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

Sello Ralethe and **Jan Buys**
Department of Computer Science, University of Cape Town, South Africa
rltsel002@myuct.ac.za, jbuys@cs.uct.ac.za

Abstract

Question answering in low-resource languages faces critical challenges when models encounter questions beyond their knowledge boundaries, often producing confident but incorrect answers. We propose Knowledge-Enhanced Reinforcement Learning for Question Answering (KERLQA), a novel approach that combines knowledge graph integration with reinforcement learning to enable principled abstention decisions. Unlike existing refusal-tuned methods that make binary decisions based solely on internal confidence, KERLQA implements a three-way decision process: answer with internal knowledge, answer with external knowledge assistance, or abstain. Using a composite reward function that jointly optimizes for correctness, appropriate abstention, and efficient knowledge utilization, we train policies via PPO and DPO with dynamic calibration for low-resource settings. Experiments on CommonsenseQA and OpenBookQA across English and four South African languages show KERLQA achieves improved F1 scores, with up to 6.2 point improvements in low-resource languages. Our analysis reveals that KERLQA reduces false positive abstention rates by 30% while expanding the boundary of answerable questions through external knowledge integration.

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., 2024; Chen et al., 2023). These challenges are 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 would be preferable, particularly in low-resource settings where knowledge gaps are more prevalent.

Recent advances in knowledge-enhanced question answering have shown promise for low-resource languages through external knowledge integration (Yasunaga et al., 2021; Zhang et al., 2022). However, these approaches lack mechanisms for handling uncertainty and determining when to abstain from answering, leading to potential hallucinations when models encounter questions beyond their knowledge boundaries.

In this paper, we introduce Knowledge-Enhanced Reinforcement Learning for Question Answering (KERLQA), which combines graph neural network-based knowledge integration with reinforcement learning to enable abstention decisions. Our approach uses a GNN architecture for joint reasoning over question context and relevant knowledge sources, augmented with reinforcement learning techniques, specifically Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023), to optimize decision-making behavior.

A key innovation in KERLQA is how it fundamentally reformulates the abstention problem compared to existing refusal-tuned approaches. While previous methods like honesty-SFT (Yang et al., 2024) and preference-based tuning (Cheng et al., 2024) frame abstention as a binary decision based solely on internal model confidence, KERLQA introduces a three-way decision process: answer using internal knowledge, answer with knowledge graph assistance, or abstain entirely. This approach addresses an important limitation of existing methods, which is their inability to expand their knowledge boundaries beyond what was learned during pre-training. By dynamically integrating external knowledge through reinforcement learning, KERLQA can both sharpen the boundary of what it knows and expand that boundary when reliable external information is available.

Our approach is particularly significant for lan-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1834–1846

guages with diverse morphological structures, as it enables effective knowledge transfer while respecting language-specific reasoning patterns. By integrating reinforcement learning with knowledge enhancement, KERLQA learns to calibrate its confidence across languages, addressing the problem of overconfident hallucinations in low-resource settings through a principled abstention mechanism.

We demonstrate KERLQA's effectiveness on English and four low-resource South African languages (isiZulu, isiXhosa, Sepedi, and SeSotho), using the multilingual mT5 language model. Our experimental results show that KERLQA consistently outperforms both knowledge augmentation without abstention and existing approaches to QA with an abstention mechanism, with improvements of up to 5.3 percentage points in low-resource settings. The statistical significance of these improvements ($p < 0.01$) confirms the robustness of our approach.

Our main contributions are: (1) We introduce a novel three-way decision process for question answering that fundamentally differs from existing binary abstention approaches by incorporating external knowledge in the decision; (2) We develop a composite reward function and dynamic calibration mechanism that addresses abstention bias in low-resource languages; (3) We provide comprehensive evaluation using precision, recall, and F1 metrics across five languages, demonstrating significant improvements over state-of-the-art methods; (4) We provide detailed error analysis revealing distinct patterns across languages, with insights into knowledge gaps and abstention behaviors in low-resource settings.

2 Related Work

Our work intersects with knowledge-enhanced question answering and reinforcement learning for NLP tasks, particularly in low-resource language 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, while Zhang et al. (2022) proposed GreaseLM to enhance language models with graph-based reasoning. Other approaches include GSC (Wang et al., 2022), QAT (Park et al., 2023), and FiTs (Ye et al., 2023), which address challenges in fusing representations from pre-trained language models and knowledge graphs.

