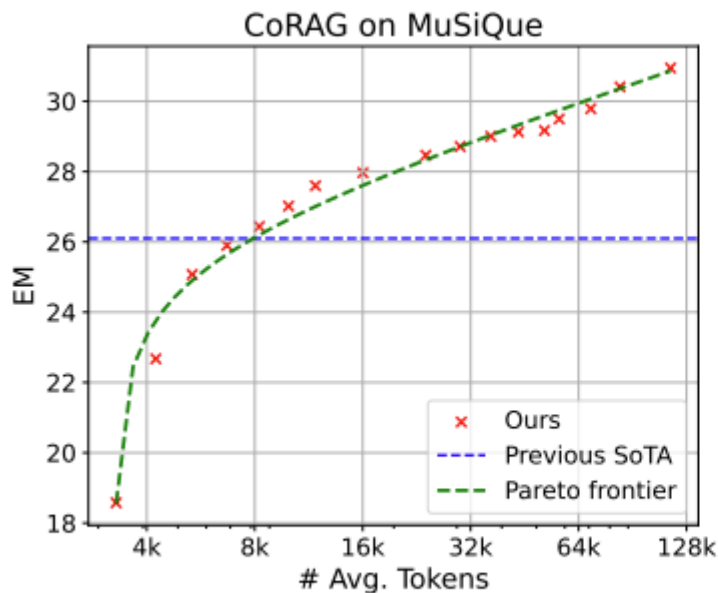


# Chain-of-Retrieval Augmented Generation

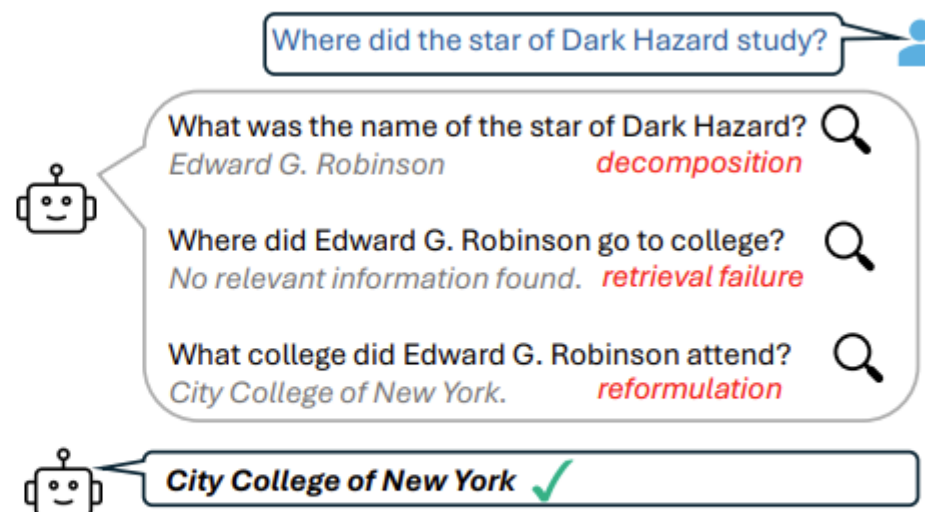
**Liang Wang**, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, Furu Wei  
Microsoft Research

# Motivation

- Break the bottleneck of imperfect retrieval results
- Explore test-time scaling for RAG

