

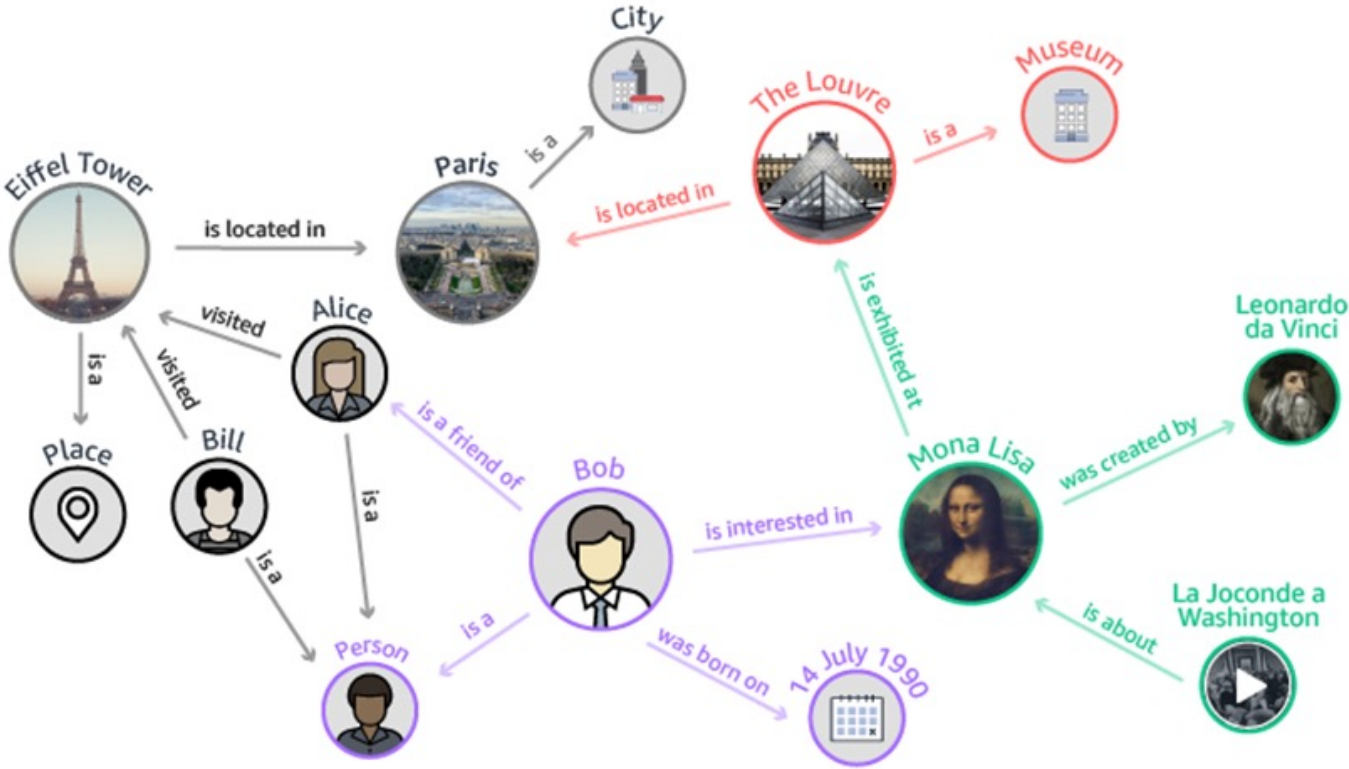
Efficient and Accurate Contextual Re-Ranking for Knowledge Graph Question Answering

LREC-COLING 2024

Kexuan Sun¹ Nicolaas Jedema² Karishma Sharma² Ruben Janssen²
Jay Pujara¹ Pedro Szekely² Alessandro Moschitti²

