
THE UNIVERSITY *of* EDINBURGH

# The Role of Information Retrieval in Deep Research

Chuan Meng

The University of Edinburgh

16 August 2026

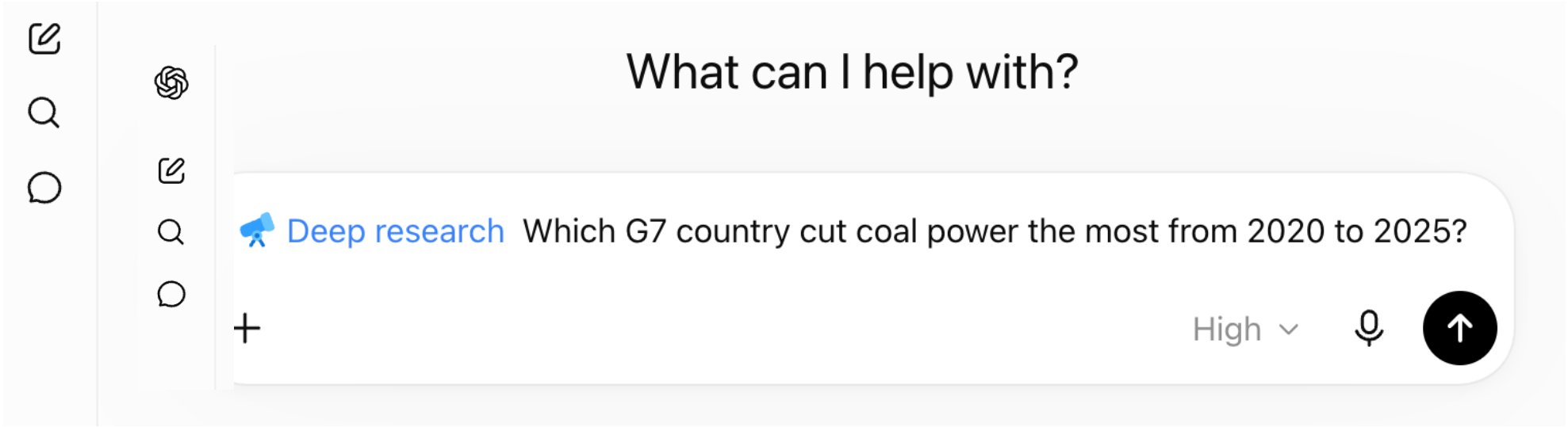
# About me

Photo with Maarten from my PhD defence (19/6/2025)

**Chuan Meng**

- Postdoc, University of Edinburgh (since 9/2025)
  - Funded by Dr. Jeff Dalton
- PhD, University of Amsterdam (10/2021 to 6/2025)
  - supervised by
    - Prof. Maarten de Rijke (ACM Fellow)
- Applied Scientist Intern, Amazon (8/2024 to 1/2025)
  - worked with Dr. Gabriella Kazai
- Research focus: **Agentic Search**
  - Search planning (When and how to search)
  - ↓
  - Query generation (What to search for)
  - ↓
  - Search result reflection (Did we find what we need)

# Background


What can I help with?
Deep research
Which G7 country cut coal power the most from 2020 to 2025?
High

# Background

- Deep research aims to answer hard, multi-hop, reasoning-intensive questions that require extensive open-web exploration.
  - It is extremely challenging [1]:
    - Humans gave up on 70% of questions after two hours of searching
    - LLMs achieve near-zero accuracy

