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

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

Fraunhofer-Institut für
Integrierte Schaltungen IIS

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

RAG over enterprise wikis is challenging

2024-10-02 Meeting Notes

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

Agenda

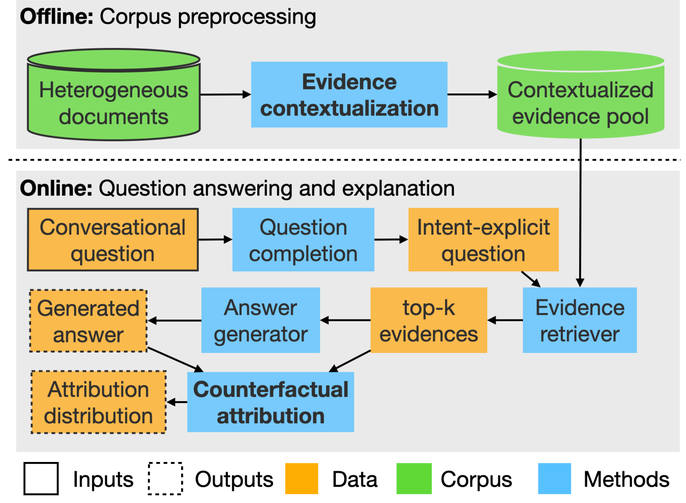
- We'll first do a basic round of RAG team updates in this month's meeting
- Everyone will report what has been done, and the to-dos

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

Question: *todo for alice in oct rag meeting?*

We propose RAGONITE, and ...



... a new benchmark: ConfQuestions

ConfQuestions is suitable for evaluating

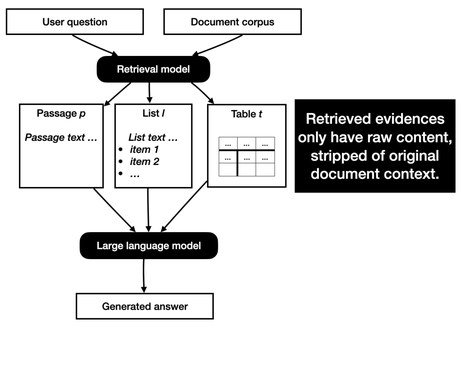
- ✓ Conversational question answering
- ✓ Complex question answering
- ✓ RAG on heterogeneous enterprise documents

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

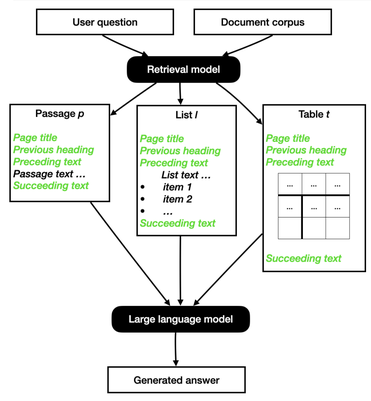
Idea sketch: Evidence contextualization

Traditional RAG on heterogeneous data



Retrieved evidences only have raw content, stripped of original document context.

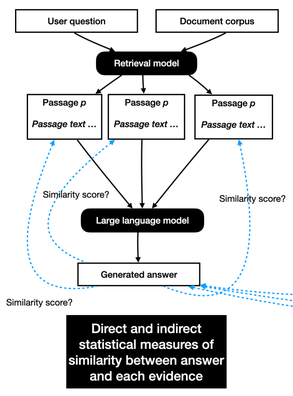
Proposed RAG with evidence contextualization



Retrieved evidences contain surrounding document context, in addition to their raw content.

