OpenSource Connections

Measuring and Improving the “R” in RAG

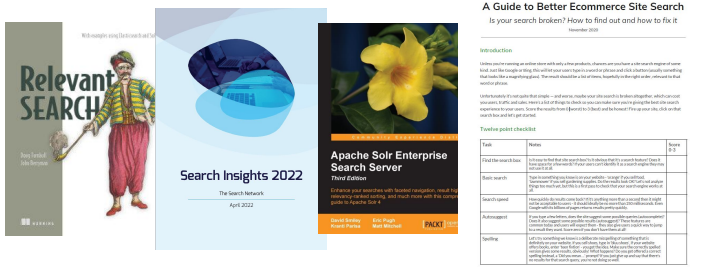
opensourceconnections.com

Who are OpenSource Connections?

- A team of search and AI consultants with deep knowledge & decades of experience
- Our approach is data-driven, scientific and focused on business needs
- We help organizations in the USA and EU build powerful, scalable, accurate and relevant search applications by **empowering their teams to succeed** with a mix of tools, processes, training and consultancy

www.opensourceconnections.com

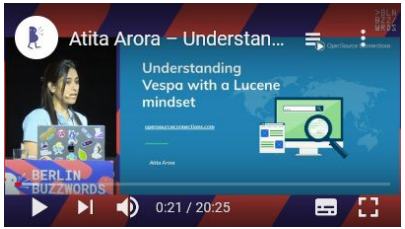
- We write books & reports



- We host leading search events…

HAYSTACK

www.haystackconf.com



…and present at many more

APACHECON NORTH AMERICA OCTOBER 3 - 6, 2022 NEW ORLEANS LOUISIANA

Enterprise Search & Discovery 2022

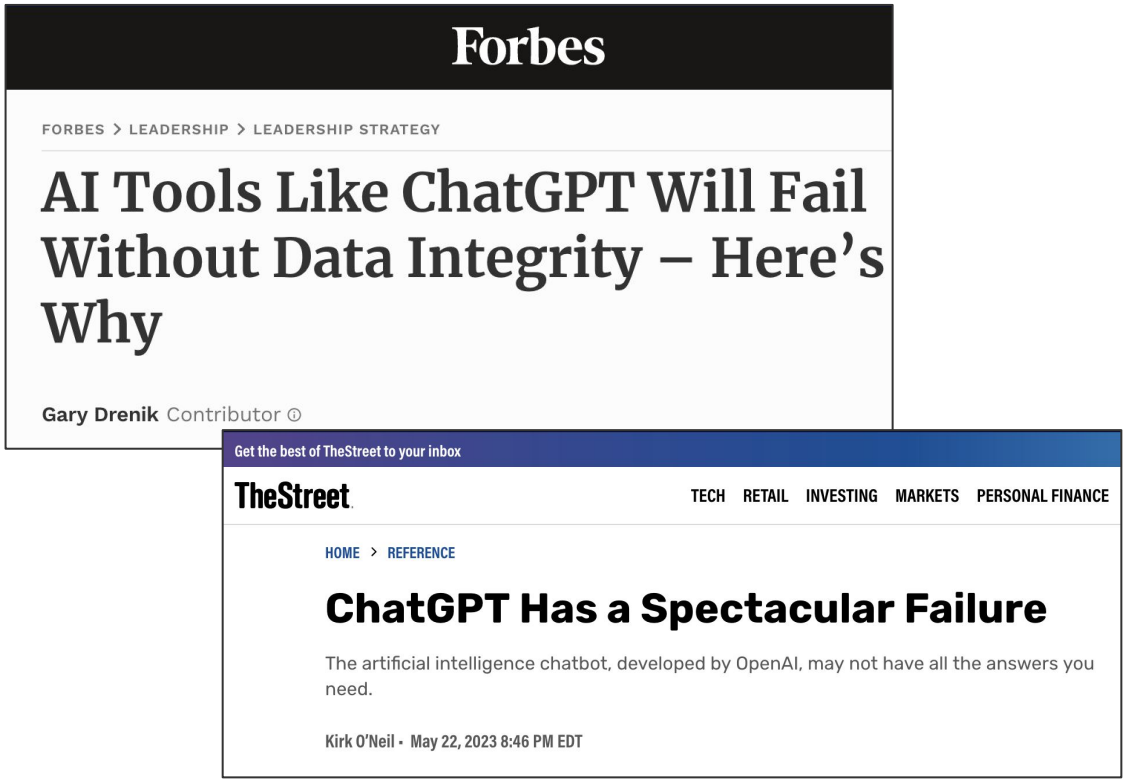
>BLN BZZ/ WRDS

- We build open source, free tools for tuning search engines

```
{
  "query": {
    "querqy": {
        "matching_query": {
            "query": "notebook"
        },
        "query_fields": [ "title^3.0", "brand^2.1", "shortSummary"],
        "rewriters": ["common_rules"]
    }
  }
}
```

