

CSCE 670 - Information Storage and Retrieval

Week 13: Wrap-Up

Yu Zhang

yuzhang@tamu.edu

Course Website: https://yuzhang-teaching.github.io/CSCE670-F26.html

Quiz 4!

- All policies are the same as Quiz 1 (number of questions, time limit, grading, etc.)
- Scope:
  - Week 10 (ColBERT, Neural Collaborative Filtering, Sequential Recommendation)
  - Week 11 (LLMs Basics, LLMs with Search Engines)
  - Week 12 (LLMs for Ranking and Recommendation)
  - Homework 3
  - None of the content in Week 13 slides will be covered.

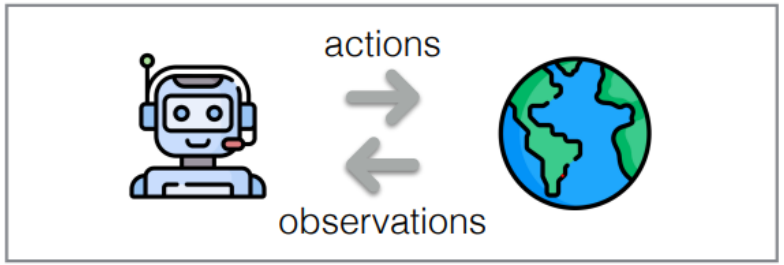
Extended Content: Agentic Search
(will not appear in quizzes or the exam)

LLM Agents: From Words to Action

- LLMs predict textual output for a given input

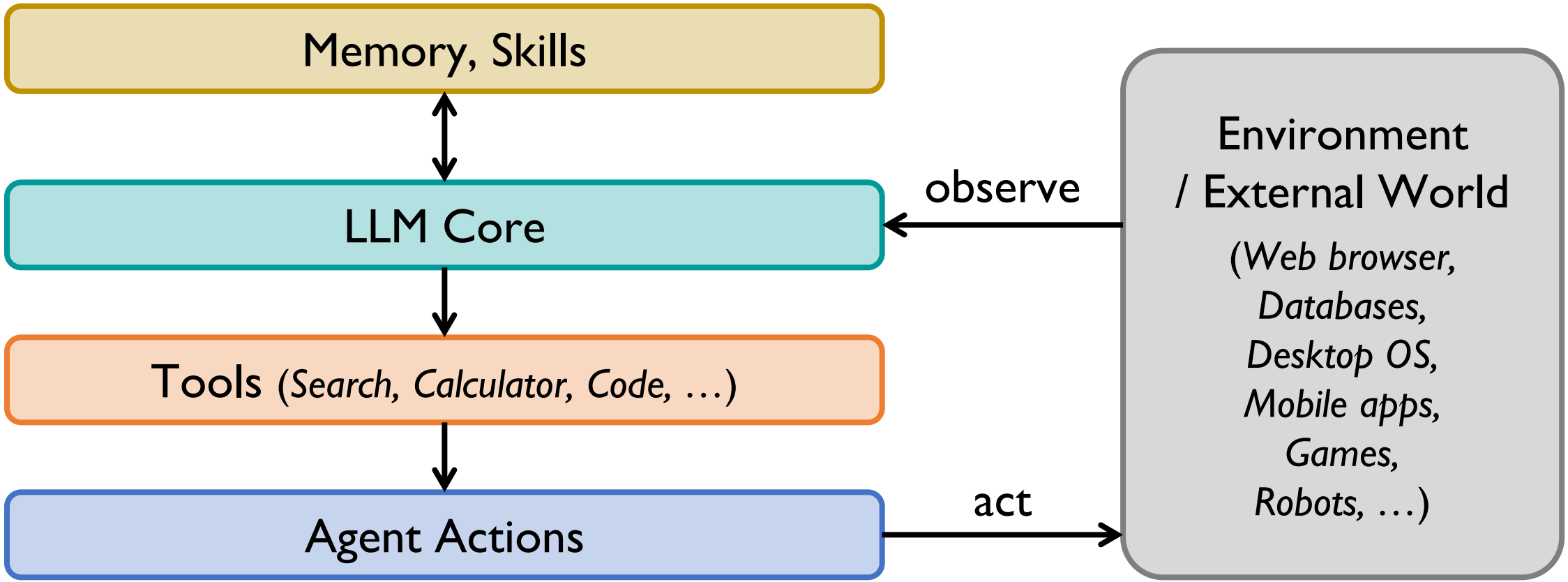
P(text)