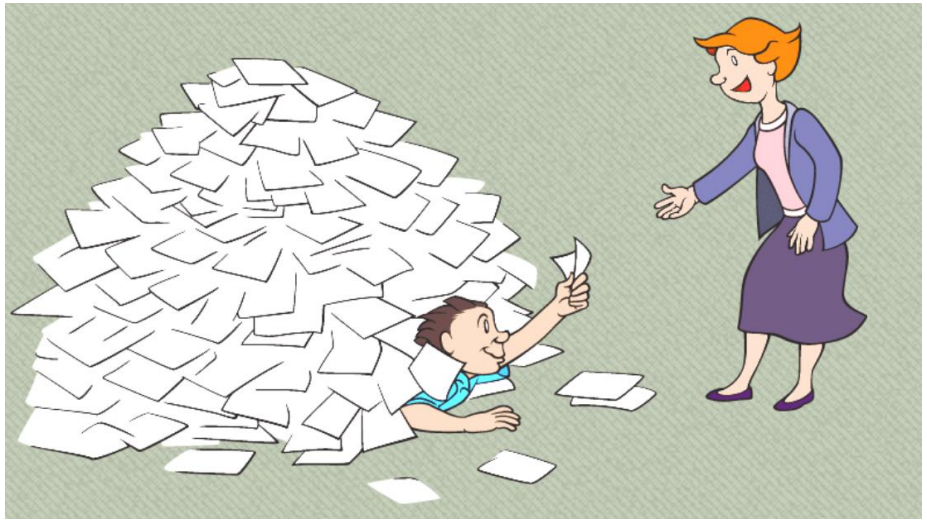
| Model | Accuracy (%) |
|---|---|
| GPT-4o | 0.6 |
| GPT-4.5 | 0.9 |
| OpenAI o1 | 9.9 |

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. 2025.

# Background

- Deep research can free up valuable time for humans who engage in intensive knowledge work (e.g., finance, science, policy, and engineering), where thorough, precise, and reliable research is critical

High-stakes purchase decisions
(e.g., cars, appliances, furniture)



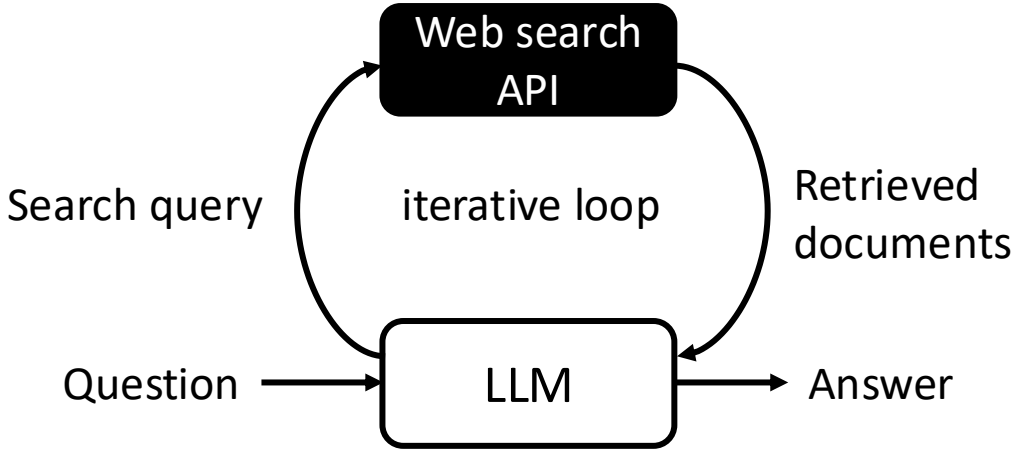
Finding needle-in-a-haystack information requiring browsing many websites

[1] OpenAI. Introducing Deep Research. 2025.

# Motivation

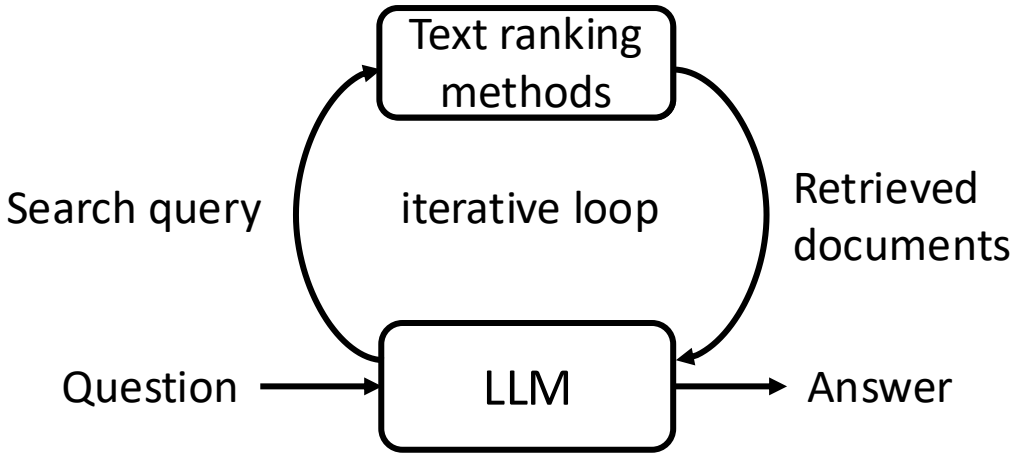
- Existing work typically uses large language model (LLM)-based agents to invoke web search APIs iteratively before producing a final answer

