Exploring Retrieval Augmented Generation Approaches for Natural Language Question Understanding

Christoph Kowalski†, Vincent Emmerling,
Amelie Robrecht-Hilbig, Stefan Kopp

TRR 318 Subproject A01
†christoph.kowalski@uni-bielefeld.de

CONSTRUCTING EXPLAINABILITY
TRR 318

Constructig Explainability

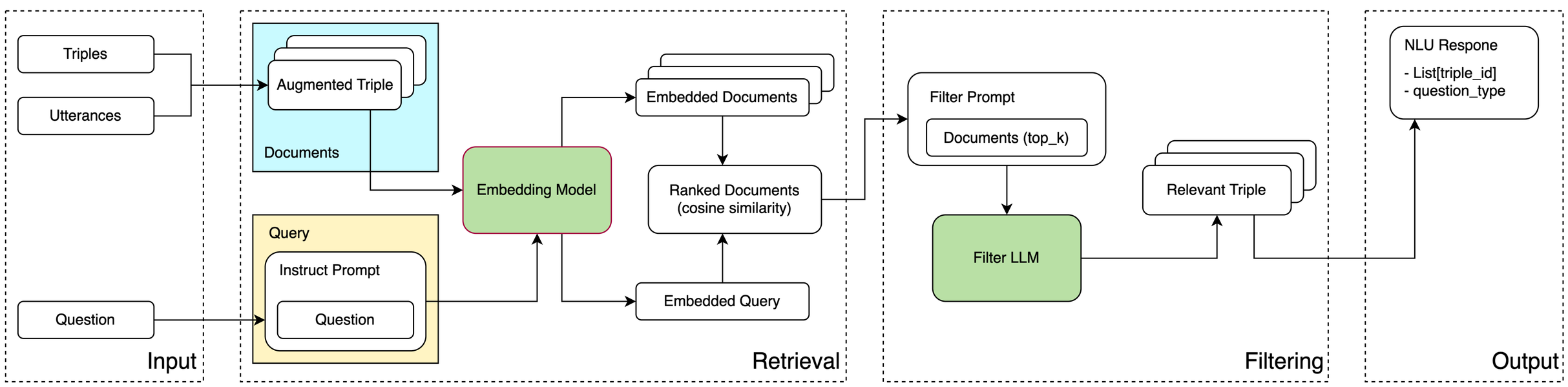
UNIVERSITÄT BIELEFELD
Faculty of Technology

Social Cognitive Systems Group

Research Objective

- Use **RAG** [1] for **Natural Language Question Understanding (NLQU)** in explanation dialogues.
- NLQU identifies the **question type** and the user's **knowledge gap**.
- Map spoken questions to **knowledge graph triples**.
- Interpret questions with **multiple references** or **misconceptions**.
- Evaluate on a **human-human explanation dialogue dataset**.

NLQU architecture



The proposed NLQU architecture based on RAG.

Datasets

- Synthetic dataset for model selection.
- Human-human dataset for testing.
- Questions from explanations of Quarto! board game.
- Spoken-language questions are more fragmented.

Question	Triples
Validation Dataset:	
How many players are there?	(Players have_amount two)
Test dataset:	
Uh, is there somehow a fixed number of these different pieces, like is there only a certain amount?	(Game_pieces have_amount sixteen)

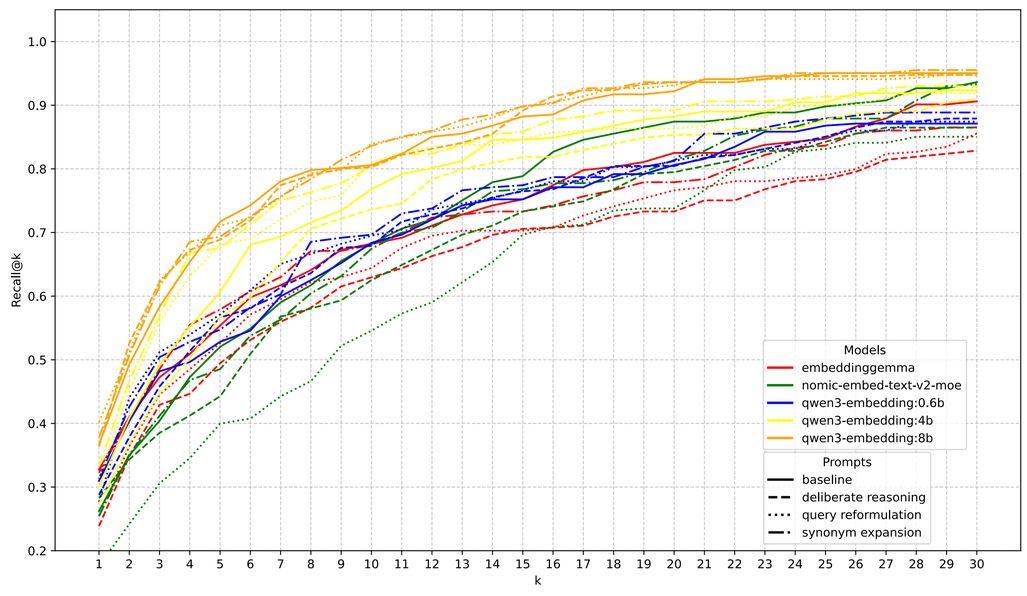
Model Selection

- Compare models for embedding and filtering.
- Evaluate different prompting strategies for Instruct and Filter prompts [2].
- Analyze different top_k filtering values.
- Baseline: LLM directly selects relevant triples from all candidates.

Model	F1	Recall	Precision	Q-Type Acc.
NLQU architecture with $k = 10$ and best embedding model (qwen3-8b):				
gpt-oss-20b	**0.758**	0.827	**0.764**	0.633
gemma-3-27b	0.616	**0.829**	0.518	0.788
mistral-small-3.2	0.689	0.773	0.661	0.587
Baseline:				
gpt-oss-20b	0.641	0.825	0.577	0.663
gemma-3-27b	0.554	0.808	0.447	0.712
mistral-small-3.2	0.497	0.798	0.389	**0.850**

Results

- Considerable drop in performance on test dataset for both architectures.
- Prompt choice matters for retrieval performance.
- Qwen Embedding models perform best



Retrieval performance of different models and prompts on the human-human test dataset

- Proposed NLQU architecture improves triple extraction and question type retrieval.
- Great improvement in question type accuracy
- Only $\approx 1\%$. F1 improvement.
- Lower gain compared to the validation results.

Architecture	F1	Recall	Precision	Q-Type Acc.
NLQU architecture	**0.621**	0.771	**0.568**	**0.705**
Baseline	0.609	**0.8**	0.549	0.571

References & Acknowledgement

[1] Aoran Gan et al. *Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey*. 2025.

[2] Shubham Vatsal and Harsh Dubey. *A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks*. 2024.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/3 2026 – 438445824, project A01.