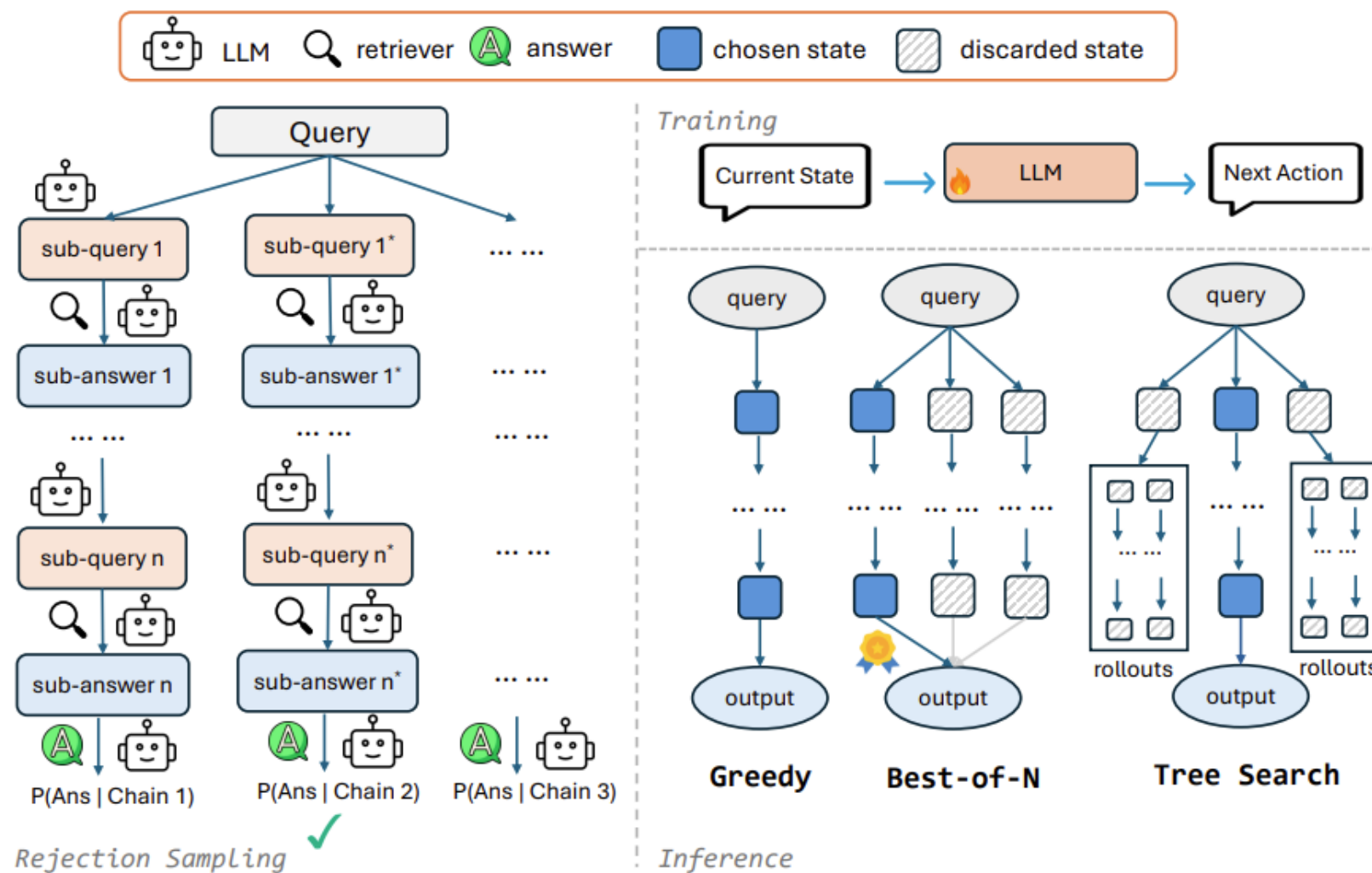


(a) Test-time scaling behavior of CoRAG.



(b) An example of CoRAG in action.

