Cross-Lingual Extractive Question Answering with Unanswerable Questions

Yuval Gorodissky[1], Elior Sulem[1], Dan Roth[2,3]
Ben-Gurion University of the Negev[1]
University of Pennsylvania[2] ,Oracle AI[3]

EMNLP 2025 Suzhou, China 中国苏州

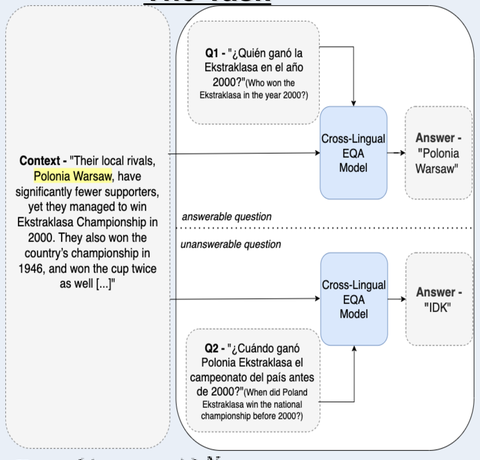
1 Motivation

- Previous cross-lingual QA assumes all questions are answerable.
- 51% of real queries lack answers in given context.
- EQA techniques can be effectively applied to downstream tasks.

2 Contributions

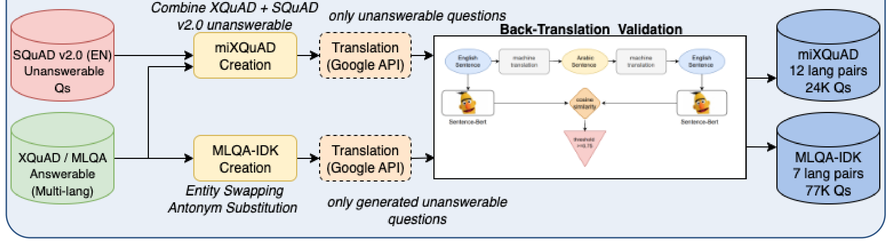
- **Enhanced G-XLT Task:** Extended to handle unanswerable questions.
- **New Datasets: miXQuAD:** 12 languages, 24K Qs **| MLQA-IDK:** 7 languages, 77K Qs.
- **Comprehensive Evaluation:** Fine-tuning, LoRA, and In-Context Learning approaches.

3 The Task



$$Given: \mathrm{D} = \{(c_i, q_i, a_i)\}_{i=1}^{N}$$
$$\text{Learn: } f : (q \in L_q, c \in L_c) \rightarrow \{s \subseteq c, \mathrm{IDK}\}$$
$$\text{Training: } L_c = L_q = \mathrm{EN}$$
$$\text{Evaluation: } L_q = \mathrm{EN}, L_c \neq \mathrm{EN} \text{ OR } L_c = \mathrm{EN}, L_q \neq \mathrm{EN}$$

4 Dataset Creation and Validation



5 Methods

Training set - SQuAD v2.0.

Full-Fine tuning

All model parameters are updated during training.

Parameter-Efficient Fine-Tuning

Only low-rank adaptation matrices are trained while freezing base model weights, enabling efficient fine-tuning of large models.

In-Context Learning

No parameter updates; models learn from 3 few-shot examples (2 answerable + 1 unanswerable) provided in the prompt.

Regular-Prompt

"Given the following passage and question, answer the question. Your answer must be in {LANG}."

Hint-Prompt

"Given the following passage and question, answer the question. Your answer must be in {LANG}.{HINT_ADD}"

Hint-Translate-Prompt

"Given the following passage and question.first translate the question to {LANG}, then answer the question, Your answer must be in {LANG}.{HINT_ADD}",

Question Answering

{HINT_ADD} - "If it cannot be answered based on the passage, reply "unanswerable""

{LANG} - replace with the appropriate language

6 Results

Model Architecture and Training Approach Effects

- Hint prompting dramatically improves unanswerable detection
- Trade-off: Small fine-tuned models excel at unanswerable, large prompted models excel at answerable
- Encoder-only models: Strong unanswerable detection, weak answering

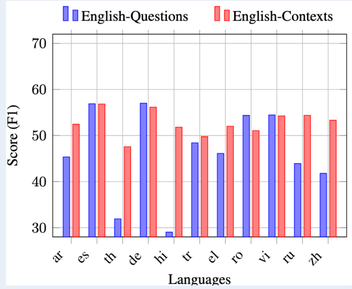
MiXQuAD

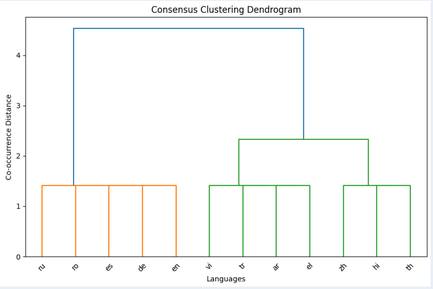
Model	Avg	Has Ans	No Ans
mT5-large	64.03	50.55	82.20
Aya-101	53.94	67.86	35.16
+Hint	61.61	66.71	54.74
+Hint-translate	61.41	66.20	54.94
+Fine tuned	**81.23**	**77.09**	**86.80**

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

Language Dependency

- English-Contexts outperforms English-Questions.
- Language families influence cross-lingual transfer performance patterns.





Evaluating Model Robustness

- Strong out-of-domain generalization on MLQA-IDK
- Effective on post-training repliQA-Trans and low-resource open-domain XTREME-UP.

Model	repliQA	XTREME-UP
mT5-large	70.61	53.30
AYA-101	68.11	**67.30**
+Hint	68.69	65.58
+Fine-tuned	**81.87**	56.41
GPT4o-mini	38.18	19.01
+Hint	67.32	47.50

Model	EN-Questions	EN-Contexts
mT5-large	51.24	65.66
AYA-101 (FT)	**69.96**	**76.23**
AYA-101 (+Hint)	52.68	65.42
GPT4o (+Hint)	47.20	41.52

MLQA-IDK