produced high combined Spearman correlation values, indicating strong overall agreement with the hybrid reference scores across the bilingual evaluation setting. Claude Sonnet 4.6 achieved the highest combined Spearman correlation with a value of 0.905, followed closely by GPT-5.4-mini with 0.902 and GPT-4o-mini with 0.896. In terms of linear association, the highest Pearson correlation coefficient was obtained by Claude Sonnet 4.6 with a value of 0.900. Moreover, the p values were below 0.001 for all models, indicating that the observed relationships between model scores and the hybrid reference scores were statistically significant.

To examine scoring behavior beyond classification-based performance, Table 4 reports the mean judge scores assigned to correct, partially correct, and incorrect responses in English and Turkish, averaged across the two prompt strategies. These mean scores provide a more direct view of how each model used the 1 to 5 scale and how distinctly it separated the three response quality categories.

Table 4. Mean judge scores for each candidate answer label and language.

Model	Language	Correct	Partially Correct	Incorrect
GPT-4o-mini	English	4.93	2.94	1.28
GPT-4o-mini	Turkish	4.65	2.82	1.46
Claude Sonnet 4.6	English	4.96	2.65	1.12
Claude Sonnet 4.6	Turkish	4.82	2.78	1.15
Claude Haiku 4.5	English	4.89	2.67	1.19
Claude Haiku 4.5	Turkish	4.80	2.71	1.32
GPT-4.1-mini	English	4.80	2.82	1.21
GPT-4.1-mini	Turkish	4.69	2.96	1.24
GPT-5.4-mini	English	4.92	2.60	1.04
GPT-5.4-mini	Turkish	4.84	3.18	1.10

As shown in Table 4, correct responses are assigned high scores, incorrect responses are assigned low scores, and partially correct responses are assigned intermediate scores across all models. This pattern indicates that the models are able to distinguish between the three response categories in terms of their scoring behavior. While the average scores for correct answers range from 4.65 to 4.96, the average scores for incorrect answers range from 1.04 to 1.46. Partially correct answers, however, show greater variability across models and languages. This suggests that the models are able to distinguish between correct and incorrect answers more consistently, whereas model and language-based differences are more pronounced in the scoring of partially correct answers.

5.2. Statistical Analysis

To further examine the consistency of model scoring behavior, additional statistical analyses were conducted with a particular focus on cross linguistic differences and prompt-related variation. As an initial step, score differences between

the English and Turkish evaluation settings were analyzed to determine whether the models exhibited systematic language dependent scoring patterns. The results of the Wilcoxon signed-rank test for paired English and Turkish scores are presented in Table 5.

Table 5. Wilcoxon signed-rank test results for the difference between English and Turkish scores.

Model	Statistic	p value	Effect Size (r)	Significant?
GPT-5.4-mini	5219.5	0.000117	-0.330	Yes
GPT-4o-mini	12362.5	0.084	+0.123	No
Claude Haiku 4.5	13633.5	0.496	-0.049	No
GPT-4.1-mini	14312.0	0.557	-0.042	No
Claude Sonnet 4.6	9673.5	0.727	-0.028	No

As shown in Table 5, a statistically significant difference between the English and Turkish scores was observed only for GPT-5.4-mini, with p = 0.000117 and an effect size of r = -0.330. For all other models, the p values remained above the significance threshold, indicating no statistically significant language related score difference. These results suggest that the effect of language on scoring behavior was model dependent and was most pronounced in GPT-5.4-mini.

To examine the effect of prompt type on model scores, paired simple-prompt and rubric-based-prompt scores were compared using the Wilcoxon signed-rank test. Prompt sensitivity was summarized separately as the mean absolute difference between the corresponding mean scores. Table 6 presents the mean sensitivity values, 95% confidence intervals, p values, effect sizes, and significance outcomes for each model and language.

Table 6. Wilcoxon signed-rank test results for prompt-type differences, with mean sensitivity, 95% CIs, and BH-adjusted p-values.