- Agents perform actions based on observations from the environment/world



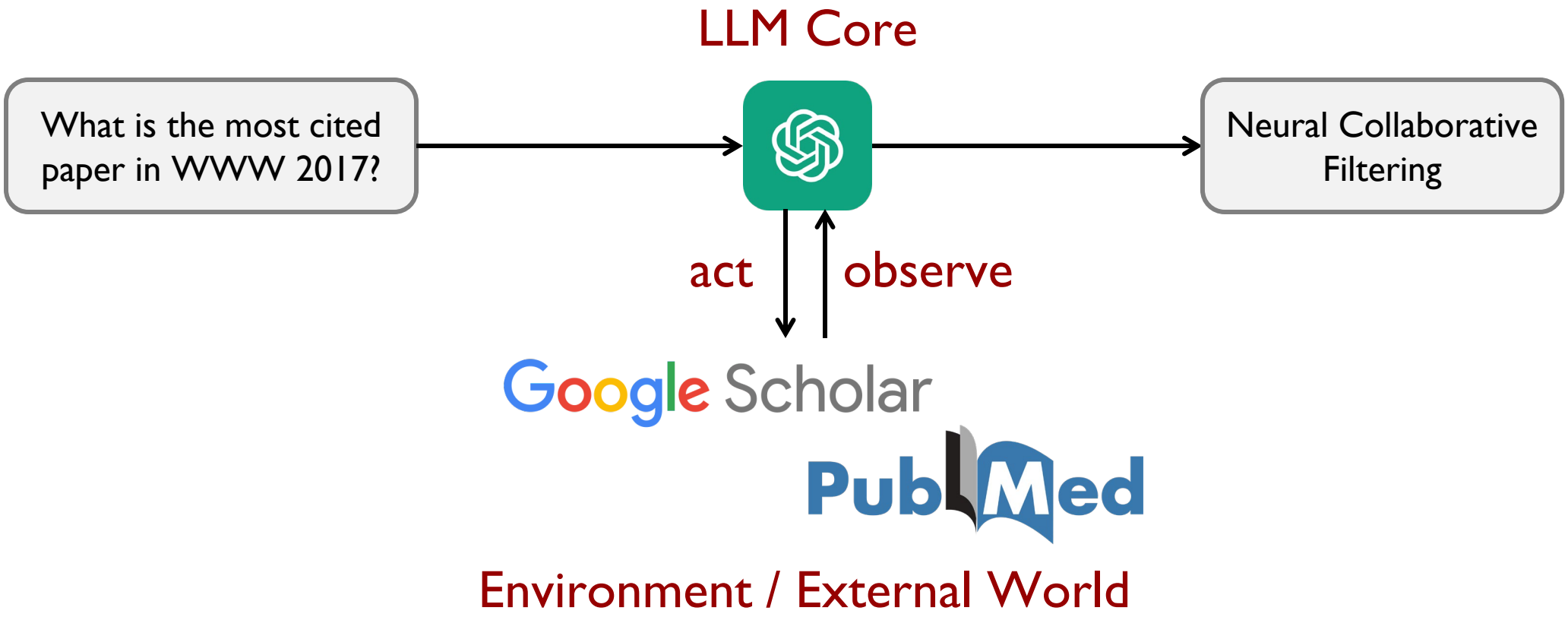
- How to combine them?
  - Lots of debate on what is an LLM agent and what is not an LLM agent
  - There is a degree of consensus that several key components are needed

LLM Agents: Key Components



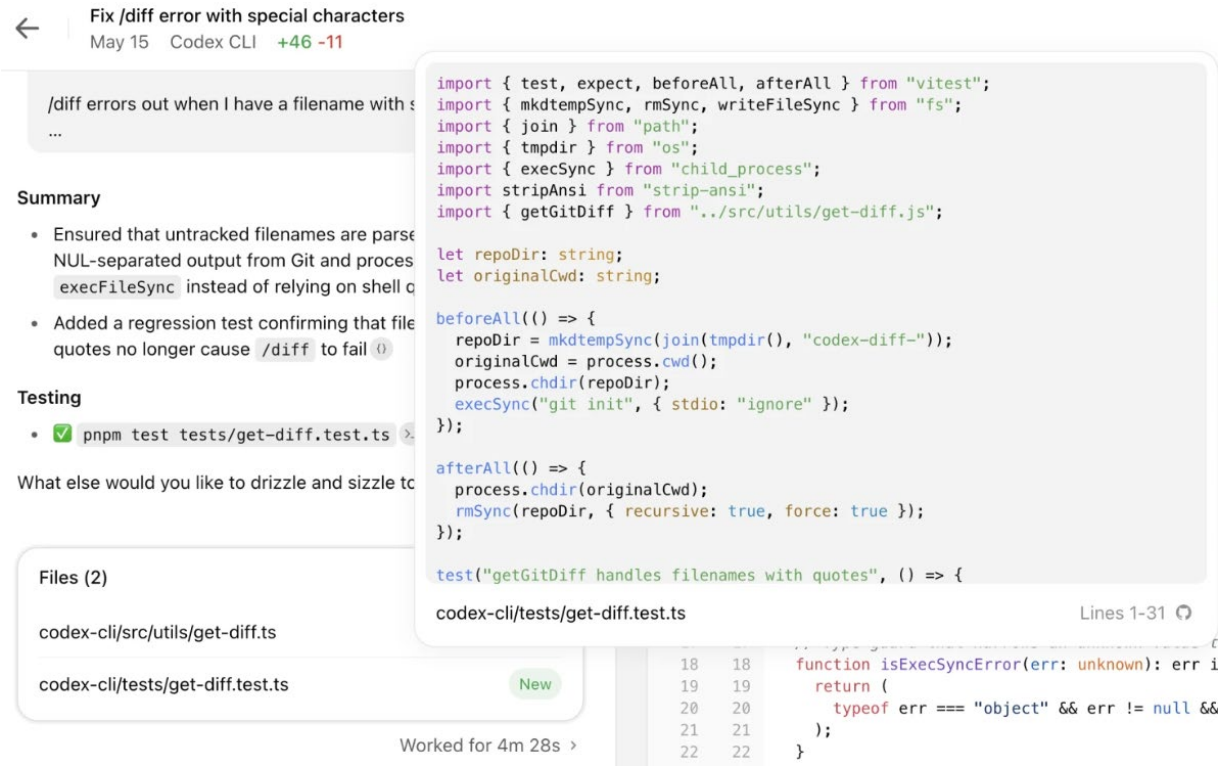
Multi-Round, Agentic RAG as a Special Case