- However, web search APIs are black boxes:
  - lacking transparency
  - preventing systematic analysis of search components
  - leaving the performance of established text ranking methods unclear in deep research



# Method

- We examine key findings and best practices for established text ranking methods in deep research [1]




- Three perspectives:
  1. Retrieval and information units (passages vs. documents)
     - Passages are more token-efficient, simplify document length normalisation…
  2. Re-ranking
     - Does re-ranking still matter when the content consumer is an LLM-based agent?
  3. Query mismatch
     - Agent queries may not match rankers' expected input: natural-language queries

**[1]** Meng et al. **Revisiting Text Ranking in Deep Research**. **SIGIR 2026**.

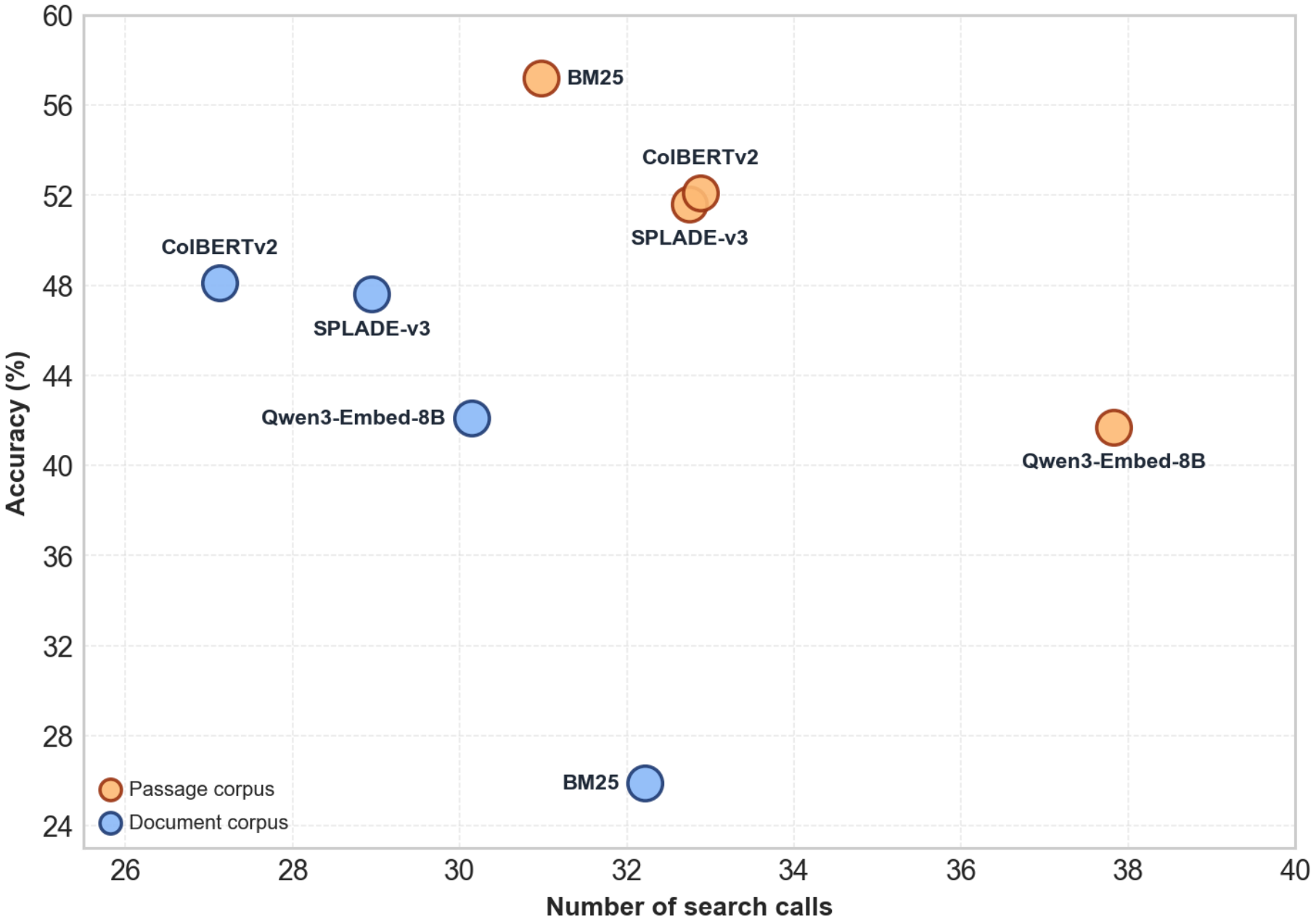
# Part 1
# Retrieval and information units (passages vs. documents)

