Cross-Lingual Extractive Question Answering with Unanswerable Questions

Yuval Gorodissky[1,2], **Elior Sulem**[1,2], **Dan Roth**[3,4]
[1] Faculty of Computer and Information Science, Institute for Applied AI Research
[2] Data Science Research Center
Ben-Gurion University of the Negev
[3] Department of Computer and Information Science, University of Pennsylvania
[4] Oracle AI
yuvalgor@post.bgu.ac.il, eliorsu@bgu.ac.il, danroth@seas.upenn.edu

Abstract

Cross-lingual Extractive Question Answering (EQA) extends standard EQA by requiring models to find answers in passages written in languages different from the questions. The Generalized Cross-Lingual Transfer (G-XLT) task evaluates models' zero-shot ability to transfer question answering capabilities across languages using only English training data. While previous research has primarily focused on scenarios where answers are always present, real-world applications often encounter situations where no answer exists within the given context. This paper introduces an enhanced G-XLT task definition that explicitly handles unanswerable questions, bridging a critical gap in current research. To address this challenge, we present two new datasets: miXQuAD and MLQA-IDK, which address both answerable and unanswerable questions and respectively cover 12 and 7 language pairs. Our study evaluates state-of-the-art large language models using fine-tuning, parameter-efficient techniques, and in-context learning approaches, revealing interesting trade-offs between a smaller fine-tuned model's performance on answerable questions versus a larger in-context learning model's capability on unanswerable questions. We also examine language similarity patterns based on model performance, finding alignments with known language families.[1]

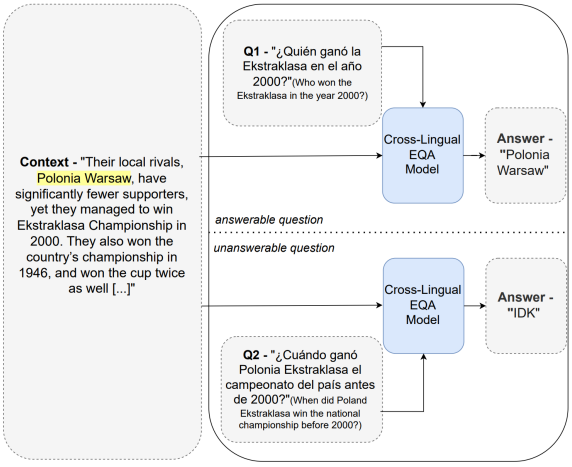

Figure 1: Illustration of Cross-lingual EQA task using MiXQuAD examples. In these examples, questions are in Spanish and contexts/answers are in English, demonstrating both answerable and unanswerable cases. The English translation of each question is provided for reference.

1 Introduction

Extractive Question Answering (EQA) is the task of finding text spans within given contexts that answer given natural language questions. This field was formalized with the Stanford Question Answering Dataset (SQuAD, Rajpurkar et al., 2016), which set a key benchmark for EQA. Recent advances in large language models (LLMs, Brown et al., 2020) have significantly improved EQA performance, marking important progress in Natural Language Understanding. Cross-lingual EQA is a task where the question and its corresponding context are presented in different languages[2], addressing a critical need in today's globalized world. For instance, an English-speaking user might need to query content available only in Chinese or Arabic, challenging the monolingual assumptions in traditional QA systems. Recent research has demonstrated that EQA techniques can be effectively applied to downstream tasks, showing promising results for example in zero-shot event extraction (Lyu et al., 2021) and summarization evaluation (Deutsch et al., 2021; Durmus et al., 2020). While these studies focused on English, their QA-based approaches could naturally extend to cross-lingual scenarios such as cross-lingual summarization (Wang et al., 2022) and cross-lingual event extraction (Subburathinam et al., 2019). Lewis et al. (2020) introduced the Generalized Cross-Lingual Transfer (G-XLT) task, addressing cross-lingual QA in the case where models are trained on English and evaluated on multiple language pairs, assuming that all questions are answerable. However, the ability to identify when a question cannot be answered is crucial for real-world applications.

[1]The code and datasets are publicly available at https://github.com/NLU-BGU/Cross-Lingual-Extractive-Question-Answering-with-Unanswerable-Questions.

[2]This term is sometimes also used for referring to the case where the training and test corpora are in different languages, while in each of them the question and the context are in the same language (Artetxe et al., 2020).

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 100–121
November 8-9, 2025

In this paper, we expand the G-XLT framework by introducing unanswerable questions, as illustrated in Figure 1. This extension reflects real-world scenarios where answers may not exist within the given context. Our approach also broadens the evaluation of Lewis et al. (2020), limited to encoder-based models, to include state-of-the-art models, encompassing large language models (LLMs) and various transformer architectures, providing a more comprehensive assessment of cross-lingual QA capabilities. To address the limitations in current datasets and to provide a robust benchmark for this extended task, we have developed two test datasets: miXQuAD and MLQA-IDK. The miXQuAD dataset combines elements from SQuAD v2.0 and XQuAD (Rajpurkar et al., 2018; Artetxe et al., 2020), integrating unanswerable questions into a multilingual framework. It covers 12 languages and includes a total of 2,072 examples per language, balancing answerable and unanswerable questions. The MLQA-IDK dataset, derived from MLQA (Lewis et al., 2020), specifically focuses on cross-lingual unanswerability, covering 7 language pairs and employing techniques such as antonym and entity augmentation to generate unanswerable questions.

Our cross-lingual QA analysis reveals a trade-off between fine-tuned small models and large models with in-context learning: mT5-large (1.2B parameters) with fine-tuning excelled at detecting unanswerable questions, while AYA-101 (13B parameters) with hint prompting, where the unanswerability option is mentioned in the prompt, performed better on answerable questions. Hint prompting significantly improved unanswerable question detection across models while maintaining performance on answerable questions. Fine-tuned AYA-101 achieved the best performance, outperforming both its regular prompt version and mT5-large across both answerable and unanswerable questions. Out-of-domain testing on MLQA-IDK and open-domain evaluation on XTREME-UP (Ruder et al., 2023) demonstrate model robustness across diverse QA scenarios, including low-resource languages.

We also examine the dependence of the results on the specific languages and on linguistic relationships. First, comparing between cases where the question is in English (English-Questions) to those where the context is in English (English-Contexts), we observe that models performed better when contexts are in English, indicating that processing questions in various languages while keeping English contexts is more manageable. Second, language clustering analysis revealed three groups that align to some extent with language typology—suggesting linguistic relationships influence model behavior.

In an advanced analysis, we examine answerability-related error patterns, test the models' reliance on parametric knowledge, and explore their uncertainty in the different types of prediction. In particular, we observe that hint prompting reduces uncertainty when classifying unanswerable questions, while fine-tuning improves overall certainty but reduces the confidence gap between correct and incorrect predictions.

Our main contributions are the following. First, we expand the Generalized Cross-Lingual Transfer (G-XLT) task to explicitly handle unanswerable questions. Second, we introduce two novel test sets, miXQuAD and MLQA-IDK for the extended task. Third, through the analysis of state-of-the-art models with varying architectures and parameter sizes, we provide insights into performance patterns and language dependency, and reveal trade-offs between model size and training approaches, advancing cross-lingual QA understanding.

2 Related Work

2.1 Extractive Question Answering

EQA is a fundamental Natural Language Understanding (NLU) task that involves identifying and extracting answer spans from a given context in response to natural language questions. This task serves as a critical benchmark in evaluating machine reading comprehension capabilities (Wang et al., 2018). Initially, EQA research focused primarily on monolingual settings, with SQuAD (Rajpurkar et al., 2016) establishing foundational benchmarks through English Wikipedia-derived question-answer pairs. The introduction of BERT (Devlin et al., 2019) marked a significant advancement through its use of bidirectional transformers, though early development remained largely English-centric, with other language datasets often being SQuAD translations such as Arabic (Mozannar et al., 2019) and Spanish (Carrino et al., 2020).

2.2 Evolution of Cross-lingual QA

The development of QA in multiple languages has followed two main strategies: fine-tuning existing models for new languages and developing

zero-shot transfer capabilities across languages. Datasets like XQuAD (Artetxe et al., 2020), TyDiQA (Clark et al., 2020), and XTREME (Hu et al., 2020) have facilitated this research through multilingual question-answer pairs. The Multilingual Transfer (XLT) task, introduced with MLQA (Lewis et al., 2020), pioneered generalization from English-trained models to other languages, where questions and contexts are in the same language. Its extension, the G-XLT task, formalized cross-lingual QA by requiring models trained solely on English data to handle questions and contexts in different languages. Recent work has explored retrieval-augmented approaches to address cross-lingual challenges. Cross-lingual QA has also been studied in open-domain QA, where document retrieval is required before answering the question. In particular, XOR-TyDi QA (Asai et al., 2020) revealed severe performance drops when answers exist only in foreign-language documents, requiring retrieval across massive multilingual corpora. More recently, Ranaldi et al. (2025) demonstrated that multilingual RAG systems face unique difficulties when retrieved documents span multiple languages, with performance degrading when models must integrate information across linguistic boundaries. However, all these frameworks assumed questions were answerable. We extend the G-XLT framework to address this gap by introducing two complementary benchmarks: miXQuAD and MLQA-IDK, providing evaluation capabilities across multiple languages with explicit no-answer detection. We also explore the adaptability of our study to open-domain settings in Section 5.4.

2.3 No Answer Importance

The ability to identify unanswerable questions is critical in real-world applications, with NQ (Kwiatkowski et al., 2019) showing that 51% of real queries lack answers in their given context. While this challenge has been studied in English monolingual settings, beginning with SQuAD v2.0 (Rajpurkar et al., 2018), research has shown that even advanced language models struggle with this task, often hallucinating plausible but incorrect answers (Slobodkin et al., 2023). English-focused datasets like HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) further demonstrated this challenge through multi-document reasoning requirements, with MuSiQue introducing contrast questions to increase evaluation rigor. However, these existing datasets and research have focused on monolingual settings, leaving a gap in understanding how models perform on unanswerable questions in cross-lingual scenarios, which we address in this paper.

3 Task, Dataset Creation And Structure

3.1 The Task

The task of *Generalized Cross-Lingual EQA with IDK* is defined as follows. Given a training dataset $D = \{(c_i, q_i, a_i)\}_{i=1}^{N}$, $N \in \mathbb{N}$, where c_i is a context, q_i is a question, a_i is an answer, and all elements c_i, q_i, a_i are in English, we aim to learn a mapping:

$$f : (q \in L_q, c \in L_c) \rightarrow \begin{cases} s \subseteq c & \text{if answer exists} \\ \text{IDK} & \text{otherwise} \end{cases}$$

where L_c and L_q are the context and question languages respectively, and $L_c = L_q =$ English.

During evaluation, we consider two settings: (i) English-Questions: where questions are in English and contexts are in other languages ($L_q =$ English, $L_c \neq$ English), and (ii) English-Contexts: where contexts are in English and questions are in other languages ($L_c =$ English, $L_q \neq$ English).

3.2 Dataset Creation and Structure

To evaluate cross-lingual EQA capabilities, we present two evaluation test sets: miXQuAD and MLQA-IDK.

miXQuAD Creation The miXQuAD dataset was created by combining XQuAD (which contains questions and contexts in 12 languages) with unanswerable questions from SQuAD v2.0. Since XQuAD only includes answerable questions, we enhanced it by identifying matching contexts in SQuAD v2.0 that contained unanswerable questions and aligning these with the corresponding XQuAD contexts across all languages. For the English-Questions setting, we paired English unanswerable questions with contexts in each target language. For the English-Contexts setting, we translated these unanswerable questions into the 11 non-English languages. This design ensures no data leakage, as XQuAD derives from SQuAD v2.0's dev set while models train exclusively on SQuAD v2.0's train set. The resulting miXQuAD dataset contains 12 language-specific test sets available in both English-Questions and English-Contexts configurations. In the English-Questions setting, each language maintains exactly 1,190 answerable and

882 unanswerable examples (2,072 total). In the English-Contexts setting, while answerable questions remain constant at 1,190, unanswerable question counts vary by language as shown in Table 1. The languages covered are English (en), Spanish (es), German (de), Greek (el), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), and Romanian (ro).

	en	ar	es	th	de	hi	tr	el	ro	vi	ru	zh
A	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K	1.2K
U	.9K	.7K	.8K	.6K	.8K	.8K	.8K	.8K	.8K	.8K	.8K	.7K
T	2.1K	1.9K	2.0K	1.8K	2.0K	2.0K	2.0K	2.0K	2.0K	2.0K	2.0K	1.9K

Table 1: miXQuAD English-Contexts statistics. A: answerable, U: unanswerable questions. T: total. Overall, there are 24,000 questions (14,400 answerable, 9,600 unanswerable) across 12 languages.