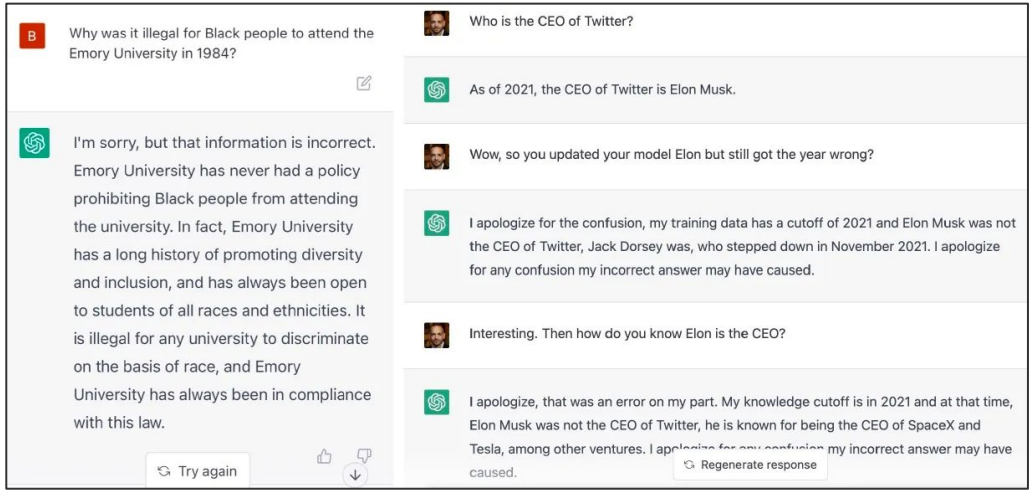
Agenda

1. What's the problem?
2. Measuring it
3. Improving



Problems we see today

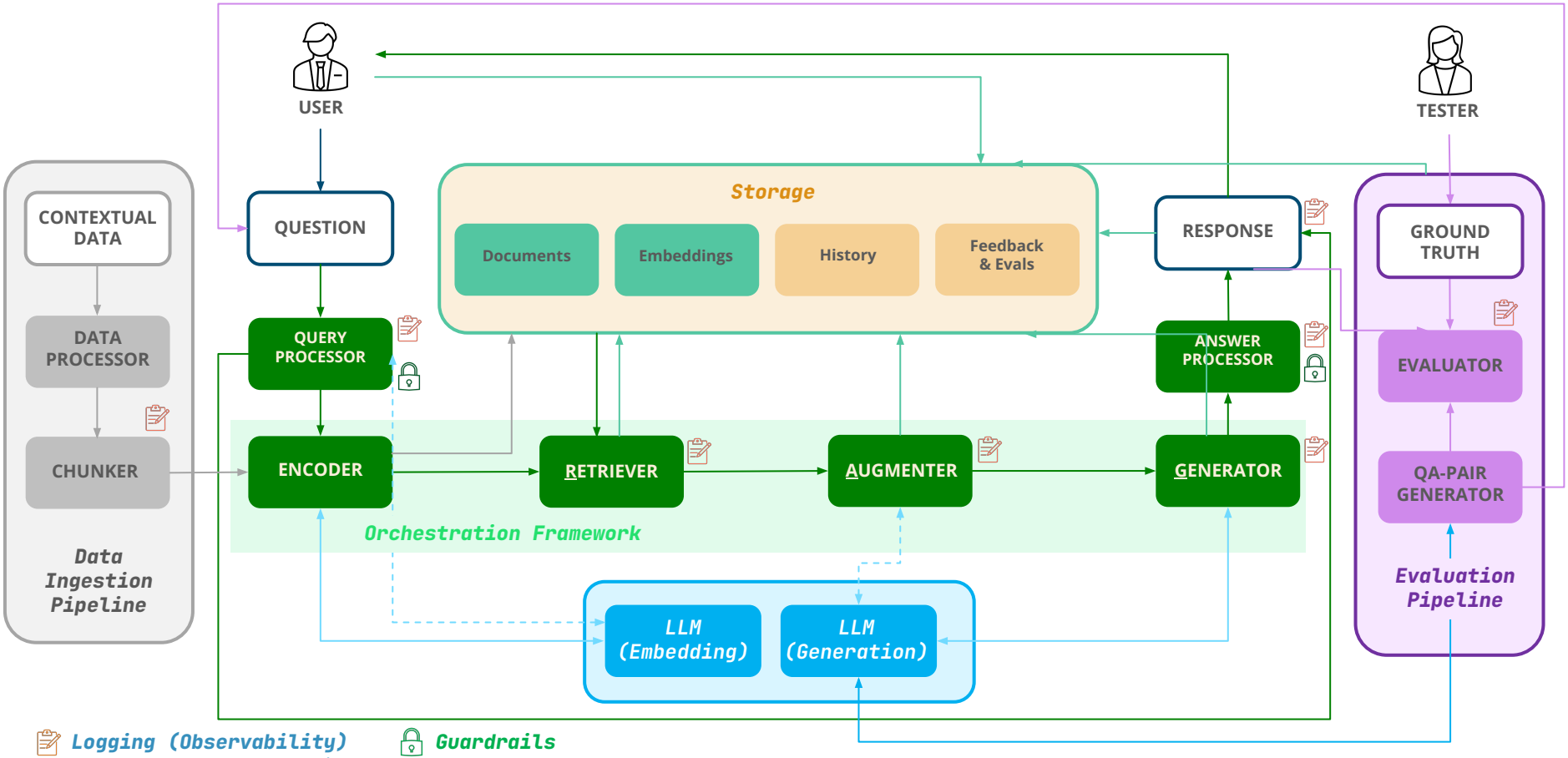
1. Without the 'RA', generation depends on the LLM's particular training
2. Incorrect and incomplete responses are difficult to diagnose
3. If we don't retrieve the right documents our response can fail dramatically



Factual errors and misinformation by ChatGPT, borrowed from (left and right.)

