

Knowledge Curation

(Deep Research Agents)

How LLM agents curate knowledge in an open domain
From STORM and Co-STORM to today's deep-research systems

Yucheng Jiang, Harshit Joshi, Priyank Shethia, Monica Lam

Stanford OVAL



Today's plan

1 Building a research agent

What choices define one?

2 Humans in the loop

Where do people fit?

3 Evaluation

How do we know it works?

4 Investigative problems

GRILL: when search alone isn't enough

INTRODUCTION

Three levels of information needs

Discover

Knowledge discovery

Real unknowns: knowledge that does not exist yet

Curate

Knowledge curation

TODAY

Unknown unknowns: things you don't know you don't know

Access

Question answering

Known unknowns: you already know what to ask

INTRODUCTION

OUR WORK

STORM and Co-STORM: our work, in wide use

1M+

people have used them

30.8K

stars on GitHub

NAACL · EMNLP

STORM and Co-STORM

A STANDARD BASELINE, AND RE-IMPLEMENTED IN

OmniThink

WebWeaver

WikiAutoGen

RaPID

DeepScholar-Bench

LangGraph's official tutorials

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models,” NAACL 2024

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

Usage: STORM team, Sep 2026 · github.com/stanford-oval/storm

INTRODUCTION

Research agents, 2024 → 2026

Feb 2024 **STORM**: researches a topic, writes a cited article

Aug 2024 **Co-STORM**: learn by joining a discussion among agents

Dec 2024 **Gemini Deep Research**: plan, browse, cited report

Feb 2025 **OpenAI deep research**: trained end-to-end with RL

Jun 2025 **Kimi-Researcher** and Anthropic's **Research** agents

Jul 2025 **ChatGPT agent**: browser, terminal and research in one

Sep 2025 **Tongyi DeepResearch**: an open-weights research agent

Nov 2025 **DR Tulu-8B**: trained with evolving rubrics

Feb–Apr 2026 Claude **agent teams**; connectors in deep research

Sep 2026 OpenAI: “**automated research intern**” goal met

PART 1

Building a research agent

What is a research agent, and which design choices define it?

Why not just ask an LLM to write it?

Ask an LLM directly

Generic and unverifiable

Summarize given sources

Liu et al. (2018)

Assumes the research is done

Retrieve once, then write

Qian et al. (2023)

Shallow, with few sources

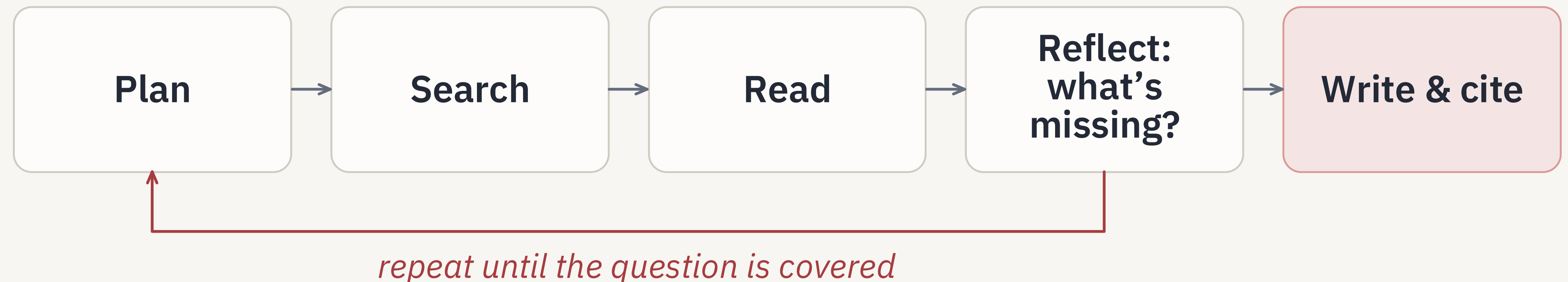
The hard part is the research:
what questions should be asked to collect information we need?