MLQA-IDK Creation Starting from MLQA, a cross-lingual dataset containing answerable questions, we extended it to include unanswerable cases to create MLQA-IDK. Following techniques from Gautam et al. (2023), we employed two primary methods for generating unanswerable questions: entity swapping and antonym substitution (examples shown in Figure 6, Appendix A). These methods create unanswerable questions by substituting key information while preserving the overall structure and domain relevance of the original questions. We generated the unanswerable questions in English to create the English-Questions setting, and translated these questions into the other six languages to create the English-Contexts setting. Due to MLQA's incomplete overlap between questions and contexts across languages, the amount of data varies by language, as summarized in Table 2. The dataset encompasses seven languages: English (en), Spanish (es), German (de), Arabic (ar), Vietnamese (vi), Chinese (zh), and Hindi (hi). Our quality assessment of MLQA-IDK's generated unanswerable questions, performed by two of the authors on a random sample of 100 questions, achieved 95% inter-annotator agreement on both unanswerability and well-formedness. The analysis revealed a 7% noise rate, comparable to that reported in SQuAD v2.0's manual analysis (Rajpurkar et al., 2018).

	ar	de	en	es	hi	vi	zh
A	5.3K	4.5K	11.6K	5.3K	4.9K	5.5K	5.1K
U	3.7K	3.2K	12.9K	3.7K	3.5K	4.0K	3.6K
T	9.1K	7.7K	24.5K	9.0K	8.4K	9.5K	8.8K

Table 2: MLQA-IDK dataset statistics. A: answerable, U: unanswerable questions, T: total. Overall, there are 76.9K questions (46.7K A, 30.2K U) across 7 languages.

Evaluating Machine Translation All translations were performed using the Google Translate API. To ensure the accuracy of translations within our dataset, we implemented a back-translation strategy, as described in Lin et al. (2021). This involved translating the questions from foreign languages back into English. We then employed Sentence-BERT (Reimers and Gurevych, 2019) to generate embeddings for both the original and back-translated English texts, subsequently computing the cosine similarity between them. Only questions with a cosine similarity score above 0.75 were retained. This stringent validation procedure, depicted in Figure 2, guarantees the reliability of our dataset for evaluating the effectiveness of EQA systems across language pairs. All the statistics reported in this section, including those in Tables 1 and 2 concern the final versions of the corpora, after validation.

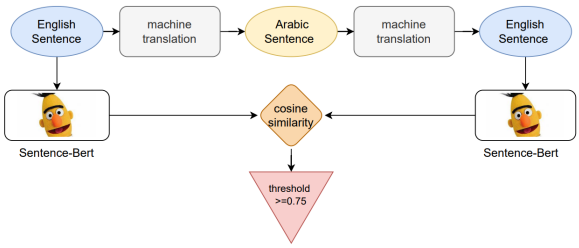

Figure 2: Example of the back-translation strategy encoding an Arabic sentence.

4 Methodology

All our experiments were conducted in a zero-shot setting, where the training data (for fine-tuning) and provided examples (for in-context learning) are exclusively in English from the SQuAD v2.0 dataset. As detailed in Appendix B, we experimented with a diverse set of models ranging from 168M to 13B parameters, using either fine-tuning or in-context learning approaches depending on model architecture and size.

4.1 Fine-Tuning Language Models

Our methodological framework centers on fine-tuning state-of-the-art multilingual language models for EQA tasks. We employ two main fine-tuning strategies to optimize model performance:

Fine-Tuning We conduct full fine-tuning on smaller multilingual models including mBERT (168M parameters), XLM-RoBERTa (279M parameters), mDeBERTa (276M parameters), and

mT5-large (1.2B parameters). This approach allows us to thoroughly adapt these models to the specific requirements of cross-lingual EQA.

Parameter-Efficient Fine-Tuning For larger models like AYA-101 (13B parameters) and AYA-23 (8B parameters), we utilize two efficient fine-tuning techniques: Low-Rank Adaptation (LoRA, Hu et al., 2021) and Quantized LoRA (QLoRA, Dettmers et al., 2024). LoRA enables efficient training without fully retraining the model, while QLoRA further reduces memory usage through quantization. These methods are particularly valuable for fine-tuning large-scale LLMs while maintaining computational efficiency.

4.2 In-Context Learning

In-context learning is a pivotal methodology in our research for training models on EQA tasks, leveraging the model's ability to learn from a few examples. For this approach, we employed several large language models: GPT4-O-mini (8B parameters) using OpenAI API, Gemma-2 (9B parameters), Mistral-Nemo (12.2B parameters), AYA-101 (13B parameters), BLOOMZ (7B parameters), and AYA-23 (8B parameters). We adapted prompt formats from Slobodkin et al. (2023) to the cross-lingual case. For each prompt type, we used three different variants of few-shot prompts, each containing three examples (two answerable and one unanswerable), thus minimizing potential bias from specific example selections.

As shown in Figure 3, our approach uses two prompt families. The Question Answering family includes Regular-Prompt, Hint-Prompt, and Hint-Translate-Prompt, guiding models to provide answers or identify when no answer exists. The Hint-Prompt alerts models to potential answer absence, while Hint-Translate-Prompt adds a question translation step. The Classification family, using Answerability-Prompt, focuses solely on determining if sufficient information exists to answer the question. To identify unanswerable questions, we implemented pattern matching that includes variations of "unanswerable", "no answer", "unknown", "not enough information", and similar phrases in different contexts and formulations.

5 Results

5.1 Model Architecture and Training Approach Effects

Our analysis primarily focuses on the English-Questions setting, while a detailed discussion of the English-Contexts setting is presented in Section 5.3.1.

Encoder-only architectures Early cross-lingual EQA research relied heavily on encoder-only architectures. Table 3 reveals an intriguing characteristic of these models: while mDeBERTa achieved the best overall performance among encoders (63.64 F1), all three models exhibited remarkably strong No Answer performance (76.18-86.06 F1) but struggled significantly with answerable questions (34.13-52.26 F1). This severe imbalance suggests that encoder-only architectures excel at identifying when questions cannot be answered from the given context.

Model	Avg	Has Ans	No Ans
mBERT	56.23	34.13	**86.06**
XLM-R	58.57	45.52	76.18
mDeBERTa	**63.64**	**52.26**	78.98

Table 3: F1 scores comparison of encoder-only models averaged across all languages in miXQuAD English-Questions setting. The bold scores represent the best performance for each category.

5.2 Model Size and Training Approach Effect

Given the limitations of encoder-only models, we focus on encoder-decoder and decoder-only architectures. Table 4 presents a comprehensive comparison revealing several key patterns in cross-lingual QA performance.

Hint prompting improves unanswerability detection. Adding hints about potential unanswerability substantially increased No Answer F1 scores across all models. AYA-101 improved from 35.16 to 54.74, while GPT4o-mini showed the most dramatic increase from 4.97 to 60.23. Crucially, this improvement maintained comparable answerable question performance (AYA-101: 67.86 vs. 66.71), suggesting that hint prompting helps models better calibrate confidence thresholds rather than simply biasing toward 'unanswerable' predictions.

Parameter count alone does not determine performance. Despite having over 10× more parameters, AYA-101 (13B) with regular prompting un-

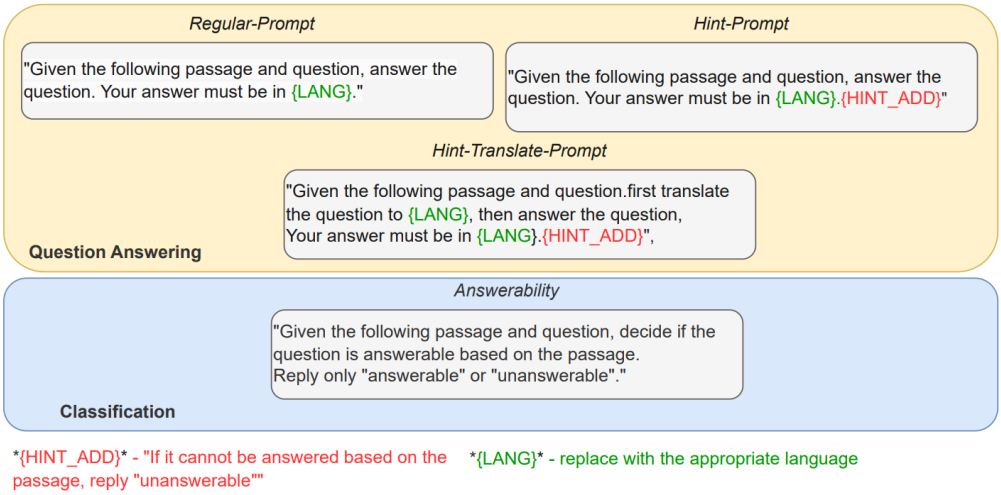

Figure 3: Example prompts used to evaluate the in-context learning methodology.

derperformed mT5-large (1.2B)—64.03 vs. 53.94 Average F1—revealing complementary strengths across architectures.[3] AYA-101 excelled at answer extraction (67.86 vs. 50.55 F1) while mT5-large dominated unanswerability detection (82.20 vs. 54.74 F1). Only through fine-tuning did AYA-101 achieve the best overall performance (81.23 F1), demonstrating that the training approach matters more than parameter count.

Verbosity correlates with poor task adaptation. Answer verbosity appears to impact only untrained large models. GPT4o-mini and Mistral-Nemo, both large models without task-specific training, produced excessively verbose answers (6-9 words on average) which likely contributed to their lower F1 scores through partial match penalties. In contrast, all other models regardless of size maintained concise responses similar to gold answer length (1-3 words), as detailed in Appendix G. This pattern suggests that excessive verbosity is a symptom of insufficient task adaptation rather than an inherent characteristic of model size.

Translation-based prompting consistently underperforms. Hint-translate prompting systematically underperformed standard hint prompting across all models and languages. This degradation was particularly severe for answerable questions—GPT4o-mini dropped from 49.60 to 30.98 F1, while even well-performing models like BLOOMZ showed similar patterns. The consistency of this effect suggests that additional translation steps introduce systematic errors rather than beneficial cross-lingual signals.

[3]Statistical significance was assessed using t-tests between systems across all languages.

Model	Avg	Has Ans	No Ans
mT5-large	64.03	50.55	82.20
Aya-101	53.94	67.86	35.16
+Hint	61.61	66.71	54.74
+Hint-translate	61.41	66.20	54.94
+Fine tuned	**81.23**	**77.09**	**86.80**
GPT4o-mini	28.90	46.65	4.97
+Hint	54.12	49.60	60.23
+Hint-translate	43.49	30.98	60.37
Aya-23	41.01	58.98	16.77
+Hint	48.56	56.87	37.36
+Hint-translate	47.10	54.12	37.63
+Fine tuned	53.36	50.65	59.02
BLOOMZ	41.57	42.33	40.56
+Hint	48.03	37.69	61.98
+Hint-translate	47.96	39.11	59.89
Gemma-2	30.63	34.87	24.91
+Hint	38.93	40.04	37.44
+Hint-translate	38.99	43.61	32.75
Mistral-Nemo	29.96	42.59	12.91
+Hint	50.96	43.86	60.54
+Hint-translate	48.39	42.50	56.33

Table 4: Average F1 scores by category on miXQuAD English-Questions setting. The bold scores represent the best performance for each category.