Jiang et al. (2022) revealed that relation features from commonsense knowledge graphs are the primary contributors to improving reasoning capacity in pre-trained language models. While these methods show promising results, they face challenges in fully utilizing external knowledge graphs and addressing the modality gap between text and knowledge graphs. Our work builds upon these insights but uniquely combines knowledge enhancement with reinforcement learning, particularly focusing on low-resource languages where knowledge projection requires careful handling. Importantly, our two-stage pruning mechanism and adaptive knowledge utilization (learned via reinforcement learning) address the robustness challenges that arise when knowledge sources are imperfect or incomplete.

2.2 Reinforcement Learning in NLP

Reinforcement Learning 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). Two key approaches in this area are Proximal Policy Optimization (PPO) (Schulman et al., 2017), which optimizes a surrogate objective function that prevents large policy updates, and Direct Preference Optimization (DPO) (Rafailov et al., 2023), which learns directly from pairwise preference comparisons.

In the context of abstention mechanisms, several approaches have been developed: Yang et al. (2023) used fine-tuning with "I don't know" responses, Zhang et al. (2024) proposed R-Tuning with rejection sampling for edge cases, Cheng et al. (2024) used DPO to learn abstention behavior implicitly, while Liang et al. (2024) used PPO with a hallucination-focused reward model to determine knowledge boundaries.

Our work differs from these approaches in two fundamental ways: (1) while prior methods rely solely on the model's internal knowledge, KERLQA dynamically integrates external knowledge through graph reasoning; and (2) KERLQA's abstention mechanism is trained to consider both answer correctness and knowledge source utilization,

learning when to rely on internal knowledge, when to leverage external knowledge, and when neither is sufficient.

Recent research has shown that models often exhibit biased abstention patterns in low-resource settings (Brahman et al., 2024), typically over-refusing or under-refusing based on surface features rather than true knowledge boundaries. Our work addresses these challenges through a novel approach that dynamically adapts abstention thresholds based on language-specific features and knowledge availability, producing more calibrated refusal behavior across diverse linguistic contexts. Importantly, our framework is designed to be robust to variations in knowledge source quality: when external knowledge is noisy or incomplete, the RL component learns to reduce reliance on external sources and increase appropriate abstention, making the approach practical for real-world deployments where perfect knowledge bases are unavailable.

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. The state s is defined as a combination of the textual representation 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 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 considering 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

KERLQA combines GNN-based knowledge integration with reinforcement learning for abstention decisions. Figure 1 gives a schematic overview of the architecture. The architecture consists of the following steps:

Text Encoding We encode the question q and answer candidates $\{a_i\}$ using the pre-trained mT5 encoder, producing dense representations h_q and $\{h_{a_i}\}$:

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

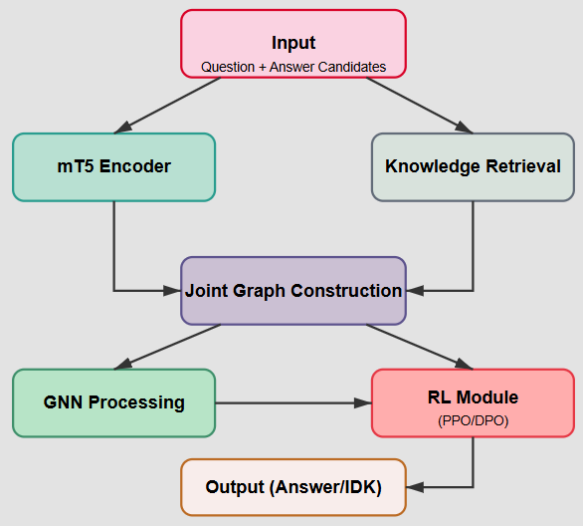

Figure 1: KERLQA architecture. The model processes inputs through both language model encoding and knowledge retrieval for graph construction, integrated through a graph neural network architecture. Reinforcement learning is used to adapt the decision making behavior.

Knowledge Retrieval We query the knowledge base for triples where the subject or object matches entities in q or $\{a_i\}$.

Joint Knowledge Graph Construction We build a heterogeneous graph (Yasunaga et al., 2021) combining text nodes, knowledge nodes, and a global context node. The graph G_W consists of nodes V_W and edges E_W, where:

$$V_W = V_{\text{text}} \cup V_{\text{knowledge}} \cup \{z\}, \quad (2)$$

with text nodes V_{text} representing question and answer embeddings, knowledge nodes $V_{\text{knowledge}}$ constructed from retrieved knowledge triples, and a context node z for global information aggregation. Edges E_W are added between nodes directly related by a knowledge triple, between nodes representing the same entity, and between z and all other nodes.

For low-resource languages, we propose a two-stage knowledge pruning mechanism (see Appendix E for details):

- **Static pruning**: Filter triples based on cross-lingual alignment confidence $\tau_{\mathcal{L}} = 0.35 + 0.4 \cdot r_{\mathcal{L}}$.
- **Dynamic pruning**: Select top-k relevant triples based on semantic similarity to the question.