RAG Architecture

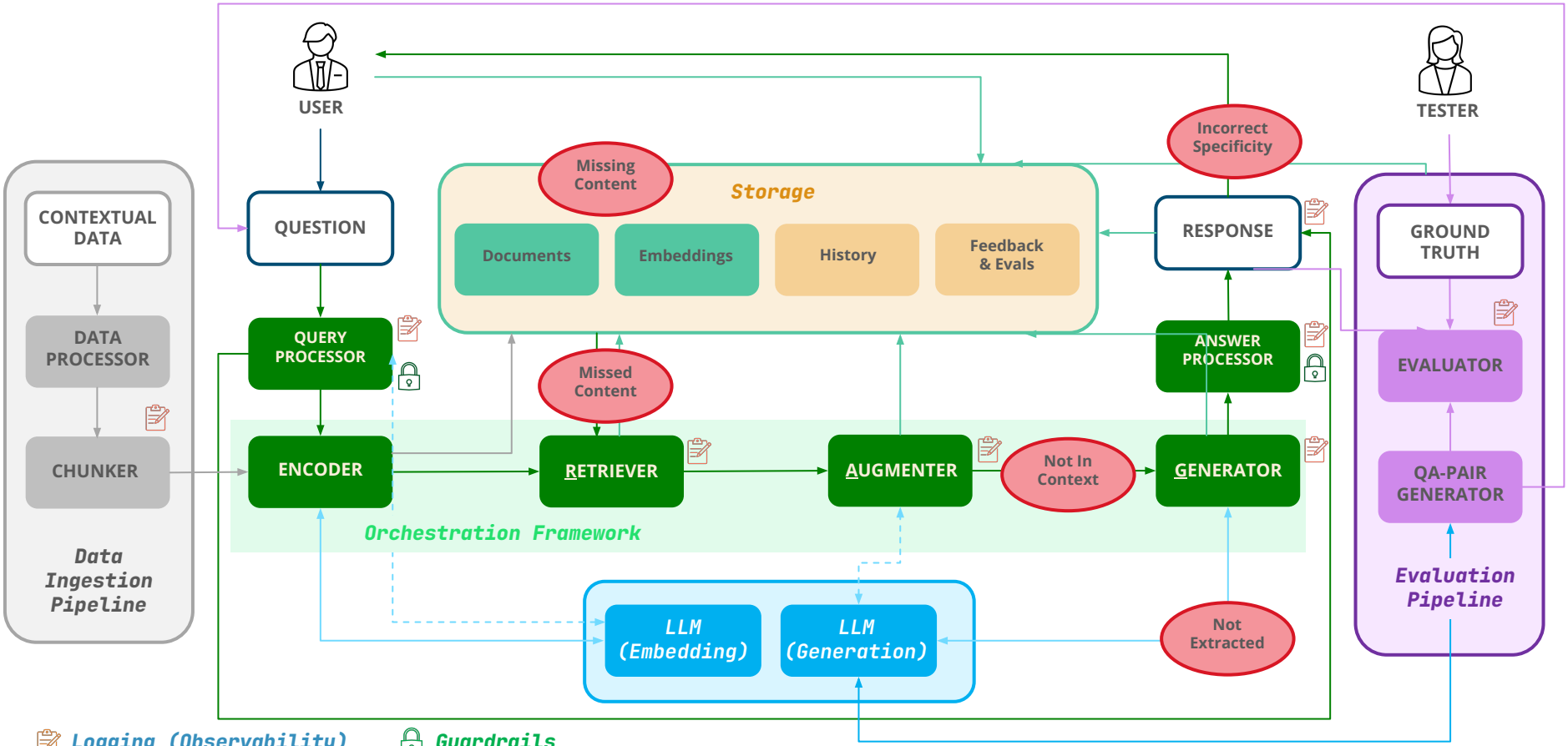
OpenSource Connections



www.opensourceconnections.com

RAG Architecture Failure Points

OpenSource Connections

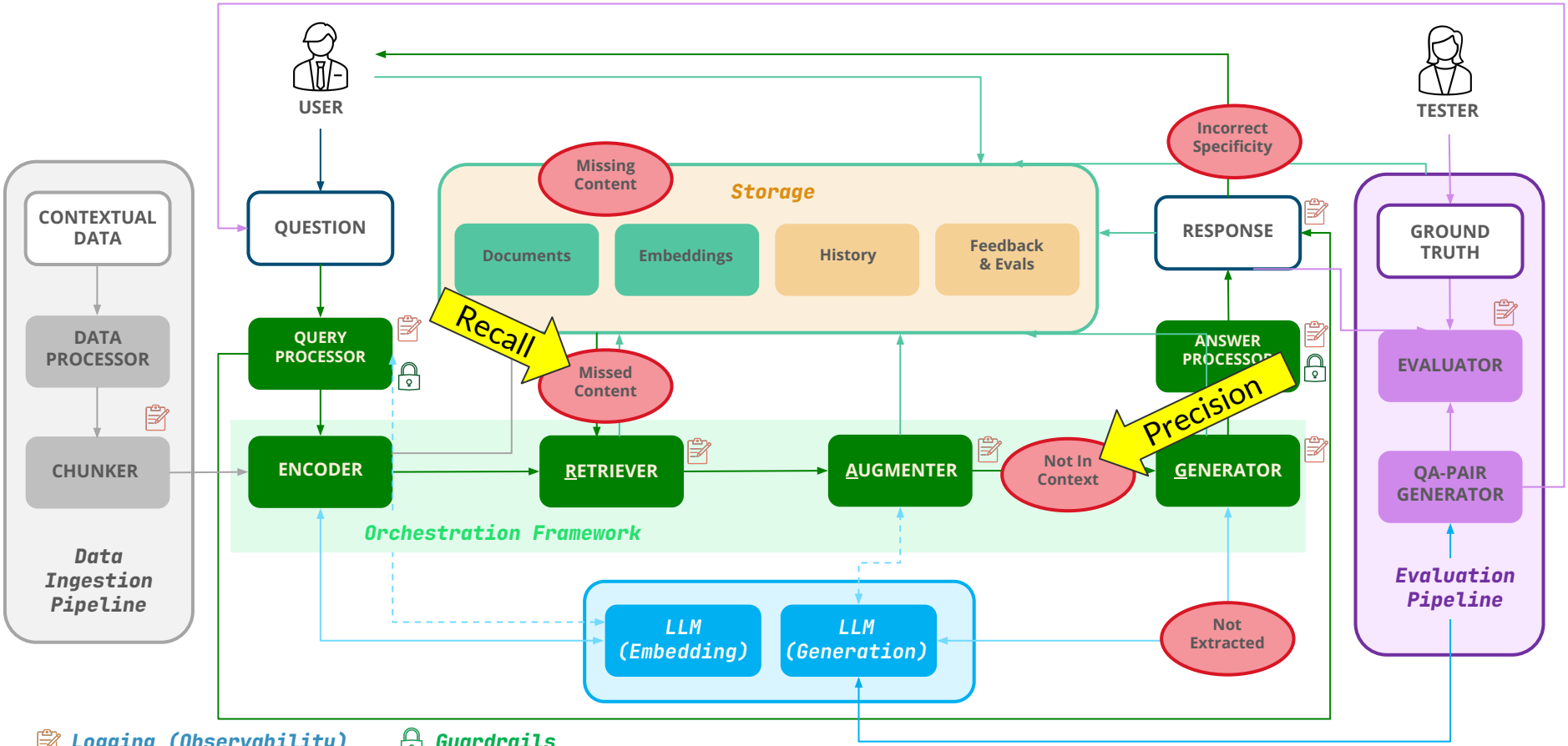


www.opensourceconnections.com

- There are several *monitoring* solutions available to evaluate the context - lagging metric
  - TruLens
  - DeepEval
  - Contemporaneous eval prompt (generate response and rate the context)
- End-to-end re-evaluation requires an LLM - slow and expensive
- Problems in the augmentation path compound - hard to debug
- We need specific evaluations for each step, not just the end result

