

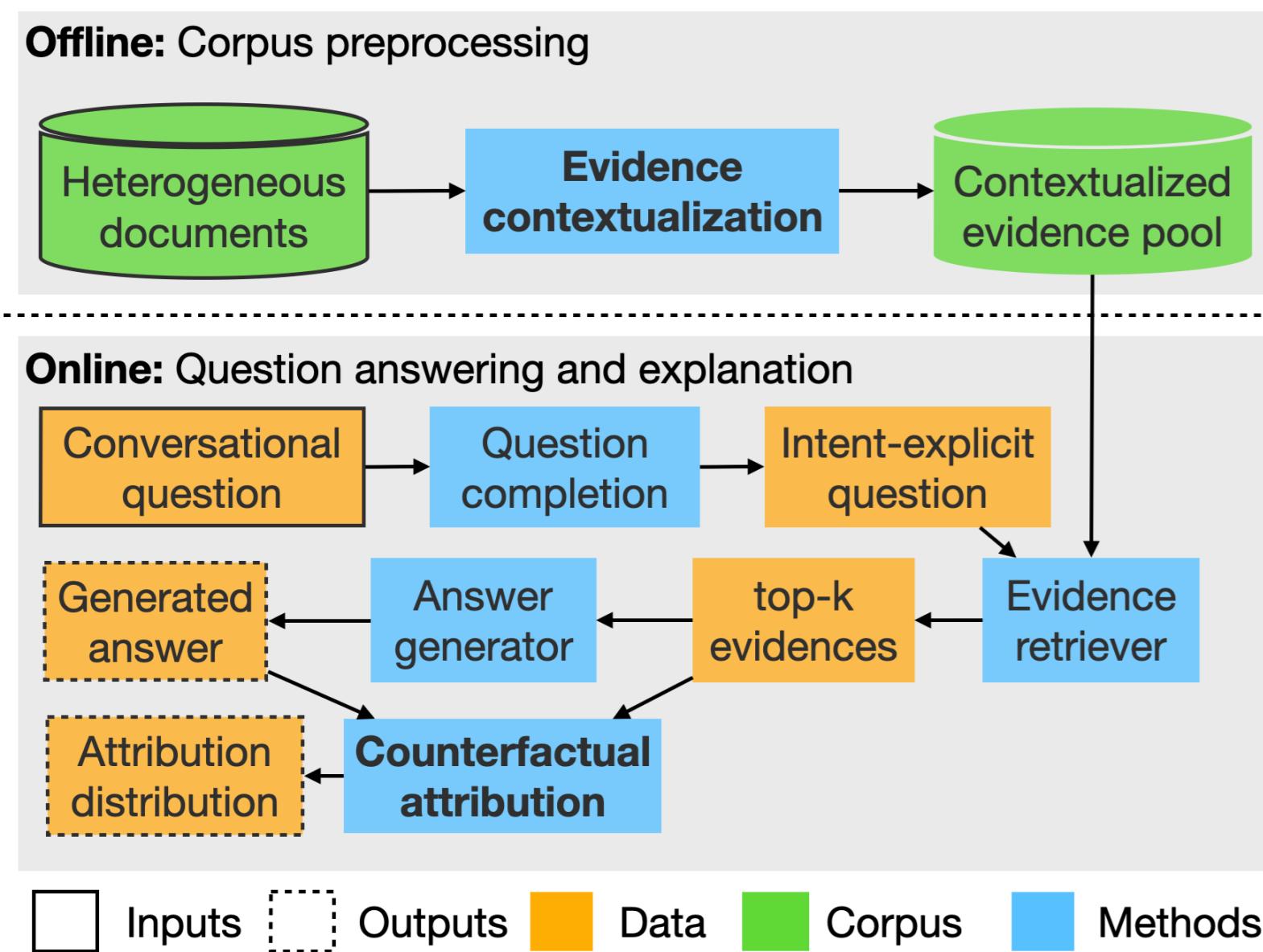
Evidence Contextualization and Counterfactual Attribution For ConvQA over Heterogeneous Data with RAG Systems

Rishiraj Saha Roy, Chris Hinze, Joel Schlotthauer, Andreas Foltyn, Luzian Hahn, and Fabian Küch

{rishiraj.saha.roy, chris.hinze, joel.Schlotthauer, andreas.foltyn, luzian.hahn, fabian.kuech}@iis.fraunhofer.de

Extended version: <https://arxiv.org/pdf/2412.10571>
All artifacts: <https://github.com/Fraunhofer-IIS/RAGonite>

System overview



Heterogeneous corpus: Confluence pages

2024-10-02 Meeting Notes

Today we will talk about the progress of the project on retrieval augmented generation.

