KERLQA: Knowledge-Enhanced Reinforcement Learning for Question Answering in Low-resource Languages

Anonymous ACL submission

Abstract

Question answering in low-resource languages poses challenges for Large Language Models due to limited training data and knowledge resources. We propose Knowledge-Enhanced Reinforcement Learning for Question Answering (KERLQA), a novel approach integrating external knowledge with reinforcement learning to optimize model behavior. KERLQA employs a graph neural network for joint reasoning over question context and knowledge sources, while introducing an abstention mechanism to address the heightened risk of hallucination in low-resource settings. This mechanism allows the model to refrain from answering when uncertain, which is particularly important for low-resource languages where knowledge gaps are more prevalent. We evaluate KERLQA on CommonsenseQA and OpenBookQA across English and four low-resource South African languages: isiZulu, isiXhosa, Sepedi, and SeSotho. Results show KERLQA outperforms baselines and state-of-the-art systems, with notable improvements in low-resource settings. Our error analysis reveals distinct patterns of knowledge gaps, reasoning failures, and abstention errors across languages, with higher abstention rates in low-resource languages confirming the model's ability to recognize and mitigate knowledge gaps.

1 Introduction

Question answering in low-resource languages presents unique challenges for language models, including limited training data, scarce knowledge resources, and complex cross-lingual transfer issues (Samuel et al., 2023; Chen et al., 2023). These challenges are particularly prevalent for languages with distinct linguistic structures and cultural contexts that differ from high-resource languages like English (Ogundepo et al., 2022). A question answering model that abstains when it does not have the necessary knowledge to answer a question would be preferable, in particular in low-resource settings where the proportion of such questions will likely be higher.

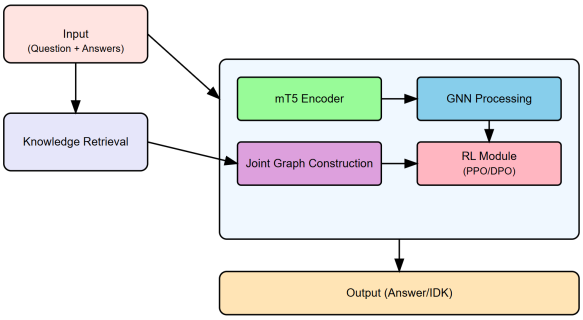

Figure 1: KERLQA architecture. The model processes inputs through both language model encoding and knowledge retrieval for graph construction, integrated through a graph neural network architecture. Reinforcement learning is used to adapt the decision making behaviour.

In this paper, we propose Knowledge-Enhanced Reinforcement Learning for Question Answering (KERLQA), a novel approach enabling language models to effectively utilize both internal and external knowledge sources while learning when to abstain from answering. KERLQA integrates external knowledge from sources such as ConceptNet (Speer et al., 2016) and DBpedia (Mendes et al., 2012) with reinforcement learning techniques to optimize model behavior. Our approach employs a graph neural network architecture for joint reasoning over question context and relevant knowledge sources based on QA-GNN (Yasunaga et al., 2021). We then employ Reinforcement Learning (RL) techniques to optimize decision-making behaviour, implementing both Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) to learn a policy that not only selects an answer from a candidate set but also decides when to abstain. Joint knowledge enhancement and reinforcement learning enable the model to learn when to rely on its own knowledge, when to seek external information, and when to abstain due to uncertainty. Figure 1 provides an overview of the KERLQA architecture.

We demonstrate KERLQA's effectiveness on English and four low-resource South African languages: isiZulu, isiXhosa, Sepedi, and SeSotho, fine-tuning the multilingual mT5 language model. Our experimental results show that KERLQA improves question answering performance compared to baselines. Furthermore, our model exhibits the ability to make informed decisions about knowledge utilization and abstention, leading to more reliable and accurate responses.

To gain deeper insights into KERLQA's performance, we conduct a comprehensive error analysis across languages and datasets. This analysis reveals distinct patterns of knowledge gaps, reasoning failures, and abstention errors, highlighting the unique challenges posed by different linguistic contexts and question types. Our error analysis revealed that knowledge gaps were 30-40% more prevalent in low-resource languages compared to English, while reasoning failures decreased by 15-25% as we moved from high to low-resource languages.

Our main contributions are: (1) We introduce KERLQA, a novel approach that combines knowledge-enhanced QA with reinforcement learning to address the challenges of low-resource languages; (2) We demonstrate KERLQA's effectiveness across five languages and two datasets, showing significant improvements over existing methods; (3) We provide a detailed error analysis that offers insights into the specific challenges of QA in low-resource settings, paving the way for future research directions.

2 Related Work

Our work intersects with several key areas in natural language processing and machine learning. We review relevant literature in two main categories: (1) Knowledge-Enhanced Question Answering, which focuses on integrating external knowledge sources to improve QA performance, and (2) Reinforcement Learning for NLP Tasks, which explores the application of RL techniques to language tasks, particularly in QA contexts.

2.1 Knowledge-Enhanced Question Answering

Recent advancements in question answering have focused on augmenting language models with external knowledge sources. Yasunaga et al. (2021) introduced QA-GNN, which uses graph neural networks to reason over knowledge graphs for QA tasks. Building on this, Zhang et al. (2022)) proposed GreaseLM, which enhances language models with graph-based reasoning. Our work differs from these approaches by integrating reinforcement learning with knowledge graph reasoning, allowing for more adaptive use of external knowledge.