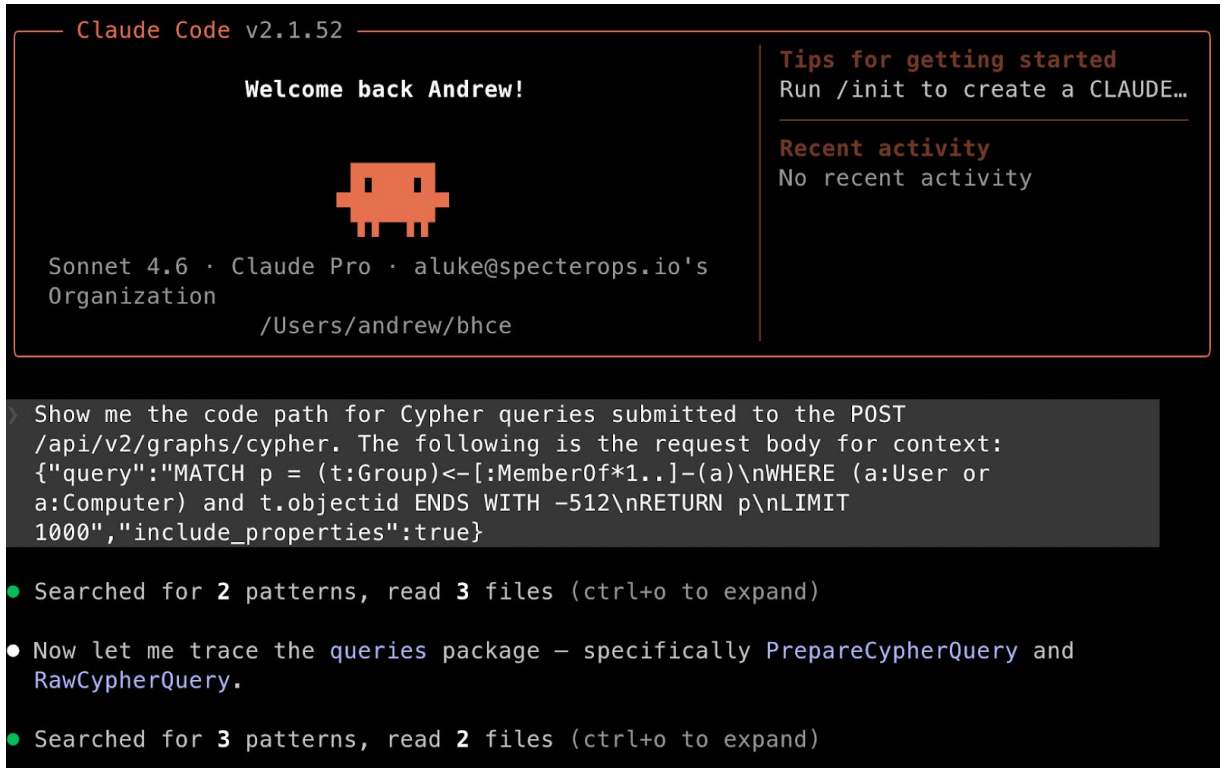
- What is an action (i.e., "tool call")?
  - In a typical RAG workflow: BM25 or neural retriever (e.g., ColBERT) or Google API



Multi-Round, Agentic RAG as a Special Case

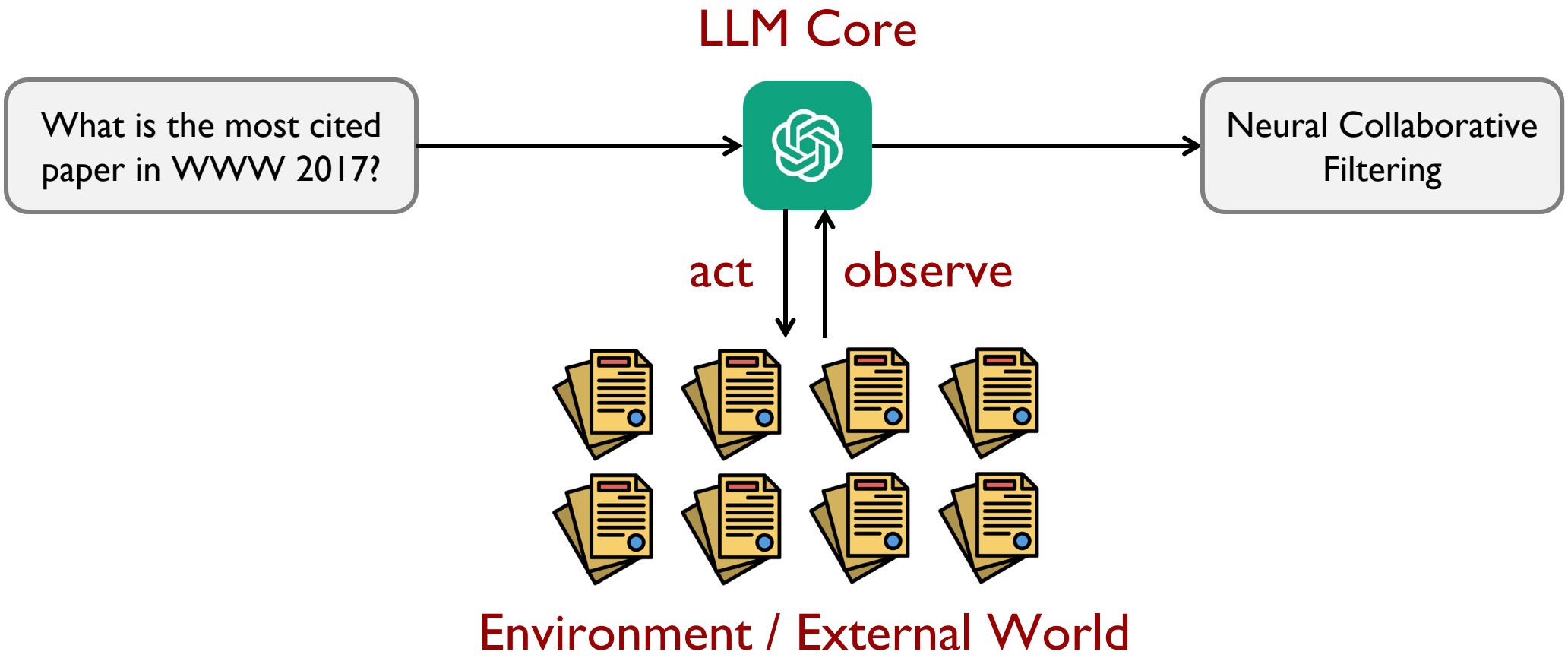
- Are there "better" options for the retrieval tools that we equip agents with?
- How do OpenAI Codex and Claude Code handle this?





Multi-Round, Agentic RAG as a Special Case

- What is an action (i.e., "tool call")?
  - How about command-line tools (e.g., grep, find, head, …)?