Agenda

- We'll first do a basic round of RAG team updates in this month's meeting
- Everyone will report what has been done, and the to-dos

Member	Task	Action items	Time needed	Notes
Bob	Basic FE and BE	Follow-up q in UI	3 days	Currently manual
Alice	Similarity function	Fine-tune with gpt4o*	1 week	Now w/ embed cos
Trudy	Verbalizations	Batch configs*	6 hours	Running superbly

* Alice and Trudy to fix long-standing embedding error with openxt strings

A new benchmark: ConfQuestions

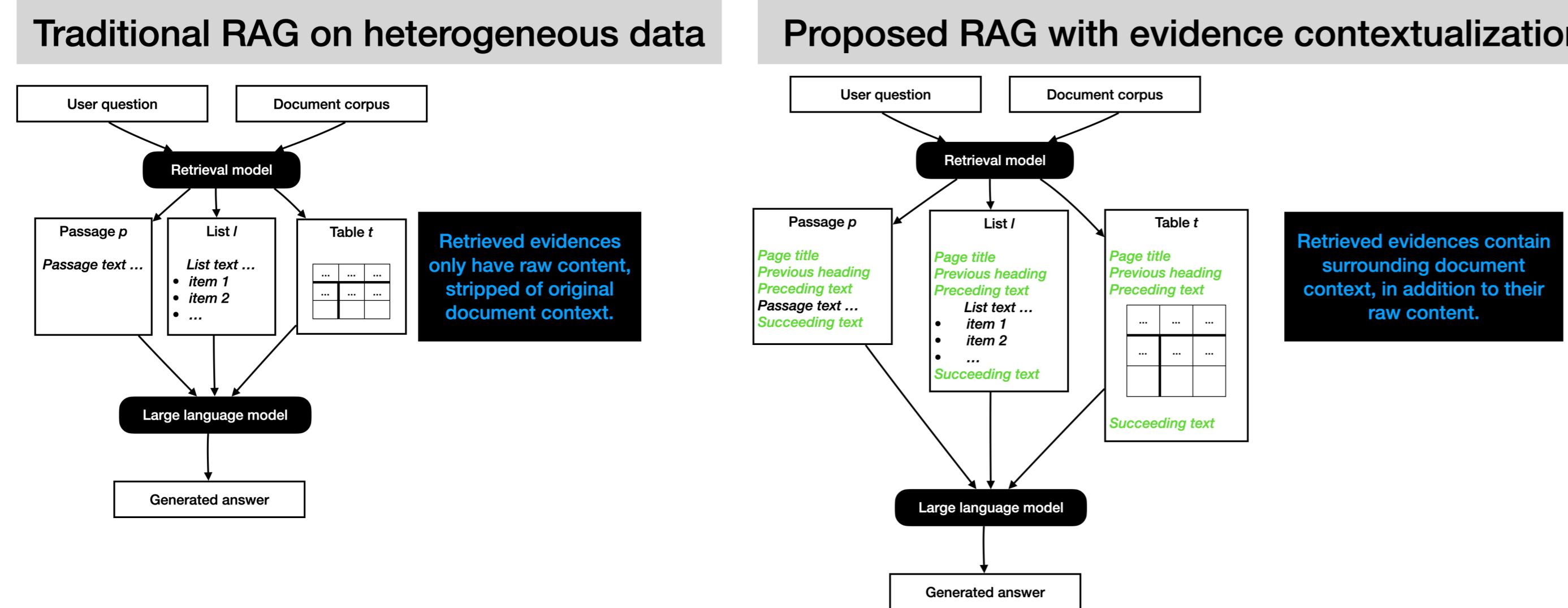
ConfQuestions is suitable for evaluating

- Conversational question answering
- Complex question answering
- RAG on heterogeneous enterprise documents