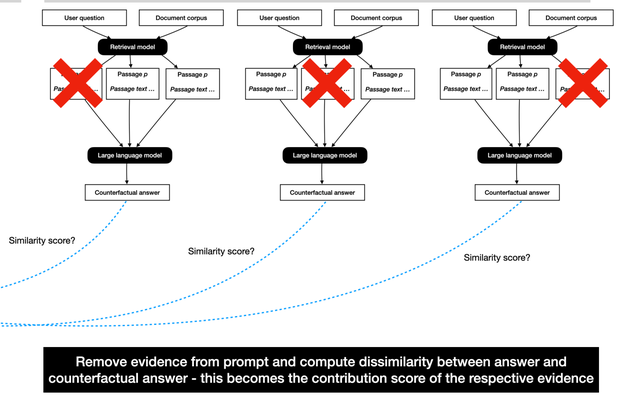
Idea sketch: Counterfactual attribution

Traditional attribution in RAG pipelines



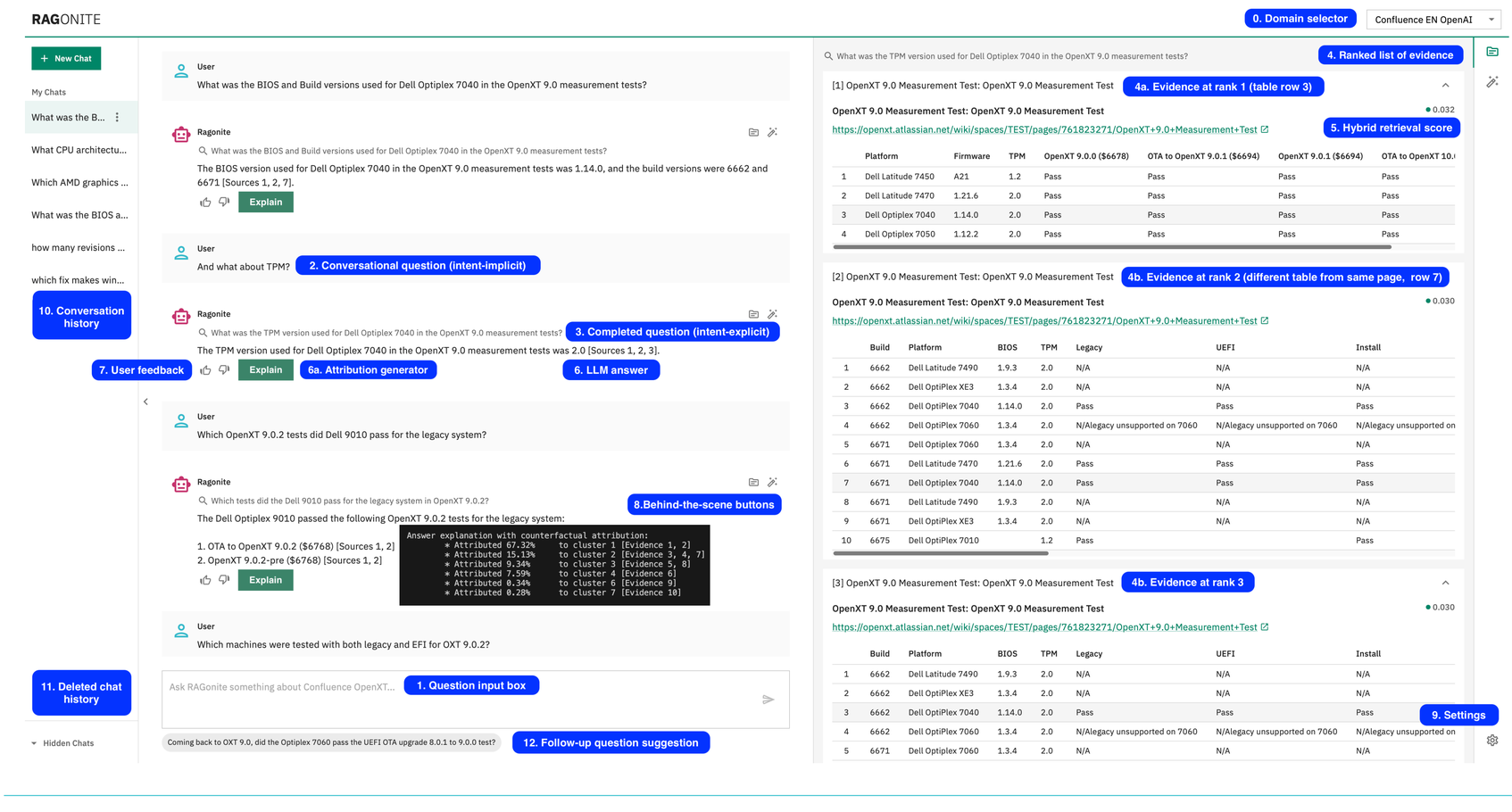
Direct and indirect statistical measures of similarity between answer and each evidence