GNN Processing Information propagates through 3 GNN message passing layers, refining node representations by incorporating neighborhood information. We follow the QA-GNN architecture (Yasunaga et al., 2021):

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

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

RL Policy Network The key innovation in KERLQA is the RL-trained policy network π_θ that integrates the aggregated GNN encoding h_{KG} and outputs a probability distribution over the extended action space $A' = \{a_1, ..., a_n, \text{“I don't know”}\}$.

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

Output Action Selection The answer to question q is predicted by selecting $\arg\max_{a \in A'} \pi_\theta(a|s)$.

Detailed model hyperparamters are given in Appendix B.

3.3 RL Decision Process

We use RL for *classification* rather than generation as in LLM post-training. Unlike existing refusal-tuned approaches that make binary decisions (answer or abstain) based solely on internal confidence, KERLQA implements a three-way decision process:

- **Direct answering**, when h_q and h_a provide sufficient signal (high internal confidence).
- **Knowledge-enhanced answering**, when external knowledge from z increases confidence above threshold.
- **Informed abstention**, when neither internal nor external knowledge is sufficient.

This framing allows us to directly optimize the abstention-accuracy trade-off through a composite reward function while maintaining computational efficiency.

Reward Design The terms in our reward function correspond to the three-way decision process:

$$\begin{aligned} r(s,a) = &\alpha \cdot \mathbb{1}[\text{correct answer}] \\ &+ \beta_1 \cdot \mathbb{1}[\text{abstain when unanswerable}] \\ &- \beta_2 \cdot \mathbb{1}[\text{abstain when answerable}] \\ &+ \gamma_1 \cdot \mathbb{1}[\text{used KG appropriately}] \\ &- \gamma_2 \cdot \mathbb{1}[\text{used KG unnecessarily}] \end{aligned} \quad (5)$$

The reward function explicitly rewards appropriate knowledge utilization (γ terms) alongside correct answering and calibrated abstention. The hyperparameters were set through a grid search on the English validation set (see Appendix B for details). We set $\alpha = 1.5$, $\beta_1 = 0.7$, $\beta_2 = 0.5$, $\gamma_1 = 0.6$, and $\gamma_2 = 0.4$.

Language-Adaptive Calibration We observe an abstention bias in low-resource languages. To address this, we adjust β_2 based on language resource level $r_\mathcal{L}$:

$$\beta_{2,\mathcal{L}} = \beta_2 \cdot (1 + 0.5 \cdot (1 - r_\mathcal{L})) \quad (6)$$

This increases the penalty for unnecessary abstention as resources decrease, counteracting overconservative behavior. Details are given in Appendix D.

3.4 Reinforcement Learning Training

We implement two RL algorithms. PPO maximizes the clipped objective:

$$L_{PPO}(\theta) = \mathbb{E}[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)], \quad (7)$$

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), with $\epsilon = 0.2$ and a learning rate 5×10^{-5}.

For the DPO objective we construct preference pairs where correct answers or appropriate abstentions are preferred. 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:

$$L_{DPO}(\theta) = -\mathbb{E}[\log(\sigma(r_\theta(x, y^+) - r_\theta(x, y^-)))], \quad (8)$$

where y^+ is the preferred action and y^- is a rejected option.

An end-to-end example of the model training process is given in Appendix A.

Method	English			isiZulu			isiXhosa			Sepedi			SeSotho		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
mT5	67.1	67.1	67.1	57.1	57.1	57.1	55.8	55.8	55.8	56.2	56.2	56.2	55.1	55.1	55.1
mT5+QA-GNN	70.3	70.3	70.3	61.9	61.9	61.9	58.5	58.5	58.5	60.0	60.0	60.0	57.9	57.9	57.9
Honesty-SFT	58.2	82.1	68.1	47.3	73.2	57.5	45.1	70.8	55.1	46.8	71.5	56.6	44.9	69.3	54.5
R-Tuning	59.1	80.5	68.2	48.7	71.9	58.1	46.2	69.3	55.4	47.5	70.1	56.6	45.8	68.2	54.8
Idk-DPO	60.3	83.7	70.1	50.2	75.4	60.3	47.8	72.1	57.5	49.1	73.8	59.0	47.3	71.5	56.9
Self-Aware PPO	61.8	85.2	71.6	51.5	77.8	62.0	48.9	74.5	59.1	50.3	76.2	60.6	48.7	73.9	58.7
RLQA (PPO)	59.4	83.9	69.5	48.1	74.3	58.4	45.7	71.2	55.7	47.2	72.8	57.3	45.3	70.4	55.1
RLQA (DPO)	60.7	84.1	70.5	49.8	75.1	59.9	47.1	72.0	57.0	48.5	73.5	58.4	46.9	71.8	56.7
KERLQA (PPO)	**69.3**	**88.7**	**77.8**	**56.2**	**81.9**	**66.7**	**52.3**	**78.5**	**62.8**	**54.8**	**80.3**	**65.1**	**51.8**	**77.2**	**62.0**
KERLQA (DPO)	68.5	87.9	77.0	55.7	81.2	66.1	51.4	77.1	61.7	54.1	79.8	64.5	50.9	76.3	61.1