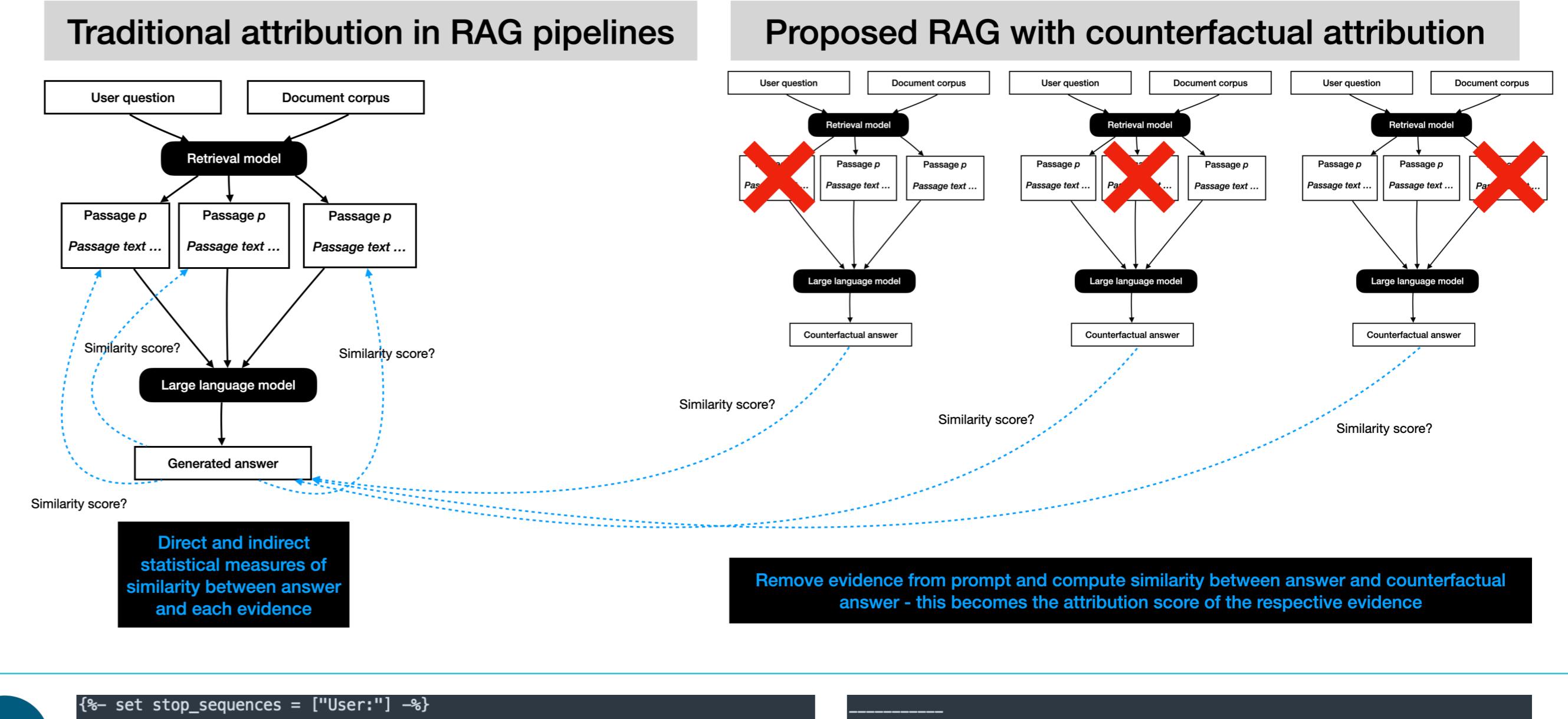
ConfQuestions contains

- 50 human conversations with 300 questions
- Corresponding human-generated completed questions
- Gold URL and answers: entities, phrases, lists, passages
- Each question in English and German
- Metadata: Answer-source, Complexity type
- 215 Confluence pages as corpus

Idea sketch: Evidence contextualization



Idea sketch: Counterfactual attribution



LLM prompts

1

```
## document_database.prompt_prefix.en|default("You are a helpful assistant.")
}} The following dialog was conducted with the user:
{{% for message in history %}}
{{< message.author >}: {{ message.content }}}
{{%- endfor %}}
User: {{ query }}
Rephrase the user's last statement into an independent question or statement.
Question/Statement: |
```

2

```
## set_stop_sequences = ["User:"]
## document_database.prompt_prefix.en|default("You are a helpful assistant.")
}} You are specialized in answering conversational questions in a retrieval-augmented generation setting. Be concise and concise answer to the input question as possible (less than 50 words if possible), using the retrieved evidences in this prompt as sources for answering. There is no need to provide additional information beyond the requested information. If the requested information cannot be found in the retrieved pool of evidence, "Please use only the information presented in the evidences, and mark the sources used in your answer within square brackets, like [Source 2] or [Source 3]. Please do not use your parametric memory and world knowledge. These are the available document evidences:
Assistant: |
```

