Evaluation of Retrieval Augmented Generation

::: MASTERSEMINAR SUCHMASCHINEN & RAG, CIS, SOMMERSEMESTER 2025 :::

Overview

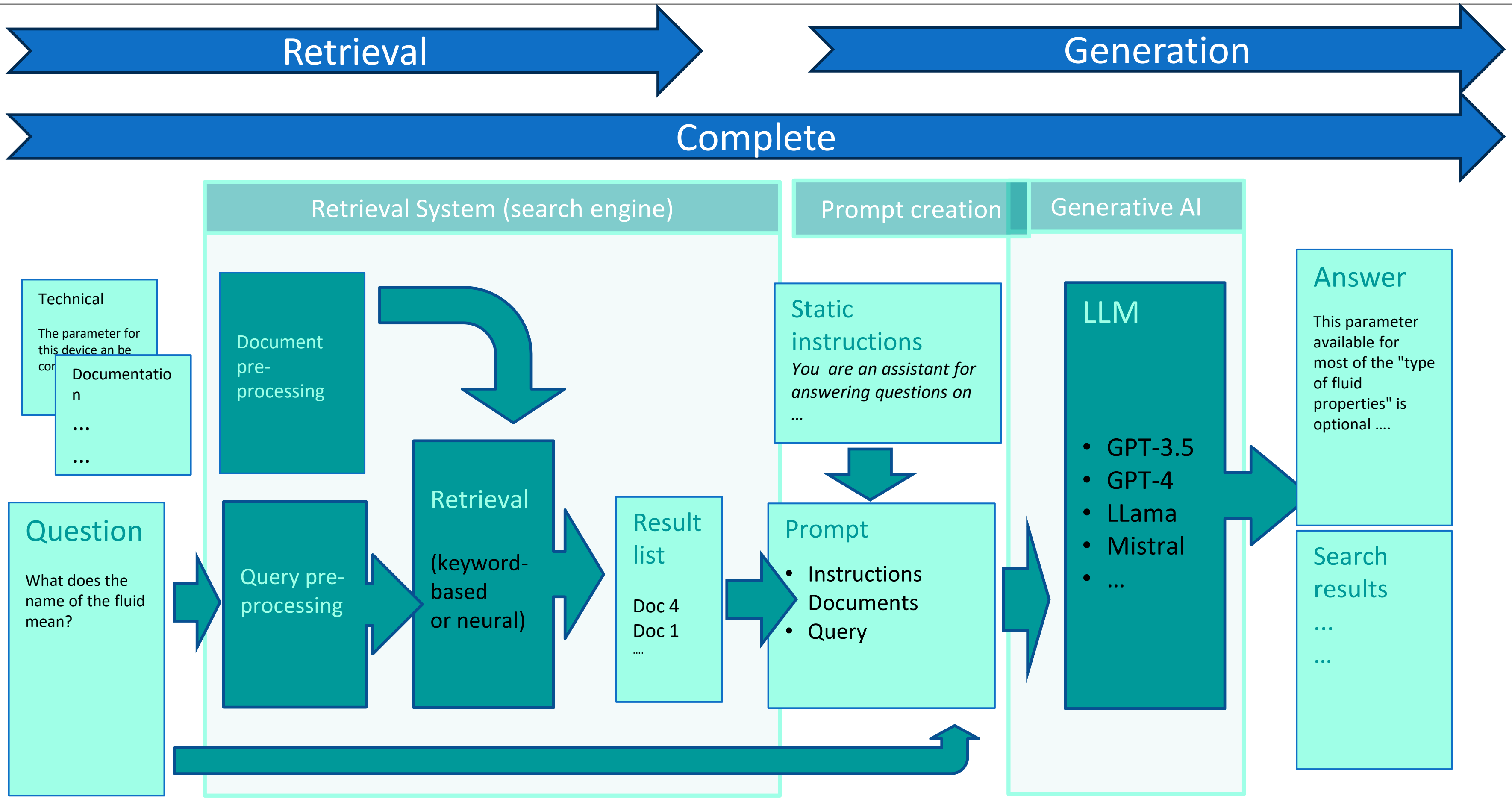
- Introduction

- Recap: Evaluation measures for the retrieval component

- Evaluation measures using a ground truth – Bleu, Rouge, Meteor, Bert Score

- Evaluation measures without ground truth -

Recap: RAG architecture



Retrieval

Generation

Complete

Retrieval System (search engine)

Prompt creation

Generative AI

Technical

The parameter for this device an be cor

Documentation

...

...

Question

What does the name of the fluid mean?

Document pre-processing

Query pre-processing

Retrieval

(keyword-based or neural)

Result list

Doc 4

Doc 1

....

Static instructions

You are an assistant for answering questions on ...

Prompt

- Instructions
- Documents
- Query

LLM

- GPT-3.5
- GPT-4
- LLama
- Mistral
- ...

Answer

This parameter available for most of the "type of fluid properties" is optional ....