Model	Language	Mean Sensitivity	95% CI	p value	p (BH)	Effect Size	Significant?
GPT-4.1-mini	English	0.230	[0.177, 0.287]	< 0.001	< 0.001	0.884	Yes
GPT-4o-mini	Turkish	0.230	[0.177, 0.290]	< 0.001	< 0.001	0.793	Yes
GPT-5.4-mini	Turkish	0.163	[0.113, 0.210]	< 0.001	< 0.001	0.754	Yes
Claude Haiku 4.5	Turkish	0.100	[0.070, 0.133]	0.028	0.0569	-0.400	No
Claude Sonnet 4.6	English	0.033	[0.013, 0.057]	0.002	0.0039	1.000	Yes
GPT-4.1-mini	Turkish	0.173	[0.127, 0.227]	0.050	0.083	0.310	No
GPT-4o-mini	English	0.117	[0.080, 0.157]	0.139	0.1737	0.257	No
GPT-5.4-mini	English	0.133	[0.087, 0.183]	0.137	0.1737	-0.286	No
Claude	Turkish	0.087	[0.053, 0.121]	0.705	0.783	0.077	No

Sonnet 4.6	h		0.120]		8		
Claude Haiku 4.5	English	0.057	[0.033, 0.083]	0.819	0.8185	0.059	No

As shown in Table 6, prompt sensitivity varied across both models and languages. The highest mean sensitivity was observed for GPT-4.1-mini in English and GPT-4o-mini in Turkish, both with a value of 0.230, followed by GPT-4.1-mini in Turkish with 0.173 and GPT-5.4-mini in Turkish with 0.163. After applying the Benjamini–Hochberg correction, statistically significant prompt effects were observed for GPT-4.1-mini in English, GPT-4o-mini in Turkish, GPT-5.4-mini in Turkish, and Claude Sonnet 4.6 in English. Among the Claude models, Claude Sonnet 4.6 exhibited a statistically significant but very small sensitivity in English, whereas Claude Haiku 4.5 in Turkish was not significant after correction. Overall, these results indicate that prompt formulation affected the scoring behavior of some models more strongly than others, and that this effect varied depending on both the model and the evaluation language.

To examine how scoring behavior changed across prompt types, prompt sensitivity was quantified as the mean absolute difference between the corresponding mean scores obtained under the simple and rubric-based prompt conditions for each model in the English and Turkish evaluation settings. Figure 5 presents these prompt sensitivity values and provides a visual comparison of how strongly each model’s scoring behavior varied across the two prompt conditions.

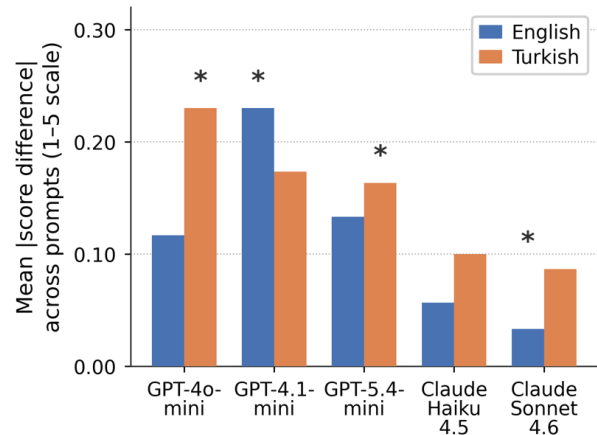


Figure 5. Prompt sensitivity across English and Turkish, measured as the mean absolute score difference between the simple and rubric prompts (1–5 scale). Error bars: 95% bootstrap CIs. Asterisks (*) denote Benjamini–Hochberg-adjusted significance across ten comparisons (Wilcoxon signed-rank test, $p < 0.05$).

As shown in Figure 5, prompt sensitivity differed substantially across models and languages. Among the OpenAI models, GPT-4o-mini exhibited the highest sensitivity in Turkish, whereas GPT-4.1-mini showed the highest sensitivity in English; both effects were statistically significant after Benjamini–Hochberg correction. GPT-5.4-mini also displayed noticeable prompt-related variation, with a statistically significant effect in Turkish but not in English. In contrast, the Claude models were generally less sensitive to prompt type. Claude Haiku 4.5 showed moderate sensitivity in Turkish in terms of mean score difference, but this effect was not statistically significant after correction. Claude Sonnet 4.6

exhibited the lowest sensitivity overall, although its small English prompt effect was statistically significant. These results indicate that prompt formulation influenced the scoring behavior of some models more strongly than others, particularly for GPT-4o-mini in Turkish, GPT-4.1-mini in English, and GPT-5.4-mini in Turkish, while Claude Sonnet 4.6 produced comparatively more stable evaluations across prompt conditions.