Contextualization, attribution, and overall RAG results on ConfQuestions

Metric →	Retrieval precision@1						Answer relevance					
	NONE	+TTL	+HDR	+BEF	+AFT	+ALL	NONE	+TTL	+HDR	+BEF	+AFT	+ALL
All questions (600)	0.398	0.483	0.448	0.453	0.460	0.528	0.388	0.477	0.435	0.445	0.404	0.529
Simple questions (300)	0.413	0.470	0.477	0.450	0.483	0.510	0.423	0.537	0.482	0.502	0.458	0.593
Complex questions (300)	0.383	0.497	0.420	0.457	0.437	0.547	0.352	0.417	0.388	0.388	0.350	0.465
Answer in passage (200)	0.365	0.410	0.390	0.415	0.420	0.445	0.430	0.547	0.460	0.515	0.500	0.603
Answer in list (200)	0.340	0.475	0.440	0.435	0.490	0.560	0.328	0.422	0.390	0.400	0.340	0.507
Answer in table (200)	0.490	0.565	0.515	0.510	0.470	0.580	0.405	0.460	0.455	0.420	0.372	0.477
English questions (300)	0.413	0.500	0.467	0.480	0.483	0.563	0.407	0.530	0.472	0.483	0.432	0.575
German questions (300)	0.383	0.467	0.430	0.427	0.437	0.493	0.368	0.423	0.398	0.407	0.377	0.483

† NONE=No context; TTL = Page title; HDR = Previous heading; BEF = Evidence before; AFT = Evidence after; ALL = All context

Data ↓ / Attribution →	Naive	CFA	CFA w/ Clusters
All questions (364)	0.772	0.791	0.799
Simple questions (175)	0.771	0.811	0.817
Complex questions (189)	0.772	0.772	0.783
Answer in passage (98)	0.816	0.847	0.806
Answer in list (123)	0.756	0.780	0.821
Answer in table (143)	0.755	0.762	0.776
English questions (192)	0.807	0.765	0.786
German questions (172)	0.733	0.820	0.814

Answer explanation with counterfactual attribution:
* Attributed 67.32% to cluster 1 [Evidence 1, 2]
* Attributed 15.13% to cluster 2 [Evidence 3, 4, 7]
* Attributed 9.34% to cluster 3 [Evidence 5, 8]
* Attributed 7.59% to cluster 4 [Evidence 6]
* Attributed 0.34% to cluster 6 [Evidence 9]
* Attributed 0.28% to cluster 7 [Evidence 10]

Find all info here ↗



Turns →	1	2	3	4	5	6-10
Retr. P@1	0.660	0.530	0.500	0.430	0.490	0.560
Ans. relevance	0.670	0.495	0.465	0.440	0.555	0.550
Corpus →	Passages	Lists	Tables	All		
Retr. P@1	0.363	0.422	0.335	0.528		
Ans. relevance	0.289	0.407	0.333	0.529		
Completion →	LLM	Human				
Retr. P@1	0.528	0.658				
Ans. relevance	0.529	0.627				
Linearizer† →	VBL	PIPE	MD	HTML	TXT	
Retr. P@1	0.528	0.392	0.382	0.368	0.372	
Ans. relevance	0.529	0.364	0.363	0.366	0.361	
Indexing →	Row	Table	Both			
Retr. P@1	0.405	0.492	0.528			
Ans. relevance	0.362	0.514	0.529			
LLM →	GPT-4o	Llama3.1				
Retr. P@1	0.528	0.480				
Ans. relevance	0.529	0.449				
Embeddings →	BGE	OpenAI				
Retr. P@1	0.528	0.525				
Ans. relevance	0.529	0.538				
Ranking →	Lexical	Dense	Hybrid			
Retr. P@1	0.430	0.465	0.528			
Ans. relevance	0.417	0.443	0.529			
Reranking →	BGE+RRF	RRF				
Retr. P@1	0.528	0.417				
Ans. relevance	0.529	0.507				

† VBL=Verbalization; PIPE=Piped; MD=Markdown; HTML=HTML format; TXT=Plaintext

Take-home message: Contextualizing evidence with surrounding document text, and explaining answers with counterfactual attribution, are vital RAG add-ons