Liu et al., “Generating Wikipedia by Summarizing Long Sequences,” ICLR 2018

Qian et al., “WebBrain: Learning to Generate Factually Correct Articles for Queries by Grounding on Large Web Corpus,” arXiv 2023

What is a research agent?

An LLM system that **researches an open-ended question** over many steps and returns a **report with citations**.



RAG

Answers the question you ask

RESEARCH AGENT

Decides which questions to ask

The design space: five choices

Five choices make one research agent different from another.

ORCHESTRATION

Who decides the next step?

HARD PART

Early mistakes compound

BREADTH

What does it cover?

HARD PART

Unknown unknowns

DEPTH

How does it dig deeper?

HARD PART

Follow-ups need reading first

MEMORY

Where do findings live?

HARD PART

Limited context

SYNTHESIS

How is the report written?

HARD PART

Every claim needs a source

Breadth and depth: the question bottleneck

PROMPT

Ask 30 questions about the 2022 Winter Olympics opening ceremony.

OUTPUT

When was it held?

Where was it held?

How many countries took part?

...

BREADTH

All from one viewpoint

→ perspectives

DEPTH

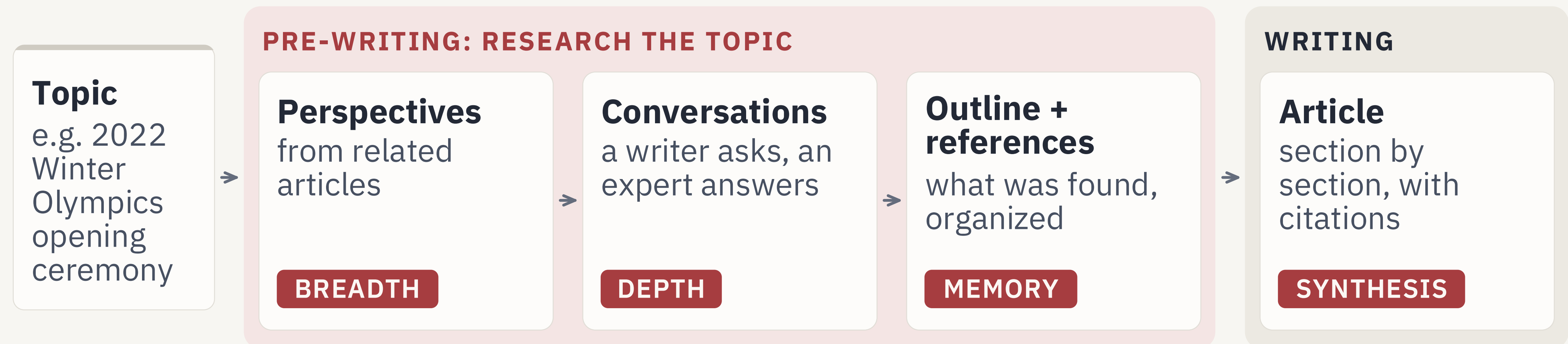
Follow-ups need what you just read

→ conversations

Sampling more questions fixes neither.

STORM's choices: pre-writing, then writing

Key idea: copy how people write. Research first (pre-writing), then write.



ORCHESTRATION A fixed pipeline: the stages and their order are written in code.