Proposed RAG with counterfactual attribution



Remove evidence from prompt and compute dissimilarity between answer and counterfactual answer - this becomes the contribution score of the respective evidence

RAGONITE screenshot



ConfQuestions excerpt


```
"conv_id": "1",
"turns": [
    {
        "turn_id": "1",
        "q_type": "complex",
        "q_en": "What was the BIOS and Build versions used for Dell
            Optiplex 7040 in the OpenXT 9.0 measurement tests?",
        "q_de": "Welches BIOS und welche Build versionen wurden für den
            Dell Optiplex 7040 im Zuge der OpenXT 9.0 measurement tests
            eingesetzt?",
        "completed_q_en": "What was the BIOS and Build versions used
            for Dell Optiplex 7040 in the OpenXT 9.0 measurement tests?"
            ,
        "completed_q_de": "Welches BIOS und welche Build versionen
            wurden für den Dell Optiplex 7040 im Zuge der OpenXT 9.0
            measurement tests eingesetzt?",
        "a_url": [
            "https://openxt.atlassian.net/wiki/spaces/TEST/pages/
                761823271/OpenXT+9.0+Measurement+Test"
        ],
        "a_source": "table",
        "a": "Build version: 6662 and BIOS version: 1.14.0"
    },
    {
        "turn_id": "2",
        "q_type": "complex",
        "q_en": "And what about TPM?",
        "q_de": "Und was ist mit TPM?",
        "completed_q_en": "What was the TPM version used for Dell
            Optiplex 7040 in the OpenXT 9.0 measurement tests?",
        "completed_q_de": "Welche TPM version wurde für den Dell
            Optiplex 7040 im Zuge der OpenXT 9.0 measurement tests
            genutzt?",
        "a_url": [
            "https://openxt.atlassian.net/wiki/spaces/TEST/pages/
                761823271/OpenXT+9.0+Measurement+Test"
        ],
        "a_source": "table",
        "a": "Version 2.0"
    },
    {
        "turn_id": "3",
        "q_type": "complex",
        "q_en": "Which OpenXT 9.0.2 tests did Dell 9010 pass for the
            legacy system?",
        "q_de": "Welche OpenXT 9.0.2 Tests hat der Dell 9010 für
            veraltete (legacy) Systeme bestanden?",
        "completed_q_en": "Which OpenXT 9.0.2 measurement tests did
            Dell Optiplex 9010 pass for the legacy system?",
        "completed_q_de": "Welche OpenXT 9.0.2 Tests hat der Dell 9010
            für veraltete (legacy) Systeme bestanden?",
        "a_url": [
            "https://openxt.atlassian.net/wiki/spaces/TEST/pages/
```


Evaluation results

Retrieval P@1 for CONTEXTUALIZATION methods					
None	w/ Page title	w/ Preceding heading	w/ Preceding text	w/ Succeeding text	w/ All context
0.398	0.483	0.448	0.453	0.460	**0.528** ★

Answer relevance for CONTEXTUALIZATION methods					
None	w/ Page title	w/ Preceding heading	w/ Preceding text	w/ Succeeding text	w/ All context
0.388	0.477	0.435	0.445	0.404	**0.529** ★

Accuracy for ATTRIBUTION methods		
Standard	CF	CF with Clusters
0.772	0.791	**0.799** ★

Scan the QR code for all info on project... ☝

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