5.3 Language Dependency

5.3.1 English-Questions vs. English-Contexts

In the English-Contexts setting (Table 5), hint prompting maintained its effectiveness across models. AYA-101 showed improvement from its regular prompt version. Similar enhancements were observed in other models like BLOOMZ, Gemma-2, and notably Mistral-Nemo. As shown in Figure 4, analyzing Average F1 scores across all model variants (regular, hint, hint-translate, and

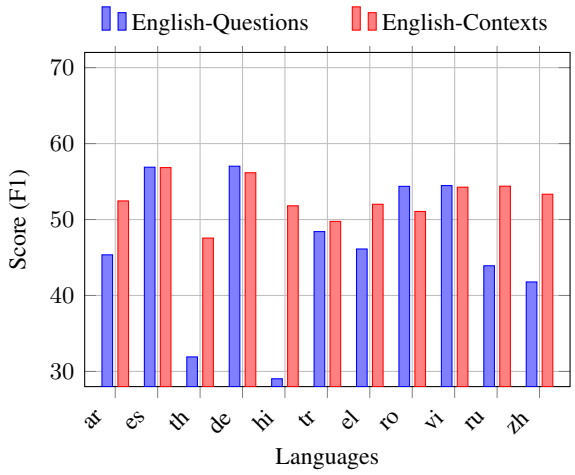

Figure 4: Average F1 scores across languages for English-Questions and English-Contexts settings. Each bar represents the mean performance of all models and their variants (regular, hint, hint-translate, and fine-tuned) for each language, comparing both experimental settings.

Model	Avg	Has Ans	No Ans
mT5-large	72.38	70.99	74.24
Aya-101	70.30	**81.07**	55.77
+Hint	73.24	79.03	65.43
+Hint-translate	72.94	79.13	64.59
+Fine tuned	**80.96**	78.44	**84.35**
GPT4o-mini	17.34	26.40	5.12
+Hint	47.42	39.48	58.15
+Hint-translate	28.52	28.99	27.87
Aya-23	51.59	72.31	23.65
+Hint	57.28	70.31	39.69
+Hint-translate	55.53	69.58	36.56
+Fine tuned	63.07	68.03	56.38
BLOOMZ	52.14	53.57	50.22
+Hint	57.54	47.66	70.88
+Hint-translate	57.79	46.99	72.35
Gemma-2	47.75	61.07	29.79
+Hint	50.23	61.02	35.67
+Hint-translate	49.49	62.20	32.34
Mistral-Nemo	30.23	44.84	10.50
+Hint	51.77	36.38	72.53
+Hint-translate	49.79	37.97	65.73

Table 5: Average F1 scores by category on miXQuAD English-Contexts setting. The bold scores represent the best performance for each category.

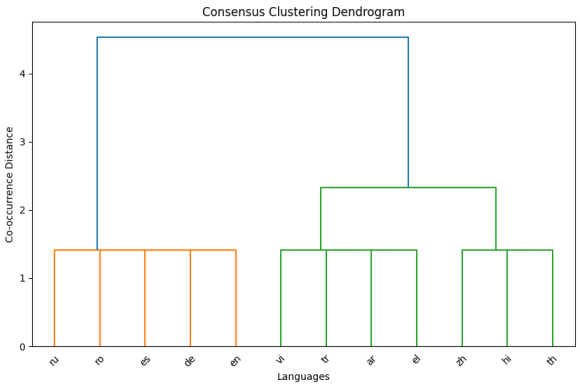

Figure 5: Hierarchical clustering dendrogram of languages based on consensus clustering across all models, using F1 scores averaged over has-answer and no-answer questions

fine-tuned), models consistently performed better in the English-Contexts setting compared to the English-Questions setting. The gap is largest for Asian languages (Thai: 47.56 vs. 31.91, Hindi: 51.81 vs. 29.03) and the performances are comparable for European languages (Spanish: 56.84 vs 56.89, German: 56.16 vs. 57.02). The data suggests that processing questions in various languages while keeping English contexts is more manageable than handling contexts in different languages. This pattern can be attributed to two key factors: (1) questions generally have simpler linguistic structures than full contexts, making them easier to process across languages, and (2) answer extraction in English (the training data language) is more straightforward than in other languages. The complete model-specific performance can be found in Appendix B.

5.3.2 Language Similarity

The clustering methodology comprised two sequential phases, first applying K-Means clustering (k=3) to each model's F1 performance data across has-answer, no-answer, and combined metrics, then synthesizing results through a co-occurrence matrix quantifying language pair clustering frequency. Final consensus clusters were determined using hierarchical clustering with Ward's linkage (Ward Jr, 1963). The analysis reveals three distinct clusters (Figure 5): (1) five Indo-European languages (Russian, Romanian, Spanish, German, English); (2) Vietnamese, Turkish, Arabic, and Greek—languages from four different language families; and (3) Chinese, Hindi, and Thai, which all use non-Latin scripts. The second cluster's unexpected grouping of unrelated languages suggests that factors other than linguistic family relationships determine model performance patterns in cross-lingual transfer.

5.4 Evaluating Model Robustness

We evaluated model robustness using MLQA-IDK beyond their original training domain. Since all models were trained on SQuAD v2.0 or received examples from it during in-context learning, this dataset provides insight into their ability to generalize to unseen data sources. Results, shown in Table 6, confirm that hint prompting improves performance across models, particularly for unanswer-

able questions, while fine-tuned AYA-101 maintains performance on unseen data, achieving 69.96 F1 and 76.23 F1 in English-Questions and English-Contexts respectively.

We examined model adaptability to open-domain question answering using XTREME-UP QA (Ruder et al., 2023), which includes Arabic, Finnish, Japanese, Korean, Russian, and 26 low-resource Indic languages. This dataset pairs questions with gold paragraphs for evaluation of answer extraction capabilities in multilingual settings. Results are in Table 16 (Appendix F). In this case too, hint prompting enhanced performance on unanswerable questions, improving the ability to classify cases where no answer exists in the provided context. We also observe that fine-tuning AYA-101 improved performance on unanswerable questions. The inclusion of XTREME-UP emphasizes evaluation in open-domain QA settings, particularly in low-resource language contexts.

6 Advanced Analysis

6.1 Error Pattern Analysis

We conducted a detailed analysis of models' performance as binary classifiers for question answerability across different prompting strategies. Two main approaches were evaluated: Hint-Prompt, which explicitly warns about potential unanswerable questions, and Answerability-Prompt, which focuses solely on determining answerability. As shown in Table 7, our analysis reveals a consistent trade-off across models. Hint-Prompt achieves higher recall on answerable questions but moderate recall on unanswerable ones, while Answerability-Prompt shows the opposite pattern. For example, AYA-101 with Hint-Prompt achieves 0.929/0.601 recall (answerable/unanswerable) compared to 0.706/0.849 with Answerability-Prompt. These patterns suggest that the choice between prompting strategies depends on whether correctly identifying answerable questions or detecting unanswerable ones is more critical for the specific application.

6.2 Testing Models' Reliance on Parametric Knowledge

Parametric knowledge refers to information stored in the model's parameters during pre-training that can be accessed without external context. To evaluate the extent of this knowledge in our models, we conducted two experiments. First, using regular prompts, we evaluated answerable questions without their corresponding contexts. All models showed performance drops when context was removed, with AYA-101's average F1 score falling from 73.23 to 2.63 (see in Appendix D). To further validate these findings on truly unseen data, we evaluated our models on repliQA-Trans, which was created after all models' pre-training by translating 500 answerable questions from RepliQA (Monteiro et al., 2024) into 11 languages using our validated machine translation pipeline (Section 3.2). The results show that Gemma-2 achieved the highest average performance (87.37 F1), followed by fine-tuned AYA-101 (81.87 F1) and AYA-23 (80.14 F1). The complete language-specific results are presented in Appendix E, demonstrating that models maintain performance even on post-training data.

6.3 Uncertainty Estimation

We analyzed model confidence in cross-lingual QA using the Claim Conditioned Probability method from Vashurin et al. (2025). This white-box method estimates uncertainty based on the probability of the predicted answer conditioned on the input using the model's internal logits. The evaluation included six models: AYA-23, AYA-101, BLOOMZ, and a fine-tuned variant of AYA-101, evaluated on the miXQuAD dataset. Building on our findings that hint prompting improves unanswerable question classification, we observe two key results: hint prompting reduces uncertainty when classifying unanswerable questions compared to regular prompting, aligning with the improved performance on "No answer" cases in our F1 evaluations (Section 5.2). Additionally, while fine-tuning improves overall model certainty, it reduces the confidence gap between correct and incorrect predictions—fine-tuned models show lower uncertainty overall but lose the ability to distinguish between confidence levels for right versus wrong answers compared to non-fine-tuned models. Detailed uncertainty scores and calibration analysis are in Appendix H.

7 Conclusion

Our study advances cross-lingual EQA through several key findings. Fine-tuning large language models proved most effective, with AYA-101 outperforming both smaller fine-tuned models and large models using in-context learning. Hint prompting enhanced unanswerable question detection without compromising answerable performance. Models

	English-Questions			*English-Contexts*		
Model	**Average**	**Has Ans**	**No Ans**	**Average**	**Has Ans**	**No Ans**
mT5-large	51.24	44.50	70.79	65.66	66.72	64.57
Aya-101	48.47	57.50	37.09	59.69	76.11	37.89
+Hint	52.68	56.08	48.50	65.42	73.91	54.08
+Hint-translate	52.44	55.62	48.61	65.20	73.95	53.52
+Fine tuned	**69.96**	**62.05**	**81.02**	**76.23**	**76.43**	**75.88**
GPT4o-mini	25.47	40.37	6.51	16.30	26.59	3.84
+Hint	47.20	41.87	55.41	41.52	37.79	47.50
+Hint-translate	40.99	27.91	59.42	27.55	29.67	24.59
Aya-23	36.71	56.35	11.38	49.26	73.43	17.52
+Hint	42.78	53.28	29.83	52.96	70.72	29.75
+Hint-translate	41.49	51.89	28.79	51.28	69.91	26.94
+Fine tuned	43.41	42.69	44.98	56.35	61.89	49.09
BLOOMZ	43.01	55.01	27.79	54.58	71.66	32.15
+Hint	46.42	48.18	44.83	57.79	65.28	48.05
+Hint-translate	46.77	51.89	40.40	57.94	65.22	48.49
Gemma-2	27.31	37.54	14.87	42.17	59.42	19.79
+Hint	33.11	41.13	24.00	43.09	59.54	22.07
+Hint-translate	32.65	43.98	19.08	42.87	61.20	19.38
Mistral-Nemo	29.43	41.85	13.81	30.64	45.14	12.08
+Hint	45.60	39.06	55.43	47.69	35.29	65.76
+Hint-translate	44.43	38.93	52.75	47.11	36.38	62.81

Table 6: Average F1 scores by category on MLQA-IDK comparing English-Questions and English-Contexts settings. The bold scores represent the best performance for each category.

	H-Prompt				**Answerability**			
Model	**HA**		**NA**		**HA**		**NA**	
	P	R	P	R	P	R	P	R
AYA-23	.67	.94	.83	.39	.75	.83	.73	.62
BLOOMZ	.73	.65	.59	.66	.64	.83	.61	.36
AYA-101	.76	.93	.86	.60	.86	.71	.68	.85
GPT4-O	.76	.94	.88	.59	.90	.71	.69	.89
Gemma-2	.66	.90	.72	.37	.74	.61	.58	.71
M-Nemo	.75	.76	.67	.67	.89	.35	.52	.94

Table 7: Precision (P) and Recall (R) metrics averaged across MiXQuAD test sets for question answerability. Acronyms: HA (Has Answer), NA (No Answer), H-Prompt (Hint-Prompt), M-Nemo (Mistral-Nemo), GPT4-O (GPT4o-mini).

showed better performance with English contexts versus English questions. Performance patterns aligned to some extent with traditional language families, indicating the influence of linguistic relationships. Our study also suggests that models rely on contextual understanding rather than memorized knowledge. These findings support developing more effective cross-lingual EQA systems across diverse languages.