# Retrieval units (passages vs. documents)

- Retrievers
  - Lexical-based sparse: BM25 [1]
  - Learned sparse: SPLADE-v3 (110M) [2]
  - Single-vector dense: Qwen3-Embedding (8B) [3]
  - Multi-vector dense: ColBERTv2 (110M) [4]

[1] Robertson et al. Okapi at TREC-3. 1994.
[2] SPLADE-v3: New baselines for SPLADE. 2024
[3] Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. 2025
[4] Santhanam et al. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. 2022.

# Retrieval units (passages vs. documents)




Passages vs. documents

- Context-window hit rate
  - Docs~20%, Psg~90%
- Passages allow more search iterations before hitting context-window limits, resulting in higher answer accuracy

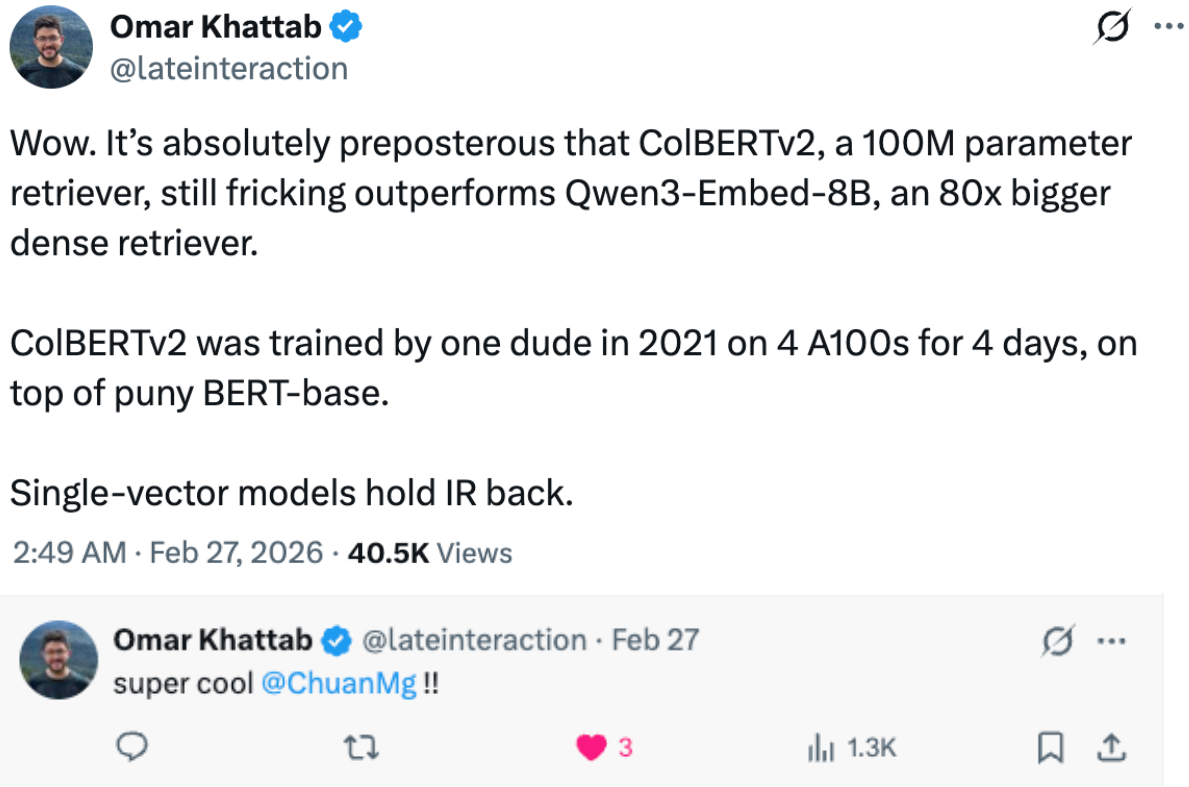
Retriever comparison

- Lexical BM25 (1994) performs best
- ColBERT ≳ SPLADE, both outperforming single-vector Qwen3-Embeding-8B