Wang et al. (2022) introduced GSC (Graph-based Semi-parametric Contextualizer) for QA, while Park et al. (2023) proposed QAT (Question Answering Transformer) which uses meta-path tokens for knowledge integration. Ye et al. (2023) proposed FiTs (Fine-grained Two-stage training), a framework designed to address the challenges of fusing representations from pre-trained language models and knowledge graphs in knowledge-aware question answering. Jiang et al. (2022) conducted a deep empirical analysis of knowledge-enhanced commonsense reasoning. Their work revealed that relation features from commonsense knowledge graphs are the primary contributors to improving the reasoning capacity of pre-trained language models, rather than node features.

While these methods have shown promising results, they face challenges in fully utilizing external knowledge graphs and addressing the modality gap between text and KGs. Our work builds upon these insights in several ways. While we draw inspiration from these works, our approach uniquely combines these ideas with reinforcement learning. KERLQA employs a dynamic, reinforcement learning-based method to determine how and when to utilize relational knowledge. This is particularly important in low-resource language settings, where the relevance and reliability of knowledge graph relations may vary.

2.2 Reinforcement Learning in NLP

Reinforcement Learning (RL) has been increasingly applied to question answering tasks with large language models. Recent work has focused on Reinforcement Learning from Human Feedback (RLHF) to align model outputs with human preferences (Ouyang et al., 2022). RLHF combines traditional RL techniques with human preferences to guide model behavior, allowing for more nuanced and context-aware responses in language tasks. Two key developments in this area are the use of Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO).

PPO (Schulman et al., 2017) is a policy gradient method that has gained popularity in RL for NLP tasks due to its stability and effectiveness. It optimizes a surrogate objective function that prevents large policy updates, which can lead to performance collapse (Schulman et al., 2017). In the context of instruction following, which includes question answering tasks, PPO has been used to train language models to generate more accurate and relevant responses. It allows for fine-tuning model behavior based on specified reward functions, such as answer correctness or relevance (Ouyang et al., 2022). DPO (Rafailov et al., 2023) is a more recent development in RL for language models, designed to align model outputs with human preferences. Unlike traditional RL methods that require explicit reward modeling, DPO learns directly from pairwise preference comparisons. In QA tasks, DPO can be used to fine-tune models to generate answers that are not only correct but also preferred by humans in terms of clarity, conciseness, or other desirable attributes (Rafailov et al., 2023).

Our work extends these RL approaches to the domain of knowledge-enhanced QA, particularly for low-resource languages. We integrate PPO and DPO with knowledge graph utilization strategies, enabling models to learn when to rely on external knowledge, when to abstain from answering, and how to optimize its responses based on both correctness and human-like preferences.

In the context of abstention, Yang et al. (2023) constructed an honesty alignment dataset by replacing incorrect or uncertain responses with "I don't know" and fine-tuning on this data. Cheng et al. (2024) and Brahman et al. (2024) employed Direct Preference Optimization to encourage models to answer questions they know and refuse those they don't. Liang et al. (2024) used Proximal Policy Optimization with a reward model trained on hallucination scores to determine knowledge boundaries.

Our work builds upon these approaches but makes a distinct contribution by focusing specifically on learning when to abstain from answering. Unlike previous methods that primarily aim to improve overall QA performance or align with general human preferences, our approach explicitly trains the model to recognize its own knowledge limitations and uncertainties across various languages and contexts. This is particularly important in low-resource settings where the risk of hallucination is higher due to limited training data and knowledge resources.

3 KERLQA

3.1 Problem Formulation

Given a question q and a set of candidate answers $A = \{a_1, \ldots, a_n\}$, our goal is to learn a policy $\pi_\theta(a|s)$ that either selects an answer from A or opts to abstain, which is encoded through adding an additional answer option. The state s is defined as a combination of the textual representation (from the language encoder) and the knowledge graph context. We model this task as a Markov Decision Process (MDP) with:

- **State Space** $\mathcal{S}$: The concatenation of the question encoding, candidate answer encodings, and the current knowledge graph state.
- **Action Space** $\mathcal{A}$: The extended set $A' = A \cup \{\text{“I don't know”}\}$.
- **Reward Function** $r(s, a)$: A composite reward that considers answer correctness, the appropriateness of abstaining, and the efficiency of external knowledge utilization.
- **Transition Function**: Deterministic, as each question is processed independently.

3.2 Model Architecture

The question q and answer candidates $\{a_1, \ldots, a_n\}$ are encoded into dense vectors with a multilingual language model such as mT5:

$$h_q = \text{mT5}_{\text{enc}}(q), \quad h_{a_i} = \text{mT5}_{\text{enc}}(a_i). \tag{1}$$

To enable knowledge enhancement we retrieve knowledge triples with subjects or objects matching question entities or answer candidates from an external knowledge base. We construct a heterogeneous graph $G_W = (V_W, E_W)$ similar to QA-GNN (Yasunaga et al., 2021):

$$V_W = V_{\text{text}} \cup V_{\text{knowledge}} \cup \{z\}, \tag{2}$$

where text nodes V_{text} represent question and answer candidate embeddings, knowledge nodes $V_{\text{knowledge}}$ are constructed from the extracted knowledge triples (and also encoded with mT5), and the context node z aggregates global interactions between questions and answers. Edges are added based on the knowledge triples and between nodes representing the same entity. Additional edges are added between nodes representing the same entity to reinforce entity consistency.