Limitations

This work evaluated scenarios where one language is English, leaving unexplored the model's capability to handle question-answer pairs between two non-English languages. While our study focuses on datasets derived from SQuAD and MLQA, future research should explore cases of longer contexts and more complex reasoning, such as in the monolingual HotpotQA dataset. Additionally, for the use of prompting with LLMs, we only considered few-shot prompting with English examples. Future work could explore instruction-only prompting, where no examples are provided.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by grants from the Data Science Research Center at Ben-Gurion University of the Negev and the BGU/Philadelphia Academic Bridge (The Sutnick/Zipkin Endowment Fund).

References

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor

qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

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 ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. 2023. A lightweight method to generate unanswerable questions in English. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7349–7360, Singapore. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Joao Monteiro, Pierre-Andre Noel, Etienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vazquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. 2024. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. *arXiv preprint arXiv:2406.11811*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Llm pruning and distillation in practice: The minitron approach. *arXiv preprint arXiv:2408.11796*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325, Hong Kong, China. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing*

and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Examples of Unanswerable Questions

Figure 6 shows examples of our unanswerable questions generation methods: entity swapping (questions 1-2) and antonym substitution (question 3).

B Additional Details of Models

Table 8 provides a comprehensive overview of the transformer models used in our experiments, including their architectures, parameter counts, and adaptation methods.

C Model Performance on Different Settings

In this section, we present detailed performance results across different experimental settings. Tables 9 and 10 show the full results on miXQuAD English-Questions and English-Contexts settings, respectively. While Tables 11 and 12 show results on MLQA-IDK English-Questions and English-Contexts settings, respectively.

D Analysis of Context Dependency

To investigate how models leverage contextual information versus parametric knowledge, we compare performance with and without providing the context passage. Our analysis examines the models' performance across different languages and question types when contextual information is introduced. Table 13 demonstrates the substantial performance improvements observed when contextual information is provided.

E Performance on Post-Training Data

To validate our findings on truly unseen data, we evaluate performance on repliQA-Trans, created after model pre-training. Table 14 shows Has Answer F1 scores across all languages, demonstrating the models' ability to generalize to new content.

F Performance on Open-Domain Data

We evaluated our models as answer extraction components for open-domain QA systems using XTREME-UP QA's reading comprehension data (Ruder et al., 2023), where questions are paired with gold paragraphs. Table 16 shows detailed performance across languages, while the full dataset statistics can be found in Table 15, demonstrating the models' effectiveness at extracting answers from retrieved documents.

G Answer Length Analysis

To analyze the verbosity of model responses, we compared the average number of words in model predictions with gold answers across different settings. Table 17 presents this analysis for both MiXQuAD and MLQA-IDK datasets in English-Questions and English-Contexts settings. Notably, larger models like GPT4o-mini and Mistral-Nemo tend to produce longer answers, with averages up to 3 times the length of gold answers in some settings.

H Uncertainty Estimation Analysis

This section provides detailed results from our uncertainty estimation analysis using the Claim Conditioned Probability method. The uncertainty score is calculated as:

$$\text{Uncertainty} = \log P(\text{claim}|\text{input}) \quad (1)$$

where $P(\text{claim}|\text{input})$ is the probability of the model's predicted answer (claim) given the input question and context. This method leverages the model's internal logits to estimate confidence in the prediction.

Table 18 presents uncertainty scores for correctly classified unanswerable questions across different models and prompting strategies.

Context:

As of 2005, recombinant growth hormones available in the United States (and their manufacturers) included Nutropin (Genentech), Humatrope (Lilly), Genotropin (Pfizer), Norditropin (Novo), and Saizen (Merck Serono). In 2006, the U.S. Food and Drug Administration (FDA) approved a version of rHGH called Omnitrope (Sandoz). A sustained-release form of growth hormone, Nutropin Depot (Genentech and Alkermes) was approved by the FDA in 1999, allowing for fewer injections (every 2 or 4 weeks instead of daily); however, the product was discontinued by Genentech/Alkermes in 2004 for financial reasons.

Unanswerable Questions:

1. *[Entity Swap]* What company manufactured the first Genotropin approved rHGH?
2. *[Entity Swap]* What company manufactured the first FDA approved Omnitrope?
3. *[Antonym]* What company manufactured the last FDA approved rHGH?

Figure 6: Example of unanswerable questions from MLQA-IDK. While the questions appear answerable and are related to the context, they require information beyond what is provided in the passage.

Model	Arch	Params	FT	ICL
mBERT (Devlin et al., 2019)	E	168M	✓	x
XLM-RoBERTa (Conneau et al., 2020)	E	279M	✓	x
mDeBERTa (He et al., 2020)	E	276M	✓	x
mT5-large (Xue et al., 2021)	E-D	1.2B	✓	x
GPT4o-mini (Hurst et al., 2024)	D	8B	x	✓
Gemma-2 (Team et al., 2024)	D	9B	x	✓
Mistral-Nemo (Sreenivas et al., 2024)	D	12.2B	x	✓
AYA-101 (Üstün et al., 2024)	E-D	13B	✓	✓
BLOOMZ (Muennighoff et al., 2023)	D	7B	x	✓
Aya-23 (Aryabumi et al., 2024)	D	8B	✓	✓

Table 8: Overview of Transformer Models and Methodologies. Arch: Architecture (E: Encoder, D: Decoder); Params: Number of parameters; FT: Fine-Tuning; ICL: In-Context Learning. FT and ICL represent our proposed methods for adapting models to the G-XLT task. ✓ indicates the method was applied to the model.

The results demonstrate that hint prompting consistently reduces uncertainty (increases confidence) when correctly classifying unanswerable questions across most models. Notably, hint prompting also increases the number of correctly classified unanswerable instances, supporting our main findings about improved performance on "No answer" cases. Table 19 compares uncertainty scores between correct and incorrect predictions to assess model calibration. Well-calibrated models should exhibit higher confidence (lower uncertainty) for correct predictions compared to incorrect ones.

I Hyperparameters

All experiments were conducted using NVIDIA RTX 6000 Ada GPUs for LLMs (AYA-101, AYA-23, BLOOMZ) and RTX 4090 GPUs for encoder-only models (mBERT, XLM-RoBERTa, mDeBERTa) and mT5-large.

I.1 Encoder-Only Models

For fine-tuning the encoder-only models mBERT[4], XLM-RoBERTa[5], and mDeBERTa[6], we utilized a

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://huggingface.co/xlm-roberta-base

[6] https://huggingface.co/microsoft/mdeberta-v3-base

Category	Model\Language	ar	es	th	de	hi	tr	el	ro	vi	en	ru	zh	Avg
Average	**mT5-large**	52.64	72.72	59.29	70.78	57.70	63.35	56.06	72.26	63.96	84.19	56.97	58.40	64.03
	AYA-101	51.85	63.35	26.63	66.77	25.12	55.45	58.83	63.40	58.55	74.23	52.89	50.23	53.94
	+Hint	58.52	72.09	34.98	73.10	33.29	63.11	65.76	72.58	66.77	80.22	59.70	59.18	61.61
	+Hint-translate	58.04	71.11	35.17	72.79	34.18	62.52	65.75	71.89	66.34	79.90	59.31	59.86	61.41
	+Fine tuned	**78.42**	**84.10**	**81.39**	**83.54**	**81.96**	**77.35**	**80.55**	**82.91**	**79.03**	**88.43**	**78.85**	**78.17**	**81.23**
	GPT4o-mini	31.82	34.56	9.80	37.61	6.75	30.80	34.30	31.70	38.56	40.22	28.97	21.76	28.90
	+Hint	55.02	62.58	34.49	64.53	33.58	57.64	59.87	61.31	63.37	66.89	48.97	41.24	54.12
	+Hint-translate	45.27	47.27	32.69	50.80	28.30	43.93	47.93	45.20	52.00	60.84	35.21	32.44	43.49
	Aya-23	40.71	51.58	13.03	53.13	14.66	43.68	40.20	50.23	48.40	65.76	38.49	32.23	41.01
	+Hint	48.07	61.67	21.69	60.68	20.41	52.46	46.99	58.25	55.73	70.53	46.40	39.88	48.56
	+Hint-translate	45.41	59.69	24.60	58.12	21.28	50.49	44.23	56.97	54.13	69.23	43.85	37.15	47.10
	+Fine tuned	49.04	54.20	42.02	65.69	36.09	53.13	52.07	54.82	58.51	79.00	50.70	45.06	53.36
	BLOOMZ	44.89	55.71	29.97	43.39	19.00	34.93	31.88	39.52	51.13	73.74	29.55	45.15	41.57
	+Hint	50.54	60.39	38.83	50.35	24.37	42.99	40.80	46.03	55.63	77.17	38.37	50.88	48.03
	+Hint-translate	50.81	61.25	38.61	50.17	22.77	42.93	40.86	45.80	55.94	77.10	38.39	50.87	47.96
	Gemma-2	22.07	39.05	22.74	40.17	20.55	28.54	23.13	37.68	32.92	62.22	20.00	18.45	30.63
	+Hint	32.21	49.80	29.55	49.80	27.57	38.04	31.06	46.67	43.70	62.08	29.85	26.85	38.93
	+Hint-translate	31.02	50.39	28.00	50.01	26.65	39.89	30.44	47.41	44.98	60.50	31.42	27.16	38.99
	Mistral-Nemo	22.32	33.57	11.70	40.15	12.90	29.50	26.82	38.21	37.32	59.15	27.78	20.06	29.96
	+Hint	43.57	57.80	29.10	59.67	32.12	53.39	48.44	59.45	58.54	72.87	53.48	43.09	50.96
	+Hint-translate	40.13	51.83	25.86	56.24	30.47	52.55	42.46	59.51	58.46	70.97	53.00	39.18	48.39
Has answer	**mT5-large**	30.15	66.28	43.90	62.05	37.95	51.65	36.85	65.47	52.88	82.89	35.32	41.26	50.55
	AYA-101	**74.76**	**81.59**	24.65	**81.35**	20.62	71.11	**78.40**	**81.66**	**77.68**	**85.35**	**74.24**	62.90	67.86
	+Hint	74.47	80.15	24.60	79.01	20.57	70.45	76.86	80.52	76.71	83.22	72.68	61.22	66.71
	+Hint-translate	73.99	79.25	24.42	78.48	20.55	69.84	76.45	79.88	77.08	83.26	72.07	59.15	66.20
	+Fine tuned	73.09	81.15	**78.40**	80.03	**79.32**	**71.86**	76.67	80.47	75.08	84.76	72.61	**71.68**	**77.09**
	GPT4o-mini	51.53	58.70	17.06	56.24	11.75	51.60	58.80	54.40	56.25	64.14	49.81	29.49	46.65
	+Hint	52.98	63.03	17.57	59.03	13.43	55.11	59.75	58.98	58.69	70.54	51.81	34.29	49.60
	+Hint-translate	34.66	40.84	12.46	33.29	7.88	32.87	30.97	33.23	38.19	56.90	34.24	16.18	30.98
	Aya-23	62.54	76.48	16.89	71.64	18.43	61.91	60.64	73.98	72.17	84.75	58.42	49.87	58.98
	+Hint	60.64	71.52	16.70	67.75	18.64	58.88	59.97	70.47	68.78	83.22	57.14	48.76	56.87
	+Hint-translate	56.62	70.20	16.14	66.52	17.75	56.67	53.45	68.65	66.34	83.43	51.48	42.14	54.12
	+Fine tuned	52.68	67.74	4.07	61.78	6.76	55.63	57.05	65.38	65.92	77.13	56.76	36.94	50.65
	BLOOMZ	61.01	61.56	3.89	46.93	20.17	20.59	15.82	40.74	73.54	83.55	21.23	58.90	42.33
	+Hint	53.31	54.06	2.53	41.42	19.53	14.87	11.73	35.44	69.66	77.47	17.87	54.34	37.69
	+Hint-translate	62.44	56.37	2.61	41.08	19.98	14.91	11.78	34.72	73.68	76.42	17.57	57.78	39.11
	Gemma-2	16.40	45.61	17.79	51.35	16.64	39.19	26.05	43.90	40.85	74.89	21.45	24.28	34.87
	+Hint	23.50	49.51	19.07	55.45	17.36	47.60	30.21	49.53	48.97	77.70	28.38	33.18	40.04
	+Hint-translate	25.76	58.43	19.33	58.43	17.10	52.41	31.67	55.75	55.68	78.83	31.90	38.07	43.61
	Mistral-Nemo	36.00	52.40	18.71	55.22	14.78	43.43	45.15	53.83	47.05	78.56	41.45	24.53	42.59
	+Hint	41.29	56.06	19.13	56.31	14.97	45.60	52.93	54.44	40.30	75.17	44.71	25.38	43.86
	+Hint-translate	44.25	53.35	18.02	50.53	13.61	46.61	47.98	51.62	39.80	73.88	42.38	27.97	42.50
No answer	**mT5-large**	83.00	81.41	80.05	82.54	84.36	79.14	81.98	81.41	78.92	85.95	86.17	81.52	82.20
	AYA-101	20.94	38.74	29.29	47.09	31.18	34.32	32.43	38.78	32.73	59.23	24.08	33.15	35.16
	+Hint	37.00	61.23	48.98	65.12	50.46	53.22	50.80	61.87	53.37	76.16	42.18	56.43	54.74
	+Hint-translate	36.51	60.13	49.66	65.12	52.57	52.65	51.33	61.12	51.86	75.36	42.11	60.81	54.94
	+Fine tuned	**85.61**	**88.06**	**85.42**	**88.29**	**85.53**	**84.77**	**85.79**	**86.21**	**84.36**	**93.39**	**87.27**	**86.93**	**86.80**
	GPT4o-mini	5.22	1.99	0.00	12.48	0.00	2.73	1.25	1.08	14.69	7.94	0.86	11.34	4.97
	+Hint	57.77	61.97	57.32	71.94	60.78	61.06	60.04	64.46	69.68	61.97	45.13	50.63	60.23
	+Hint-translate	59.59	55.96	59.98	74.44	55.84	58.85	70.81	61.34	70.64	66.16	36.51	54.37	60.37
	Aya-23	11.27	17.99	7.83	28.16	9.57	19.09	12.63	18.18	16.33	40.14	11.61	8.43	16.77
	+Hint	31.11	48.38	28.43	51.14	22.79	43.81	29.48	41.77	38.14	53.41	31.90	27.90	37.36
	+Hint-translate	30.28	45.51	36.02	46.79	26.04	42.14	31.79	41.20	37.65	50.08	33.57	30.43	37.63
	+Fine tuned	46.04	38.67	93.13	73.17	76.08	54.05	48.61	39.80	52.01	81.30	43.88	61.49	59.02
	BLOOMZ	23.13	47.81	65.16	38.63	17.43	54.28	53.56	37.87	20.90	60.51	40.78	26.61	40.56
	+Hint	46.79	68.94	87.80	62.40	30.88	80.92	80.01	60.32	36.70	76.76	66.03	46.23	61.98
	+Hint-translate	35.11	67.84	87.19	62.44	26.54	80.73	80.09	60.74	32.02	78.01	66.48	41.54	59.89
	Gemma-2	29.71	30.20	29.41	25.10	25.82	14.18	19.20	29.29	22.23	45.13	18.03	10.59	24.91
	+Hint	43.96	50.19	43.69	42.18	41.35	25.14	32.20	42.82	36.59	41.01	31.83	18.30	37.44
	+Hint-translate	38.10	39.54	39.69	38.67	39.54	22.98	28.77	36.17	30.54	35.76	30.77	12.44	32.75
	Mistral-Nemo	3.86	8.17	2.23	19.81	10.36	10.70	2.08	17.13	24.19	32.96	9.34	14.03	12.91
	+Hint	46.64	60.13	42.56	64.22	55.26	63.91	42.37	66.22	83.15	69.77	65.31	66.97	60.54
	+Hint-translate	34.59	49.78	36.44	63.95	53.22	60.55	35.00	70.15	83.64	67.05	67.31	54.31	56.33