Table 1: Results on CommonsenseQA across languages. R: Recall, P: Precision, F1: F1-score. Models without abstention capability have R=P=F1. KERLQA achieves the best F1 scores across all languages.

Method	English			isiZulu			isiXhosa			Sepedi			SeSotho		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
mT5	78.2	78.2	78.2	57.8	57.8	57.8	56.9	56.9	56.9	57.3	57.3	57.3	56.3	56.3	56.3
mT5+QA-GNN	83.5	83.5	83.5	63.4	63.4	63.4	61.1	61.1	61.1	61.3	61.3	61.3	58.8	58.8	58.8
Honesty-SFT	69.1	86.4	76.8	49.2	75.6	59.6	47.1	72.9	57.2	48.3	74.1	58.5	46.7	71.8	56.6
R-Tuning	70.3	85.1	77.0	50.8	74.3	60.3	48.4	71.8	57.8	49.5	72.7	58.9	47.9	70.5	57.0
Idk-DPO	71.8	87.3	78.8	52.3	77.1	62.3	49.7	74.2	59.5	50.9	75.3	60.7	49.1	73.4	58.8
Self-Aware PPO	73.2	88.9	80.3	53.7	79.2	64.0	51.0	76.1	61.1	52.1	77.5	62.3	50.4	75.6	60.5
RLQA (PPO)	70.5	87.2	78.0	50.3	76.1	60.6	47.8	73.4	57.9	49.0	74.8	59.2	47.2	72.7	57.2
RLQA (DPO)	71.9	87.8	79.1	51.9	77.0	62.0	49.2	74.1	59.1	50.4	75.6	60.5	48.7	73.9	58.7
KERLQA (PPO)	**81.3**	**91.8**	**86.2**	**58.9**	**83.7**	**69.2**	**54.6**	**80.1**	**64.9**	**56.8**	**82.3**	**67.2**	**53.7**	**79.4**	**64.1**
KERLQA (DPO)	80.4	91.1	85.5	58.3	83.0	68.5	53.8	79.2	64.1	56.1	81.6	66.5	52.9	78.6	63.3

Table 2: Results on OpenBookQA across languages. KERLQA consistently outperforms all baselines across all metrics.

4 Experiments

4.1 Experimental Setup

Our approach is the first to combine knowledge enhancement with a RL-tuned abstention mechanism. Our experiments therefore compared KERLQA against two categories of baselines: (1) knowledge-enhanced QA systems without abstention capabilities, and (2) refusal-tuned approaches without knowledge enhancement. We implement all methods using mT5-large (Xue et al., 2021) as the base language model, ensuring consistency across evaluations. While larger models are available, mT5 models are still competitive and outperform LLMs on many tasks for low-resource African languages (Ojo et al., 2025; Adelani et al., 2025).

We compare the following models:

- *mT5* Base multilingual T5-large model fine-tuned on QA datasets.
- *mT5+QA-GNN*: Knowledge-enhanced model with 3-layer GNN.
- *RLQA*: mT5 with RL-based abstention but *without* knowledge enhancement.
- *KERLQA*: Full model combining GNN-based knowledge integration with RL-based abstention.

We also implement four recent refusal-tuned methods as abstention baselines:

- *Honesty-SFT* (Yang et al., 2024) uses supervised fine-tuning on datasets with "I don't know" responses;
- *R-Tuning* (Zhang et al., 2024) employs rejection sampling to identify edge cases;
- *Idk-DPO* (Cheng et al., 2024) uses preference optimization to learn abstention behavior.
- *Self-Aware PPO* (Liang et al., 2024) uses reinforcement learning with hallucination-focused rewards.

Since the papers of each of these approaches used different benchmarks and evaluation metrics, and our primary goal is to improve performance on low-resource languages, we trained each of the models on our datasets to enable consistent evaluation.

