Crossing Language Boundaries: Evaluation of Large Language Models on Urdu-English Question Answering

Samreen Kazi[1], Maria Rahim[2], Shakeel Khoja[3]
School of Mathematics & Computer Science
Institute of Business Administration (IBA), Karachi, Pakistan
[1]sakazi@iba.edu.pk, [2]mrkhowaja@iba.edu.pk, [3]skhoja@iba.edu.pk

Abstract

This study evaluates the question-answering capabilities of Large Language Models (LLMs) in Urdu, addressing a critical gap in low-resource language processing. Four models GPT-4, mBERT, XLM-R, and mT5 are assessed across monolingual, cross-lingual, and mixed-language settings using the UQuAD1.0 and SQuAD2.0 datasets. Results reveal significant performance gaps between English and Urdu processing, with GPT-4 achieving the highest F_1 scores (89.1% in English, 76.4% in Urdu) while demonstrating relative robustness in cross-lingual scenarios. Boundary detection and translation mismatches emerge as primary challenges, particularly in cross-lingual settings. The study further demonstrates that question complexity and length significantly impact performance, with factoid questions yielding 14.2% higher F_1 scores compared to complex questions. These findings establish important benchmarks for enhancing LLM performance in low-resource languages and identify key areas for improvement in multilingual question-answering systems.

1 Introduction

The rapid advancement of LLMs has revolutionized natural language processing, demonstrating remarkable capabilities in various tasks, particularly in English and other high-resource languages. However, their effectiveness in low-resource languages, such as Urdu, remains a critical area requiring systematic evaluation. As Lewis et al. (2020) demonstrated that Question Answering (QA), as a fundamental test of language understanding, serves as an excellent probe for assessing these models' cross-lingual and multilingual capabilities.

Wu and Dredze (2022) highlighted significant disparities in the performance of large language models (LLMs) between high-resource and low-resource languages. Similarly, Arif et al. (2024b) showed that while models like GPT-4 and mT5 achieve impressive results in English, their performance often degrades substantially when handling languages with limited training data or complex morphological structures. Furthermore, Daud et al. (2017), Rahim and Khoja (2024), and Kazi et al. (2023) emphasized that Urdu, spoken by approximately 170 million people worldwide, serves as a particularly intriguing case study due to its rich morphological structure, distinct script, and limited computational resources.

The challenge of cross-lingual question answering has gained increasing attention in recent years. Clark et al. (2020) focused primarily on transfer learning and fine-tuning approaches. However, the emergence of large-scale multilingual models has opened new possibilities for zero-shot and cross-lingual applications. Conneau et al. (2020) demonstrated the potential of cross-lingual representation learning, while Pfeiffer et al. (2020) explored adapter-based approaches for cross-lingual transfer.

The development of Urdu-specific resources has also seen notable progress. Kazi and Khoja (2021) created UQuAD1.0, providing crucial benchmarks for evaluating model performance. These resources, combined with advances in multilingual model architectures, create an opportunity to systematically assess how well current LLMs handle cross-lingual and multilingual QA tasks involving Urdu. Kazi and Khoja (2024) proposed a context-aware QA framework tailored to Urdu, utilizing sliding window score specifically designed

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 1–11
January 20, 2025.

for comprehension of long-context dependencies. Their methodology sets a benchmark that aligns with this study's focus on evaluating cross-lingual model performance for low-resource languages.

Arif et al. (2024a) have shown that models with fewer parameters but more language-specific training often outperform larger, general-purpose models in Urdu NLP tasks. This finding raises important questions about the trade-offs between model size and language-specific optimization, as discussed by Chen et al. (2023). Furthermore, Wang et al. (2024) suggest that carefully designed prompting strategies can significantly impact cross-lingual performance.

The relationship between script systems and model performance presents another crucial consideration. Unlike languages that use Latin script, Rahman et al. (2023) note that Urdu's Nastaliq script introduces additional complexity in text processing and token alignment. Wang et al. (2019) demonstrated that script differences can significantly impact model performance in cross-lingual tasks, making this an important factor in our evaluation.

Our work makes several key contributions to this developing field:

- We present the first comprehensive evaluation of LLMs' question answering capabilities across monolingual, cross-lingual, and mixed-language settings involving Urdu.
- We analyze performance patterns across different question types and lengths, providing insights into the models' handling of varying complexity levels.
- We identify and quantify specific challenges in cross-script processing and boundary detection, offering valuable insights for future model development.
- We establish benchmark results for four major LLMs (GPT-4, mBERT, XLM-R, and mT5) in Urdu QA tasks, providing a foundation for future research.