Information is propagated in the Graph Neural Network (GNN) via standard message passing:

$$h_v^{(l+1)} = \text{GNN}\Big(h_v^{(l)}, h_u^{(l)} : u \in \mathcal{N}(v)\Big), \tag{3}$$

where $\mathcal{N}(v)$ denotes the neighbors of node v.

The final policy is obtained by fusing representations from the language encoder, the graph neural network, and the knowledge aggregation module:

$$\pi_\theta(a|s) = \text{softmax}\Big(\text{MLP}\big([h_q; h_a; z]\big)\Big). \quad (4)$$

During training, we first perform supervised fine-tuning on mT5 only using QA pairs. Then QA-GNN is trained, using the GNN to encode the extracted knowledge base information.

3.3 Reinforcement Learning Framework

We apply two RL strategies to optimize the answer-selection policy, each providing a different perspective on handling abstentions.

Proximal Policy Optimization (PPO) After initial supervised fine-tuning on QA pairs, we apply PPO to further refine $\pi_\theta(a|s)$. PPO maximizes the following clipped objective:

$$L_{PPO}(\theta) = \mathbb{E}[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)], \quad (5)$$

where $r_t(\theta)$ is the ratio of the new to old policy probabilities, A_t is the advantage estimate computed using temporal-difference methods, and ϵ is a clipping parameter (set via grid search on validation data).

Our reward function balances correct answering, appropriate abstention, and efficient use of external knowledge:

$$\begin{aligned} r = {} & \alpha \cdot \nVdash[\text{correct}] \\ & + \beta_1 \cdot \nVdash[\text{abstain on unanswerable}] \\ & - \beta_2 \cdot \nVdash[\text{abstain on answerable}] \\ & + \gamma_1 \cdot \nVdash[\text{used KG appropriately}] \\ & - \gamma_2 \cdot \nVdash[\text{used KG unnecessarily}], \end{aligned} \quad (6)$$

where $\nVdash[\text{condition}]$ is an indicator function that equals 1 when the condition is true and 0 otherwise. The parameters α, β_1, β_2, γ_1, and γ_2 are tunable hyperparameters that control the relative importance of each reward component. The intuition behind our reward structure is to encourage the model to answer correctly when it can; abstain when the question is truly unanswerable; use external knowledge when necessary; and avoid unnecessary abstention or knowledge use. In scenarios where the KG is not used (because the question is answerable solely from internal knowledge), the reward function naturally guides the model away from unnecessary KG retrieval.

Direct Preference Optimization (DPO) For DPO, the reward model $r_\theta(x, y)$ is learned implicitly through preference pairs, thereby avoiding the need for explicit reward engineering. For each question where the baseline mT5 answered correctly, the correct answer is marked as the "chosen" action and the others as "rejected." For questions where the baseline failed, the abstention action (i.e., *"I don't know"*) is marked as "chosen." Although the DPO formulation does not explicitly incorporate a term for KG usage, the preference pairs are derived from baseline performance that includes KG integration. Consequently, if incorporating KG information improves performance, the resulting preference pairs will indirectly favour actions that use the KG appropriately. Conversely, if KG usage is unnecessary, the model will learn to minimize its use.

The DPO objective is then:

$$L_{DPO}(\theta) = -\mathbb{E}[\log(\sigma(r_\theta(x, y) - r_\theta(x, y')))], \quad (7)$$

where y is the chosen answer and y' is a rejected option.

4 Experiment Results and Analysis

4.1 Experimental Setup

We evaluate KERLQA on CommonsenseQA (Talmor et al., 2019) and OpenBookQA (Mihaylov et al., 2018) datasets. For isiZulu and Sepedi, we use manually translated test sets obtained from Ralethe and Buys (2025). For broader coverage, we also used machine translation to obtain translations in isiXhosa and SeSotho, utilizing Tencent's Multilingual Machine Translation System for WMT22 Large-Scale African Language Translation (Jiao et al., 2022). To assess the impact of using automatic translations we also evaluate our QA models using machine translations into Sepedi and isiZulu, using SeamlessM4T (Barrault et al., 2023) for the latter. Results from these experiments are given in Appendix C, showing a small drop in accuracy but similar trends overall.

We utilize ConceptNet (Speer et al., 2016) as our primary knowledge source. For the four South African languages (isiZulu, isiXhosa, Sepedi, and SeSotho), we incorporate projected knowledge bases derived using LeNS-Align (Ralethe and Buys, 2025). LeNS-Align projects English ConceptNet triples into these target languages through a combined process of lexical alignment, named-entity recognition, and semantic alignment. This ap-

Method	English		isiZulu		isiXhosa		Sepedi		SeSotho	
	Acc. ↑	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**
mT5	67.11	-	57.10	-	55.76	-	56.15	-	55.12	-
mT5+QA-GNN	70.32	-	61.87	-	58.52	-	60.02	-	57.94	-
RLQA (PPO)	69.21	20.52	58.10	28.62	56.51	30.61	57.82	29.33	56.19	31.77
RLQA (DPO)	69.36	19.14	59.34	26.36	57.02	28.36	58.30	29.41	57.25	32.41
KERLQA (PPO)	77.33	14.35	63.67	25.81	59.33	27.14	62.13	27.31	59.12	29.17
KERLQA (DPO)	76.61	14.11	63.21	26.88	59.11	30.23	62.11	28.58	58.19	29.87

Table 1: Test Accuracy (Acc.) and Abstention Rate (AR) results on CommonsenseQA for different methods across languages. Accuracy is calculated over all questions, including abstentions.

proach yields approximately 670k triples per language, with human evaluations indicating an accuracy exceeding 85% (Ralethe and Buys, 2025).