Finally, to determine whether the five judge models differed significantly in their scoring behavior, a Friedman test was conducted separately for each language, using the paired scores assigned to the same candidate responses across models. This analysis complements the preceding Wilcoxon tests by shifting the focus from language and prompt effects to differences among the judge models themselves. The results are presented in Table 7.

Table 7. Friedman test results for differences among the five judge models, reported separately for each language

Language	Friedman χ^2	df	p value
English	74.63	4	< 0.001
Turkish	23.68	4	< 0.001

As shown in Table 7, the Friedman test revealed statistically significant differences among the judge models in both English ($\chi^2(4) = 74.63$, $p < 0.001$) and Turkish ($\chi^2(4) = 23.68$, $p < 0.001$). These results confirm that the choice of judge model has a significant effect on the assigned scores in both languages. The larger test statistic obtained for English indicates that inter-model differences were more pronounced in the English evaluation setting, whereas the models produced comparatively more homogeneous scores in Turkish. Taken together with the classification and correlation results, this finding reinforces that judge-model selection is a decisive factor in multilingual LLM-as-a-Judge evaluation.

6. Discussion

The findings from this study provide promising evidence that large language models can be used as automatic evaluators in question-answering assessment tasks. The fact that Spearman’s correlation coefficients were high for all five models examined indicates that the models generally exhibited a ranking behavior consistent with the hybrid reference scoring framework. This result suggests that the LLM-as-a-Judge approach may serve as a viable evaluation method not only in high-resource languages such as English but also in morphologically rich languages such as Turkish, under controlled experimental conditions.

However, the performance of the models should not be evaluated solely based on correlation values. The findings reveal that, despite high rank correlation, there are significant differences among the models in terms of classification performance and macro-F1 scores. This indicates that, although the models can rank responses relatively accurately, they are not equally successful when converting them into discrete categories such as “correct,” “partially correct,” and “incorrect.” Therefore, in LLM-as-a-Judge systems, correlation metrics must be evaluated alongside classification-based metrics such as accuracy, macro-F1, and Cohen’s kappa coefficient. This finding also shows that the thresholding and calibration strategies used to convert model scores into class labels can have a decisive impact on the results. In this study, the proxy score for partially correct responses was set to 3 to ensure consistency with the score-to-label conversion rule. Nevertheless, because the “partially correct” category is more

ambiguous than the clearly correct and incorrect categories, future analyses should examine alternative score-to-label thresholding schemes to test the robustness of classification results.

In terms of model comparisons, Claude Sonnet 4.6 demonstrated the most balanced performance. This model showed high classification accuracy in both English and Turkish results, exhibited only limited differences between the languages, and demonstrated a relatively stable structure in the face of prompt changes. GPT-4o-mini, on the other hand, achieved the highest Spearman correlation on the English dataset but showed a more pronounced decline in Turkish. This result should be interpreted with caution because GPT-4o-mini was used both to generate candidate responses and as one of the judge models, which may introduce same-model or self-preference bias. This finding indicates that, in multilingual evaluation scenarios, model selection should not be based solely on the highest performance in a single language, but should also take into account cross-lingual consistency and robustness. However, it should be noted that these comparisons are limited by the dataset used, the candidate response generation method, the prompt design, and the scoring protocol.

Another notable finding of the study is the weaker class-level discrimination observed for GPT-5.4-mini. Despite its competitive Spearman correlation values, this model showed lower classification performance, particularly for the “partially correct” class. The lower macro-F1 score suggests that a substantial portion of the classification errors was associated with this intermediate category. This suggests that while the model preserves the relative ranking of responses, it may not use the scoring scale with sufficient calibration for fine-grained classification. Therefore, although GPT-5.4-mini appears suitable for ranking-based evaluations, it should be used with caution in three-class classification tasks. In particular, when evaluating the “partially correct” class, it may be necessary to calibrate the score ranges and class thresholds on a per-model basis.