Our evaluation framework includes five experimental settings: (E1) full Urdu prompts, (E2) Urdu questions with English context, (E3) English questions with Urdu context, (E4) full English prompts, and (E5) mixed-language prompts. This setup allows us to examine various cross-lingual comprehension and generation challenges.

The findings reveal significant performance gaps, with models experiencing noticeable degradation in Urdu and cross-lingual settings. GPT-4, for instance, achieves an F_1 score of 89.1% in English but drops to 76.4% in Urdu, with further declines in cross-lingual tasks. These results underscore the complexities of multilingual model development and the need for progress in low-resource languages like Urdu.

This study contributes valuable insights into LLMs' cross-lingual limitations, emphasizing the ongoing need for robust multilingual modeling, especially for morphologically complex languages.

The remainder of this paper is organized as follows: Section 2 provides a review of related work, highlighting key advancements and challenges in multilingual NLP and cross-lingual question answering. Section 3 gives details of the methodology, including models selected and prompting techniques. Section 4 describes the datasets used and experiments done. Section 5 presents the results and discussion, focusing on performance gaps, question type analysis, and error patterns. Section 6 outlines the limitations of the current study.

2 Related Work

The exploration of large language models (LLMs) in multilingual contexts, particularly for low-resource languages like Urdu, has garnered significant attention in recent years. This literature review examines key studies that have contributed to understanding and advancing LLMs' capabilities in cross-lingual question answering (QA) and related tasks.Cross-lingual QA involves answering questions in one language based on context provided in another, posing unique challenges for LLMs. Zhou et al. (2021) investigated zero-shot cross-lingual transfer for multilingual QA over knowledge graphs, highlighting the difficulties LLMs face when transferring knowledge across languages without fine-tuning . Similarly, Riabi et al. (2020) proposed synthetic data augmentation to en-

hance zero-shot cross-lingual QA performance, demonstrating that generating synthetic data in target languages can improve model accuracy without additional annotated data . The scarcity of high-quality datasets in Urdu has been a significant barrier to developing effective NLP models. To address this, Arif et al. (2024a) introduced UQA, a corpus for Urdu QA generated by translating the Stanford Question Answering Dataset (SQuAD2.0) using the EATS technique, which preserves answer spans in translated contexts . Additionally, Kazi and Khoja (2021) developed UQuAD1.0, an Urdu QA dataset combining machine-translated SQuAD data with human-generated samples, providing a substantial resource for training Urdu QA models . Evaluating LLMs on low-resource languages like Urdu has revealed performance disparities compared to high-resource languages. A study by Arif et al. (2024b) assessed general-purpose models such as GPT-4-Turbo and Llama-3-8b against specialized models fine-tuned on specific tasks, focusing on classification and generation tasks in Urdu. The findings indicated that models with fewer parameters but more language-specific data performed better than larger models with less language-specific data, underscoring the importance of tailored training for low-resource languages Prompting techniques play a crucial role in zero-shot learning scenarios, where models are expected to perform tasks without task-specific training. Agarwal et al. (2022) explored zero-shot cross-lingual open-domain QA, emphasizing the impact of prompt design on model performance across languages. Their work suggests that carefully crafted prompts can enhance LLMs' ability to generalize across languages, even in the absence of fine-tuning. Despite advancements, challenges persist in developing LLMs for low-resource languages. The limited availability of high-quality training data, coupled with inherent linguistic complexities, hampers model performance. Future research should focus on creating comprehensive multilingual datasets, developing effective cross-lingual transfer learning techniques, and designing models that can adapt to the nuances of low-resource languages like Urdu. In summary, while significant progress has been made in cross-lingual QA and the development of resources for low-resource languages, ongoing efforts are essential to bridge the performance gap between high-resource and low-resource languages in NLP applications.

3 Methodology

This study investigates the performance of large language models (LLMs) on Urdu Question Answering (QA) using zero-shot and cross-lingual prompts. We evaluate multiple models, explore various prompt settings, and assess model responses to identify the strengths and limitations of LLMs in a low-resource language context.