Direct Corpus Interaction (DCI) [Li et al., NeurIPS 2026]

Beyond Semantic Similarity: Rethinking Retrieval for Agentic Search via Direct Corpus Interaction

Zhuofeng Li[1,*,†] **Haoxiang Zhang**[2,3*] **Cong Wei**[2,*] **Pan Lu**[4,*] **Ping Nie**[2,†]
Yi Lu[2] **Yuyang Bai**[1] **Shangbin Feng**[5] **Hangxiao Zhu**[1] **Ming Zhong**[6] **Yuyu Zhang**[7]
Jianwen Xie[8] **Yejin Choi**[4] **James Zou**[4] **Jiawei Han**[6] **Wenhu Chen**[2] **Jimmy Lin**[2]
Dongfu Jiang[2,✉,†] **Yu Zhang**[1,✉]

[1]Texas A&M University [2]University of Waterloo [3]UC San Diego [4]Stanford University
[5]University of Washington [6]University of Illinois Urbana-Champaign [7]Verdent AI [8]Lambda

https://github.com/DCI-Agent/DCI-Agent-Lite

Examples of Command-Line Tools: find and head

- Each paper is a text file in papers/
- A quick preview helps the agent decide what to inspect next.

bash ~/corpus

```
$ find papers/ -name "*.txt"
papers/paper_01.txt
papers/paper_02.txt
papers/paper_03.txt

$ head -n 3 papers/paper_02.txt
Title: Learning to Use Search Tools
Abstract:
We study agents that answer questions using tools.
```

Find files

Search by filename. *.txt matches text files.

Preview a file

Read the first 3 lines. Start with the title and abstract.

Examples of Command-Line Tools: grep

- Search for a phrase, then read the surrounding evidence.
- A keyword match identifies a candidate. Context explains what it says.

```
bash  ~/corpus

$ grep -r -l "reinforcement learning" papers/
papers/paper_03.txt
papers/paper_02.txt


$ grep -n -C 1 "reinforcement learning" papers/paper_02.txt
7-The agent chooses when to call a search engine.
8:We train the policy with reinforcement learning.
9-The reward combines answer correctness and search cost.
```

Candidate files

-r search subfolders
-l filenames only

Nearby evidence

-n line numbers
-C 1 context on both sides

Examples of Command-Line Tools: Pipes

- The | symbol passes one command's output to the next command.
- These tools do not seem much more powerful than Boolean retrieval?

```
bash  ~/corpus

$ grep -r -l "reinforcement learning" papers/ | head -n 1
papers/paper_03.txt


$ grep -r -l "reinforcement learning" papers/ | wc -l
2
```

Limit output

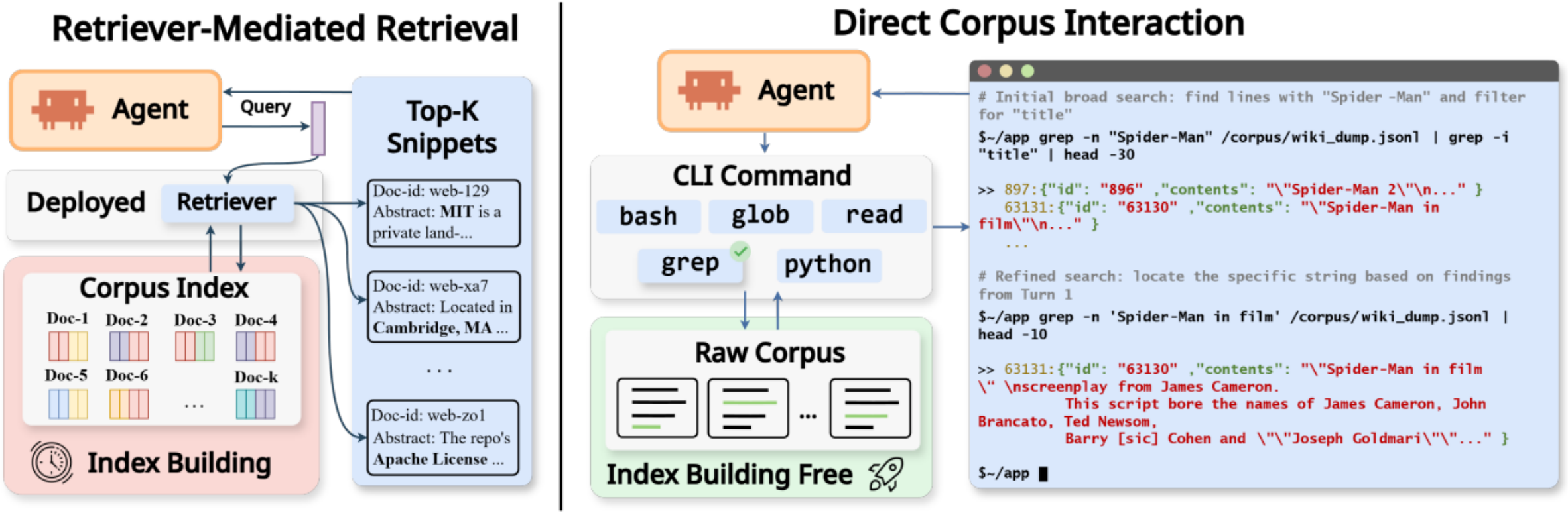
Keep only the first matching filename.

Count files

wc -l counts lines. Each filename occupies one line.