The base language model is mT5-large (Xue et al., 2021). We use the adapted QA-GNN architecture with 3 message-passing layers. PPO and DPO are implemented using HuggingFace's PPOTrainer and DPOTrainer (Huang et al., 2023). Hyperparameter tuning details are given in Appendix B.

Additionally, we also train models for English based on RoBERTa-Large (Liu et al., 2019) to enable comparison to other recent knowledge-enhanced QA approaches. We train QA baselines based on mT5 only and QA-GNN without RL training, and also compare to performing RL training but without knowledge enhancement (RLQA).

4.2 Evaluation Metrics

We evaluate our models using the following metrics:

- **Accuracy:** The fraction of total questions answered correctly, where abstentions are treated as incorrect:

$$\text{Accuracy} = \frac{\text{Correct Answers}}{\text{Total Questions}}.$$

- **Precision:** The fraction of correctly answered questions among the attempted questions (i.e., excluding abstentions):

$$\text{Precision} = \frac{\text{Correct Answers}}{\text{Attempted Questions}}.$$

- **Abstention Rate (AR):** The fraction of questions where the model opts to abstain:

$$\text{AR} = 1 - \frac{\text{Attempted Questions}}{\text{Total Questions}}.$$

4.3 Results

Results on CommonsenseQA (Table 1) show that QA accuracy is highest on English, while isiZulu has the highest accuracy among the low-resource languages. KERLQA (with PPO and DPO) shows substantial improvements over the mT5 baseline for all languages. Notably, abstention rates are generally higher for low-resource languages, indicating increased model uncertainty in these contexts. OpenBookQA results (Table 2) exhibit similar trends: Abstention rates are generally slightly lower than on CommonsenseQA, suggesting models find this dataset somewhat easier to navigate.

The performance improvement from mT5 to mT5+QA-GNN confirms that the QA-GNN models are able to effectively leverage external knowledge to mitigate gaps in the training data and the pretrained model's knowledge. On low-resource languages this is the case even though the knowledge graphs were automatically projected from English and therefore contain some noise. By connecting questions to language-agnostic concepts, the model may also better leverage understanding gained from high-resource language pretraining, boosting its performance in low-resource settings.

The results also show that reinforcement learning leads to improved accuracy over approaches without RL, both in settings with knowledge enhancement (KERLQA over mT5+QA-GNN) and without knowledge enhancement (RLQA over mT5), despite abstention being an additional option. KERLQA achieves the best overall performance. The effectiveness of these RL techniques can be attributed to their ability to adaptively adjust confidence thresholds, balance exploration and exploitation, and optimize based on reward signals that align with desired outcomes. Comparing PPO and DPO, we observe that PPO slightly outperforms DPO in terms of accuracy across most scenarios. However, DPO often achieves lower abstention rates, particularly for low-resource languages. This pattern suggests that PPO's approach to policy optimization may be more adept at handling the complexities inherent in multilingual QA

Method	English		isiZulu		isiXhosa		Sepedi		SeSotho	
	Acc. ↑	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**	**Acc. ↑**	**AR ↓**
mT5	78.23	-	57.83	-	56.87	-	57.32	-	56.32	-
mT5+QA-GNN	83.48	-	63.42	-	61.13	-	61.33	-	58.76	-
RLQA (PPO)	79.33	16.32	58.25	29.95	57.12	32.86	57.22	31.91	56.12	32.78
RLQA (DPO)	79.85	18.33	59.74	29.04	56.45	29.86	56.93	29.35	56.35	29.93
KERLQA (PPO)	86.42	10.12	64.32	24.54	61.11	27.10	63.52	26.69	60.05	28.72
KERLQA (DPO)	84.79	11.36	64.81	26.14	60.24	29.54	62.85	28.33	59.95	29.94

Table 2: Test Accuracy (Acc.) and Abstention Rate (AR) results on OpenBookQA for different methods across languages. Accuracy is calculated over all questions, including abstentions.

Method	CSQA		OBQA	
	Prec.	**Acc.**	**Prec.**	**Acc.**
QAT (Park et al., 2023)	75.4	75.4	86.9	86.9
FIT (Ye et al., 2023)	75.6	75.6	86.0	86.0
GRT (Zhao et al., 2024)	76.1	76.1	87.3	87.3
KERLQA (PPO) **ours**	**78.2**	76.6	**88.6**	87.9

Table 3: Performance Comparison between KERLQA (PPO) with KG-augmented QA systems on CommonsenseQA (CSQA) and OpenbookQA (OBQA), using RoBERTa-Large (Liu et al., 2019). Precision excludes abstained questions while accuracy includes them.

tasks, while DPO shows promise in managing uncertainty in low-resource contexts.

The performance gap between KERLQA and other methods is particularly notable in English, especially for OpenBookQA. This could be due to KERLQA's enhanced ability to leverage the richer knowledge resources available in English. The smaller performance gains in low-resource languages suggest that while KERLQA improves performance, it remains constrained by limited knowledge resources in these languages.