4.2 Datasets and Evaluation

We evaluate on the CommonsenseQA (Talmor et al., 2019) and OpenBookQA (Mihaylov et al., 2018) datasets across English and four low-resource South African languages. We use man-

Method	en	zu	xh	nso	st
mT5	0.0	0.0	0.0	0.0	0.0
mT5+QA-GNN	0.0	0.0	0.0	0.0	0.0
Honesty-SFT	23.5	29.8	32.1	30.9	32.4
R-Tuning	22.8	28.4	30.7	29.6	31.2
Idk-DPO	23.1	28.9	31.3	30.2	31.8
Self-Aware PPO	22.3	28.1	30.5	29.4	31.1
RLQA (PPO)	25.2	31.8	33.9	32.7	34.2
RLQA (DPO)	24.1	30.6	32.8	31.9	33.5
KERLQA (PPO)	19.7	27.3	29.8	28.6	30.4
KERLQA (DPO)	19.4	27.6	30.2	29.1	30.8

Table 3: Abstention rates (%) on CommonsenseQA on English (en), isiZulu (zu), isiXhosa (xh), Sepedi (nso) and SeSotho (st). KERLQA shows lower abstention rates while maintaining higher precision.

ually translated test sets for isiZulu and Sepedi, and machine translated test sets for isiXhosa and SeSotho from Ralethe and Buys (2025). The machine translations were produced using Tencent's Multilingual Machine Translation System (Jiao et al., 2022). The impact of machine translation quality is evaluated in Appendix C.

We utilize ConceptNet (Speer et al., 2016) as our primary knowledge source for knowledge-enhanced models, with knowledge projected from English to the South African languages using LeNS-Align (Ralethe and Buys, 2025). This yields approximately 670k triples per language with human-evaluated accuracy exceeding 85%.

We evaluate the question answering using precision, recall and F1 score (harmonic mean of precisions and recall). In our setup precision is the proportion of attempted questions answered correctly, and recall is the proportion of all questions answered correctly (counting abstentions as incorrect). Additionally we compute the Abstention Rate (AR), which is the proportion of questions where the model abstains.

4.3 Main Results

Tables 1 and 2 present the main results on CommonsenseQA and OpenBookQA across all languages. KERLQA consistently outperforms all baselines. The improvement from mT5 to mT5+QA-GNN confirms the effectiveness of knowledge integration, with average recall gains of 3.2% on CommonsenseQA and 5.3% on OpenBookQA. Comparing RLQA with mT5 shows that RL-based abstention alone provides modest improvements in precision at the cost of recall. However, KERLQA's combination of knowledge enhancement and RL-based abstention achieves the best F1 scores across all languages, demonstrating the synergistic benefit of our approach.

Compared with other recent refusal-tuned approaches, our RLQA models perform better than Honesty-SFT, comparable to R-Tuning and Idk-PPO, and a bit lower than Self-Aware PPO. However by incorporating knowledge enhancement KERLQA performs substantially better than Self-Aware PPO. Additionally, Table 3 shows that KERLQA achieves lower abstention rates than all refusal-tuned baselines while maintaining higher precision. This suggests that external knowledge not only improves answer quality but also increases model confidence in a calibrated manner. The performance gap between English and low-resource languages is smallest for KERLQA (11.1 F1 points) compared to refusal-tuned baselines (average 13.8 F1 points), indicating that knowledge enhancement particularly benefits low-resource settings.

To assess the reliability of our results, we performed statistical significance testing using stratified bootstrap resampling (Berg-Kirkpatrick et al., 2012) with 10,000 iterations, stratifying by question difficulty (based on baseline accuracy). Table 4 shows that all F1 score improvements of KERLQA over the strongest baseline methods (mT5+QA-GNN and Self-Aware PPO) are statistically significant across all languages ($p < 0.05$). The difference between PPO and DPO variants of KERLQA is significant for English ($p = 0.041$) but not for low-resource languages ($p > 0.05$), suggesting both RL approaches are equally effective in resource-constrained settings. Additionally, we computed 95% confidence intervals for F1 scores using the same bootstrap procedure as shown in Table 5.

4.4 Performance on Answerable vs. Unanswerable Questions

In order to analyse to what extend the model is able to learn to leverage the KG knowledge to answer questions related to knowledge in the KG, we partitioned the test sets based on knowledge availability. We categorize questions as KG-Answerable if: (1) At least one entity from the question or correct answer appears in the knowledge graph; and (2) A path of length ≤ 3 exists between question and answer entities. This categorization was validated on 200 manually labeled examples with 91% accuracy.

Table 6 shows that KERLQA significantly outperforms all baselines on KG-Answerable questions, with 11.2% higher recall than Self-Aware

Comparison	en	zu	xh	avg(nso,st)
KERLQA vs mT5+QA-GNN	<0.001**	<0.001**	0.002**	0.003**
KERLQA vs Self-Aware PPO	0.008**	0.014*	0.021*	0.019*
KERLQA(PPO) vs KERLQA(DPO)	0.041*	0.127	0.089	0.156

Table 4: p-values from bootstrap significance tests on CommonsenseQA. *: $p < 0.05$, **: $p < 0.01$. KERLQA significantly outperforms all baselines. PPO vs DPO differences are only significant for English.