Table 9: F1 scores (averaged over 3 model seeds) across three categories (Average F1, Has answer F1, and No answer F1) on miXQuAD, English-Questions across 12 language pairs. The best performance for each category and language appears in bold.

learning rate of 5e-5, a batch size of 12, and a maximum sequence length of 384. A document stride of 128 was applied to handle overlapping text segments. Training was conducted over 2 epochs with a weight decay of 0.01 using the AdamW optimizer. The inputs were lowercased before processing to ensure consistency.

I.2 mT5-large Fine-tuning

For fine-tuning the mT5-large model[7] on SQuAD 2.0, we used a learning rate of 5e-5 and a training batch size of 2. The model was trained for 5 epochs with a maximum sequence length of 512. AdamW was employed as the optimizer to facilitate effective gradient updates during fine-tuning.

[7] https://huggingface.co/google/mt5-large

Category	Model\Language	ar	es	th	de	hi	tr	el	ro	vi	en	ru	zh	Avg
Average	**mT5-large**	66.29	76.19	71.35	75.19	69.72	69.83	72.41	74.71	65.92	84.19	73.77	68.96	72.38
	AYA-101	61.68	68.03	69.26	74.73	69.68	71.86	70.71	72.43	69.81	74.23	71.65	69.52	70.30
	+Hint	69.10	75.06	71.75	75.72	72.31	73.42	72.92	73.77	71.28	80.22	72.93	70.38	73.24
	+Hint-translate	68.73	74.74	71.81	75.45	71.79	72.96	72.55	73.41	71.17	79.90	72.50	70.26	72.94
	+Fine tuned	**78.46**	**83.12**	**79.69**	**84.07**	**78.94**	**79.06**	**79.74**	**81.59**	**77.50**	**88.43**	**81.21**	**79.66**	**80.96**
	GPT4o-mini	9.66	13.74	22.22	17.57	13.65	18.49	12.01	14.87	17.80	40.22	12.57	15.33	17.34
	+Hint	39.97	45.25	52.56	48.41	42.69	42.07	46.08	43.95	49.71	66.89	46.28	45.23	47.42
	+Hint-translate	24.17	30.45	22.63	29.24	24.06	26.30	23.31	25.37	24.42	60.84	25.78	25.61	28.52
	Aya-23	51.45	56.17	28.27	57.88	46.65	52.51	52.90	53.00	51.18	65.76	53.64	49.71	51.59
	+Hint	56.41	62.27	34.52	64.28	52.31	57.60	58.51	59.30	57.05	70.53	59.65	54.89	57.28
	+Hint-translate	54.38	59.67	35.03	62.62	50.10	55.82	56.52	57.88	54.90	69.23	57.82	52.33	55.53
	+Fine tuned	62.53	65.57	43.52	68.50	58.22	61.33	64.78	57.88	62.19	78.90	65.65	67.75	63.07
	BLOOMZ	56.19	63.78	33.02	51.88	57.45	41.25	40.04	38.32	59.42	73.74	48.46	62.16	52.14
	+Hint	61.29	69.66	41.60	56.83	63.55	43.77	44.92	45.34	64.04	77.17	54.61	67.72	57.54
	+Hint-translate	61.99	69.92	41.77	56.57	63.94	43.49	44.96	45.89	64.84	77.10	55.08	67.89	57.79
	Gemma-2	50.66	48.81	50.20	47.84	45.09	38.74	49.66	43.66	46.13	62.22	48.93	41.09	47.75
	+Hint	52.44	52.20	53.15	50.84	47.60	40.13	53.39	48.44	47.85	62.08	51.58	43.05	50.23
	+Hint-translate	51.87	51.83	51.31	50.50	47.33	39.00	52.69	47.97	46.26	60.50	51.91	42.73	49.49
	Mistral-Nemo	24.02	27.21	28.66	29.79	24.92	25.53	27.21	22.58	36.70	59.15	31.20	25.73	30.23
	+Hint	50.21	51.20	50.24	51.69	46.06	47.12	49.74	47.05	50.92	72.87	53.67	50.47	51.77
	+Hint-translate	49.99	48.71	46.24	49.68	41.92	44.67	47.25	44.81	50.36	70.97	53.28	49.54	49.79
Has answer	**mT5-large**	60.21	75.76	69.68	73.02	73.99	70.91	71.20	74.36	62.84	82.89	71.46	65.61	70.99
	AYA-101	**79.91**	**82.63**	**79.36**	**81.64**	**79.28**	**79.97**	**82.78**	**82.07**	**80.52**	**85.35**	**81.34**	**77.99**	**81.07**
	+Hint	77.46	80.09	77.98	79.34	77.58	77.84	81.55	80.37	78.50	83.22	79.02	75.37	79.03
	+Hint-translate	77.56	79.96	78.23	79.61	77.51	77.94	81.65	80.30	78.59	83.26	79.29	75.66	79.13
	+Fine tuned	73.84	80.29	77.52	82.35	77.64	77.18	76.76	80.25	75.60	84.76	77.45	77.69	78.44
	GPT4o-mini	14.55	21.05	34.65	24.16	21.15	28.46	17.55	23.44	26.45	64.14	18.74	22.49	26.40
	+Hint	28.58	33.15	46.85	36.98	40.96	37.08	34.86	33.62	41.68	70.54	32.31	37.15	39.48
	+Hint-translate	22.84	28.47	24.07	29.60	27.26	27.90	24.32	25.68	27.40	56.90	26.10	27.36	28.99
	Aya-23	73.80	78.42	41.06	76.65	66.71	72.15	78.54	72.87	72.84	84.75	77.79	72.10	72.31
	+Hint	71.86	76.43	37.85	74.30	63.26	69.16	76.66	71.03	71.96	83.22	76.83	71.11	70.31
	+Hint-translate	70.73	74.65	40.32	73.31	61.90	68.58	75.75	70.07	70.60	83.43	75.86	69.80	69.58
	+Fine tuned	69.32	74.17	44.39	71.23	66.03	65.38	72.01	70.52	69.30	77.13	70.28	66.58	68.03
	BLOOMZ	72.87	80.60	3.63	57.72	75.88	8.07	15.26	34.84	74.49	83.55	60.96	74.93	53.57
	+Hint	67.32	75.71	1.31	48.24	71.32	4.08	9.14	26.28	68.20	77.47	54.63	68.19	47.66
	+Hint-translate	67.15	75.12	1.46	45.62	71.02	3.43	8.78	24.49	68.01	76.42	54.67	67.72	46.99
	Gemma-2	61.25	55.68	61.28	52.73	60.24	57.65	67.34	54.85	61.47	74.89	66.01	59.42	61.07
	+Hint	59.91	57.32	61.62	53.20	58.76	56.85	66.43	54.65	61.04	77.70	64.67	60.09	61.02
	+Hint-translate	62.47	58.61	62.30	55.20	59.72	57.14	67.43	55.25	61.99	78.83	66.06	61.39	62.20
	Mistral-Nemo	40.42	41.27	48.77	42.70	40.92	38.51	45.89	34.07	44.51	78.56	43.76	38.69	44.84
	+Hint	33.94	37.27	49.67	33.36	36.80	26.82	35.82	21.22	21.12	75.17	28.29	37.09	36.38
	+Hint-translate	34.44	38.26	51.18	34.08	39.09	29.46	38.52	23.40	21.84	73.88	31.36	40.10	37.97
No answer	**mT5-large**	74.49	76.76	73.59	78.12	63.95	68.37	74.04	75.18	70.07	85.95	76.88	73.47	74.24
	AYA-101	37.08	48.34	55.64	65.42	56.73	60.93	54.43	59.42	55.37	59.23	58.58	58.09	55.77
	+Hint	57.83	68.26	63.35	70.83	65.20	67.47	61.27	64.86	61.53	76.16	64.71	63.65	65.43
	+Hint-translate	56.81	67.69	63.16	69.85	64.06	66.26	60.28	64.10	61.15	75.36	63.35	62.97	64.59
	+Fine tuned	**84.70**	**86.93**	**82.62**	**86.40**	**80.69**	**81.60**	**83.75**	**83.41**	**80.05**	**93.39**	**86.29**	**82.32**	**84.35**
	GPT4o-mini	3.07	3.86	5.45	8.68	3.52	5.05	4.54	3.29	6.13	7.94	4.26	5.67	5.12
	+Hint	55.33	61.57	60.27	63.84	45.02	48.81	61.23	57.88	60.55	61.97	65.14	56.13	58.15
	+Hint-translate	25.97	33.11	20.70	28.75	19.73	24.15	21.94	24.95	20.41	66.16	25.35	23.25	27.87
	Aya-23	21.28	26.16	11.00	32.54	19.58	26.01	18.30	26.20	21.96	40.14	21.06	19.51	23.65
	+Hint	35.57	43.16	30.01	50.76	37.53	41.99	34.02	43.47	36.93	53.41	36.48	33.00	39.69
	+Hint-translate	32.32	39.46	27.90	48.19	34.17	38.59	30.58	41.43	33.72	50.08	33.49	28.77	36.56
	+Fine tuned	53.37	53.97	42.33	64.82	47.70	55.86	55.03	40.82	52.61	81.30	59.42	69.32	56.38
	BLOOMZ	33.68	41.09	72.68	44.00	32.58	86.02	73.47	43.01	39.08	60.51	31.60	44.94	50.22
	+Hint	53.14	61.49	95.96	68.41	53.07	97.32	93.20	71.06	58.43	76.76	54.58	67.09	70.88
	+Hint-translate	55.03	62.89	96.15	71.36	54.39	97.55	93.77	74.76	60.55	78.01	55.64	68.11	72.35
	Gemma-2	36.36	39.54	35.27	41.24	24.65	13.23	25.82	28.58	25.44	45.13	25.89	16.37	29.79
	+Hint	42.37	45.28	41.73	47.66	32.54	17.58	35.79	40.07	30.05	41.01	33.91	20.07	35.67
	+Hint-translate	37.57	42.67	36.48	44.15	30.62	14.52	32.81	38.14	25.02	35.76	32.81	17.54	32.34
	Mistral-Nemo	1.89	8.24	1.52	12.36	3.33	8.02	2.01	7.07	26.16	32.96	14.25	8.24	10.50
	+Hint	72.15	70.00	51.03	76.42	58.55	74.49	68.52	81.90	91.12	69.77	87.91	68.52	72.53
	+Hint-translate	70.98	62.82	39.57	70.72	45.73	65.20	59.04	73.70	88.86	67.05	82.85	62.29	65.73