Table 3 compares our approach to other recent QA models using knowledge enhancement, evaluated on the English datasets only and using RoBERTa-Large (Liu et al., 2019) as the backbone model instead of mT5. The results demonstrate that KERLQA achieves higher precision than previous approaches while maintaining competitive accuracy. This indicates that while KERLQA may abstain from answering some questions, it exhibits higher confidence and accuracy on the questions it chooses to answer.

4.4 Abstention Behaviour Analysis

Analysis of abstention patterns reveals that KERLQA primarily abstains on questions in three key scenarios: when the required knowledge is not present in the external knowledge bases, when multiple answer options appear plausible given the available information, and when the question intent is ambiguous or requires complex reasoning. These scenarios reflect situations where the model recognizes its limitations or uncertainty in providing accurate responses.

While the introduction of abstention creates an apparent asymmetry in evaluation, our dual-metric approach provides a comprehensive and fair assessment. The precision metric allows us to evaluate KERLQA's decision-making capability in comparable terms to previous approaches, while accuracy provides a conservative estimate of overall performance.

The higher precision demonstrates that KERLQA's abstention mechanism successfully identifies cases where the model lacks sufficient confidence or knowledge to provide a reliable answer. This capability is particularly valuable in real-world applications where incorrect answers may be more costly than abstentions.

KERLQA demonstrates lower abstention rates compared to RLQA across all languages, suggesting that knowledge enhancement can effectively enhance model confidence while increasing accuracy at the same time, leading to more robust and reliable QA systems. The higher abstention rates in low-resource languages compared to English demonstrate the impact of data scarcity on model behavior. DPO tends to results in higher abstention rates that PPO, particularly in low-resource languages, suggesting that the choice of RL algorithm can influence a model's abstention strategy. The results demonstrates KERLQA's improved ability to align confidence (lower abstention) with actual performance (higher accuracy). Furthermore, KERLQA exhibits an enhanced capability in distinguishing between questions it can answer correctly and those it should avoid, contributing to its overall performance across languages and datasets.

	English		isiZulu		isiXhosa		Sepedi		SeSotho	
Error Type	OB	CS	OB	CS	OB	CS	OB	CS	OB	CS
Knowledge Gap	18%	22%	27%	30%	29%	32%	28%	31%	30%	34%
Reasoning Failures	25%	21%	23%	20%	21%	19%	22%	19%	20%	18%
Abstention Errors	10%	8%	15%	18%	17%	20%	16%	20%	18%	21%

Table 4: Error Distribution Across Languages and Datasets. OB: OpenBookQA, CS: CommonsenseQA. Percentages represent the proportion of errors within each category and do not sum to 100% as they are calculated relative to the total number of questions, not just erroneous responses.

4.5 Performance on Filtered Datasets

As an additional experiment to assess KERLQA's ability to leverage external knowledge bases, we created filtered versions of both QA datasets, including only questions where the entities mentioned in the questions and answers occur in our knowledge bases. This approach simulates scenarios where the model's knowledge aligns closely with the task requirements. On these filtered sets, KERLQA demonstrated higher accuracy and lower abstention rates compared to the full datasets. On the filtered English CommonsenseQA, KERLQA (PPO) achieved an accuracy of 79.84% (an increase of 2% from the full dataset) with an abstention rate of 11.74% (a decrease of 1.42% from the full dataset). Similar improvements were observed for OpenBookQA.

5 Error Analysis

To gain deeper insights into KERLQA's performance and limitations, we conducted an error analysis on a randomly selected subset of 100 questions per language-dataset pair, for a total of 1000 analyzed questions. We used KERLQA (PPO) for this analysis, as it showed the best overall performance. Our analysis focused on three main error types: (1) Knowledge Gap: instances where the model lacked the necessary background knowledge to answer correctly; (2) Reasoning Failures: Cases where the model failed to make correct logical inferences; (3) Abstention Errors: Instances where the model incorrectly chose to abstain or failed to abstain when it should have.

5.1 Error Analysis Process

Figure 2 illustrates our error analysis process, showing how KERLQA processes a sample question and where different types of errors can occur. This example demonstrates how Knowledge Gap, Reasoning Failure, and Abstention Error can occur in a single question-answering scenario. It highlights the challenges KERLQA faces in balancing the use of retrieved knowledge, inference capabilities, and decision-making about when to answer or abstain.

5.2 Overall Error Distribution

Table 4 presents an overview of the error distributions across languages and datasets for KERLQA. We observe that the knowledge gap is higher in low-resource languages across both datasets, and higher for CommonsenseQA (CS) compared to OpenBookQA (OB) for all languages. Reasoning failures are generally higher in OpenBookQA compared to CommonsenseQA, and decrease as we move from high-resource (English) to low-resource languages. Abstention errors increase in frequency from high-resource to low-resource languages, and are generally higher in CommonsenseQA compared to OpenBookQA for low-resource languages.

5.3 Language-Specific Error Patterns

Our error analysis, presented in Table 4, reveals distinct error patterns across languages. As the highest-resource language in our study, English has the lowest knowledge gap error rate. Reasoning failures are most prevalent, suggesting that when knowledge is available, the challenge shifts to correct reasoning. isiZulu and isiXhosa have intermediate resource availability; here the Knowledge Gap is significantly higher than English across both datasets. Reasoning Failures for these languages are lower, but still substantial. As the lowest-resource languages in our study, Sepedi and SeSotho exhibits the highest rate of knowledge gap errors across both datasets. However, these languages had the lowest rate of reasoning failures, possibly due to increased abstention in uncertain cases rather than improved reasoning capabilities.

