Brainstorming

Instruction: Please write any ideas, topics, or issues that you believe are worth the group discussing. It can be a high-level concept, a question, or an unpopular opinion. Please do not delete existing points. If you already have thoughts on an existing point, feel free to write under the point to expand it. Organizers between the break will collect and organize the points for group discussion.

---

1. Effectiveness in LLMs:
    - Definition of effectiveness in RAG evaluation
    - Should the evaluation of retrieved documents be different than generated documents?

Rubric for evaluation:

- Rubric/questions to evaluate RAG/LLM
    - Granularity of nuggets. Yes/no? Simple facts? Complex but still facts? Something bigger?
    - How to build the rubric? Human or LLM?
- Is this the “same thing” as grading (eg exams)? Can we get any ideas from education? Can we export ideas to education?
    - Students get feedback from their teachers, should feedback be given back to systems? (I guess this is RL)

Evaluation criteria:

- How to merge different criteria in the evaluation? Treat them the same or differently?
    - Examples of criteria: faithfulness, factuality, etc.

2. Benchmark for LLMs:
    - Benchmark for RAG evaluation
    - Scaling issue with RAG benchmark

3. Future of IR/RAG:
    - RAG is dead — forget retrieval, just use LC-LLMs: make our datasets to be true long contexts

4. Personalization:
    - Evaluation of personalization in RAG
5. Role of users:
    - Role of users in RAG evaluation
    - Who are the users? Humans or agents?

6. Evaluating reasoning:
    - Measures for reasoning? How to evaluate the quality of the reasoning part?