Search results

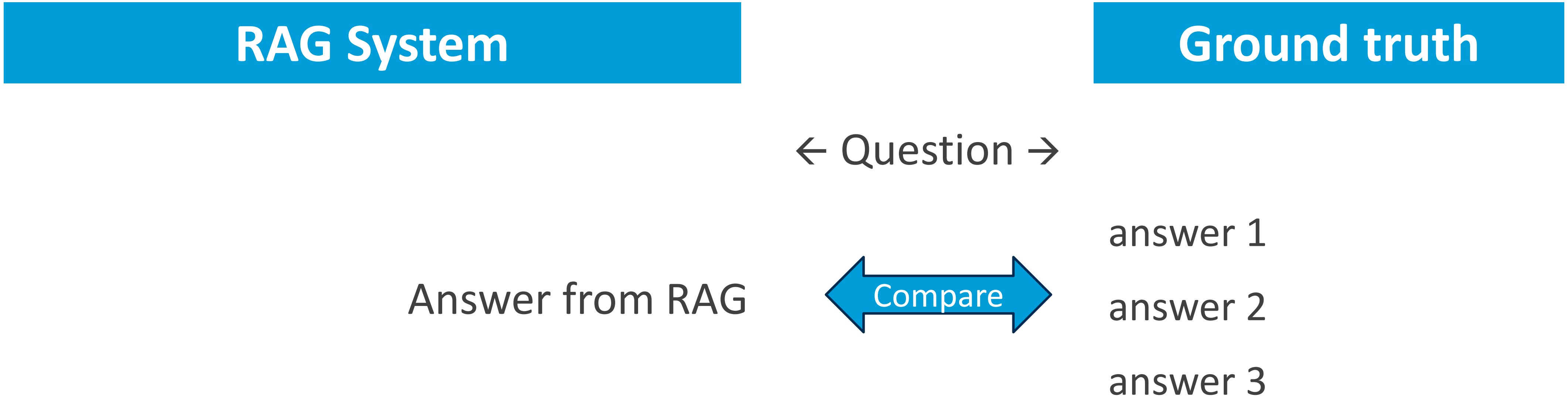
...

...

Retrieval measures

- Recall / Precision / F1
- Precision at k
- Mean reciprocal rank
- NDCG

Bleu, Rouge, Meteor, Bert



RAG System

Ground truth

← Question →

answer 1

Answer from RAG

Compare

answer 2

answer 3

Scores without embeddings

Bleu – precision based

Weighted geometric mean of modified n-gram precisions.

N-gram Precision: The precision is calculated as the number of matching n-grams in the output and reference text, divided by the total number of n-grams in the output.

Brevity Penalty: penalty for outputs shorter than the references to avoid rewarding overly short sequences. If the output length is less than the reference length, the BLEU score is multiplied by the ratio of these lengths.

Cumulative BLEU Score: To calculate the final BLEU score, the modified n-gram precisions are combined geometrically, and then a brevity penalty is applied.

Rouge – recall based

ROUGE-N: is a variant of ROUGE, which considers recall of n-grams. An n-gram is a contiguous sequence of n items from a given sample of text or speech. The recall is calculated as the total number of matching n-grams in the machine output and reference text, divided by the total number of n-grams in the reference text.

ROUGE-L: Another variant of ROUGE, ROUGE-L, measures the longest common subsequence (LCS) between the system and reference summaries. The LCS is the longest sequence of words that are the same between the system and reference summaries and appear in the same order.

Meteor – + linguistics

Exact Matching: Exact word-to-word correspondence between the output and the reference output.

Stem Matching: If exact matching fails, METEOR checks for matches in word stems.

Synonym and Paraphrase Matching: If stem matching fails, it checks for matches in synonyms and paraphrases using WordNet.

Alignment: METEOR creates an alignment between the words and phrases in the machine and reference translations to identify matching and non-matching spans.

Penalty Calculation: METEOR applies two penalties: a penalty for unmatched words and a penalty for non-sequential matches.

The final METEOR score is a weighted harmonic mean of precision and recall, penalized by the amount of fragmentation in the alignment.

Scores using embeddings

BERT score (2020)

Semantic similarity between ground truth answer and found answer.

- match most similar word embedding between the two strings

- calculate average word embedding similarity

- uses transformer embeddings

Variants and other metrics:

- Bleurt

- BARTScore

LLM as a judge - RAGAS score

Score name	Description	Area	Question	Context	Answer	Ground truth answer
Faithfulness	How close is the answer content to the context?	Generation	no	yes	yes	no
Answer relevancy	How relevant is the answer with respect to question + context	Generation	yes	yes	yes	no
Context recall	Does the context relate to the ground truth answer?	Retrieval	no	yes	no	yes
Context precision	Is the best context ranked highest with respect to the ground truth answer?	Retrieval	no	yes	no	yes
Context relevancy	Is the context relevant for the question	Retrieval	yes	yes	no	no
Answer sem. similarity	Measures similarity with the ground truth	End2End	no	no	yes	yes
Answer correctness	Similarity and factual correctness	End2End	no	no	yes	yes

LLM as judge – other frameworks

- DeepEval

- Promptfoo (general framework for reliability and security)

LLM / RAG evaluation – testing dimensions

Quality

- RAGAS tests and related tests
- Language support
- Expert-reviewed data?

Performance

- Specific performance tests
    - End-to-end, retrieval, generative models
- Definition of performance requirements
- Tests on different platforms

Security and reliability

- Denial of service attacks
- Malicious output
- Data leakage