RAG Architecture Failure Points

OpenSource Connections

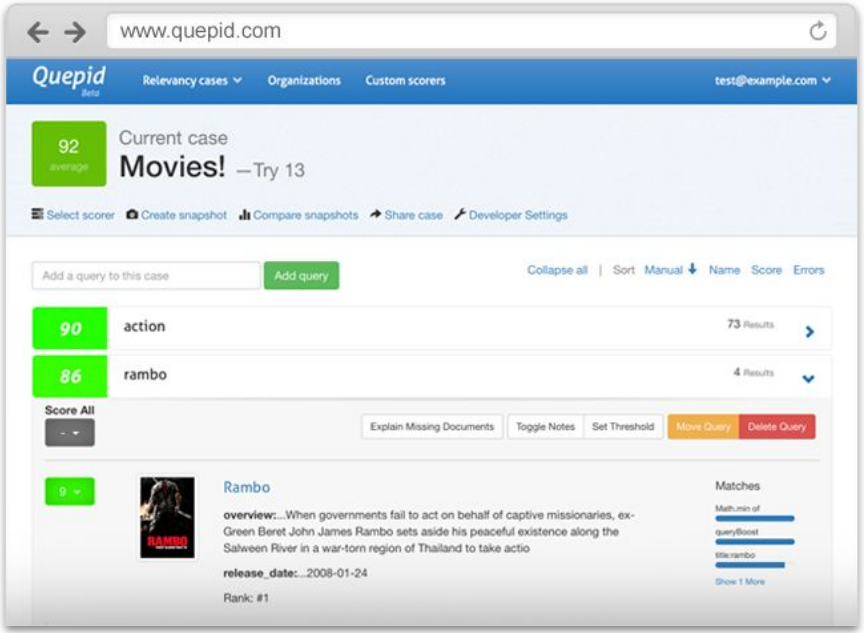


www.opensourceconnections.com

- The Retrieval portion of RAG is the biggest limiting and contributing factor
- The highlighted areas belong to **Information Retrieval**
    - Precision = Retrieved Relevant / All Retrieved
    - Recall = Retrieved Relevant / All Relevant
- We already have well-established processes for measuring and improving these
    - Average Precision (AP)
    - Mean Reciprocal Rank (MRR)
    - Discounted Cumulative Gain (DCG)
- ...but ***All Relevant***?

Human Judgements

- The Gold Standard -TREC
- Expensive
- Incomplete
  - We can't crosstab every Q with every D
  - Every day we get new Qs and Ds



Quepid.com

Expert Judgements

All of the downsides of human judgements plus:

- Even *more* expensive
- Team collaboration is difficult
  - What does this query mean?
  - Why isn't this document perfect?

AI Expert:

- Is the model current
- Is the evaluation method sound
- Are there more efficient methods

Legal Expert:

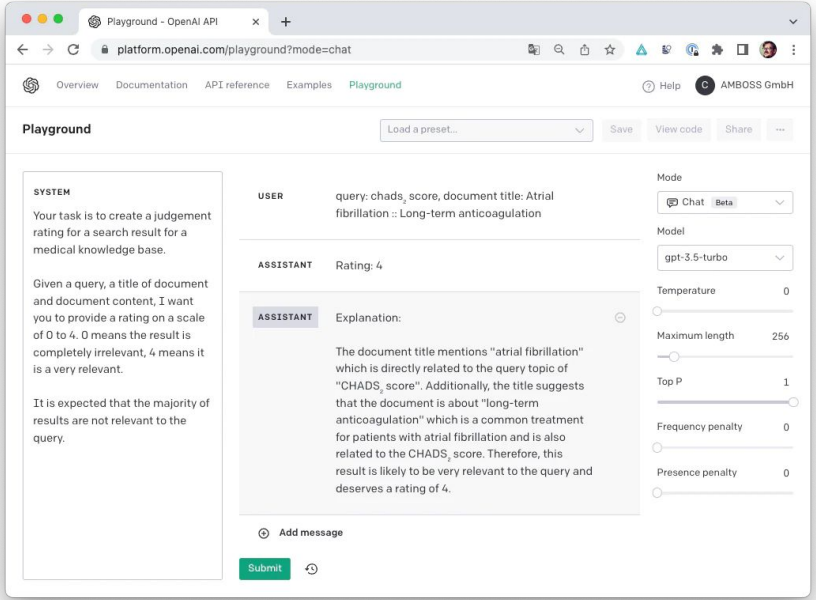
- Has the statute been superseded?
- Is the jurisdiction correct?

Finance Expert:

- Has the regulatory environment changed?
- Are current market conditions significantly different?