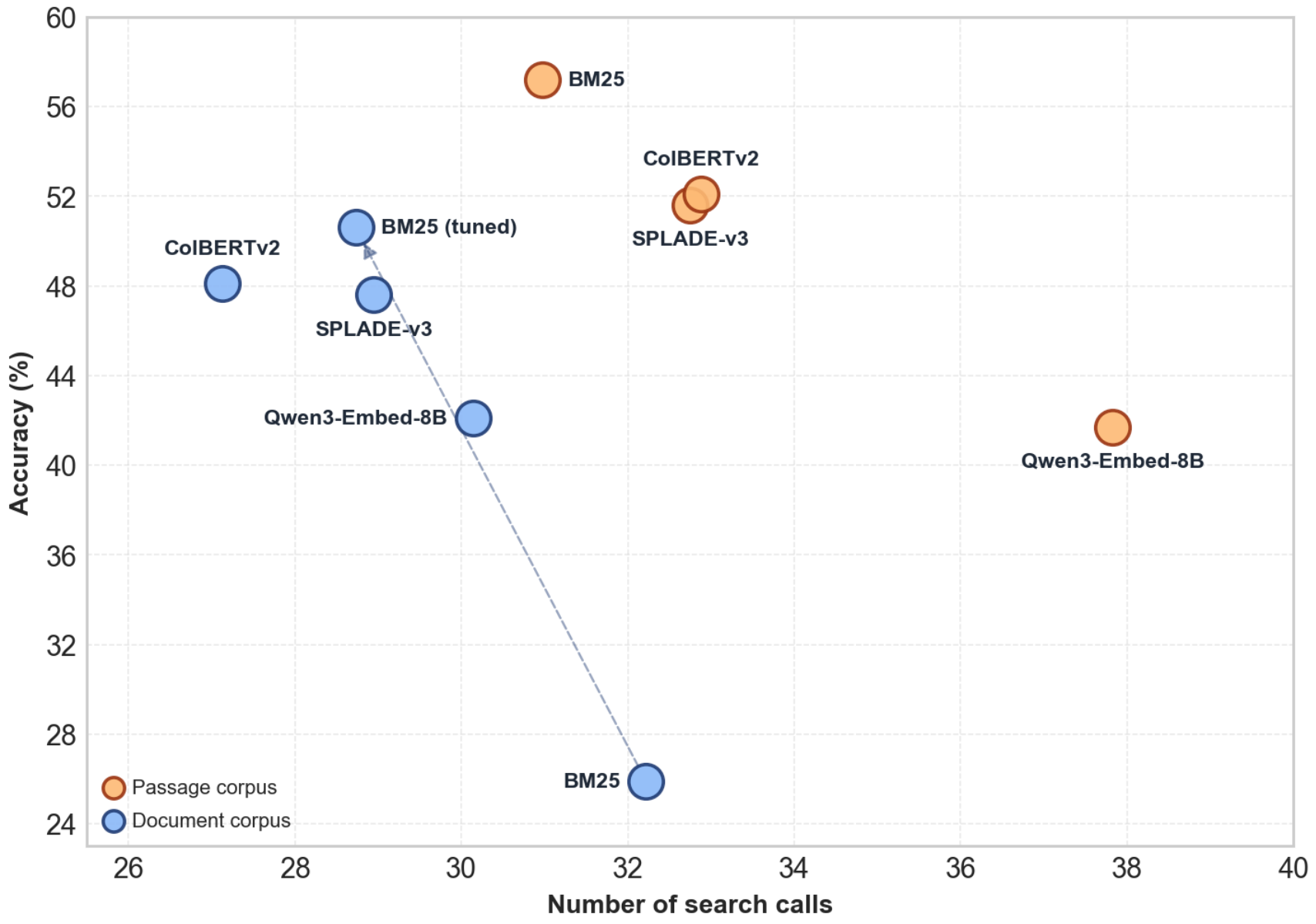
# Retrieval units (passages vs. documents)