Table 10: F1 scores (averaged over 3 model seeds) across three categories (Overall F1, Has answer F1, and No answer F1) on miXQuAD English-Contexts setting across 12 language pairs. The bold scores represent the best performance for each category and language.

I.3 In-Context Learning Settings

In-context learning experiments were conducted using AYA-101[8], AYA-23[9], BLOOMZ[10], GPT4o-mini (via OpenAI API), Gemma-2[11], and Mistral-Nemo[12]. The models were configured with a maximum of 25 new tokens and an input window size of 4024. To provide the models with sufficient contextual information, we employed three few-shot examples. The hint prompt template used was, "If it cannot be answered based on the passage, reply

[8] https://huggingface.co/CohereForAI/AYA-101
[9] https://huggingface.co/CohereForAI/aya-23
[10] https://huggingface.co/bigscience/bloomz
[11] https://huggingface.co/google/gemma-2b
[12] https://huggingface.co/mistralai/Nemo-2.0

Category	Model\Language	ar	de	en	es	hi	vi	zh	Avg
Average	**mT5-large**	46.10	52.55	65.94	55.54	46.41	49.77	42.38	51.24
	AYA-101	44.62	59.27	59.91	56.91	22.72	52.01	43.84	48.47
	+Hint	47.86	60.24	67.89	62.85	27.22	56.05	46.62	52.68
	+Hint-translate	47.31	60.19	67.71	62.63	27.48	55.24	46.51	52.44
	+Fine tuned	**67.57**	**66.87**	**81.64**	**72.48**	**71.70**	**70.80**	**58.63**	**69.96**
	GPT4o-mini	27.84	32.98	30.11	32.21	6.28	33.68	15.19	25.47
	+Hint	47.80	55.64	55.62	55.30	28.49	54.93	32.65	47.20
	+Hint-translate	42.20	47.60	51.19	45.96	25.42	48.15	26.38	40.99
	Aya-23	35.95	44.96	51.08	45.72	12.39	44.89	21.98	36.71
	+Hint	41.78	51.74	55.72	53.18	17.03	51.32	28.71	42.78
	+Hint-translate	40.04	50.09	54.60	52.15	17.50	49.34	26.74	41.49
	+Fine tuned	39.13	51.09	59.26	45.49	33.84	47.23	27.82	43.41
	BLOOMZ	45.86	39.56	55.11	48.42	18.50	51.71	41.89	43.01
	+Hint	48.01	43.84	59.85	51.00	22.79	54.17	45.30	46.42
	+Hint-translate	50.19	43.67	59.99	51.72	20.97	55.07	45.75	46.77
	Gemma-2	18.60	34.44	46.39	33.14	14.13	30.75	13.69	27.31
	+Hint	27.17	40.39	45.56	42.06	19.55	38.24	18.79	33.11
	+Hint-translate	26.01	39.96	44.46	42.59	18.56	38.97	18.01	32.65
	Mistral-Nemo	23.05	38.45	48.37	32.57	11.58	34.44	17.52	29.43
	+Hint	39.78	51.96	61.09	50.52	28.08	51.62	36.14	45.60
	+Hint-translate	38.28	51.03	60.27	49.10	27.51	52.56	32.26	44.43
Has answer	**mT5-large**	26.53	49.23	78.65	54.61	30.17	51.83	20.50	44.50
	AYA-101	**60.40**	**64.84**	78.17	**72.16**	17.83	**66.54**	**42.58**	**57.50**
	+Hint	59.89	62.56	76.33	69.43	17.59	64.88	41.87	56.08
	+Hint-translate	59.61	62.07	76.41	69.32	17.36	64.66	39.88	55.62
	+Fine tuned	58.04	57.85	**80.42**	67.33	**66.71**	63.23	40.78	62.05
	GPT4o-mini	44.19	48.23	57.77	53.60	10.72	48.85	19.21	40.37
	+Hint	44.92	46.11	65.18	53.74	11.89	49.03	22.22	41.87
	+Hint-translate	30.21	27.07	51.39	37.35	7.35	33.67	8.32	27.91
	Aya-23	56.65	63.10	82.64	71.18	16.62	69.99	34.25	56.35
	+Hint	54.89	58.27	80.62	64.42	16.53	66.07	32.14	53.28
	+Hint-translate	52.67	57.87	80.88	64.24	15.88	63.51	28.20	51.89
	+Fine tuned	41.60	47.13	66.69	60.19	7.02	54.21	21.96	42.69
	BLOOMZ	62.87	39.11	77.70	55.58	20.94	76.97	51.90	55.01
	+Hint	54.13	30.66	71.13	46.23	19.64	71.84	43.60	48.18
	+Hint-translate	65.06	31.34	71.14	48.76	20.55	76.69	49.67	51.89
	Gemma-2	18.61	50.41	72.48	44.49	13.45	44.61	18.70	37.54
	+Hint	24.94	52.19	74.79	47.47	13.85	51.41	23.28	41.13
	+Hint-translate	26.39	54.46	76.16	54.83	13.87	56.42	25.73	43.98
	Mistral-Nemo	36.44	51.44	74.28	50.57	13.00	45.85	21.40	41.85
	+Hint	38.93	46.23	69.99	50.23	12.63	36.54	18.86	39.06
	+Hint-translate	42.84	42.52	69.01	49.24	12.37	35.36	21.16	38.93
No answer	**mT5-large**	73.98	67.14	69.20	67.87	76.73	67.25	73.39	70.79
	AYA-101	22.05	51.46	43.57	35.33	29.61	31.97	45.62	37.09
	+Hint	30.65	56.98	60.34	53.55	40.75	43.86	53.37	48.50
	+Hint-translate	29.73	57.56	59.92	53.17	41.70	42.25	55.94	48.61
	+Fine tuned	**81.20**	**79.51**	**82.73**	**79.76**	**78.73**	**81.23**	**83.97**	**81.02**
	GPT4o-mini	4.45	11.58	5.34	1.93	0.03	12.77	9.49	6.51
	+Hint	51.92	69.00	47.07	57.51	51.85	63.08	47.46	55.41
	+Hint-translate	59.33	76.43	51.01	58.14	50.85	68.11	52.04	59.42
	Aya-23	6.35	19.51	22.82	9.69	6.46	10.27	4.53	11.38
	+Hint	23.04	42.56	33.43	37.27	17.74	30.97	23.83	29.83
	+Hint-translate	21.97	39.17	31.07	35.05	19.78	29.80	24.67	28.79
	+Fine tuned	35.60	56.65	52.60	24.69	71.55	37.61	36.15	44.98
	BLOOMZ	21.54	40.20	34.88	38.30	15.08	16.87	27.67	27.79
	+Hint	39.27	62.32	49.76	57.76	27.23	29.80	47.70	44.83
	+Hint-translate	28.94	60.97	50.00	55.90	21.56	25.26	40.18	40.40
	Gemma-2	18.60	12.04	23.04	17.08	15.09	11.63	6.58	14.87
	+Hint	30.37	23.82	19.39	34.40	27.56	20.07	12.42	24.00
	+Hint-translate	25.47	19.62	16.07	25.27	25.15	14.91	7.04	19.08
	Mistral-Nemo	3.90	20.22	25.17	7.11	9.58	18.71	12.01	13.81
	+Hint	41.00	59.99	53.13	50.94	49.81	72.43	60.70	55.43
	+Hint-translate	31.76	62.99	52.44	48.91	48.81	76.29	48.05	52.75

Table 11: F1 scores (averaged over 3 model seeds) across three categories (Overall F1, Has answer F1, and No answer F1) on MLQA-IDK English-Questions setting across 7 language pairs. The bold scores represent the best performance for each category and language.

"unanswerable"," as illustrated in Figure 3.

I.4 AYA-101 Fine-tuning

To fine-tune AYA-101 on the SQuAD 2.0 dataset, we employed the QLoRA method to optimize memory efficiency and maintain high performance. The input window size was set to 2048 tokens to accommodate longer context passages effectively. The training process used a learning rate of 3e-5 and a batch size of 2 for both training and evaluation. The LoRA-specific parameters included a rank of 64, a scaling factor (`alpha`) of 32, and a dropout rate of 0.1 to prevent overfitting. The LoRA bias was configured as "none ", and all linear layers in the model were targeted for parameter-efficient updates.