3.1 Models Selected

We selected the following LLMs for evaluation, focusing on their capacity for multilingual understanding:

- **GPT-4**: Known for its strong multilingual capabilities, particularly with zero-shot and few-shot prompts (OpenAI, 2023).
- **mBERT**: Multilingual BERT, pre-trained on 104 languages, commonly used for low-resource languages (Devlin et al., 2019).
- **XLM-R**: Cross-lingual XLM-RoBERTa, trained on 100 languages with enhanced performance in cross-lingual tasks (Conneau et al., 2020).
- **mT5**: A multilingual version of T5, which has demonstrated effectiveness in question-answering tasks across languages (Xue et al., 2020).

These models were selected based on their established performance in multilingual NLP tasks and availability for zero-shot or cross-lingual QA tasks.

3.2 Prompting Techniques

We employed a zero-shot prompting approach where models are given questions in Urdu without prior fine-tuning. The models are tested on their ability to understand and respond accurately in Urdu. Different prompt formats are tested to understand how prompt structure influences model performance:

- **Original Urdu Prompts**: Both the context and question are presented in Urdu, allowing us to evaluate the models' zero-shot capabilities in handling native Urdu input.

- **Translated Prompts**: Questions and context are translated between Urdu and English to create various cross-lingual scenarios, including:

  - **Urdu Question, English Context**: Tests comprehension when the question is in Urdu but context is in English.
  - **English Question, Urdu Context**: Tests understanding when the question is in English and context in Urdu.

- **Full Urdu Prompt**: Both the question and context are in Urdu.

- **Full English Prompt**: For comparison, we also provide English questions and contexts.

- **Mixed-Language Prompts**: Combining languages within the prompt to evaluate models' ability to bridge language gaps in real-time.

3.3 Evaluation Metrics

To assess model performance, we utilized the following evaluation metrics, which are standard in question-answering tasks:

- **Exact Match (EM)**: Measures the percentage of responses that exactly match the ground-truth answers, ensuring a strict assessment of accuracy.

- **F_1 Score**: Calculated based on the overlap of predicted answers with ground-truth answers, accounting for partial matches to capture nuanced correctness.

- **ROUGE-L**: Measures the longest common subsequence between the predicted and actual answer, providing insights into answer relevance.

4 Experimental Details

4.1 Data

In this study, we utilize the UQuAD1.0 (Kazi and Khoja, 2021) and SQuAD 2.0 (Rajpurkar et al., 2018) datasets to evaluate question-answering performance in Urdu and English, respectively. UQuAD1.0, specifically tailored for the Urdu language, comprises approximately 49,000 question-answer pairs, including 45,000 machine-translated pairs derived from SQuAD and 4,000 manually curated pairs to ensure linguistic and cultural relevance to Urdu. The manually curated QA pairs consists of diverse array of question types, categorized by cognitive difficulty as shown in Table 1. Since UQuAD1.0 is an extractive machine reading comprehension dataset, it exclusively includes questions with answers directly found as spans of text in the context, thereby excluding yes/no questions.

Statistic	**Value**
QA Pairs	4,000
Data Sources	Urdu Wikipedia, O-level content
Unique Paragraphs	1,972
Average Sentences per Paragraph	6.33
Average Paragraph Length	168.11 tokens 582.45 characters
Average Question Length	12.92 tokens or 43.70 characters
Average Answer Length	3.48 tokens 14.27 characters
Question Types	What When, Where, Who
Topics Covered	Politics, Religion, Education Miscellaneous

Table 1: Statistics of the Crowdsourced UQuAD1.0 Dataset

For English, we use SQuAD 2.0, an extensive dataset with over 130,000 question-answer pairs, including over 50,000 unanswerable questions crafted to challenge model comprehension.

Since UQuAD is a direct translation of SQuAD, it allows controlled cross-lingual experiments with consistent question-answer pairs in Urdu and English. This dual data set approach allows us to measure the zero-shot capabilities of the models in both low-resource (Urdu) and high-resource (English) contexts, providing a broad assessment of linguistic adaptability and cross-lingual understanding. Both datasets consists of:

- **Context**: A passage of text.
- **Question**: Question based on the passage.
- **Answer**: A text span from the passage.

