Large language model

PROGRESS REPORT
19/12/2024

Overview

1. Results

Metric aggregation

$\boldsymbol{x}$
User query

$\rightarrow$

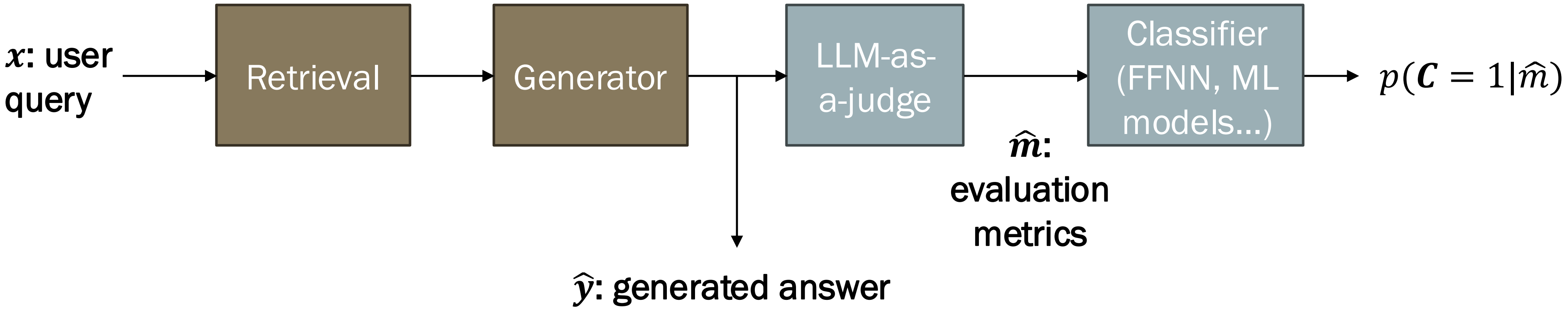
$\boldsymbol{m} = (\boldsymbol{m_1}, \dots \boldsymbol{m_n})$
Inference metric

$\rightarrow$

$$\boldsymbol{\delta} = \sum_{i=1}^{n} \boldsymbol{\alpha_i m_i}$$
Decision rule

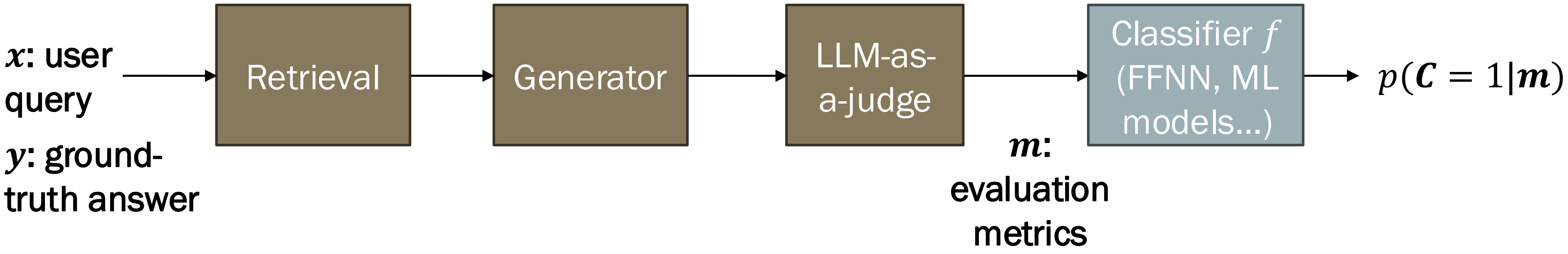
- Use only one metric $\Longrightarrow$ Not whole phenomenon captured
- Hence why, aggregated metric
  - Weighted average ?
  - Ranking ?

Pipeline in inference



- C confidence metric , $C = 0$ if there is an hallucination.

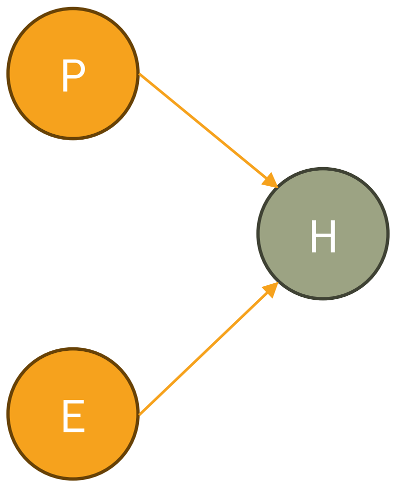
Pipeline: training



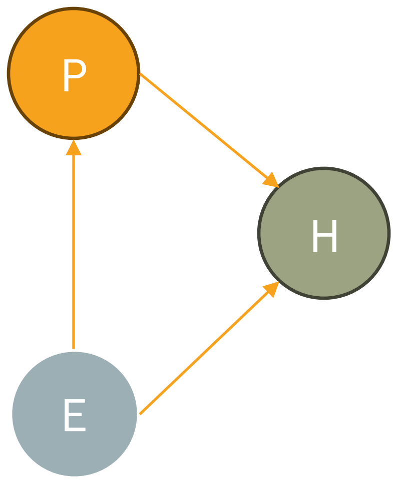
- C confidence metric, $C = 0$ if there is an hallucination.
  - On what data ?
    - $(m_i, C_i)_{i=1}^{N}$? $\Longrightarrow$ MC Sample of m_i on $(x_i)_{i=1}^{N}$ $\Longrightarrow$ Logistic regression.
    - Few-shot learning or use dataset for hallucination detection RAGTruth [arxiv.org:2401.00396] and the learned coefficients $(\alpha_i^{RAGTruth})_{i=1}^{n} \approx (\alpha_i^{data})_{i=1}^{n}$ (OOD generalization).

Causal perspectives

E: External context or knowledge, P: parametric knowledge, H: hallucination

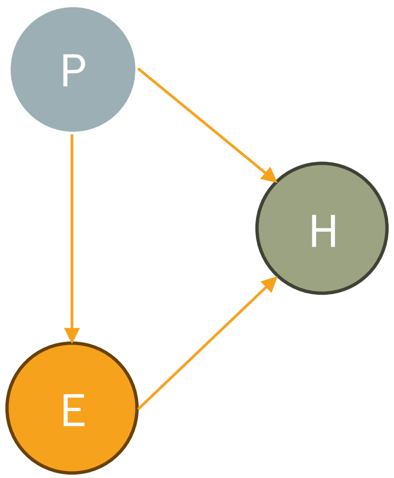


MixPE
which combines both P and E directly using uncertainty or sampling techniques



P confounded by E:
which relies on the LLM's hidden states for hallucination detection

Eigenscore
ReDeEP



E confounded by P :
by leveraging external context and model responses

RAGAS metrics
LMvLM...

Challenges faced

- Find annotated dataset

- Better incorporate lingo terminology
  - LoRA on a fine-tuning dataset

- Better define « hallucination »
  - Number-based in particular…

- Topological structure of metrics

- Find insightful benchmarks as methods are recent