Several trends emerge across the language spectrum. There is an inverse relationship between Knowledge Gap and Reasoning Failures: as we move from high-resource to low-resource languages, Knowledge Gap errors increase while Rea-

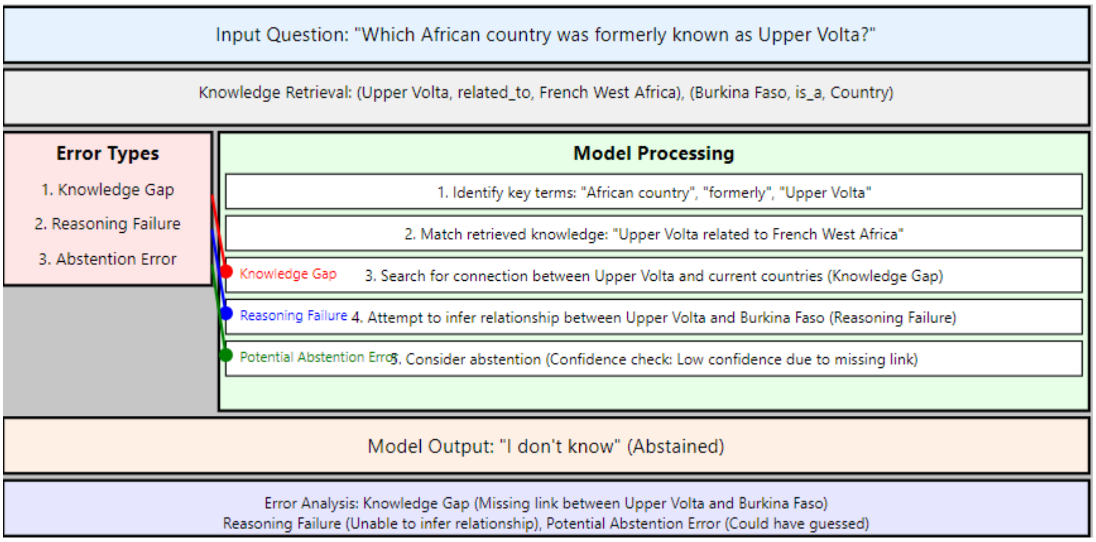
Input Question: "Which African country was formerly known as Upper Volta?"
Knowledge Retrieval: (Upper Volta, related_to, French West Africa), (Burkina Faso, is_a, Country)
Error Types
1. Knowledge Gap
2. Reasoning Failure
3. Abstention Error
Model Processing
1. Identify key terms: "African country", "formerly", "Upper Volta"
2. Match retrieved knowledge: "Upper Volta related to French West Africa"
Knowledge Gap
3. Search for connection between Upper Volta and current countries (Knowledge Gap)
Reasoning Failure
4. Attempt to infer relationship between Upper Volta and Burkina Faso (Reasoning Failure)
Potential Abstention Error
5. Consider abstention (Confidence check: Low confidence due to missing link)
Model Output: "I don't know" (Abstained)
Error Analysis: Knowledge Gap (Missing link between Upper Volta and Burkina Faso)
Reasoning Failure (Unable to infer relationship), Potential Abstention Error (Could have guessed)

Figure 2: Error analysis illustrating the process of how KERLQA handles the question “Which African country was formerly known as Upper Volta?”

soning Failures decrease. Abstention Errors increase as language resources decrease, reflecting KERLQA’s growing uncertainty in low-resource settings. CommonsenseQA show higher rates of Knowledge Gap errors compared to OpenBookQA across all languages.

5.4 Abstention Analysis

Our examination of KERLQA’s abstention behaviour reveals important insights into the model’s decision-making under uncertainty. False positives, where the model abstained but the correct answer had the highest probability among non-abstention options, occurred in 6% of questions for English and 9–12% for low-resource languages. False negatives, where the model attempted an incorrect answer when abstention had a higher probability, happened in 4% of questions for English and 7–9% for low-resource languages.

Abstention rates increase as language resources decrease. In low-resource languages, abstention errors were more often false positives, suggesting a tendency towards over-caution. This behavior aligns with our goal of reducing misinformation in scenarios where knowledge is scarce. We observe inconsistencies in 5% of similar question pairs across languages, where KERLQA abstained in one language but attempted to answer in another. This suggests that KERLQA’s abstention mechanism is sensitive to subtle linguistic differences, which could be beneficial for capturing language-specific nuances and may indicate areas for improvement in cross-lingual consistency.

6 Conclusion

We introduced Knowledge-Enhanced Reinforcement Learning for Question Answering (KERLQA), a novel approach designed to improve question answering performance in low-resource languages. By integrating external knowledge sources with reinforcement learning techniques, KERLQA demonstrates advancements in addressing the challenges posed by limited language resources. Results on English and four low-resource South African languages show that KERLQA outperforms existing baseline models and state-of-the-art KG-augmented QA systems across all languages, with particularly notable improvements in low-resource settings. The incorporation of reinforcement learning enables making more informed decisions about knowledge utilization and abstention. KERLQA contributes to ongoing efforts to bridge the gap between high-resource and low-resource language capabilities in question answering tasks.