Method	English F1	isiZulu F1
mT5+QA-GNN	70.3 [69.8, 70.8]	61.9 [61.2, 62.6]
Self-Aware PPO	71.6 [71.1, 72.1]	62.0 [61.3, 62.7]
KERLQA (PPO)	**77.8** [77.3, 78.3]	**66.7** [66.0, 67.4]

Table 5: F1 scores with 95% confidence intervals on CommonsenseQA. KERLQA's improvements are well outside the confidence intervals of baselines.

	KG-Answerable		KG-Unanswerable	
Method	R	AR	R	AR
mT5+QA-GNN	78.2	0.0	51.3	0.0
Self-Aware PPO	71.5	18.3	43.8	31.2
RLQA (PPO)	70.8	19.7	42.1	35.8
KERLQA (PPO)	**82.7**	12.1	**48.9**	32.5

Table 6: Recall (R) and Abstention Rate (AR) on English CommonsenseQA split by knowledge availability. KG-Answerable is the set of questions where relevant knowledge exists in the KG (ConceptNet). KG-Unanswerable is the set of questions requiring knowledge not in the KG.

PPO. KERLQA shows more conservative abstention on KG-Answerable questions (12.1%) compared to baselines, indicating effective knowledge utilization. On KG-Unanswerable questions, all RL-based methods show similar recall, but KERLQA has lower abstention rates, suggesting better calibration of internal knowledge boundaries.

4.5 Error Analysis

4.5.1 Analysis Methodology

We conducted a comprehensive error analysis on 500 randomly sampled errors from each language-dataset combination (5,000 total). For each error, two annotators independently categorized it into one of three types:

1. **Knowledge Gap**: The required information is not present in either the model's parameters or the knowledge graph
2. **Reasoning Failure**: The information exists but the model fails to make correct inferences
3. **Abstention Error**: The model abstains when it could have answered correctly (false positive) or attempts an incorrect answer when it should abstain (false negative)

Inter-annotator agreement was high (Cohen's κ = 0.83). Disagreements were resolved through discussion. An example of the error analysis process is given in Figure 2.

4.5.2 Error Distribution Analysis

Table 7 presents our the error analysis results. Knowledge gaps increase monotonically with decreasing language resources (22% $\rightarrow$ 34% on CommonsenseQA); Reasoning failures show slight inverse correlation with resources, possibly due to increased abstention filtering out complex cases. False positive abstentions are 2 to 3 times higher in low-resource languages despite dynamic calibration. False negatives remain relatively stable across languages, suggesting consistent confidence calibration for clearly wrong answers.

4.5.3 Qualitative Comparison with Refusal-Tuned Approaches

To understand how KERLQA differs from pure refusal-tuned methods, we manually analyzed 100 questions where KERLQA succeeds but Self-Aware PPO fails, categorizing them as follows:

- **Explicit knowledge retrieval (42%):** Questions requiring specific facts present in ConceptNet but not in model parameters. Example: "What material are mosaics typically made from?" $\rightarrow$ KERLQA retrieves (mosaic, made_of, tile);
- **Confidence transformation (31%):** Cases where Self-Aware PPO correctly identifies uncertainty and abstains, but KERLQA finds supporting evidence to answer correctly; and
- **Multi-hop reasoning (27%):** Questions requiring inference across multiple knowledge triples that neither model's parameters nor single facts can answer.

This analysis confirms that KERLQA's three-way decision process enables qualitatively different behavior compared to binary abstention mechanisms.

	English		isiZulu		isiXhosa		Sepedi		SeSotho	
Error Type	CS	OB	CS	OB	CS	OB	CS	OB	CS	OB
Knowledge Gap	22%	18%	30%	27%	32%	29%	31%	28%	34%	30%
Reasoning Failures	21%	25%	20%	23%	19%	21%	19%	22%	18%	20%
Abstention Errors										
False Positive	5%	6%	11%	9%	13%	10%	12%	10%	14%	11%
False Negative	3%	4%	7%	6%	7%	7%	8%	6%	7%	7%

Table 7: Error distribution across languages and datasets (CS: CommonsenseQA, OB: OpenBookQA). Percentages are relative to total questions. False Positive: unnecessary abstention; False Negative: failed abstention.