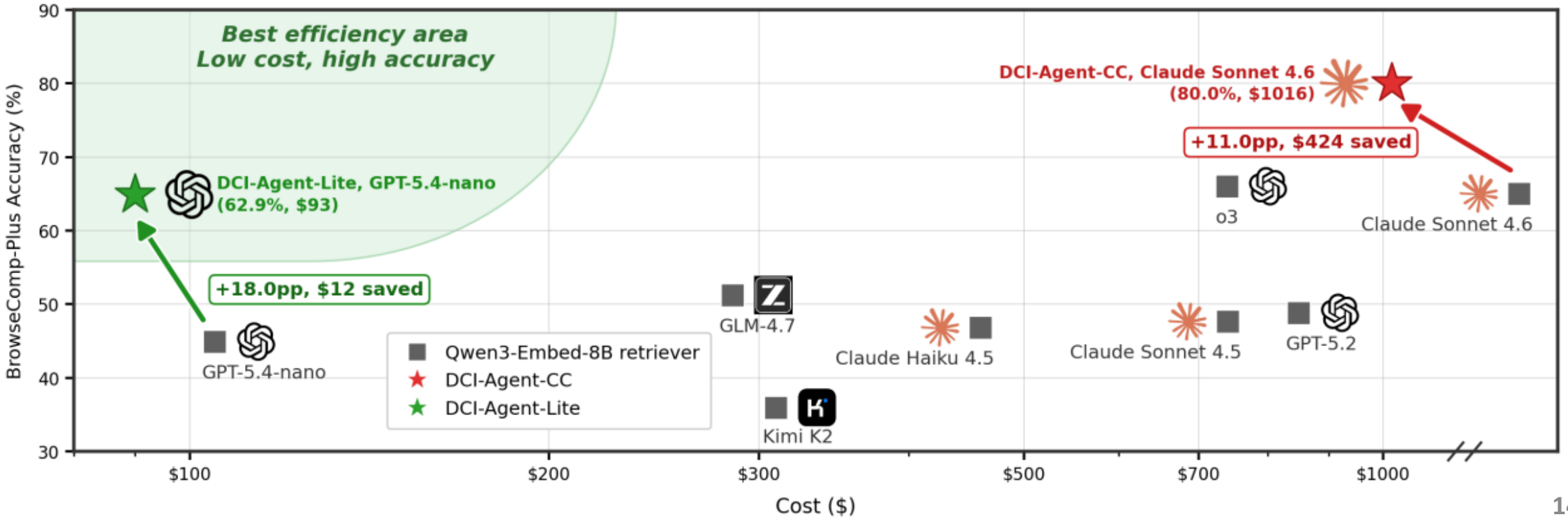
Command-line tools are all you need!

- A sufficiently capable LLM can reliably chain seemingly simple tools (e.g., grep, find, head, …) to perform complex tasks.
- These tools operate directly on text, offering greater flexibility than embedding-based retrieval.
  - And they do NOT require a prebuilt index.



Performance on an Agentic RAG Benchmark

- We implemented DCI in two harnesses: Pi (DCI-Agent-Lite) and Claude Code (DCI-Agent-CC).
- Dataset: BrowseComp-Plus, 830 questions over 100K+ documents, averaging 20+ search calls per question.



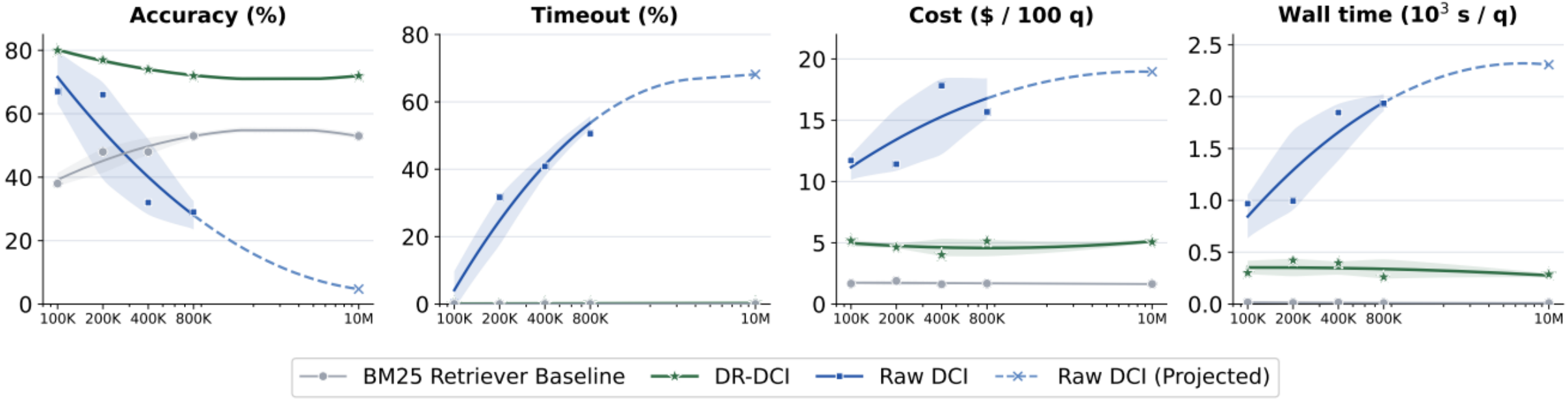
Performance on Document Retrieval Benchmarks

- Evaluation metric: NDCG@10

Method	BRIGHT				BEIR		Summary	
	Bio.	Earth.	Econ.	Robotics	ArguAna	SciFact	Avg.	Δ Avg.
Sparse & Dense Retrieval								
BM25	18.9	27.2	14.9	13.6	31.5	15.8	20.3	↓ 26.7
OpenAI-text-emb-3-large	23.3	26.7	19.5	12.8	58.1	58.1	33.1	↓ 13.9
GTE-Qwen2-7B-Instruct	30.6	36.4	17.8	13.2	62.7	75.3	39.3	↓ 7.7
Rank-R1-14B	31.2	38.5	21.2	22.6	31.3	72.2	36.2	↓ 10.8
Rank1-32B	49.7	35.8	22.0	22.5	57.6	74.8	43.7	↓ 3.3
ReasonRank-32B	58.2	48.9	<u>36.6</u>	33.9	28.7	<u>75.5</u>	47.0	–
DCI Agents								
DCI-Agent-Lite (GPT-5.4 nano)	<u>60.0</u>	<u>50.8</u>	32.3	<u>42.4</u>	<u>81.9</u>	72.7	<u>56.7</u>	↑ 9.7
DCI-Agent-CC (Sonnet 4.6)	**77.1**	**69.0**	**46.8**	**56.8**	**85.3**	**75.7**	**68.5**	↑ 21.5