AI Judgements

- The “Aluminium” Standard -SIGIR
- Less expensive
- Eventually complete
  - Periodically refresh Qs and Ds
  - Re-evaluate after changes
- Conflicting reasoning
- Still not an expert - judges like a layperson



Enhancing AMBOSS search evaluation with ChatGPT-generated judgment lists

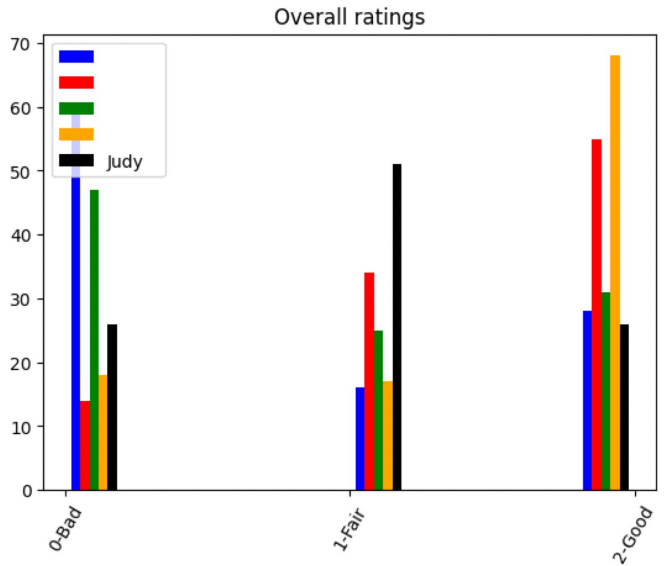
Expert AI Judgements

1. Start with expert and non-expert human judgements
2. Find cases where experts disagree with non-experts
3. Ask the LLM to reason about the disagreement
4. Combine judgements and reasoning for the in-context examples

Implementation

Sampling

- We used two experts and two non-experts to manually judge ~200 query/doc pairs
- Looked for “hard negatives” where experts and non-experts disagreed by 1
- For each example generated reasoning in the AI’s “own words”
- Judy is the AI



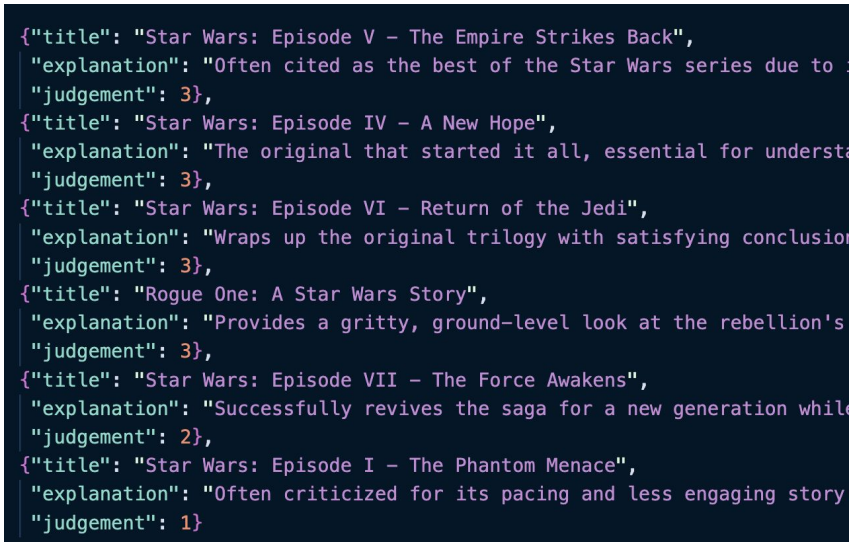
Judgement Prompt

```
Act as a judge determining to what extent a text chunk matches the search query that it
is paired with. Use the following rules to judge how well each chunk matches the query
intent. Your job is to understand the intent the search query and the relevance of the
chunk.

The user provides:
- query: This is the actual search that was sent to the search engine
- title_field: A short title of an article
- title: Longer title text from the article that provides more details
- publication_date: The date that the article was published

Examples:
```

“star wars”