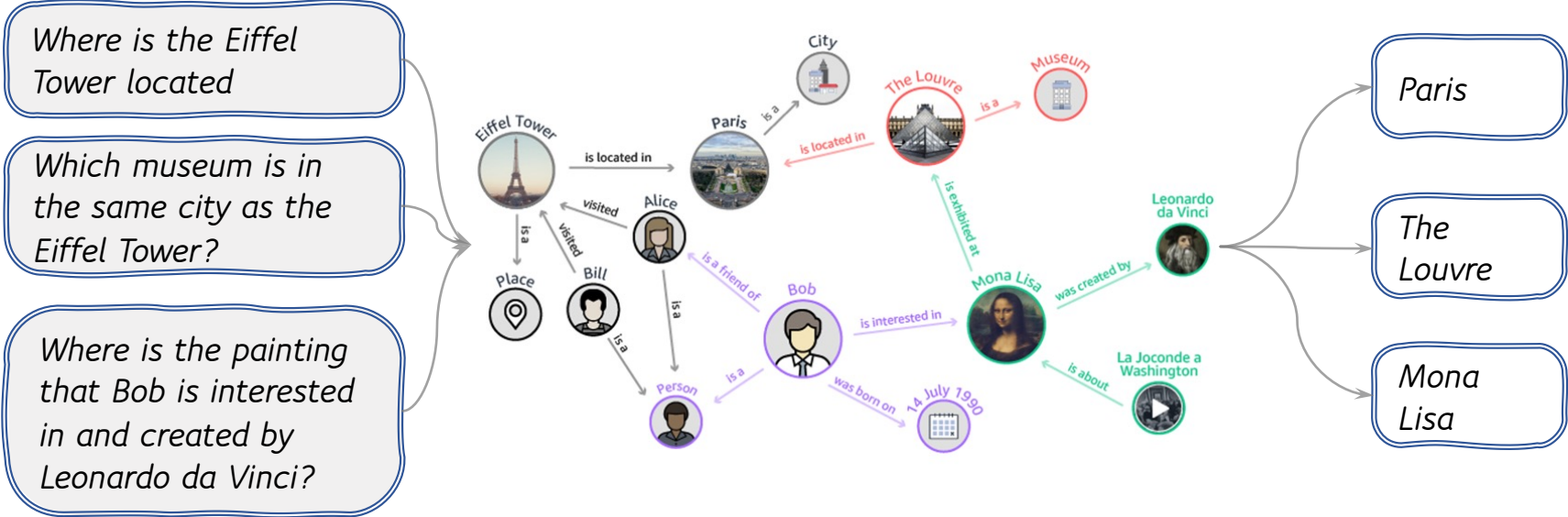
1 University of Southern California
2 Amazon AGI

What is a Knowledge Graph?



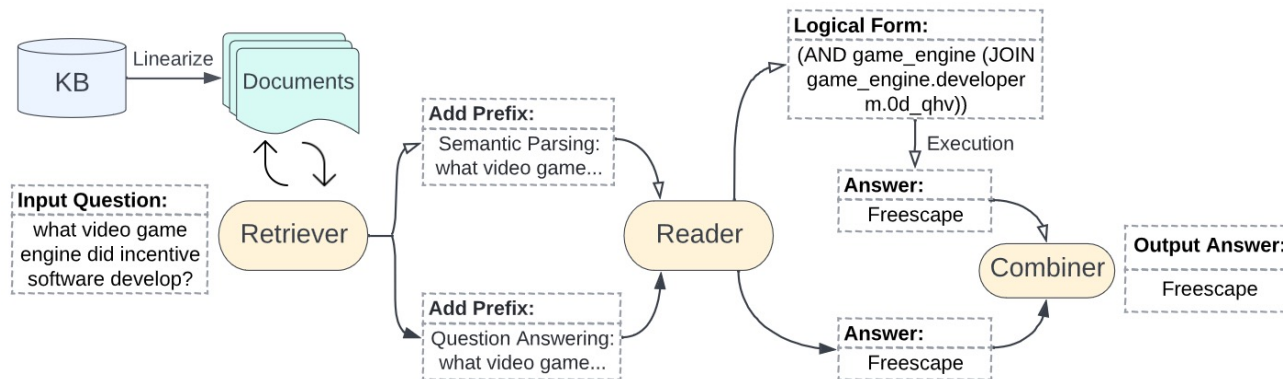
What is Knowledge Graph Question Answering (KGQA)?