4.2 Experiments

In this study, we used LLM to assess their performance in QA tasks, specifically focusing on their capabilities in a zero-shot cross-lingual environment for Urdu. Due to the limited availability of cross-lingual datasets tailored for QA in low-resource languages, our approach provides insights into the effectiveness of LLMs in handling QA tasks without extensive fine-tuning. For our experiments, temperature settings were not applicable since our task focused on answer span extraction rather than text generation. Span extraction relies on direct probability distributions over possible token positions, making temperature parameters unnecessary for this specific application. Each experimental configuration is assigned a unique identifier (E1, E2, etc.) to facilitate reference throughout the study, as shown in Table 8. The prompt settings are named as follows:

- **E1 - Full Urdu prompt:** In this setting, both the context and the question are provided in Urdu, using UQuAD1.0 exclusively. This prompt tests the model's ability to interpret and respond in Urdu, providing insights into its performance in low-resource language settings.
- **E2 - Urdu Question, English Context:** Here, the question is given in Urdu from UQuAD1.0, while the context is provided in English from SQuAD 2.0. This cross-lingual prompt evaluates the model's capacity to bridge language gaps, understanding a question in Urdu and finding answers in English.
- **E3 - English Question, Urdu Context:** For this setting, the question comes from SQuAD 2.0 in English, while the context is provided in Urdu from UQuAD1.0. This approach tests the model's ability to interpret context in Urdu while understanding and responding to an English question, further assessing its cross-lingual adaptability.
- **E4 - Full English Prompt:** Both the context and question are in English, sourced entirely from SQuAD 2.0. This monolingual English prompt acts as a baseline for evaluating model performance in a high-resource language environment.
- **E5 - Mixed Language Prompt:** In this prompt setting, context and question data are mixed between Urdu and English, combining inputs from both UQuAD1.0 and SQuAD 2.0. This configuration tests the model's adaptability to handle code-switching, evaluating its ability to seamlessly interpret and respond within a mixed linguistic framework.

Table 2 presents the performance comparison across different models and prompt settings, demonstrating each model's capacity to handle both monolingual and mixed-language inputs. Notably, GPT-4 consistently outperformed other models across all settings, showing robust exact match (EM), F_1, and ROUGE-L scores. The model performed particularly well in fully English settings (E4), achieving the highest overall scores. However, performance decreased for the same models when the prompts were fully in Urdu (E1) or in a mixed-language setting (E5). This underscores the challenges models face when processing low-resource languages directly without fine-tuning.

Table 3 provides a closer examination of cross-lingual scenarios, where the question and context are presented in different languages. Here, GPT-4 again leads in terms of F_1 and ROUGE-L scores, but its performance drops significantly in cross-lingual settings compared

to fully monolingual English prompts. For example, when tested with Urdu questions and English contexts (E2), as well as English questions and Urdu contexts (E3), we observed a reduction in F_1 scores by 3.8% and 4.7%, respectively. This indicates that even sophisticated models face difficulties bridging language gaps without fine-tuning, likely due to limited exposure to certain linguistic nuances during pretraining. Through this setup, we aim to provide a comprehensive evaluation of each model's strengths and limitations in handling both monolingual and cross-lingual prompts in Urdu. These prompt settings and naming conventions will be used consistently throughout the discussion sections, offering a structured view of model performance across varied linguistic scenarios.

5 Discussion

This section discusses the findings from results, focusing on performance gaps, question type analysis, error patterns, prompt setting impacts, and model-specific observations.

Language Performance Gap: An analysis of the language performance gap shows a marked decrease in model accuracy when transitioning from English to Urdu prompts. On average, EM scores dropped by 18.5%, F_1 scores by 12.7%, and ROUGE-L scores by 13.3% when shifting from English to Urdu. This significant drop highlights the models' limitations in handling low-resource languages, as well as the need for more language-specific training data to mitigate these gaps. The language performance gap is most apparent in mBERT and XLM-R, which are pretrained on a wide variety of languages but still struggle with Urdu-specific constructs and contextual understanding.

Question Type Analysis: UQuAD1.0, being an extractive machine reading comprehension dataset, exclusively contains questions with answers that are direct spans from the context. However, the models displayed varying levels of effectiveness across different question types. Factoid questions (e.g., Who, What, When, Where) showed a 14.2% higher F_1 score on average compared to complex questions (e.g., Why, How). This difference suggests that factoid questions are less context-dependent and simpler for models to answer accurately, whereas complex questions introduce greater ambiguity and require deeper comprehension of the context. Furthermore, response times were 42% longer on average for complex questions, indicating the additional processing needed to handle these more demanding queries. This variance in question type performance underlines the importance of training models specifically on complex question structures. Additionally, the exclusive focus on extractive questions in UQuAD1.0 suggests the need for expanded datasets that capture a broader range of question-answering scenarios in Urdu.

