Automation is risky

Many methods for automatically evaluating RAG results have been proposed. But it's risky to fully rely on automated RAG evaluation, because these methods can often fail.

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- Large language model as judge
- User feedback
- String similarity
- Semantic similarity

What if the article is wrong?

Faithful to the correct part of the article?

Users don't want to do your QA

What if there's more than one right answer?

What if there's more than one way to write the answer?

The stakes are high

When **search** returns poor results, people just modify their query and then search again.

When **RAG** returns a wrong result, people believe the authoritative-sounding answer generated by the large language model. And then they sue when the answer gets them in trouble!

Air Canada ordered to pay customer who was misled by airline's chatbot

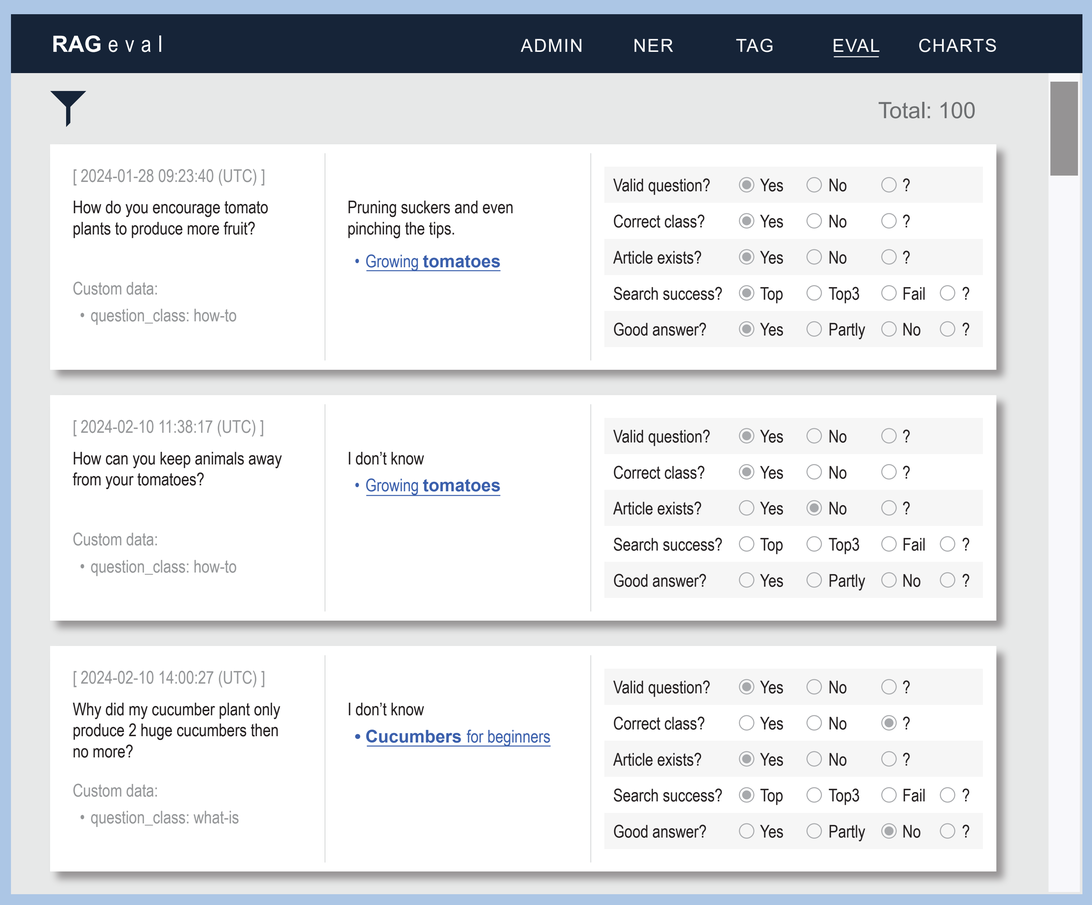
Company claimed its chatbot "was responsible for its own actions" when giving wrong information about bereavement fare

https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit

EVALUATING RETRIEVAL-AUGMENTED GENERATION (RAG)

ibm.biz/RAG-evaluation

A streamlined, human-in-the-lead approach



Our team created a web app we use to evaluate results returned by our RAG solutions:

- For each user question, we assess returned answers according to multiple criteria
- We meet regularly to review and discuss results
- We fix content gaps, search failures, writing problems
- As we manually classify, tag, and evaluate results, that manual work is used to create training data
- Over time, the web app automatically classifies, tags, and evaluates more and more results, using models trained on the data from our manual work
- We share trends seen in user questions with our larger product team to improve the user experience