Given a question, provide answers using triples from KGs



How to approach this task?

- Semantic-parsing [1,2]
 - Convert natural language questions into executable queries
- Direct answer prediction[3,4]
 - Answer generation without logical forms



[1] Retrack: A flexible and efficient framework for knowledge base question answering. Chen et al. ACL 2021.

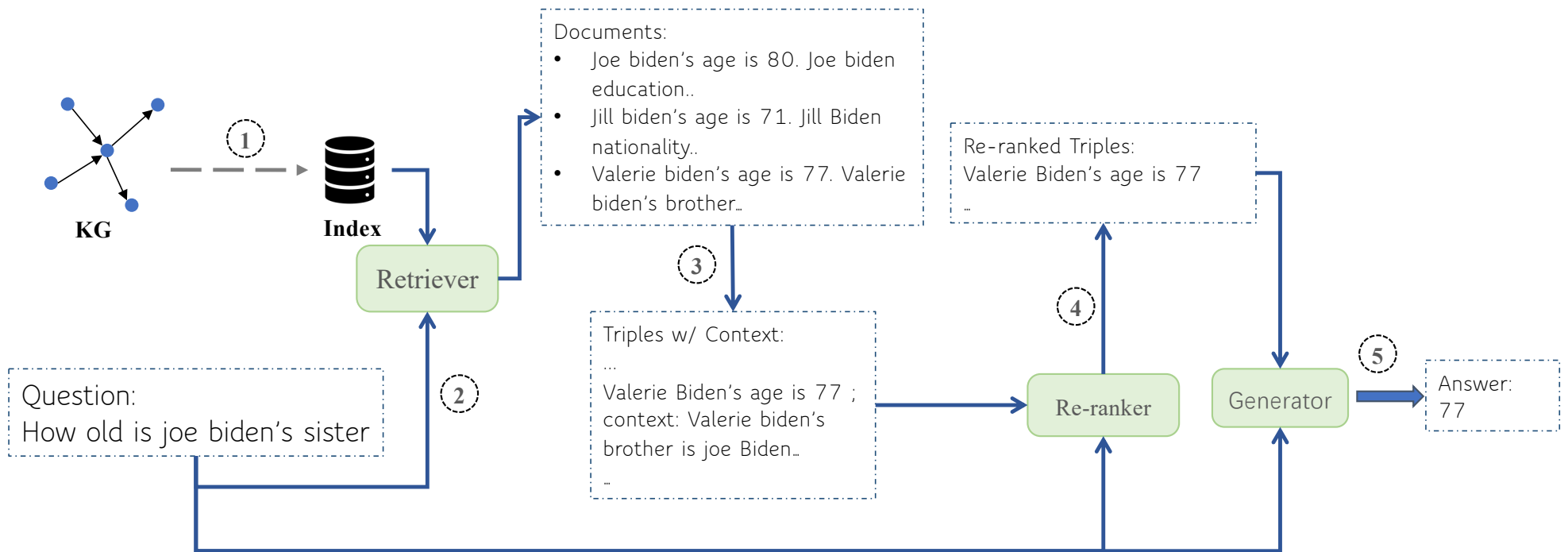
[2] Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. Ye et al. ACL 2022.

[3] Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. Oguz et al. NAACL 2022.

[4] Decaf: Joint Decoding of Answers and Logical Forms for Question Answering over Knowledge Bases. Yu et al, ICLR 2023.