Error Analysis: Error analysis in monolingual and cross-lingual settings, as shown in Tables 4 and 5, reveals common error types that impacted model performance. In monolingual settings, boundary detection was a prevalent issue, particularly in mBERT and XLM-R, with error rates of 35% and 32%, respectively. Even GPT-4, the most robust model, exhibited a 28% error rate in this category. Context understanding errors were also frequent, particularly in mBERT (31%) and XLM-R (28%), while GPT-4 and mT5 showed relatively better performance in this area.

In cross-lingual settings, translation mismatches and script issues were prominent error types, with mBERT showing the highest error rate in translation mismatch at 42%. Script issues, particularly the handling of Urdu script alongside English, posed challenges across all models, with GPT-4 handling it slightly better at 25% error rate, compared to mBERT's 33%. mT5, which is known for its multilingual training, exhibited improved handling of diverse scripts with a 29% error rate in script issues, suggesting its training benefits in multilingual environments. These findings indicate that model robustness in mixed-language environments still has room for improvement, especially in overcoming script and translation challenges.

Impact of Question Length: Table 6 examines the impact of question length on model performance, showing that all models experience a decline in accuracy as question length increases. For short questions (10 words), the Exact Match and F_1 scores are notably high across models, with GPT-4 achieving an F_1

Model	Prompt Setting	Exact Match	F_1 Score	ROUGE-L
GPT-4	Full Urdu (E1)	65.8%	76.4%	74.2%
	Full English (E4)	84.3%	89.1%	87.5%
	Mixed Lang. (E5)	71.2%	81.5%	79.8%
mBERT	Full Urdu (E1)	48.5%	61.2%	59.7%
	Full English (E4)	65.7%	75.3%	73.8%
	Mixed Lang. (E5)	52.3%	64.8%	62.9%
XLM-R	Full Urdu (E1)	53.2%	65.7%	63.9%
	Full English (E4)	69.1%	78.4%	76.5%
	Mixed Lang. (E5)	57.8%	68.9%	66.7%
mT5	Full Urdu (E1)	58.4%	67.9%	66.2%
	Full English (E4)	73.8%	80.2%	78.6%
	Mixed Lang. (E5)	61.4%	72.1%	70.3%

Table 2: Overall performance of models across different prompt settings.

Model	Question-Context Lang	Exact Match	F_1 Score	ROUGE-L
GPT-4	Urdu-English (E2)	64.5%	75.2%	73.1%
	English-Urdu (E3)	62.8%	73.9%	71.8%
mBERT	Urdu-English (E2)	45.2%	57.8%	55.9%
	English-Urdu (E3)	43.7%	56.3%	54.2%
XLM-R	Urdu-English (E2)	49.8%	62.4%	60.5%
	English-Urdu (E3)	48.1%	60.9%	58.7%
mT5	Urdu-English (E2)	54.6%	66.1%	64.3%
	English-Urdu (E3)	53.2%	65.5%	63.8%

Table 3: Cross-lingual performance for different models with varying language settings.

score of 81.5% and mT5 performing reasonably well at 78.2%. As question length increases to the medium range (11-20 words) and beyond, the Exact Match and F_1 scores drop noticeably across all models. This pattern indicates that longer questions introduce more complexity, potentially leading to greater context ambiguity or more challenging boundary detection for answer spans. The results highlight the need for models with enhanced capacity for processing and accurately interpreting extended contextual information, particularly when dealing with longer questions.

Invalid Output Analysis: Table 7 analyzes the incidence of invalid outputs, including answers that are out of context, incorrectly formatted, or missing altogether. GPT-4 exhibits a lower number of invalid outputs (43 instances), indicating its advantage in generating contextually relevant and correctly formatted answers. In contrast, mBERT and XLM-R display a significantly higher number of invalid outputs, with mBERT producing the highest number of "Wrong Format" errors (67) and "Out of Context" responses (46). mT5, while better than mBERT in maintaining context, still faces challenges in answer format consistency. Although mT5 outperforms some baseline models, it has room for improvement in reliably maintaining answer relevance and structure. These findings emphasize that even with recent advancements in LLMs, generating contextually grounded and syntactically accurate outputs remains an area for potential refinement, particularly in cross-lingual and format-sensitive applications.