The results regarding language differences and prompt sensitivity also highlight the importance of calibration when using LLM-as-a-Judge. The fact that a significant difference was found only between the English and Turkish scores for GPT-5.4-mini suggests that some models may exhibit different scoring tendencies when evaluating Turkish responses. Furthermore, the fact that prompt sensitivity is more pronounced in OpenAI models suggests that evaluation outputs may be influenced by the way prompts are phrased. For this reason, particularly in Turkish evaluation scenarios, prompt language, rubric details, and scoring criteria should be treated as fundamental components of the experimental design. It should also be noted that the simple and rubric-based prompts differed slightly in the verbal descriptions of the scoring scale, which may have contributed to the observed prompt-sensitivity effects. These results suggest that, in multilingual LLM-as-a-Judge systems, a single general prompt or fixed scoring scale may not be equally suitable for every model and language.

Overall, the findings indicate that the LLM-as-a-Judge approach shows promise for evaluating Turkish question-answering systems; however, for reliable use, model selection, language effects, prompt design, and score calibration must be evaluated together. Since the reference scoring framework was only partially based on manual human scoring, the findings should not be interpreted as direct evidence of full alignment with human judgment, but rather as preliminary evidence of agreement with a partially human-annotated and partially

proxy-based reference structure. Furthermore, although a total of 6,000 model judgments were obtained, the number of fundamental independent examples was 100 question-answer pairs. The remaining observations were repeated evaluations derived from these same underlying examples across models, languages, response categories, and prompt conditions. Consequently, observations originating from the same question may exhibit structural dependence because they share the same question, context, and reference answer. Therefore, the Wilcoxon, Friedman, and observation-level bootstrap results should be interpreted with this dependence structure in mind. Another limitation is the possible same-model bias caused by using GPT-4o-mini both as the candidate-response generator and as one of the judge models. Future studies should separate the response-generation and evaluation roles more strictly by using a generator model that is not included among the judge models or by generating candidate responses from multiple independent models. In future work, the number of examples and the coverage of human annotations will be increased, and multiple independent annotators will be included to provide stronger evidence of reliability. Future studies should also examine model behavior across different task types and develop language-specific calibration strategies aimed specifically at more reliably distinguishing “partially correct” responses. Mixed-effects models or question-level grouped bootstrap approaches could also be used to draw more precise conclusions. The 100 examples were selected through deterministic sequential truncation rather than random or stratified sampling. Consequently, the sample may not fully represent the broader distribution of question types, topics, and difficulty levels in the complete aligned SQuAD-SQuAD-TR pool. Future studies should therefore evaluate the robustness of the findings using larger randomly selected or stratified samples.

7. Conclusion

This study systematically investigated the feasibility of using large language models as automatic evaluators in English and Turkish question-answering assessments. The findings reveal that all models examined exhibited a high degree of rank correlation with the hybrid reference scores and that the LLM-as-a-Judge approach offers a promising evaluation method even in a morphologically complex language such as Turkish, under certain experimental conditions.

The results show that model selection is a determining factor in multilingual evaluation performance. Claude Sonnet 4.6 emerged as the most balanced model in terms of classification performance, cross-language consistency, and relative stability against prompt changes. While GPT-4o-mini produced the highest rank correlation in the English evaluation, Claude Sonnet 4.6 and GPT-5.4-mini shared the highest Spearman correlation in the Turkish evaluation, and Claude Sonnet 4.6 achieved the highest correlation in the combined English-Turkish results. However, the fact that GPT-5.4-mini struggled to distinguish the “partially correct” category clearly demonstrates that high correlation values alone are not a sufficient indicator of reliability.

These findings demonstrate that, for LLM-as-a-Judge systems to be used reliably, correlation-based metrics must be evaluated in conjunction with classification performance, language effects, prompt sensitivity, and score calibration. The study provides an empirical foundation based on preliminary findings regarding the use of multilingual LLM evaluators for Turkish question-answering evaluation and offers guiding

insights for future research with larger datasets and broader human annotation coverage. Since the reference scoring framework was only partially based on manual human scoring, the findings should not be interpreted as direct evidence of full alignment with human judgment, but rather as preliminary evidence of agreement with a partially human-annotated and partially proxy-based reference structure. Future work may extend this analysis by increasing the number of examples, expanding the coverage of human annotations, and including multiple independent annotators to provide stronger evidence of reliability.