KGQA with Contextual Re-ranker

The key is to identify the most important triples from the whole KG



Context-aware Re-ranker

Extra contextual information that is helpful for better-ranking triples

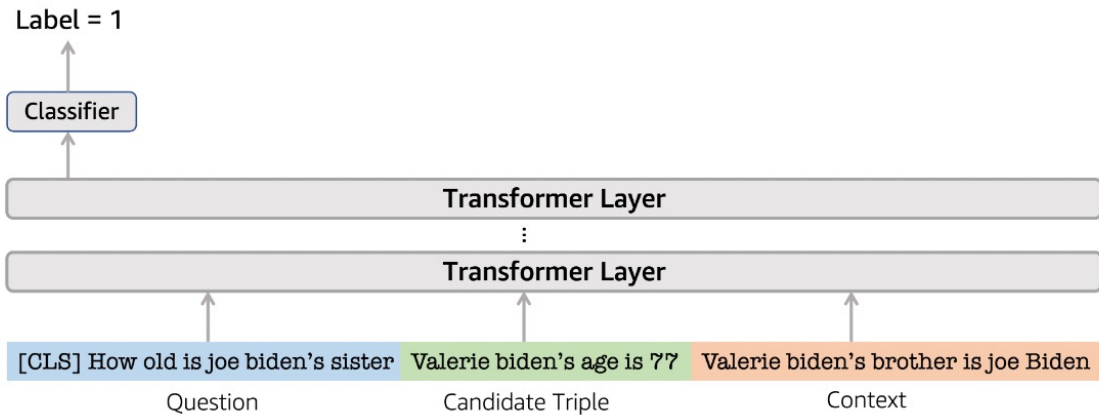
How old is joe biden's sister?



- Triple 1: Joe biden's age is 80
- Triple 2: Hunter biden's age is 52
- Triple 3: Jill Biden's age is 71
- Triple 4: valerie biden's age is 75

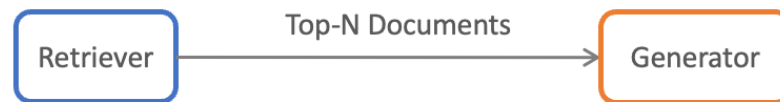


- Triple 4: valerie biden's age is 75.
Context: Valerie Biden's sibling is Joe Biden, Valerie Boden's brother is Joe Biden..
- Triple 1: Joe biden's age is 80.
Context: Joe biden's birthday is Nov 20, 1942

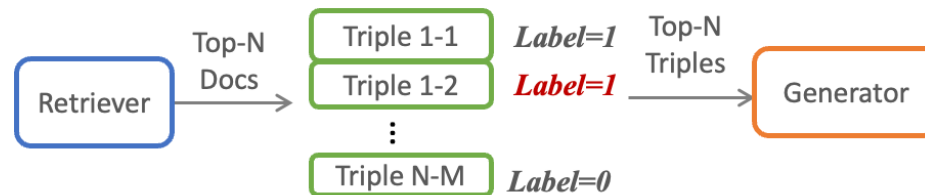


How to train the re-ranker?

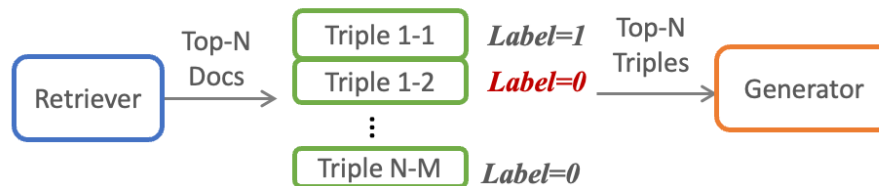
Fine-grained labeling strategies could more accurately find key information