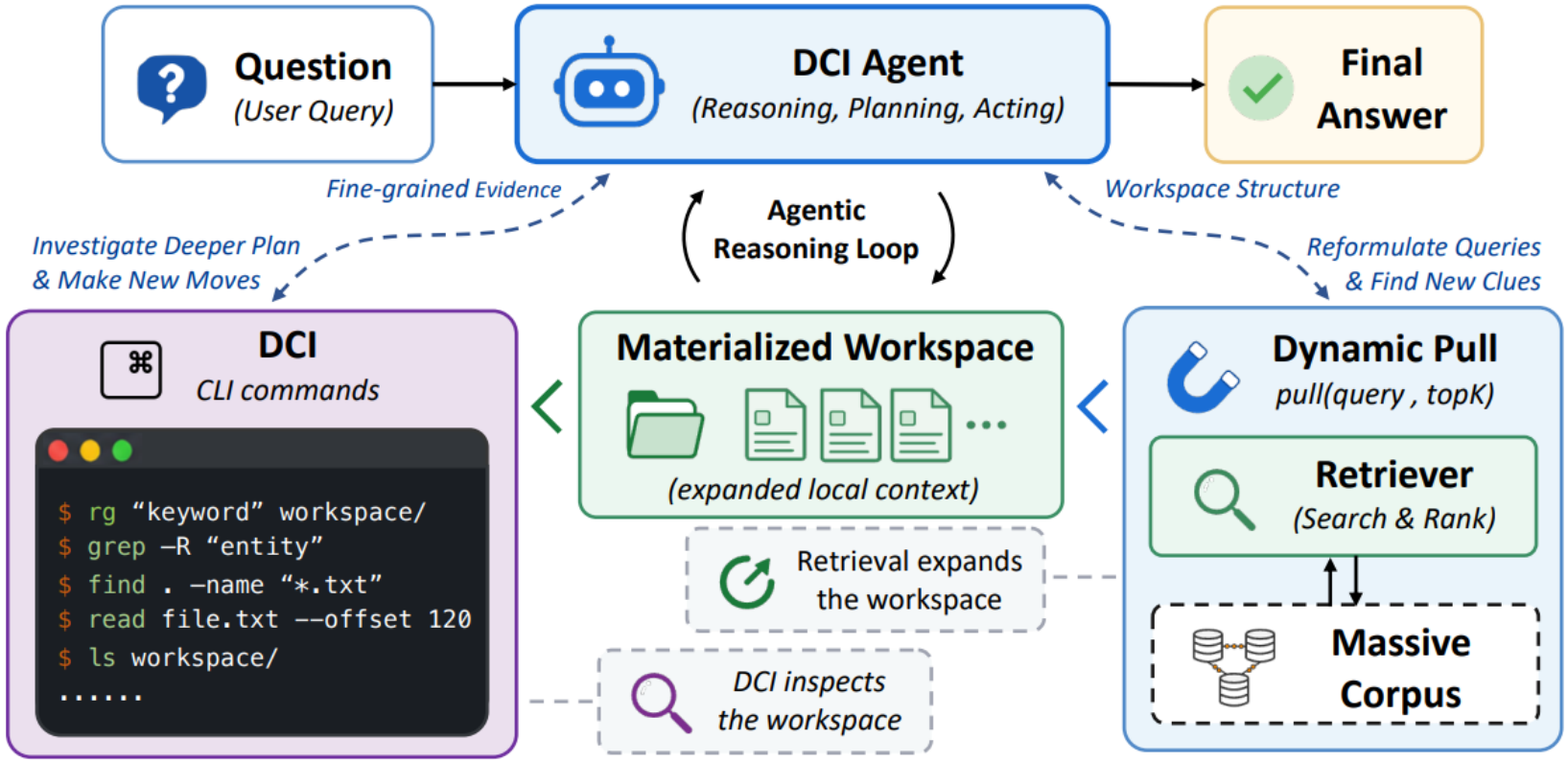
Potential Limitations of DCI

- The corpus must be stored locally.
- The corpus cannot be too large.
  - As the corpus grows from 100K to 800K documents, DCI's accuracy drops from 67% to 29%, average time per question doubles from 16 to 32 minutes, and over half of tool calls fail.



How can we scale DCI to larger corpora? [Lu et al., arXiv 2026]

- Retrieve: pull relevant documents into a local workspace (e.g., using BM25)
- Investigate: use terminal tools to search, compare, and verify evidence
- Expand: retrieve more documents as new clues emerge



Logistics for Final Exam

Final Exam (Overall Information)