- **Agent-issued queries typically follow a web-search style with keywords, phrases, and quotation marks for exact matching**
  - Sparse retrievers are a natural fit
  - Multi-vector retrievers preserve token-level signals, making them more robust

| Agent | Search query |
|---|---|
| gpt-oss-20b | “90+7” attendance 61700<br>“Man United” “4-1” “90+4”<br>“assist” “4–1” “Premier League” “2020” |
| GLM-4.7-Flash | “90’+4” football match attendance<br>“61,888” football match Stockholm Vienna Prague<br>“goal in the 6th minute” football |

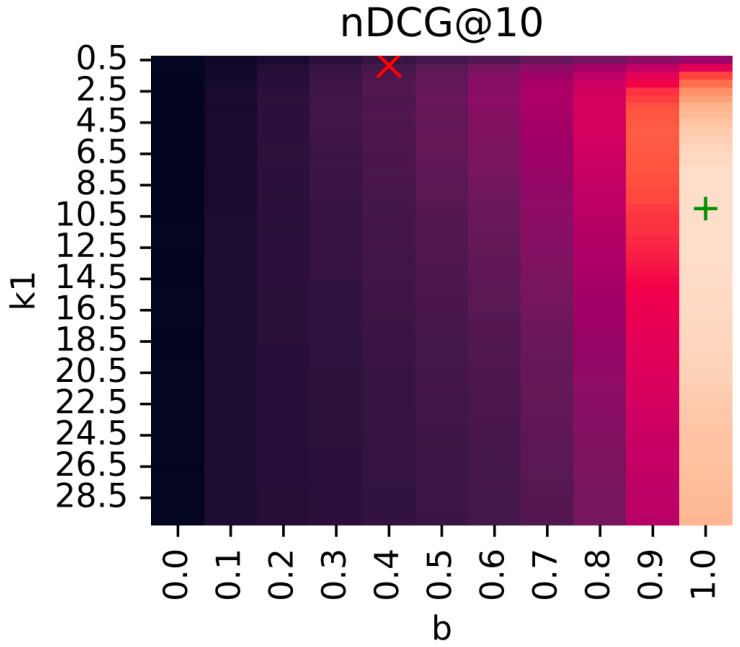


# Retrieval units (passages vs. documents)




Why BM25 perform poorly on documents?

- Default hyperparameters (k1 = 0.9, b = 0.4) [1] limit performance
- Proper hyperparameters DOUBLE performance




[1] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. 2025.