Limitations

While KERLQA demonstrates promising results for question answering in low-resource languages, there are some limitations. The model’s reliance on projected knowledge bases from English to low-resource languages introduces potential errors in the knowledge representation. Limited coverage in the knowledge bases will also directly influence the model’s performance. In order to evaluate on some of the languages we relied on the machine translation systems for the translations of CommonsenseQA and OpenbookQA. As such, the accuracy of the translations potentially had an impact on our reported results.

References

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Y. Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *CoRR*, abs/2308.11596.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Raghavi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. *CoRR*, abs/2407.12043.

Andong Chen, Yuan Sun, Xiaobing Zhao, Rosella P. Galindo Esparza, Kehai Chen, Yang Xiang, Tiejun Zhao, and Min Zhang. 2023. Improving low-resource question answering by augmenting question information. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10413–10420. Association for Computational Linguistics.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can AI assistants know what they don't know? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shengyi Huang, Tianlin Liu, and Leandro von Werra. 2023. The n implementation details of rlhf with ppo. *Hugging Face Blog*. https://huggingface.co/blog/the_n_implementation_details_of_rlhf_with_ppo.

Jinhao Jiang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. $great truths are always simple: $ A rather simple knowledge encoder for enhancing the commonsense reasoning capacity of pre-trained models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1730–1741. Association for Computational Linguistics.

Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, Jen-tse Huang, and Shuming Shi. 2022. Tencent's multilingual machine translation system for WMT22 large-scale african languages. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1049–1056. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *CoRR*, abs/2401.15449.

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

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1813–1817. European Language Resources Association (ELRA).

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. Africlirmatrix: Enabling cross-lingual information retrieval for african languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8721–8728. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jinyoung Park, Hyeong Kyu Choi, Juyeon Ko, Hyeon-Jin Park, Ji-Hoon Kim, Jisu Jeong, Kyung-Min Kim, and Hyunwoo J. Kim. 2023. Relation-aware language-graph transformer for question answering. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intel-*

ligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 13457–13464. AAAI Press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Sello Ralethe and Jan Buys. 2025. Cross-lingual knowledge projection and knowledge enhancement for zero-shot question answering in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10111–10124. Association for Computational Linguistics.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2023. Can llms augment low-resource reading comprehension datasets? opportunities and challenges. *CoRR*, abs/2309.12426.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. 2022. GNN is a counter? revisiting GNN for question answering. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *CoRR*, abs/2312.07000.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics.

Qichen Ye, Bowen Cao, Nuo Chen, Weiyuan Xu, and Yuexian Zou. 2023. Fits: Fine-grained two-stage training for knowledge-aware question answering. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13914–13922. AAAI Press.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ruilin Zhao, Feng Zhao, Liang Hu, and Guandong Xu. 2024. Graph reasoning transformers for knowledge-aware question answering. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19652–19660. AAAI Press.

A KERLQA: End-to-End Process

Here we give an end-to-end description of KERLQA, illustrating the flow of information using an example question.

Let's consider an example question in isiZulu:

> q: "Iyiphi indlela yokuhamba ebaluleke kakhulu eNingizimu Afrika?"
> (English: "What is the most important mode of transportation in South Africa?")

A = {a_1: "Izimoto", a_2 : "Izitimela", a_3: "Izindiza", a_4: "Amabhasi", a_5: "Angazi"}

(English: Cars, Trains, Airplanes, Buses, I don't know)

1. **Input Processing:** The question q and answer set A are tokenized and encoded using mT5's tokenizer.
2. **Knowledge Retrieval:** KERLQA queries external knowledge bases (e.g., ConceptNet, DBpedia) to retrieve relevant knowledge triplets. For example:
   - k_1: (South Africa, has_transportation, cars)
   - k_2: (South Africa, has_transportation, trains)
   - k_3: (cars, used_for, commuting)
3. **Joint Graph Construction:** KERLQA con-

Language	Dataset	Manual Translation		Machine Translation	
		Accuracy	Abstention	Accuracy	Abstention
isiZulu	CommonsenseQA	63.67	25.81	60.17	27.03
	OpenBookQA	64.32	24.54	61.56	26.87
Sepedi	CommonsenseQA	62.13	27.31	59.23	28.08
	OpenBookQA	63.52	26.69	60.19	27.11

Table 5: Comparison of KERLQA (PPO) performance on manually translated and machine-translated test sets

structs a working graph $G_W = (V_W, E_W)$ as follows:

- **Nodes (V_W):**
  - v_q: Derived from encoding the question.
  - v_{a_i}: Derived from encoding each answer option.
  - v_{k_j}: Constructed from the retrieved knowledge triples (e.g., k_1, k_2, k_3).
  - z: A dedicated context node that aggregates global information.
- **Edges (E_W):**
  - Edges are added between nodes that are directly related by a knowledge triple (e.g., an edge between v_{k_1} and v_{a_1}).
  - Additional edges are inserted between nodes representing the same entity (e.g., between v_q and a relevant v_{a_i}).
  - The context node z is connected to all other nodes to facilitate global information propagation.

4. **Node Relevance Scoring:** For each node v in a subset V_{sub} (e.g., relevant to the question), KERLQA computes a relevance score:

$$\rho_v = f_{head}\Big(f_{enc}\big([\,\text{text}(z);\ \text{text}(v)]\big)\Big),$$

where z is the QA context node.