(a) Classic “retrieve then generate” pipeline without Ranking

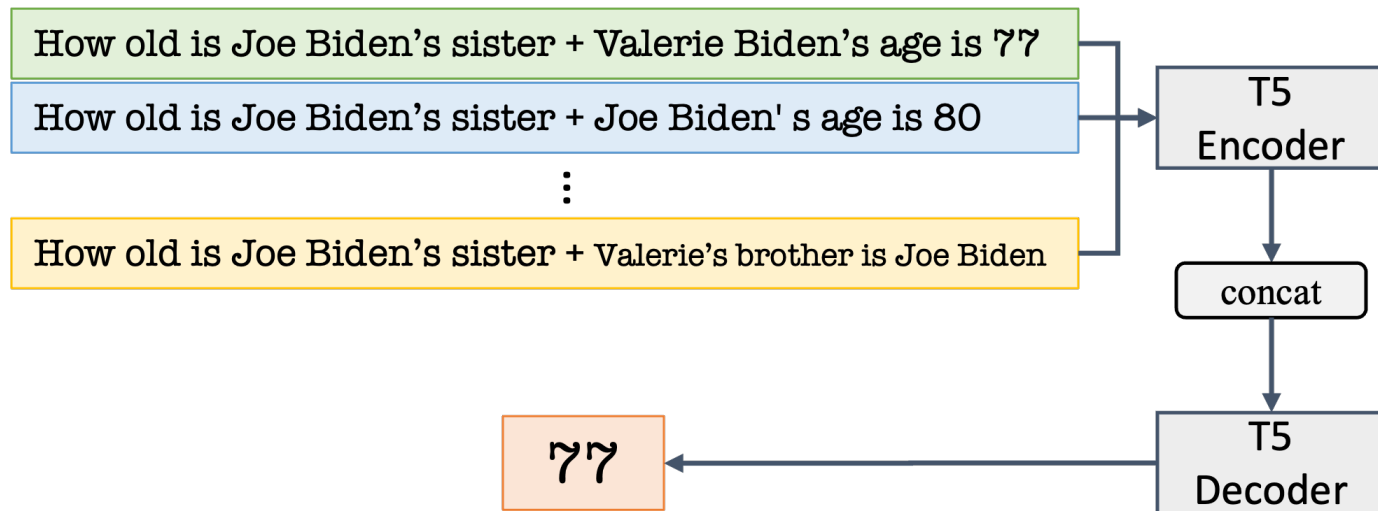


(b) Triple Ranking with document-level labeling strategy



(c) Contextual Triple Ranking with triple-level labeling strategy

How to generate answers?



Experimental Evaluation

Dataset (KG: Freebase)

- WebQSP [1]
- FreebaseQA [2]

Evaluation metric:

- Hit@1 / Exact Match

Retriever

- DPR [4]
- BM25 [3]

Re-ranker

- Pre-trained Answer sentence selection models (ELECTRA-Large) [5]

[1] The Value of Semantic Parse Labeling for Knowledge Base Question Answering. Yih, et al. ACL 2016.

[2] FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase. Jiang et al. NAACL 2019.

[3] The probabilistic relevance framework: BM25 and Beyond. Robertson, et al. Found. Trends Inf. Retr.

[4] Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin, et al. EMNLP 2020.

[5] Answer sentence selection using local and global context in transformer models. Lauriola et al. ECIR 2021.

Experimental Evaluation

A contextual ranker results in up to 6.5% improvement with the same generator!

- Achieved competitive performance with smaller-sized generators*

Model	FreebaseQA		WebQSP	
	Hit@1	LF?		Hit@1
FILM [120]	63.3	-		-
CBR-SUBG [32]	52.1	Yes		72.1
PullNet [113]	-	Yes		67.8
ReTrack [19]	-	Yes		<u>74.7</u>
DecAF (large, 100) [133]	79.0	Yes		80.7
Unik-QA [93] (base)	-	No		76.7
Unik-QA [93] (large)	-	No		79.1
DecAF - Answer only (large, 100)	79.0	No		74.7
Ours (base, 50)	80.9	No		71.8
Ours (large, 50)	84.3	No		76.9
Ours (base, 100)	80.2	No		77.2
Ours (large, 100)	<u>84.2</u>	No		<u>77.8</u>