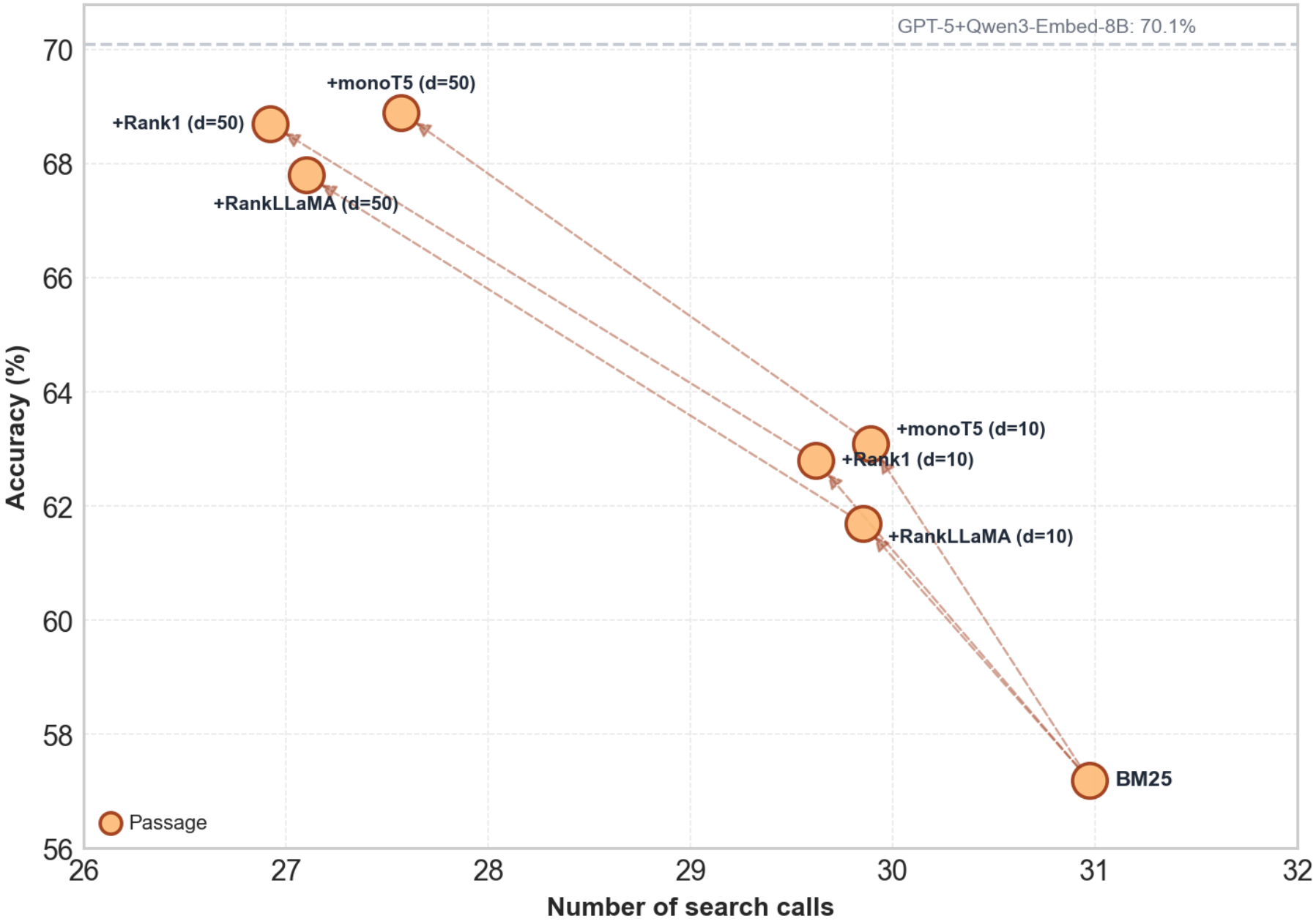
# Part 2
# Re-ranking in deep research

# Re-ranking

- Re-rankers
  - monoT5-3B [1]
  - RankLLaMA-7B [2]
  - Rank1-7B  [3]

[1] Document Ranking with a Pretrained Sequence-to-Sequence Model. 2020.
[2] Fine-Tuning LLaMA for Multi-Stage Text Retrieval. 2023.
[3] Rank1: Test-Time Compute for Reranking in Information Retrieval. 2025

# Re-ranking




Re-ranking improves effectiveness and reduces search iterations

- BM25+monoT5 (d=50) (acc. 0.689) approaches GPT-5 (acc. 0.701)
- deeper depths amplify gains