Impact of Prompt Settings: The impact of different prompt settings on model performance is also evident in these results. Mixed-language prompts (E5) consistently performed worse than monolingual settings, with an average F_1 score reduction of 5.2%. This decline is most notable in mBERT, which struggled to adapt to mixed-language prompts, underscoring the model's limitations in fluidly transitioning between languages. Cross-lingual se-

tups, such as Urdu questions with English context (E2) and English questions with Urdu context (E3), resulted in F_1 score reductions of 3.8% and 4.7%, respectively. These declines indicate that cross-lingual comprehension remains challenging for all models, even those like GPT-4 that are reputed for cross-lingual capabilities.

Model-Specific Observations: GPT-4 demonstrated superior overall performance, with the smallest language gap in F_1 score drop (15.2%) for Urdu and the most consistent cross-lingual performance. Its strong showing in complex question answering indicates an advanced capacity for nuanced comprehension, setting it apart as the most effective model in this study. mBERT, on the other hand, displayed moderate performance with a significant language gap, particularly struggling in mixed-language settings. This model excelled in answering factoid questions but faced higher variance in answer boundaries, making it less suitable for tasks requiring precise boundary detection. XLM-R maintained a good balance between languages, showing robustness in cross-lingual settings compared to mBERT, and demonstrated consistent performance across question types, though it still trailed behind GPT-4 and mT5. mT5 exhibited competitive performance, particularly in handling multilingual prompts. Its cross-lingual capabilities, though not on par with GPT-4, were stronger than mBERT, particularly in handling script diversity and translation mismatches. The model's lower variance in handling Urdu and English contexts highlights its potential as a viable option for multilingual applications, especially in low-resource settings.

Summary of Findings: Overall, the results highlight GPT-4's superior performance across various prompt configurations and error categories, establishing it as the most robust model for both monolingual and cross-lingual QA tasks. While mT5 shows promise, particularly for multilingual contexts, it falls short of GPT-4 in certain nuanced aspects. The limitations of XLM-R and mBERT, particularly in handling cross-lingual prompts and complex questions, point to potential areas for model refinement. Future research could focus on developing pretraining and fine-tuning strategies specifically tailored to improve LLM performance in low-resource, cross-lingual QA tasks, addressing issues such as translation alignment, script handling, and complex question comprehension. Future work could explore additional prompting strategies such as Few-Shot learning and Chain of Thought (CoT) reasoning, which could potentially enhance model performance, particularly for complex questions and cross-lingual scenarios. These approaches might help bridge the performance gap observed between factoid and complex questions.

6 Limitations

This study faced several limitations in evaluating zero-shot question answering in Urdu. The UQuAD1.0 dataset, being partially machine-translated, fell short in fully capturing native Urdu linguistic patterns. The analysis framework did not fully address Urdu's morphological complexities and code-switching tendencies. While zero-shot methods met our experimental needs, they limited the exploration of models' potential achievable with fine-tuning. Additionally, the prompt templates and error analysis framework showed limitations in handling certain question types and Urdu-specific model errors. Our current approach could be enhanced through several methodological extensions. The exploration of advanced prompting strategies, such as Few-Shot learning and Chain of Thought (CoT) reasoning, could potentially improve model performance for complex questions and cross-lingual scenarios.

References

Shubham Agarwal et al. 2022. Zero-shot cross-lingual open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1245.

Muhammad Arif et al. 2024a. Uqa: A corpus for urdu question answering. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1497–1504.

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024b. Generalists vs. specialists: Evaluating large language models for urdu. *arXiv preprint arXiv:2407.04459*.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the po-

tential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, 47:279–311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1. 0: development of an urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2023. A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, 56(Suppl 2):2509–2569.

Samreen Kazi and Shakeel Ahmed Khoja. 2024. Context-aware question answering in Urdu. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 233–242, Trento. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Maria Rahim and Shakeel Ahmed Khoja. 2024. Sawaal: A framework for automatic question generation in urdu. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 139–148.

