QA dataset for evaluation

Alexandr Prozorov for Scibot team

Why?

Need input from experiments - most **popular** questions or documents

↓

generate QA pairs

↓

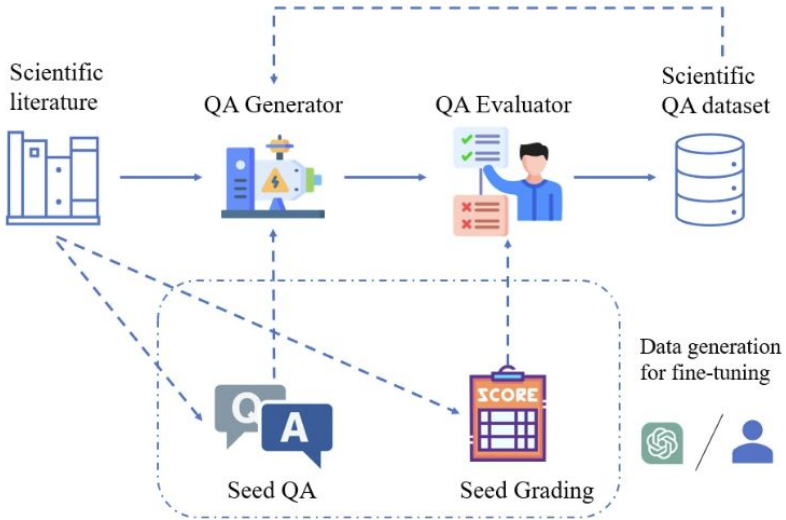
questions used in LLM **prompt** as example

↓

semi-manually **evaluation**

↓

Improvement - **multiple-choice question**

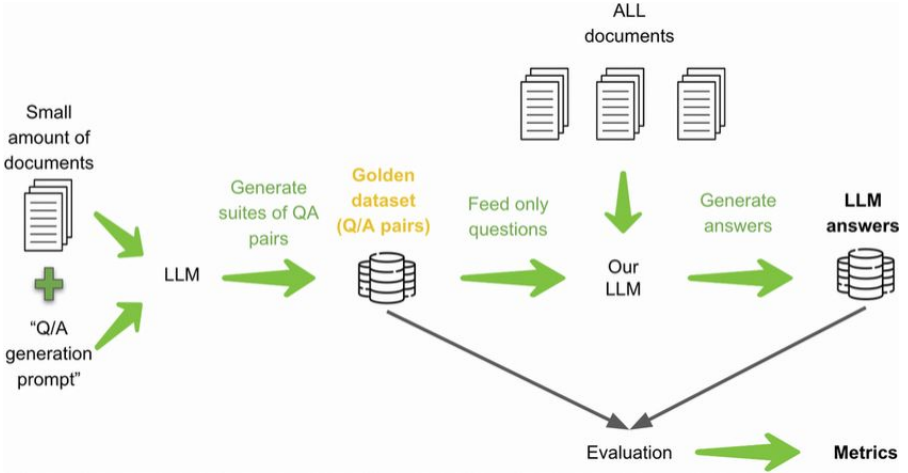


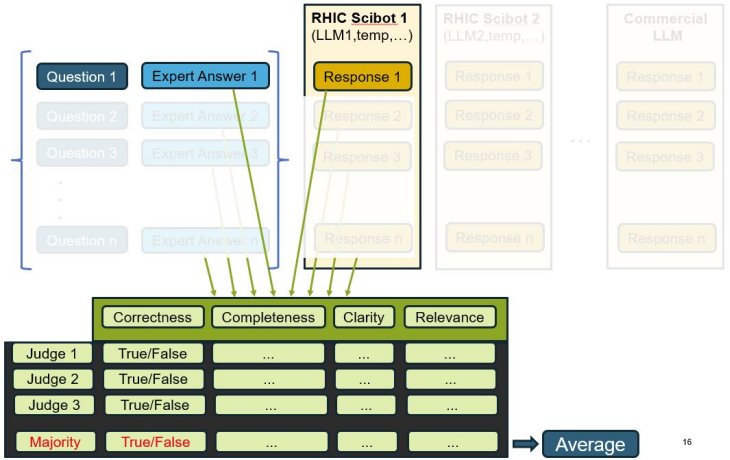
arxiv.org/abs/2405.09939v1

Can also be used in far **future** for fine-tuning LLM

Evaluation process - Scores

- **Subject Matter Expert** Rating (Human)
- **Lexical** metrics (ROUGE, BLEU..)
- **Semantic** similarity, BERTScore-like rating
- G-Eval: Ensemble of **LLM-as-a-Judge**
- Final - **weighted** average of all the above





Thank you for your attention!

Prompt example

Here is a scientific paper:

{*text*}

Here is a two-step task for you.

Step 1: Read given scientific paper and extract a list of 15 keywords focusing on the important terms and concepts within the paragraph. Avoid generic or broad words.

Step 2: Generate 10 scientific Q&A pairs as diverse as possible based on facts and knowledge presented in given paper, focusing on keywords you generated.

Keep the following requirements in mind: Avoid ask simple or definitional questions. Please assume that there is no corresponding paper to refer to when asking questions, so ensure that the questions and answers are self-contained (do not refer to external sources like figures or tables, or use demonstrative pronouns such as "this"). Incorporate specific data and insights from the paper to create detailed and informative answers. Ensure that the answers are concise, accurate, and directly related to the corresponding questions. Please present the generated keywords and question-answer pairs in the following format:

Keywords: [keyword 1], [keyword 2], ..., [keyword 15]

Q1: [Question 1] **A1**: [Answer 1]

Question example

Question: How are space charge effects considered in the STAR experiment?

Answer: Space charge effects in STAR primarily impact tracking drift detectors like the TPC. These effects are modeled using HIJING simulations, from which distortion maps are generated and validated using real collision data and known laser positions. Asymmetric distortions can arise in the TPC for asymmetric colliding systems, while space charge effects are typically negligible in fixed-target datasets. During offline reconstruction, the data is corrected by calibrating against the best correlation with luminosity as measured by the RICH scalers.

Question example

Question: What is the TOF resolution in STAR?

Answer The STAR TOF system provides a time resolution better than 80 picoseconds under optimal conditions, typically around 85–90 ps. It uses Multi-gap Resistive Plate Chambers (MRPCs) and the resolution depends on factors like channel calibration, run conditions, and detector stability.

Data processing (pdf) and deduplication

AdaParse: An Adaptive Parallel PDF Parsing and Resource Scaling Engine

https://arxiv.org/abs/2505.01435

LSHBloom: Memory-efficient, Extreme-scale Document Deduplication

https://arxiv.org/abs/2411.04257