# Part 3
# Query mismatch bridging

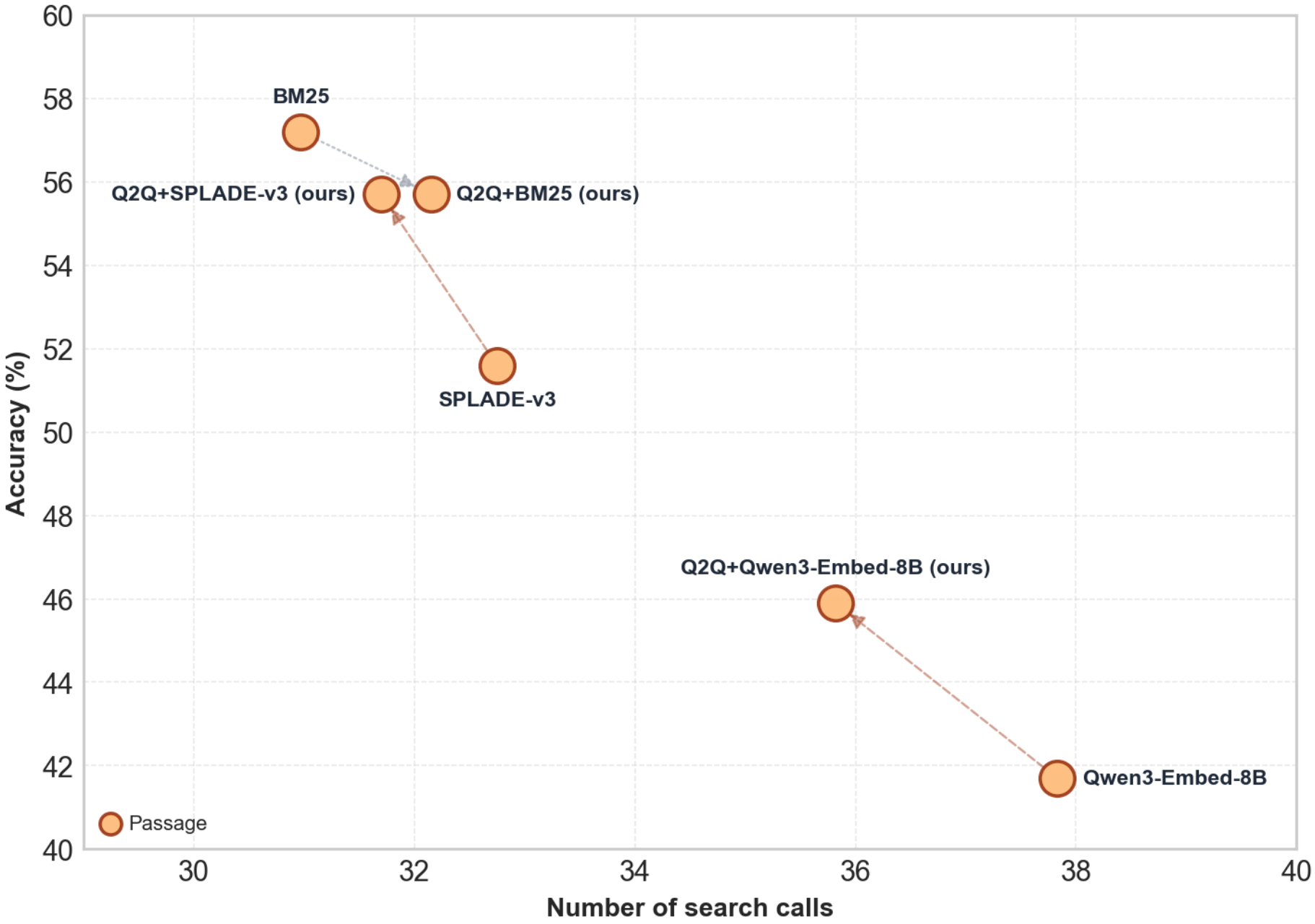
# Query mismatch bridging

- We propose **Query-to-Question (Q2Q)** method that translates agent-issued queries into natural language questions
- Challenge: Using only the raw agent query tends to shift search intent
- Solution: Add the agent's recent reasoning trace to preserve the intended search intent

The agent's recent reasoning trace:
"…Let's recall some football matches that had an attendance of exactly 61,880…

| | Search query |
|---|---|
| Raw query | "61,880" football attendance |
| Raw query→Q2Q | What is the football attendance number 61,880? |
| Raw query+reasoning→Q2Q | What football match had an attendance of 61,880? |

# Query mismatch bridging




- Q2Q significantly enhances neural retrievers
- BM25 performs worse with natural-language questions

# Summary

- Main Takeaways: **Traditional and established IR methods are not obsolete in deep research**
  1. Agent-issued queries follow a web-search style; sparse/multi-vector retrievers effective
     - *BM25 performs best*
  2. Passage-level units are more effective and token-efficient than document-level units
  3. Re-ranking improves performance and reduces search iterations; the deeper, the better
     - *20B agent with BM25 + monoT5 matches GPT-5 + Qwen3-Embedding-8B*
  4. Translating agent-issued queries into natural language questions enhances neural ranking

- Resource
  - Code, passage corpus, indexes, and agent traces are open-source

- Future work
  - Improve LLM-based agents to issue more effective search queries
  - Train text ranking methods on synthetic deep research data

Paper Code

**Thank you!** **Chuan Meng** @ChuanMg