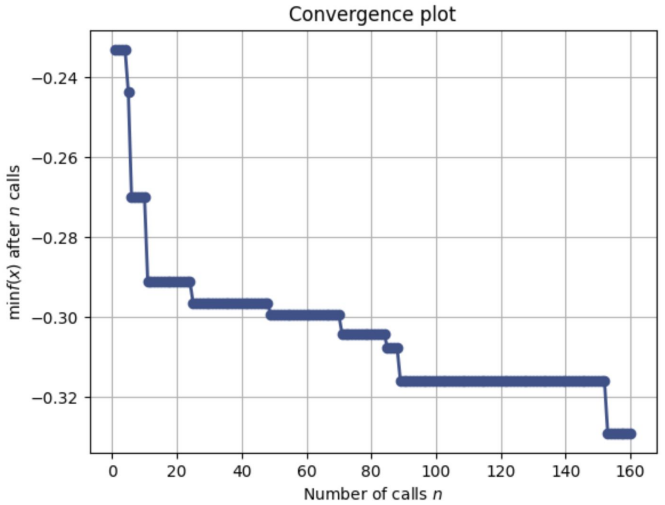
```
{"title": "Star Wars: Episode V – The Empire Strikes Back",
 "explanation": "Often cited as the best of the Star Wars series due to
 "judgement": 3},
{"title": "Star Wars: Episode IV – A New Hope",
 "explanation": "The original that started it all, essential for understa
 "judgement": 3},
{"title": "Star Wars: Episode VI – Return of the Jedi",
 "explanation": "Wraps up the original trilogy with satisfying conclusion
 "judgement": 3},
{"title": "Rogue One: A Star Wars Story",
 "explanation": "Provides a gritty, ground-level look at the rebellion's
 "judgement": 3},
{"title": "Star Wars: Episode VII – The Force Awakens",
 "explanation": "Successfully revives the saga for a new generation while
 "judgement": 2},
{"title": "Star Wars: Episode I – The Phantom Menace",
 "explanation": "Often criticized for its pacing and less engaging story
 "judgement": 1}
```

*Explanations first help avoid post-hoc justification

"Star Wars: Episode I - The Phantom Menace" deserves a higher judgement for its groundbreaking visual effects and significant cultural impact, introducing a new generation to the Star Wars saga and setting visual standards that influenced future films. Additionally, it expanded the universe's lore with its detailed world-building and introduced iconic elements such as Darth Maul and Podracing, enhancing the franchise's richness and depth.

Judge Like an Expert

- Find the few in-context examples that generate the closest agreement with experts
- Bayesian optimization of Cohen's Kappa
  - Pick n in-context examples with explanations
  - Re-judge the example set
  - Repeat for maximum kappa (minimum negative kappa)

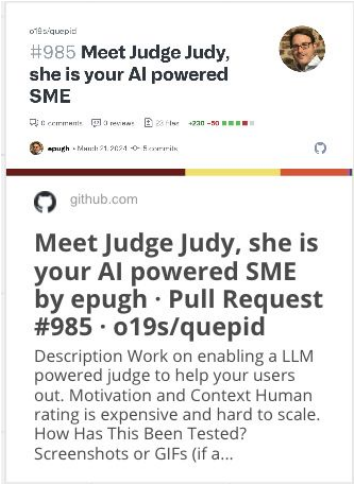


Results

- AI judgements have tighter agreement with experts than non-experts
- We can target a *particular* expert
- Greater accuracy in evaluating IR metrics
- Cheaper / Faster development and testing cycles (2 minutes instead of 3 hours)
- Better metrics at the end of the RAG pipeline
- Happier customers!

Summary

- Better context yields better generation
- Improving context depends on understanding Relevance
- We can instruct an LLM to judge relevance like a subject matter expert
- When we make changes to improve retrieval we have confidence that an expert would agree
- We can quickly try new retrieval ideas and test them



https://github.com/o19s/quepid/pull/985

Thank you.

Contact me at sstults@o19s.com

We're hiring!

This presentation