8. References

- [1] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, and H. Liu, “From generation to judgment: Opportunities and challenges of LLM-as-a-Judge,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 2757–2791. doi: 10.18653/v1/2025.emnlp-main.138.
- [2] A. Türkmen and K. E. Keleş, “Whom to trust? Analyzing the divergence between user satisfaction and LLM-as-a-Judge in e-commerce RAG systems,” in *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, 2026. pp. 189–195. doi: 10.18653/v1/2026.mme-main.12.
- [3] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, “Limitations of the LLM-as-a-Judge approach for evaluating LLM outputs in expert knowledge tasks,” in *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*, 2025. doi: 10.1145/3708359.3712091.
- [4] C. Toraman et al., “TurkBench: A benchmark for evaluating Turkish large language models,” in *Proceedings of the Second Workshop on Natural Language Processing for Turkic Languages*, 2026, pp. 126–154. doi: 10.18653/v1/2026.sigturk-1.12.
- [5] J. Yu, S. Sun, X. Hu, J. Yan, K. Yu, and X. Li, “Improve LLM-as-a-Judge ability as a general ability,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 14099–14115. doi: 10.18653/v1/2025.emnlp-main.712.
- [6] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, and T. Zhao, “An empirical study of LLM-as-a-Judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 5880–5895. doi: 10.18653/v1/2025.findings-acl.306.
- [7] C. Yamasaki, K. Sugahara, Y. Nagi, and K. Okamoto, “Function-based labels for complementary recommendation: Definition, annotation, and LLM-as-a-Judge,” *Pattern Recognition Letters*, vol. 200, pp. 8–15, 2026. doi: 10.1016/j.patrec.2025.11.042.
- [8] T. Khan, M. Rafiq, and A. Michalas, “AI on Trial: LLM-as-a-Judge for private and reliable clinical decision-making,” in *2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops)*, 2025, pp. 256–262. doi: 10.1109/ACSACW69556.2025.00033.
- [9] B. Bellard, X. C. Liu, R. Lohani, S. Raza, P. E. Velmovitsky, Q. Pham, and S. Saha, “Development and evaluation of an agentic RAG with embedded LLM-as-a-Judge for heart failure management: A pilot study,” in *2025 IEEE International Conference on Big Data (BigData)*, 2025, pp. 7770–7774. doi: 10.1109/BigData66926.2025.11402164.
- [10] R. Mahadik, A. Mandal, S. Tripathi, and S. Jaiswal, “LLM evaluations for emotional expressiveness and factual consistency in medical dialogue systems using LLM-as-a-Judge,” in *2025 International Conference on Responsible, Generative and Explainable AI (ResGenXAI)*, 2025. doi: 10.1109/ResGenXAI64788.2025.11344029.
- [11] M. I. H. Abir, N. Ferdous, A. Tasnim, and N. Mustaqeem, “Assessing large language models on zero-shot poetry summarization using ROUGE, BERT, and LLM-as-a-Judge metrics,” in *2026 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, 2026. doi: 10.1109/ICECTE69292.2026.11429378.
- [12] G. Crupi, R. Tufano, A. Velasco, A. Mastropaolo, D. Poshyvanyk, and G. Bavota, “On the effectiveness of LLM-as-a-Judge for code generation and summarization,” *IEEE Transactions on Software Engineering*, vol. 51, no. 8, pp. 2329–2345, 2025. doi: 10.1109/TSE.2025.3586082.
- [13] D. Chen, R. Chen, S. Zhang, Y. Wang, Y. Liu, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun, “MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark,” in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 2024, pp. 6562–6595. doi: 10.5555/3692070.3692324.

Özgeçmişler



Dr. Ilkay Yelmen is an Assistant Professor in the Department of Software Engineering, Faculty of Engineering and Natural Sciences, Işık University, Türkiye. He received his B.Sc. degree in Software Engineering from Istanbul Aydın University in 2013, his M.Sc. degree in Computer Engineering in 2016, and his Ph.D. degree in Computer Engineering in 2023 from the same university. His academic and research interests focus on computer science, including machine learning, deep learning, natural language processing and large language models.