Abdur Rahman, Arjun Ghosh, and Chetan Arora. 2023. Utrnet: High-resolution urdu text recognition in printed documents. In *International Conference on Document Analysis and Recognition*, pages 305–324. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic data augmentation for zero-shot cross-lingual question answering. *arXiv preprint arXiv:2010.12643*.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2024. Large language models are good multi-lingual learners: When llms meet cross-lingual prompts. *arXiv preprint arXiv:2409.11056*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Shijie Wu and Mark Dredze. 2022. Performance prediction for cross-lingual transfer learning. *arXiv preprint arXiv:2203.07706*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834.

Appendix

Error Type	GPT-4	mBERT	XLM-R	mT5
Boundary Detection	28%	35%	32%	30%
Context Understanding	22%	31%	28%	27%
Answer Format	18%	24%	21%	19%
No Answer	32%	10%	19%	15%

Table 4: Error analysis in monolingual settings for each model.

Error Type	GPT-4	mBERT	XLM-R	mT5
Translation Mismatch	35%	42%	38%	33%
Script Issues	25%	33%	30%	29%
Context Loss	22%	15%	18%	20%
Other	18%	10%	14%	12%

Table 5: Error analysis in cross-lingual settings for each model.

Question Length	Exact Match	F_1 Score	ROUGE-L
Short ($\leq$10 words)	72.3%	81.5%	79.8%
Medium (11-20)	65.8%	76.4%	74.2%
Long (>20)	58.9%	70.5%	68.7%

Table 6: Impact of question length on model performance.

Model	Total Invalid	No Answer	Wrong Format	Out of Context
GPT-4	43	12	18	13
mBERT	158	45	67	46
XLM-R	127	36	54	37
mT5	102	27	44	31

Table 7: Analysis of invalid outputs for each model.

Setting & Prompt Template with Example
E1 - Full Urdu Prompt { { "role": "user", "Prompt": "اس سوال کا جواب دیے گئے سیاق وسباق کے مطابق دیں۔ جواب صرف اردو میں لکھیں۔" **Context**: بیونسے جیزل نولس-کارٹر ایک امریکی گلوکارہ، نغمہ نگار، ریکارڈ پروڈیوسر اور اداکارہ ہیں۔ وہ ہیوسٹن، ٹیکساس میں پیدا ہوئیں اور وہاں ہی بڑی ہوئیں۔ **Question**: بیونسے نے کس شہر اور ریاست میں پرورش پائی؟ **Answer**: }, { "role": "system", "Prompt": "You are a proficient language model trained to understand Urdu. Provide concise answers based on the given context." } }
E2 - Urdu Question, English Context { { "role": "user", "Prompt": "Answer the following question based on the English context provided. Provide only the answer in Urdu. **Context**: Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she rose to fame in the late 1990s. **Question**: بیونسے نے کس شہر اور ریاست میں پرورش پائی؟ **Answer**: }, { "role": "system", "Prompt": "Ensure the answer is in Urdu, derived from the English context provided." } }
E3 - English Question, Urdu Context { { "role": "user", "Prompt": "Answer the following question based on the Urdu context provided. Provide only the answer in English. **Context**: بیونسے جیزل نولس-کارٹر ایک امریکی گلوکارہ، نغمہ نگار، ریکارڈ پروڈیوسر اور اداکارہ ہیں۔ وہ ہیوسٹن، ٹیکساس میں پیدا ہوئیں اور وہاں ہی بڑی ہوئیں۔ **Question**: In what city and state did Beyoncé grow up? **Answer**: }, { "role": "system", "Prompt": "Answer the question in English using information from the Urdu context." } }
E4 - Full English Prompt { { "role": "user", "Prompt": "Answer the question based on the provided context. Only answer in English. **Context**: Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she rose to fame in the late 1990s. **Question**: In what city and state did Beyoncé grow up? **Answer**: }, { "role": "system", "Prompt": "Ensure the answer is concise and derived directly from the English context." } }
E5 - Mixed Language Prompt { { "role": "user", "Prompt": "Answer the following question based on the mixed language context provided. **Context**: Beyoncé Giselle Knowles-Carter ایک امریکی گلوکارہ، نغمہ نگار، اور اداکارہ ہیں۔ وہ Houston, Texas میں پیدا ہوئیں اور وہاں ہی بڑی ہوئیں۔ **Question**: In what city and state did Beyoncé grow up? **Answer**: }, { "role": "system", "Prompt": "Interpret the mixed language prompt and provide a relevant answer." } }

Table 8: Prompt Templates Examples