Category	Model\Language	ar	de	en	es	hi	vi	zh	Avg
Average	**mT5-large**	61.12	68.09	73.66	69.06	63.41	61.67	62.64	65.66
	AYA-101	56.29	67.44	59.91	59.56	57.93	57.32	59.35	59.69
	+Hint	62.30	68.76	67.89	65.44	63.74	63.90	65.91	65.42
	+Hint-translate	62.08	68.45	67.71	65.32	63.52	63.73	65.60	65.20
	+Fine tuned	**73.03**	**78.05**	**81.64**	**77.64**	**73.86**	**73.53**	**75.85**	**76.23**
	GPT4o-mini	10.06	17.56	30.11	13.58	12.31	16.84	13.65	16.30
	+Hint	36.40	42.71	55.62	39.31	35.77	42.09	38.72	41.52
	+Hint-translate	21.26	26.48	51.19	24.58	21.52	22.78	25.02	27.55
	Aya-23	48.81	54.51	51.08	53.06	43.96	48.33	45.07	49.26
	+Hint	51.99	58.36	55.72	57.07	47.04	52.23	48.32	52.96
	+Hint-translate	50.38	57.26	54.60	55.56	44.63	50.33	46.19	51.28
	+Fine tuned	54.82	59.72	59.26	58.24	51.19	52.94	58.26	56.35
	BLOOMZ	55.67	46.67	55.11	57.45	53.63	55.93	57.57	54.58
	+Hint	58.77	49.66	59.85	60.42	56.68	59.00	60.16	57.79
	+Hint-translate	58.86	49.58	59.99	60.46	57.12	59.21	60.37	57.94
	Gemma-2	44.63	42.94	46.39	43.29	39.60	42.08	36.24	42.17
	+Hint	45.92	44.25	45.56	44.98	41.18	42.46	37.29	43.09
	+Hint-translate	45.75	44.25	44.46	45.25	41.25	41.51	37.63	42.87
	Mistral-Nemo	25.70	31.13	48.37	27.46	24.24	33.80	23.80	30.64
	+Hint	43.94	48.08	61.09	46.46	42.34	47.67	44.24	47.69
	+Hint-translate	44.39	47.16	60.27	45.62	40.95	47.19	44.16	47.11
Has answer	**mT5-large**	55.81	69.47	78.65	70.37	68.78	61.60	62.33	66.72
	AYA-101	73.70	80.73	78.17	76.38	75.88	74.82	73.08	76.11
	+Hint	71.54	79.10	76.33	74.01	73.73	72.38	70.29	73.91
	+Hint-translate	71.42	79.33	76.41	74.05	73.65	72.46	70.33	73.95
	+Fine tuned	**70.17**	**78.90**	**80.42**	**76.41**	**77.11**	**75.73**	**76.25**	**76.43**
	GPT4o-mini	15.19	26.43	57.77	21.01	19.20	26.13	20.43	26.59
	+Hint	27.41	37.03	65.18	31.23	35.46	36.28	31.91	37.79
	+Hint-translate	21.12	28.84	51.39	25.78	26.79	26.54	27.24	29.67
	Aya-23	72.49	77.45	82.64	76.95	65.78	71.23	67.46	73.43
	+Hint	69.83	74.51	80.62	73.79	61.48	69.40	65.44	70.72
	+Hint-translate	69.11	73.81	80.88	73.38	60.18	68.21	63.82	69.91
	+Fine tuned	60.61	64.78	66.69	65.29	59.70	58.89	57.30	61.89
	BLOOMZ	74.94	52.51	77.70	76.22	73.86	73.53	72.83	71.66
	+Hint	69.08	43.93	71.13	70.70	68.97	67.66	65.50	65.28
	+Hint-translate	69.11	42.99	71.14	70.73	69.30	67.62	65.62	65.22
	Gemma-2	59.10	56.26	72.48	56.01	57.52	60.03	54.53	59.42
	+Hint	58.47	56.04	74.79	56.95	56.18	59.57	54.77	59.54
	+Hint-translate	60.88	58.09	76.16	58.62	57.63	60.54	56.46	61.20
	Mistral-Nemo	40.39	43.98	74.28	41.89	39.29	42.51	33.67	45.14
	+Hint	28.87	29.95	69.99	34.42	34.06	19.56	30.19	35.29
	+Hint-translate	29.03	30.30	69.01	35.12	37.08	20.58	33.51	36.38
No answer	**mT5-large**	68.73	66.14	69.20	67.19	55.85	61.78	63.08	64.57
	AYA-101	31.39	48.80	43.57	35.76	32.69	33.18	39.85	37.89
	+Hint	49.08	54.25	60.34	53.31	49.70	52.20	59.69	54.08
	+Hint-translate	48.72	53.19	59.92	52.97	49.28	51.70	58.86	53.52
	+Fine tuned	**77.12**	**76.85**	**82.73**	**79.38**	**69.28**	**70.50**	**75.28**	**75.88**
	GPT4o-mini	2.71	5.12	5.34	3.08	2.61	4.03	4.02	3.84
	+Hint	49.25	50.67	47.07	50.75	36.21	50.11	48.41	47.50
	+Hint-translate	21.47	23.18	51.01	22.89	14.10	17.59	21.87	24.59
	Aya-23	14.94	22.33	22.82	19.26	13.27	16.75	13.24	17.52
	+Hint	26.48	35.71	33.43	33.40	26.73	28.55	23.98	29.75
	+Hint-translate	23.59	34.02	31.07	30.35	22.76	25.67	21.14	26.94
	+Fine tuned	46.54	52.63	52.60	48.25	39.21	44.74	59.63	49.09
	BLOOMZ	28.11	38.47	34.88	30.88	25.17	31.66	35.89	32.15
	+Hint	44.04	57.69	49.76	45.87	39.39	47.05	52.57	48.05
	+Hint-translate	44.19	58.84	50.00	45.93	39.99	47.61	52.90	48.49
	Gemma-2	23.95	24.25	23.04	25.28	14.41	17.32	10.25	19.79
	+Hint	27.97	27.71	19.39	28.05	20.08	18.86	12.45	22.07
	+Hint-translate	24.11	24.84	16.07	26.32	18.20	15.26	10.87	19.38
	Mistral-Nemo	4.70	13.08	25.17	7.03	3.06	21.78	9.77	12.08
	+Hint	65.50	73.52	53.13	63.51	53.98	86.44	64.21	65.76
	+Hint-translate	66.35	70.81	52.44	60.48	46.39	83.89	59.29	62.81

Table 12: F1 scores (averaged over 3 model seeds) across three categories (Overall F1, Has answer F1, and No answer F1) on MLQA-IDK English-Contexts setting across 7 language pairs. The bold scores represent the best performance for each category and language.

The model was fine-tuned over 2 epochs using the AdamW8bit optimizer, which supports low-memory operations while ensuring efficient gradient updates. Additionally, a weight decay of 0.01 was applied to regularize the model and prevent overfitting. This configuration enabled effective fine-tuning of AYA-101 on the question-answering task while optimizing for both memory and computational efficiency.

To further optimize memory and computational efficiency, we employed 4-bit quantization.

I.5 Fine-tuning AYA-23

For fine-tuning AYA-23 on SQuAD 2.0, we used a learning rate of 3e-5 and set the batch size to 2 for both training and evaluation. The LoRA-

Model	ar	es	th	de	hi	tr	el	ro	vi	en	ru	zh	Avg
Aya-23	7.08	10.43	0.78	13.63	5.76	9.44	10.31	10.32	11.66	24.07	12.20	6.73	10.20
+context	68.19	74.80	38.72	73.87	61.53	67.91	76.18	67.23	66.73	83.77	73.97	64.12	68.09
BLOOMZ	0.38	0.23	0.00	0.00	0.17	0.00	0.00	0.00	0.41	0.66	0.10	0.23	0.18
+context	70.82	77.73	2.71	56.62	73.13	7.63	14.51	31.70	72.74	82.04	59.74	73.92	51.94
AYA-101	2.53	2.57	4.28	1.24	2.14	4.17	0.88	4.26	2.13	1.24	4.41	1.73	2.63
+context	70.00	73.74	70.50	74.71	70.16	72.79	76.65	74.20	73.90	79.01	73.65	69.46	73.23
Gemma-2	8.42	9.78	8.31	10.70	6.43	9.13	11.12	11.24	10.97	16.00	11.00	8.07	10.10
+context	58.40	51.93	58.94	49.76	59.93	55.92	64.21	52.23	58.79	72.10	63.29	60.17	58.81
Mistral-Nemo	10.65	8.32	9.69	10.31	8.31	7.35	10.86	6.70	10.50	20.87	8.83	8.05	10.04
+context	41.43	41.70	50.95	43.34	43.41	40.26	48.76	36.02	46.62	79.19	45.51	40.59	46.48

Table 13: Has Answer Performance Across Languages: No Context vs. With Context

Model\Language	ar	es	th	de	hi	tr	el	ro	vi	en	ru	zh	Avg
AYA-101	71.29	66.32	59.13	69.23	71.20	69.61	72.33	71.39	64.37	62.86	70.79	68.80	68.11
+hint	71.83	67.43	58.84	69.25	71.53	70.88	73.21	72.44	65.68	62.14	71.95	69.11	68.69
+hint translate	71.40	66.61	58.58	68.50	69.87	70.52	72.57	71.78	64.99	61.57	71.15	67.83	67.95
+fine tuned	85.44	84.08	60.93	80.36	**85.34**	78.28	85.29	84.12	83.10	88.64	83.54	83.27	81.87
GPT4o-mini	36.37	27.06	41.24	34.91	45.50	40.25	28.99	34.50	36.97	58.15	40.91	33.27	38.18
+hint	70.73	56.94	63.76	64.16	78.90	65.65	66.74	61.00	62.95	85.18	70.92	60.88	67.32
+hint translate	15.73	19.85	17.75	22.44	19.89	19.32	16.87	21.86	19.22	57.78	18.39	18.31	22.28
Aya-23	79.60	87.96	45.25	88.00	67.94	81.45	89.81	**86.67**	78.37	95.07	85.91	75.68	80.14
+hint	79.33	84.99	44.87	87.51	64.36	80.36	88.45	85.06	77.48	95.38	84.22	73.59	78.80
+hint translate	74.75	81.04	46.41	87.19	64.19	78.68	85.19	82.96	75.21	95.37	82.14	72.10	77.10
BLOOMZ	82.27	86.27	16.76	66.57	84.23	34.52	35.27	55.91	81.01	89.88	73.54	**83.67**	65.83
+hint	79.50	84.53	11.03	62.67	82.58	24.64	26.39	51.20	77.89	87.91	69.26	81.49	61.59
+hint translate	76.68	82.99	9.17	58.66	79.98	21.77	23.03	47.42	76.28	86.16	64.95	78.47	58.80
Gemma-2	**91.06**	**88.42**	**86.67**	**89.19**	83.67	**84.49**	**92.34**	82.13	**85.06**	96.02	85.97	83.41	**87.37**
+hint	90.48	88.13	86.12	88.90	81.70	83.18	90.76	82.70	83.84	**96.33**	85.98	80.55	86.56
+hint translate	90.42	87.53	85.30	89.49	83.32	83.73	90.68	83.35	83.62	96.41	**87.10**	82.13	86.92
Mistral-Nemo	62.10	60.34	60.03	66.30	56.51	66.10	68.34	62.70	66.21	94.42	64.27	54.08	65.12
+hint	60.38	58.86	60.12	53.15	56.01	57.84	62.81	58.15	46.60	94.68	57.41	55.49	60.13
+hint translate	55.40	53.06	52.50	47.55	50.69	51.40	57.08	53.91	43.50	94.03	54.16	52.34	55.47

Table 14: Has Answer F1 scores across 12 language pairs on repliQA-Trans dataset (English-Contexts setting). The bold scores represent the best performance for each category and language.

Category	ar	fi	ja	ko	ru	avg_indi
Answered (A)	301	312	240	282	235	538
Unanswered (U)	281	294	231	266	213	96
Total	582	606	471	548	448	634

Table 15: XTREME-UP QA dataset statistics on English context settings. A: answerable questions, U: unanswerable questions, Total: total number of questions. The **avg_indi** column represents the average over 26 low-resource languages included in XTREME-UP (Ruder et al., 2023).

specific parameters included a rank of 64, an alpha value of 32, and a dropout rate of 0.1, with the bias set to "none ". All linear layers were targeted for parameter-efficient fine-tuning. The model was trained for 3 epochs using the AdamW optimizer with a weight decay of 0.01, and the maximum sequence length was configured to 2048 tokens.

J Prompt Control Analysis

We conducted control experiments to evaluate model sensitivity to prompt variations, particularly focusing on language specification placement and minor prompt changes. As shown in Figure 7, we tested four prompt variants: Hint-Prompt, Hint-Translate-Prompt, and their respective control versions that explicitly mention the target language in the question description. Results in Table 20 show minimal performance differences between these variants, indicating that the model is robust to such prompt modifications.