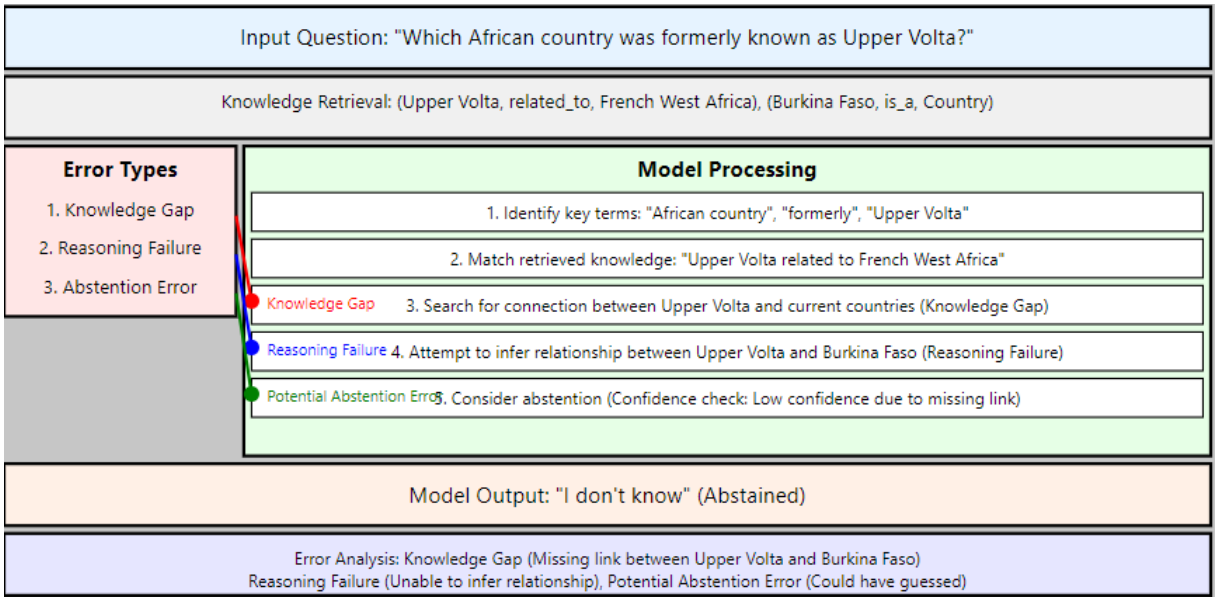

Figure 2: Error analysis illustrating the process of how KERLQA handles the question "Which African country was formerly known as Upper Volta?"

5 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 and implementing language-specific calibration mechanisms, KERLQA demonstrates significant advancements in addressing the challenges posed by limited language resources. Our evaluation across English and four South African languages shows that KERLQA consistently outperforms existing baseline models and state-of-the-art QA systems, with particularly notable improvements in low-resource settings. The incorporation of reinforcement learning enables making more informed decisions about knowledge utilization and abstention, while our dynamic reward calibration mechanism effectively addresses the abstention bias observed in low-resource languages. Our comparison with recent refusal-tuned approaches demonstrates that while teaching models when to abstain is valuable, combining abstention learning with dynamic knowledge integration delivers substantially better performance, establishing a new paradigm for developing trustworthy multilingual question answering systems.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850). Sello Ralethe is supported by the Hasso Plattner Institute for Digital Engineering, through the HPI Research School at the University of Cape Town.

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 translations of CommonsenseQA and OpenbookQA. As such, the accuracy of the translations potentially had an impact on our reported results.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

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 Hanna Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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.

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. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. 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 Intelligence, 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*, pages 10111–10124, Abu Dhabi, UAE. Association for Computational Linguistics.

Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.

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.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

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.

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 constructs 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:

$$r = \alpha \cdot \mathbb{1}[\text{correct}] + \gamma_1 \cdot \mathbb{1}[\text{used KG and needed}]$$

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

$$L_{PPO}(\theta) = \mathbb{E}\Big[\min\Big(r_t(\theta)A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\Big)\Big]. \tag{9}$$

    - For DPO, the loss function is:

$$L_{DPO}(\theta) = -\mathbb{E}\Big[\log\Big(\sigma\big(r_\theta(x, a_{\text{chosen}}) - r_\theta(x, a_{\text{rejected}})\big)\Big)\Big] \tag{10}$$

      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.

B Hyperparameter Tuning

B.1 Reward Function Parameters

The PPO reward function in KERLQA 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

The optimal hyperparameters were $\alpha = 1.5$, $\beta_1 = 0.7$, $\beta_2 = 0.5$, $\gamma_1 = 0.6$, and $\gamma_2 = 0.4$. For low-resource languages, we dynamically adjusted β_2 as described in §3.3.