5. **Graph Neural Network Processing:** The graph is processed through L layers of message passing using a GNN architecture inspired by QA-GNN. In our implementation, we use a 3-layer GNN where each node's representation is updated as:

$$h_v^{(\ell+1)} = \text{GRU}\Big(h_v^{\ell},\ \text{AGG}\Big(\\ \big\{\text{ReLU}\big(W_r^{\ell}\, h_u^{\ell} + b_r^{\ell}\big) : u \in \mathcal{N}(v)\big\}\Big)\Big),$$

with AGG being an aggregation function (e.g., mean pooling), and W_r^{ℓ}, b_r^{ℓ} learnable parameters.

6. **Answer Scoring:** KERLQA computes a score for each answer option:

$$\text{score}(a_i) = \text{MLP}\Big([\,h_q;\ h_{a_i};\ h_{KG}\,]\Big),$$

where h_q and h_{a_i} are the final representations of the question and answer a_i, and h_{KG} is the aggregated representation of the knowledge graph nodes.

7. **Policy Decision:** The RL policy $\pi_\theta(a|s)$ determines the probability of selecting each answer:

$$\pi_\theta(a|s) = \text{softmax}(W[h_q; h_a; h_{KG}] + b)$$

8. **Action Selection:** An answer is selected based on the policy probabilities. Let's say the model chooses a_2: "Izitimela" (Trains).

9. **Reward Calculation:** Assuming a_2 is the correct answer, the reward is calculated:

$$r = \alpha \cdot \nVdash[\text{correct}] + \gamma_1 \cdot \nVdash[\text{used KG and needed}]$$

10. **Learning Update:**
    - For PPO, the objective function is optimized:

$$L_{PPO}(\theta) = \mathbb{E}\Big[\min\Big(r_t(\theta)A_t,\ \text{clip}\big(r_t(\theta),\\ 1-\epsilon, 1+\epsilon\big)A_t\Big)\Big].$$

    - For DPO, the loss function is:

$$L_{DPO}(\theta) = -\mathbb{E}\Big[\log\Big(\sigma\big(r_\theta(x, a_{\text{chosen}})\\ - r_\theta(x, a_{\text{rejected}})\big)\Big)\Big],$$

where the chosen action is either the correct answer or "I don't know" (if the baseline failed), and a_{rejected} represents other answer options. Although the DPO formulation does not explicitly include a term for KG usage, the preference pairs are derived from a baseline that integrates KG information, thereby indirectly incorporating KG effects.

11. **Model Update:** The model parameters θ are updated based on the gradient of the loss function:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_\theta L(\theta),$$

where η is the learning rate.

To demonstrate that KERLQA works, we can analyze its expected behavior over many training iterations:

- The expected reward $\mathbb{E}[r]$ will increase as the model learns to balance answering, abstaining, and utilizing external knowledge
- As training progresses, we expect:
  - $P(\text{correct})$ to increase
  - $P(\text{idk}|\text{unanswerable})$ to increase
  - $P(\text{idk}|\text{answerable})$ to decrease
  - $P(\text{KG used}|\text{needed})$ to increase
  - $P(\text{KG used}|\text{not needed})$ to decrease
- This will lead to an overall increase in $\mathbb{E}[r]$, demonstrating that KERLQA is effectively learning to answer questions, abstain when appropriate, and utilize external knowledge efficiently.

Figure 1 illustrate the key components and flow of the KERLQA system. The diagram provides a visual overview of how KERLQA integrates question answering, knowledge enhancement, and reinforcement learning to improve performance on low-resource language tasks:

1. Input: The question and answer options (including “I don’t know”) are provided as input
2. mT5 Encoder: The input is encoded using the mT5 language model.
3. Knowledge Retrieval: Relevant knowledge is retrieved from external sources.
4. Joint Graph Construction: A graph is constructed using the input, mT5 encoding, and retrieved knowledge
5. Node Relevance Scoring: The relevance of each node is scored using mT5.
6. GNN Processing: The graph is processed using Graph Neural Networks.
7. RL Module: Either Proximal Policy Optimization (PPO) or Direct Preference Optimization (DPO) is applied
8. Output: The final answer or “I don’t know” is produced.

B Hyperparameter Tuning

B.1 Reward Function Parameters

The PPO reward function in KERLQA (6) combines multiple indicator functions, each with their own hyperparameter. We conducted extensive grid search over these parameters using the English CommonsenseQA validation set. The search ranges were:

- $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$: Weight for correct answers
- $\beta_1 \in \{0.3, 0.5, 0.7, 1.0\}$: Weight for appropriate abstention
- $\beta_2 \in \{0.3, 0.5, 0.7, 1.0\}$: Penalty for unnecessary abstention
- $\gamma_1 \in \{0.2, 0.4, 0.6, 0.8\}$: Weight for appropriate KB use
- $\gamma_2 \in \{0.2, 0.4, 0.6, 0.8\}$: Penalty for unnecessary KB use

C Impact of Translation Quality on Performance

While the main results reported in Table 1 and Table 2 for isiZulu and Sepedi are based on manually translated test sets, we also conducted experiments using machine-translated versions to assess the impact of translation quality on KERLQA’s performance. The results in Table 5 demonstrate a consistent pattern of higher performance for manually translated test sets compared to machine-translated ones across both languages and datasets. These findings underscore that manually curated datasets are important for accurately assessing model capabilities in low-resource languages. However, when evaluating all models on the automatically translated datasets for isiZulu and Sepedi, the same relative trends in model performance still holds.