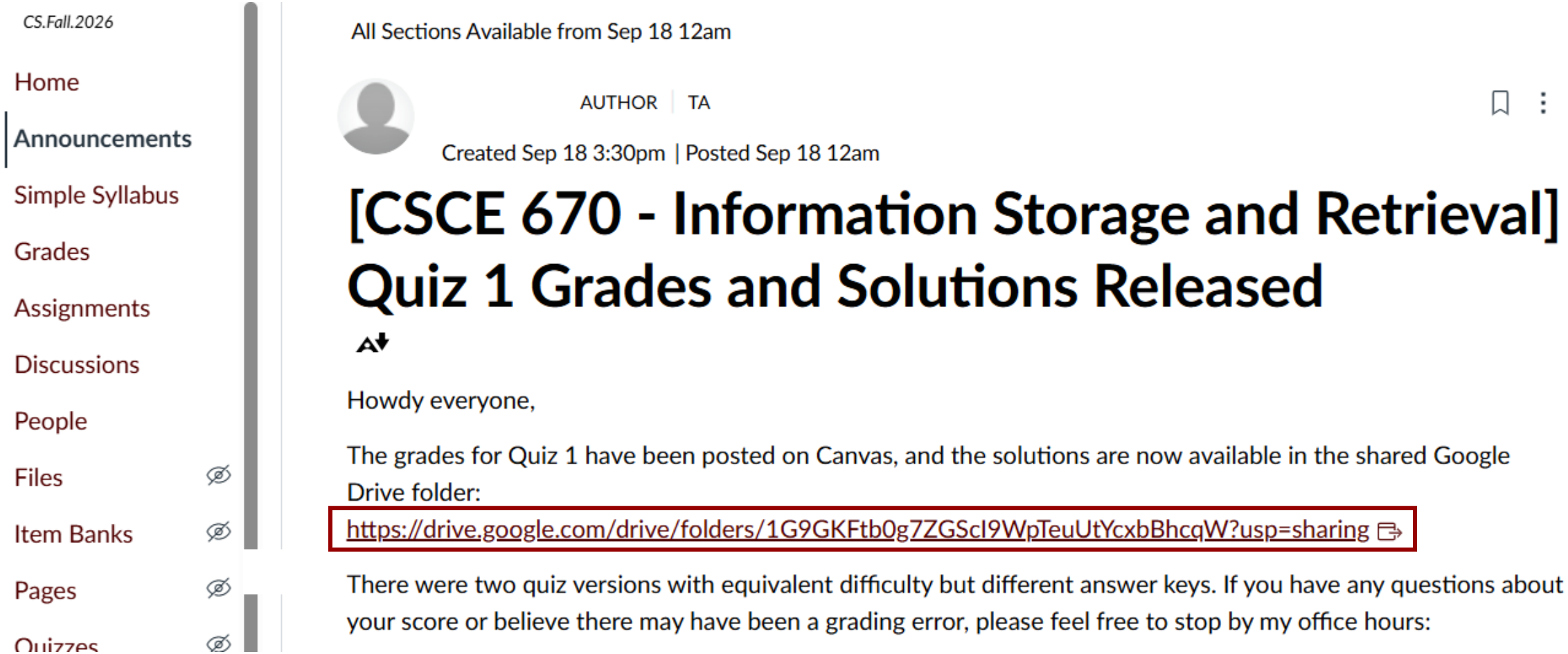
- Time: 8am – 10am on Dec 9, 2026 (Wednesday)
- Location: In our regular classroom
- Scope: Comprehensive! Covering all materials from Week 1 to Week 12, BUT
  - Slides with a gray title will NOT be on the final exam
  - Questions related to homework assignments will NOT be on the final exam

Final Exam (Format)

- 30 multiple-choice questions, 1% each
  - In total, the final exam accounts for 30% of your overall grade.
- 15 of the questions will be VERY SIMILAR to those that appeared in Quiz 1-Quiz 4
  - Test your "few-shot" problem-solving skills
  - Carefully review all quizzes (both versions), but you can skip the last question, as the final exam does NOT cover the homework assignments
- The other 15 questions will not have appeared on any quizzes, but their difficulty will be comparable to the quiz questions
  - Test your "zero-shot" problem-solving skills
  - Conceptual questions will make up a slightly larger proportion than they did on the quizzes.

Where can you find the quiz questions and answers?

- Canvas → "Announcements" tab → Quiz XXX Grades and Solutions Released → Google Drive link



Reminder: Time Allocation

- Quizzes: 40 minutes, 7 questions
- Final Exam: 120 minutes, 30 questions!

- If you found the quiz time tight, you will need to solve problems faster on the final than you did on the quizzes!
  - Or my recommended strategy: your speed on seen questions should be twice as fast as on unseen questions (i.e., 40 minutes for 15 seen questions; 80 minutes for 15 unseen questions)

Final Exam (Other Details)

- Cheat Sheet: You can bring one cheat sheet
  - 8.5" x 11" standard sheet of paper with anything on it
  - Front and back
  - Either printed or handwritten is okay

- Laptops, books, and notes other than the cheat sheet are NOT allowed
- Calculators are NOT required, and the questions will NOT involve calculations that cannot be done easily by hand.

Final Exam (Conflict)

- Please refer to Student Rule 7 (https://student-rules.tamu.edu/rule07/) and Rule 8 (https://student-rules.tamu.edu/rule08/)

- You need to submit a written request and supporting documentation in advance, unless an unforeseen circumstance occurs that prevents prior notice (e.g., sudden illness on the day of the exam)

Logistics for Project Presentations

Project Presentations (Overall Information)

- Time: Nov 23, Nov 30, Dec 1, and Dec 2 in class
  - For the detailed schedule, please refer to the shared Google Sheet
- Location: Zoom (https://tamu.zoom.us/j/6411788612; you can find it on the course webpage)
- Presenters: Each group may have one presenter, multiple presenters, or the entire group
  - If you are in a group of 1 or 2 (i.e., you present on Nov 23), you will have 8 minutes for your presentation.
  - If you are in a group of 3 or 4, you will have 9 minutes for your presentation.
  - (For prototype projects, please plan for 5-6 minutes of presentation and 3-4 minutes of live demo)

Project Presentations (Overall Information)

- Time: Nov 23, Nov 30, Dec 1, and Dec 2 in class
  - For the detailed schedule, please refer to the shared Google Sheet
- Location: Zoom (https://tamu.zoom.us/j/6411788612; you can find it on the course webpage)
- Presenters: Each group may have one presenter, multiple presenters, or the entire group
  - When your time is up, I will unmute myself and notify you, and you will have 1 minute to wrap up everything. If you are not finished after that extra minute, a 1% overtime penalty will be applied.

Project Presentations (Rubrics for a Prototype Project)

- Function of your system and motivation (1%)
- Existing similar systems and their limitations (1%)
- Data (1%)
  - You may include: dataset statistics, collection/annotation process, …
- Implementation (2%)
  - You may include: core algorithm, existing resources used, …
  - If you have unfinished parts, explain how to finish them before the project report due (Dec 5)
- Live demo (3%)
  - Minimum requirement: Given a sample input (e.g., query or item), show your system's output (relevant documents or recommended items)
- Finish the presentation within the allotted time (1%)

Project Presentations (Rubrics for a Research Project)

- Task background and definition (1%)
- Related work and their limitations (1%)
- Proposed solution (2%)
  - You may include: model architecture, objective function, …
- Data (1%)
  - You may include: dataset statistics, collection/annotation process, …
- Experimental results (2%)
  - You may include: metric, comparison with baseline(s), ablation study, case study, …
- Take-away messages (1%)
  - If you have unfinished parts, explain how to finish them before the project report due (Dec 5)
- Finish the presentation within the allotted time (1%)