Experimental Evaluation

- *A good labeling strategy is important for training an effective ranker!*
- *Context is helpful for the re-ranker(82.4 vs. 81.3 on FreebaseQA and 76.8 vs. 75.9 on WebQSP.)*

Ranking Method	Retriever	Re-ranker				Generator
	Hit@500	Hit@1	Hit@10	Hit@100	GT Triple Hit@100	Hit@1
No ranker	98.2	37.8	68.1	90.1	33.6	45.5
Doc-level Label	98.2	39.7	67.4	81.2	77.3	75.4
Triple-level Label	98.2	54.0	84.0	95.2	78.3	80.0

(a) Results on FreebaseQA

Ranking Method	Retriever	Re-ranker				Generator
	Hit@500	Hit@1	Hit@10	Hit@100	GT Triple Hit@100	Hit@1
No ranker	98.1	30.0	53.9	83.7	44.4	57.4
Doc-level Label	98.1	50.3	70.4	79.2	68.9	67.6
Triple-level Label	98.1	73.0	86.0	91.0	74.0	70.5

(b) Results on WebQSP.

Error Analysis

- Confusing triples
 - the selected triples, while relevant, were incorrect
- Strict evaluation
 - predictions semantically align with the gold answers but are still treated as wrong
- Incomplete labels
 - the predictions are accurate but are not included in the gold answer set.
- Complex constraints
 - Certain questions need the answer to satisfy all specified constraints
- Relative information
 - need the understanding of sequential or temporal information

Question:	where did clay matthews go to school?
Gold Answers:	University of Southern California ; Agoura High School
Predicted Answers:	Georgia Institute of Technology
Error Type:	Confusing Triples
Rationale:	<i>There are multiple people named Clay Matthews, and one of them graduated from Georgia Tech</i>
Question:	what all does google now do?
Gold Answers:	Google Maps
Predicted Answers:	Google Maps Engine
Error Type:	Strict Evaluation
Rationale:	<i>The predicted answer is correct but is marked wrong due to strict evaluation for exact match</i>
Question:	what shows are shot in new york?
Gold Answers:	Both Sides ; The Stand ; Flight of the Conchords ; Trial Heat
Predicted Answers:	The Big Short
Error Type:	Incomplete Labels
Rationale:	<i>The show 'The Big Short' was filmed in New York, but it's not listed in the gold answers</i>
Question:	What 1976-9 UK TV series, written by David Nobbs, frequently featured brief footage of a hippopotamus?
Gold Answers:	The Fall and Rise of Reginald Perrin
Predicted Answers:	Nightmare in the Park
Error Type:	Complex Constraints
Rationale:	<i>The question imposes multiple constraints that the model fails to meet correctly</i>
Question:	Who starred alongside Polly James in the first series of The Liver Birds?
Gold Answers:	Pauline Collins
Predicted Answers:	Nerys Hughes
Error Type:	Relative Information
Rationale:	<i>Nerys Hughes starred in 'The Liver Birds,' but she joined from the 2nd series. The model requires relative or temporal information to answer the question accurately</i>

Figure 4: Examples for error analysis were sampled from both the FreebaseQA and WebQSP datasets. Each example includes the raw question, the gold answers, the predicted answers from the best-performing model, the error type, and a detailed rationale for the error.

Conclusions

- We introduced a retriever-reranker-generator framework for KGQA
- We proposed to use a contextual re-ranker leveraging the rich context for ranking candidate triples from a KG
- We experimentally evaluated the framework on two benchmark datasets and showed performance gain with the re-ranker

Future Work

- Improving the context available to the re-ranker beyond the 1 hop neighborhood of retrieved triples will improve its ability to surface complex information
- Incorporating higher-quality labels, particularly when dealing with complex questions that necessitate the consideration of multiple significant entities
- Extending the re-ranker to rank generated logical forms

Thank you!