# CoRAG



# CoRAG for Multi-Hop QA

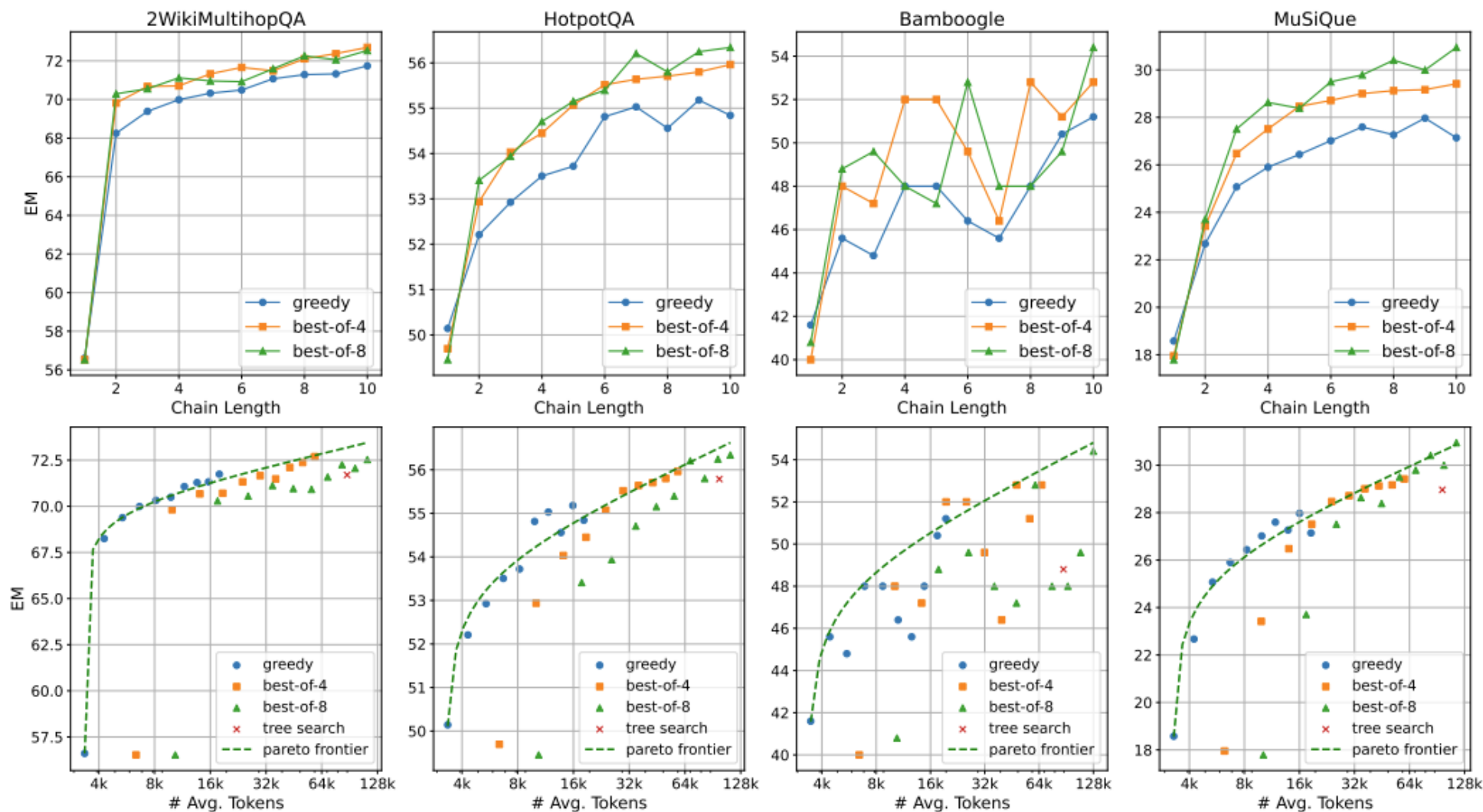
|                       | 2WikiQA     |             | HotpotQA    |             | Bamboogle   |             | MuSiQue     |             |
|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                       | EM          | F1          | EM          | F1          | EM          | F1          | EM          | F1          |
| Previous SoTA         | 58.0        | 71.4        | 50.3        | 63.5        | <b>56.0</b> | 67.8        | 26.1        | 42.0        |
| CoRAG-8B (Ours)       |             |             |             |             |             |             |             |             |
| ▷ $L=1$ , greedy      | 56.5        | 62.3        | 50.1        | 63.2        | 37.6        | 51.4        | 18.6        | 29.3        |
| ▷ $L=6$ , greedy      | 70.6        | 75.5        | 54.4        | 67.5        | 48.0        | 63.5        | 27.7        | 38.5        |
| ▷ $L=6$ , best-of-4   | 71.7        | 76.5        | 55.3        | 68.5        | 51.2        | 63.1        | 28.1        | 39.7        |
| ▷ $L=6$ , tree search | 71.7        | 76.4        | 55.8        | 69.0        | 48.8        | 64.4        | 29.0        | 40.3        |
| ▷ $L=10$ , best-of-8  | <b>72.5</b> | <b>77.3</b> | <b>56.3</b> | <b>69.8</b> | 54.4        | <b>68.3</b> | <b>30.9</b> | <b>42.4</b> |

# KILT Benchmark

Table 2: The downstream results on the *hidden test set* of the KILT benchmark. All scores are sourced directly from the official leaderboard, with the exception that “RA-DIT 65B” is from the original paper [23]. \*: “Previous Best” refers to the highest score for each task on the public KILT leaderboard as of January 10, 2025.

| System          | Entity Linking |             |             | Slot Filling |             | Open QA     |             |             | Fact        |
|-----------------|----------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|
|                 | AIDA           | WnWi        | WnCw        | T-REx        | zsRE        | NQ          | HoPo        | TQA         | FEVER       |
| KILT-RAG        | 72.6           | 48.1        | 47.6        | 59.2         | 44.7        | 44.4        | 27.0        | 71.3        | 86.3        |
| SEAL            | -              | -           | -           | 83.6         | 74.6        | 53.7        | 40.5        | 70.9        | 89.5        |
| Atlas-11B       | 90.6           | -           | -           | 85.1         | 80.8        | 61.3        | 50.6        | 84.0        | <b>93.5</b> |
| RA-DIT 65B      | 80.5           | -           | -           | 72.8         | 78.1        | 43.5        | 36.6        | 72.8        | 86.9        |
| FiD with RS     | -              | -           | -           | 85.2         | 83.7        | 61.2        | 39.1        | 84.6        | 92.2        |
| Previous Best*  | 90.6           | 87.4        | 71.2        | 87.7         | 85.3        | 62.3        | 50.6        | 84.6        | <b>93.5</b> |
| CoRAG-8B (Ours) | <b>93.9</b>    | <b>88.2</b> | <b>76.7</b> | <b>88.0</b>  | <b>87.2</b> | <b>63.1</b> | <b>60.6</b> | <b>88.3</b> | 93.1        |

# Test-time Scaling Behavior



# Resources

- Code, data, and trained models
  - Available at <https://github.com/microsoft/LMOps/tree/main/corag>
- Any questions?
  - Please contact [wangliang@microsoft.com](mailto:wangliang@microsoft.com)