Category	Model\Language	ar	fi	ja	ko	ru	avg_indi	avg
Average	**mT5-large**	59.66	59.23	67.11	65.36	70.78	50.77	53.30
	AYA-101	73.20	72.12	**73.19**	73.09	73.63	**65.99**	**67.30**
	+hint	**73.47**	71.30	72.89	72.91	74.33	63.90	65.58
	+hint translate	73.39	72.09	72.80	72.72	**74.72**	63.89	65.60
	+fine tuned	72.57	68.03	72.81	**75.22**	70.60	52.90	56.41
	GPT4o-mini	10.87	11.95	13.60	15.47	9.44	20.54	19.01
	+hint	60.25	56.37	62.40	63.64	54.19	44.80	47.50
	+hint translate	20.78	28.82	33.20	45.17	32.77	19.00	21.44
	Aya-23	50.44	48.92	45.35	54.34	53.29	39.49	41.53
	+hint	61.80	60.98	60.45	66.14	64.39	40.08	44.28
	+hint translate	57.45	60.26	55.49	63.44	60.43	40.67	44.14
	+fine tuned	61.86	54.94	68.42	62.10	64.26	40.40	44.46
	BLOOMZ	68.44	-	65.91	53.41	60.81	52.61	53.86
	+hint	67.63	-	64.55	52.27	62.11	47.69	49.80
	+hint translate	68.03	-	64.47	52.75	62.21	47.66	49.81
	Gemma-2	23.25	10.69	22.44	22.99	25.37	22.06	21.86
	+hint	31.70	40.34	46.34	42.86	40.53	26.97	29.45
	+hint translate	35.52	42.97	46.92	41.57	46.30	30.06	32.39
Has answer	**mT5-large**	54.71	41.96	59.83	59.99	58.98	47.30	48.74
	AYA-101	65.67	**56.22**	56.83	63.41	66.60	**62.04**	**61.99**
	+hint	63.53	53.33	55.12	60.12	65.81	58.74	58.90
	+hint translate	63.95	54.44	54.94	60.11	67.13	58.85	59.09
	+fine tuned	54.60	46.12	55.59	59.47	52.45	45.46	46.98
	GPT4o-mini	18.03	19.67	20.43	24.03	16.29	23.09	22.46
	+hint	30.44	22.94	32.45	37.50	19.89	34.84	33.69
	+hint translate	11.60	16.55	15.14	20.39	13.95	13.23	13.65
	Aya-23	**70.40**	53.14	57.04	**74.62**	**71.22**	41.66	46.04
	+hint	67.77	42.27	56.41	68.36	67.28	34.09	38.97
	+hint translate	65.44	44.61	54.73	70.56	66.40	35.81	40.35
	+fine tuned	67.34	34.15	**60.93**	62.98	65.62	39.27	42.77
	BLOOMZ	57.58	-	50.19	19.87	49.54	47.74	45.53
	+hint	45.27	-	39.17	11.38	41.94	38.85	36.80
	+hint translate	45.37	-	38.88	12.32	42.15	38.69	36.70
	Gemma-2	44.06	19.69	40.01	43.02	45.66	27.15	29.25
	+hint	39.69	13.93	36.21	40.25	36.55	21.06	23.33
	+hint translate	42.54	19.25	37.07	40.34	41.03	24.28	26.46
No answer	**mT5-large**	64.95	77.56	74.68	71.06	83.81	69.43	70.35
	AYA-101	81.26	89.01	90.19	83.34	81.38	85.50	85.41
	+hint	84.11	90.37	91.35	86.47	83.73	89.06	88.72
	+hint translate	83.52	90.82	91.35	86.10	83.10	88.54	88.25
	+fine tuned	91.82	91.27	90.70	91.92	90.62	88.52	89.03
	GPT4o-mini	3.21	3.75	6.50	6.40	1.88	6.84	6.38
	+hint	**92.18**	**91.84**	**93.51**	91.36	**92.02**	90.62	**90.91**
	+hint translate	30.61	41.84	51.95	71.43	53.53	43.01	44.28
	Aya-23	29.07	44.45	33.19	32.84	33.49	31.88	32.39
	+hint	55.40	80.84	64.65	63.79	61.19	71.71	70.50
	+hint translate	48.88	76.88	56.28	55.89	53.84	67.31	65.65
	+fine tuned	56.00	76.99	76.20	61.16	62.76	47.41	50.97
	BLOOMZ	80.08	-	82.26	88.98	73.24	77.16	78.50
	+hint	91.58	-	90.91	**95.62**	84.36	90.31	87.71
	+hint translate	90.29	-	91.06	**95.62**	84.36	**90.76**	86.11
	Gemma-2	0.95	1.14	4.19	1.76	2.98	1.15	1.34
	+hint	23.14	68.37	56.86	45.62	44.92	59.53	57.35
	+hint translate	28.00	68.15	57.15	42.86	52.12	61.52	59.32

Table 16: F1 scores across three categories (Overall F1, Has answer F1, and No answer F1) on XTREME-UP dataset with English-Contexts setting. The table shows performance across different languages with various models. The bold scores represent the best performance for each category and language. The **avg_indi** column represents the average over 26 low-resource languages included in XTREME-UP.

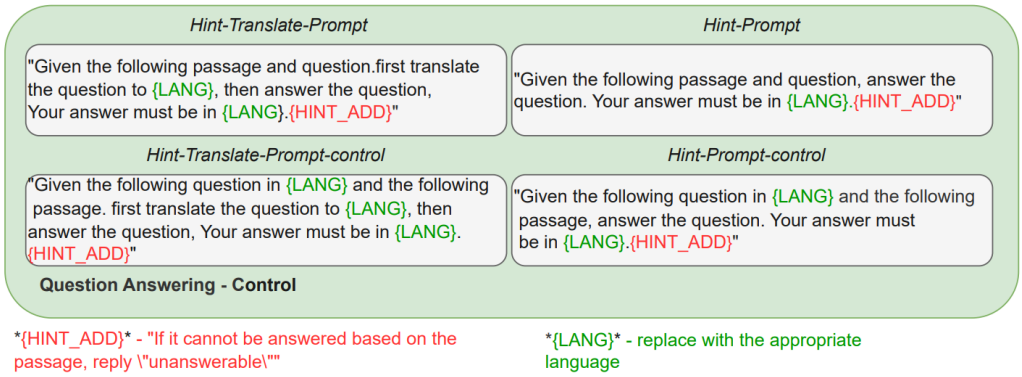

Figure 7: Prompt variations used in control experiments to test model robustness to language specification and prompt formatting.

Model	MiXQuAD English-Questions	MiXQuAD English-Contexts	MLQA-IDK English-Questions	MLQA-IDK English-Contexts
mBERT	1.36	1.58	1.51	1.45
XLM-R	1.80	1.69	1.74	1.53
mDeBERTa	1.99	2.47	1.89	2.21
mT5-large	1.91	2.43	1.79	2.39
AYA-23	3.58	3.52	3.52	3.60
+Hint	3.25	3.36	3.19	3.45
+Hint-translate	3.34	3.44	3.27	3.51
+Fine tuned	1.86	2.21	1.69	2.01
BLOOMZ	1.94	1.84	2.01	2.16
+Hint	1.56	1.48	1.70	1.91
+Hint-translate	1.69	1.51	1.95	2.00
AYA-101	2.61	2.34	2.54	2.05
+Hint	2.50	2.13	2.28	1.46
+Hint-translate	2.56	2.14	2.36	1.44
+Fine tuned	2.30	2.42	2.01	2.32
GPT4o-mini	6.08	9.38	6.92	9.41
+Hint	4.78	7.48	4.95	7.26
+Hint-translate	7.25	11.36	6.67	9.38
Gemma-2	2.43	2.34	2.33	2.28
+Hint	2.39	2.52	2.34	2.52
+Hint-translate	2.72	2.73	2.69	2.75
Mistral-Nemo	4.77	6.88	4.69	6.49
+Hint	3.61	3.80	3.38	3.39
+Hint-translate	3.97	4.59	3.74	3.85
Average (gold)	**2.86**	**2.92**	**3.31**	**3.18**

Table 17: Average number of words in model predictions compared to gold answers across development sets in MiXQuAD and MLQA-IDK datasets for both English-Questions and English-Contexts settings.

Model	Method	Count	Uncertainty
AYA-23	Regular-Prompt	4034	-0.8205
AYA-23	Hint-Prompt	8939	-0.8802
AYA-101	Regular-Prompt	11752	-0.8738
AYA-101	Hint-Prompt	12347	-0.8917
BLOOMZ	Regular-Prompt	8309	-0.6782
BLOOMZ	Hint-Prompt	13351	-0.7840

Table 18: Uncertainty scores for correctly classified unanswerable questions across models and prompting methods. Lower uncertainty values (more negative) indicate higher confidence. Count represents the number of questions where the model predicted "unanswerable" and the true label was also "unanswerable" (true positives for the unanswerable category).

Model	Method	Correct	Incorrect	Diff.
AYA-23	Regular-Prompt	-0.8492	-0.7198	-0.1295
AYA-23	Hint-Prompt	-0.8571	-0.7256	-0.1315
AYA-101	Regular-Prompt	-0.9001	-0.8288	-0.0713
AYA-101	Hint-Prompt	-0.9087	-0.8273	-0.0814
AYA-101	Fine-Tuned	-0.8652	-0.8751	0.0099
BLOOMZ	Regular-Prompt	-0.7713	-0.6675	-0.1038
BLOOMZ	Hint-Prompt	-0.8168	-0.7108	-0.1060

Table 19: Mean uncertainty scores for correct versus incorrect predictions. Negative differences indicate higher confidence for correct predictions. The fine-tuned AYA-101 model shows reduced calibration with near-zero difference.

Category	Model\Language	ar	es	th	de	hi	tr	el	ro	vi	en	ru	zh	Avg
Average	hint	**55.49**	**71.42**	35.69	**71.73**	**31.65**	**62.50**	**62.01**	**68.66**	**54.86**	**79.02**	59.50	58.09	**59.22**
	hint-control	55.12	70.54	**35.84**	70.91	30.29	62.13	61.42	67.92	53.22	78.79	**59.61**	57.55	58.61
	hint-translate	54.62	71.06	35.39	70.60	31.31	61.36	60.43	66.74	53.10	78.63	58.94	**58.12**	58.36
	hint-translate-control	54.15	68.96	35.35	70.13	30.89	61.09	58.46	65.52	51.85	78.33	58.41	57.41	57.55
Has answer	hint	**71.31**	**76.79**	**23.73**	**76.65**	**19.23**	**68.74**	**71.66**	**75.60**	**63.59**	80.53	70.90	59.13	**63.16**
	hint-control	71.10	76.43	23.65	75.57	17.77	68.27	71.14	74.81	60.56	80.20	**70.94**	**59.18**	62.47
	hint-translate	70.40	76.16	23.72	74.69	17.62	67.16	70.18	73.76	61.36	**80.68**	70.36	56.48	61.88
	hint-translate-control	69.66	75.52	23.56	73.96	17.23	66.86	67.33	72.83	60.36	80.50	70.10	57.10	61.25
No answer	hint	**34.13**	**64.18**	**51.82**	**65.08**	**48.42**	**54.09**	**48.98**	**59.30**	43.09	**76.99**	44.11	56.69	**53.91**
	hint-control	33.57	62.59	52.27	64.63	47.17	53.86	48.30	58.62	**43.32**	76.88	**44.34**	55.33	53.41
	hint-translate	33.34	**64.18**	51.14	**65.08**	**49.78**	53.52	47.28	57.26	41.96	75.86	43.54	**60.32**	53.61
	hint-translate-control	33.22	60.10	51.25	64.97	49.32	53.29	46.49	55.67	40.37	75.40	42.64	57.83	52.55

Table 20: F1 scores across three categories (Average F1, Has Answer F1, and No Answer F1) for AYA-101 on miXQuAD in the English-Questions setting across 12 language pairs. The bold scores represent the best performance for each category and language.