Project Presentations (Rubrics for a Survey Project)

- Survey scope and motivation (1%)
- Existing related surveys and their limitations (1%)
- Survey structure and approach (2%)
  - You may include: organizing principle (by tasks, methods, applications, …)
- Representative papers (2%)
- Insights and take-away messages (2%)
  - You may include: commonalities and characteristics of different works, lessons for readers, …
  - If you have unfinished parts, explain how to finish them before the project report due (Dec 5)
- Finish the presentation within the allotted time (1%)

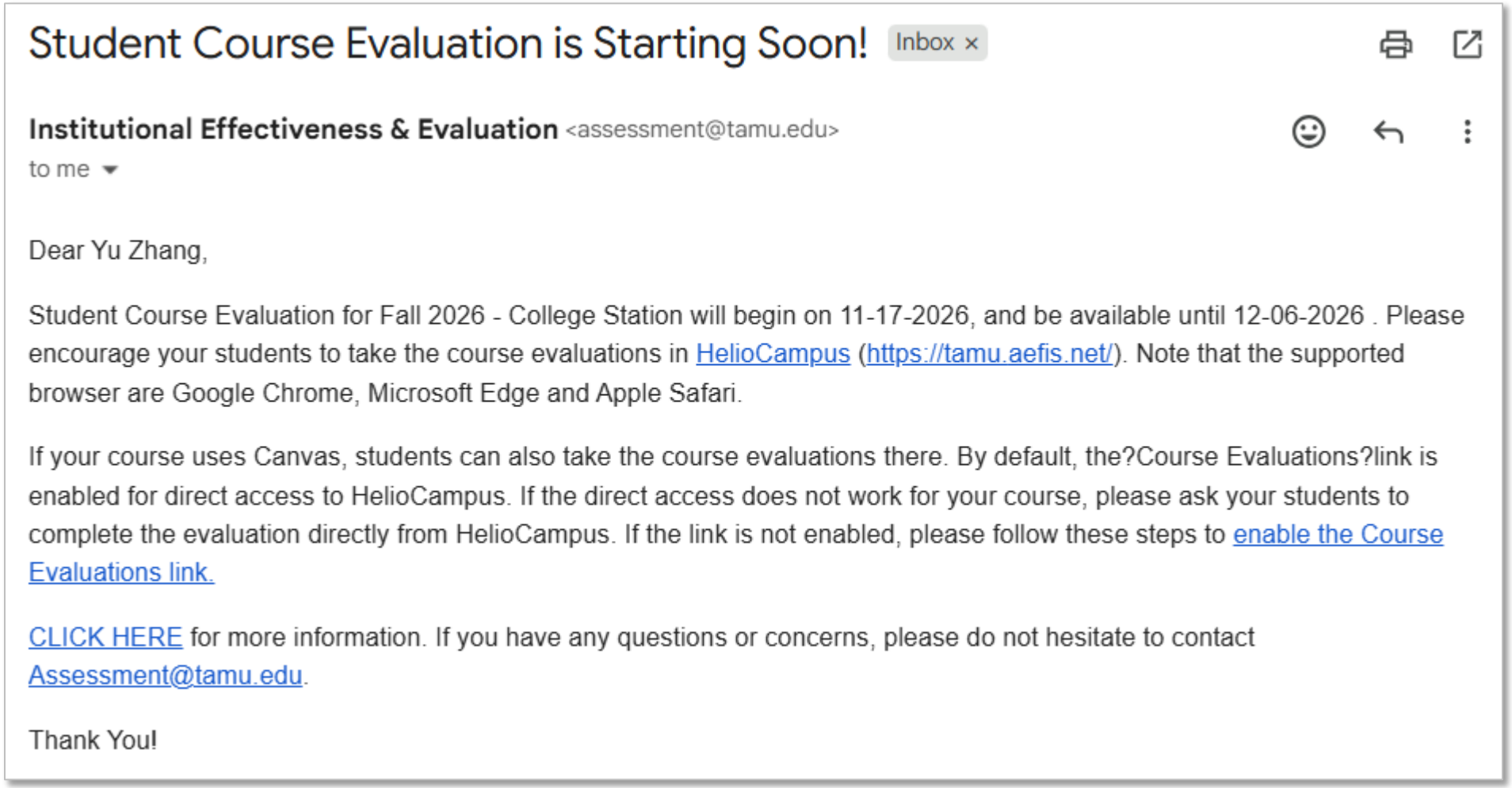
Office Hours

- The remaining office hours are scheduled as follows:
  - Additional office hour on Dec 7; you can either stop by PETR 222 or join via https://tamu.zoom.us/j/6411788612

Sun	Mon	Tue	Wed	Thu	Fri	Sat
			Nov 25 Yu (1-2 pm, PETR 222)		Thanksgiving, No TA office hour	
			Dec 2 Yu (1-2 pm, PETR 222)		Dec 4 TA (11:30-12:30 am, PETR 448)	
	Dec 7 Yu (1-2 pm, PETR 222 + Zoom)		Dec 9 Final (8-10 am, HRBB 124)			

Course Evaluation

- https://tamu.aefis.net Available until Dec 6



Student Course Evaluation is Starting Soon! Inbox ×

Institutional Effectiveness & Evaluation <assessment@tamu.edu>

to me

Dear Yu Zhang,

Student Course Evaluation for Fall 2026 - College Station will begin on 11-17-2026, and be available until 12-06-2026 . Please encourage your students to take the course evaluations in HelioCampus (https://tamu.aefis.net/). Note that the supported browser are Google Chrome, Microsoft Edge and Apple Safari.

If your course uses Canvas, students can also take the course evaluations there. By default, the?Course Evaluations?link is enabled for direct access to HelioCampus. If the direct access does not work for your course, please ask your students to complete the evaluation directly from HelioCampus. If the link is not enabled, please follow these steps to enable the Course Evaluations link.

CLICK HERE for more information. If you have any questions or concerns, please do not hesitate to contact Assessment@tamu.edu.

Thank You!



Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE670-F26.html