B.2 Model Hyperparameters

For the GNN component, we used a 3-layer architecture with hidden dimension 256 and dropout rate 0.2. The message passing employed a GraphSAGE aggregation function with mean pooling. For the PPO algorithm, we used a learning rate of 5×10^{-5}, clip parameter $\epsilon = 0.2$, and 4 PPO epochs per batch. For DPO, we used a learning rate of 2×10^{-5} and a reference model KL penalty coefficient of 0.1. All models were trained for 15 epochs with early stopping based on validation performance.

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 comparison revealed notable differences.

The results in Table 8 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.

D Dynamic Reward Calibration Details

To address the observed abstention bias in low-resource languages, we implement a dynamic reward calibration approach that adjusts reward parameters based on language resource levels.

For each language $\mathcal{L}$, we define a resource level $r_{\mathcal{L}} \in [0, 1]$ based on factors such as pre-training data volume, number of speakers, and available linguistic resources. In our implementation, we assign $r_{\text{English}} = 1.0$, $r_{\text{isiZulu}} = 0.4$, $r_{\text{isiXhosa}} = 0.3$, $r_{\text{Sepedi}} = 0.2$, and $r_{\text{SeSotho}} = 0.2$.

We then define language-specific reward parameters as follows:

$$\alpha_{\mathcal{L}} = \alpha \tag{11}$$

$$\beta_{1,\mathcal{L}} = \beta_1 \tag{12}$$

$$\beta_{2,\mathcal{L}} = \beta_2 \cdot (1 + 0.5 \cdot (1 - r_{\mathcal{L}})) \tag{13}$$

$$\gamma_{1,\mathcal{L}} = \gamma_1 \tag{14}$$

$$\gamma_{2,\mathcal{L}} = \gamma_2 \tag{15}$$

This adjustment increases the penalty for unnecessary abstention (β_2) in lower-resource languages, counteracting the model's tendency to be overly conservative. The scaling factor of 0.5 was determined through ablation studies, balancing improved coverage against potential precision losses.

For example, with base $\beta_2 = 0.5$, the language-specific values become:

$$\beta_{2,\text{English}} = 0.5 \tag{16}$$

$$\beta_{2,\text{isiZulu}} = 0.5 \cdot (1 + 0.5 \cdot (1 - 0.4)) = 0.65 \tag{17}$$

$$\beta_{2,\text{isiXhosa}} = 0.5 \cdot (1 + 0.5 \cdot (1 - 0.3)) = 0.675 \tag{18}$$

$$\beta_{2,\text{Sepedi}} = 0.5 \cdot (1 + 0.5 \cdot (1 - 0.2)) = 0.7 \tag{19}$$

$$\beta_{2,\text{SeSotho}} = 0.5 \cdot (1 + 0.5 \cdot (1 - 0.2)) = 0.7 \tag{20}$$

E Knowledge Graph Pruning Mechanism

Our knowledge graph pruning mechanism operates in two stages:

Static pruning: During knowledge base projection, we filter triples based on cross-lingual alignment confidence. For each triple (h, r, t) projected from English to target language $\mathcal{L}$, we compute:

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

$$\begin{aligned}\text{conf}_{\text{static}}(h, r, t) = \text{align}_{\text{conf}}(h) \\ \cdot \text{align}_{\text{conf}}(t) \cdot \text{rel}_{\text{conf}}(r)\end{aligned} \tag{21}$$

where $\text{align}_{\text{conf}}$ represents alignment confidence from LeNS-Align, and rel_{conf} is relation type reliability. We discard triples below a threshold $\tau_{\mathcal{L}} = 0.35 + 0.4 \cdot r_{\mathcal{L}}$, where $r_{\mathcal{L}}$ is the language resource level.

Dynamic pruning: During question answering, we further prune retrieved triples based on relevance to the current question. For each retrieved triple, we compute:

$$\begin{aligned}\text{conf}_{\text{dynamic}}(h, r, t) = \text{sim}(h_q, h_h) \cdot \text{sim}(h_q, h_t) \\ \cdot \text{conf}_{\text{static}}(h, r, t)\end{aligned} \tag{22}$$

where sim is cosine similarity between question embedding h_q and entity embeddings h_h, h_t. We retain only the top-k triples, where k is determined based on question complexity.

This two-stage pruning approach significantly improves knowledge quality, especially for low-resource languages, by removing potentially misleading triples before they can influence the reasoning process.

Ablation Study Results: When disabling the static pruning step, we observed a 2.1% drop in accuracy for English and a 3.8-4.5% drop for low-resource languages. When disabling dynamic pruning, the accuracy decreased by 1.4% for English and 2.3-3.1% for low-resource languages. This confirms the importance of both pruning stages, with an even stronger impact in low-resource settings where noise from knowledge projection is more prevalent.