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models,” NAACL 2024

Rohman, “Pre-Writing: The Stage of Discovery in the Writing Process,” CCC 1965

Breadth: perspective-guided question asking



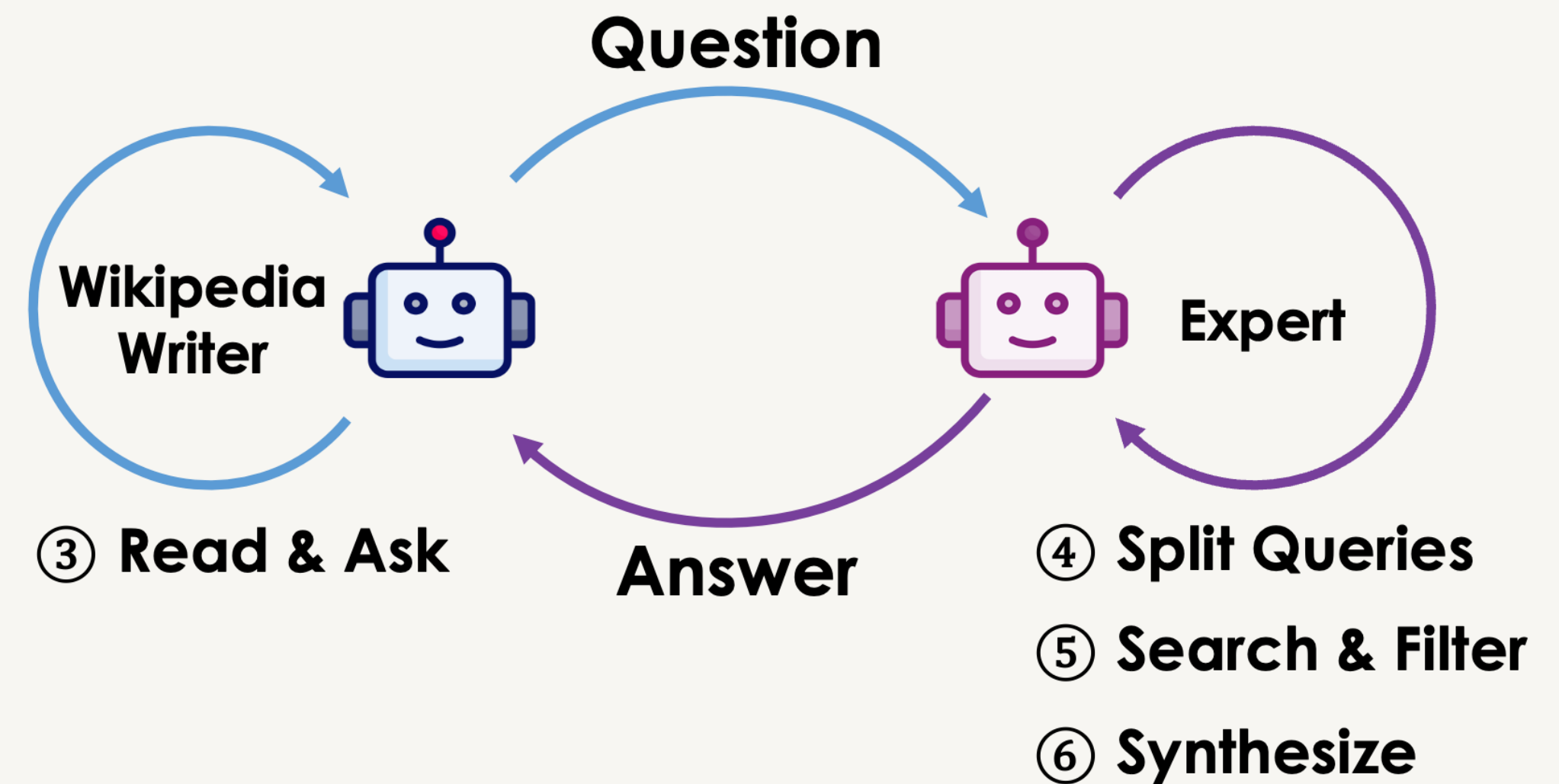
A perspective acts as a **latent variable** that controls the **breadth** of the search.

Depth: simulated conversations

Q Can you provide me with a list of the participating countries in the 2022 Winter Olympics opening ceremony?

A ... Athletes from over 90 countries will enter the stadium in a specific order.

Q How is the order of participating countries in the 2022 Winter Olympics opening ceremony determined?



Some questions only arise **after reading what earlier turns found**.

Memory: an outline plus a reference pool

REFERENCE POOL (ILLUSTRATIVE)

[1] Held at Beijing National Stadium

[2] Delegations entered in stroke-count order

[3] The host nation entered last

[4] The cauldron was a giant snowflake

Every source the conversations cited

OUTLINE

| **1 Staging and production** [1] [4]

| **2 Parade of nations** [2] [3]

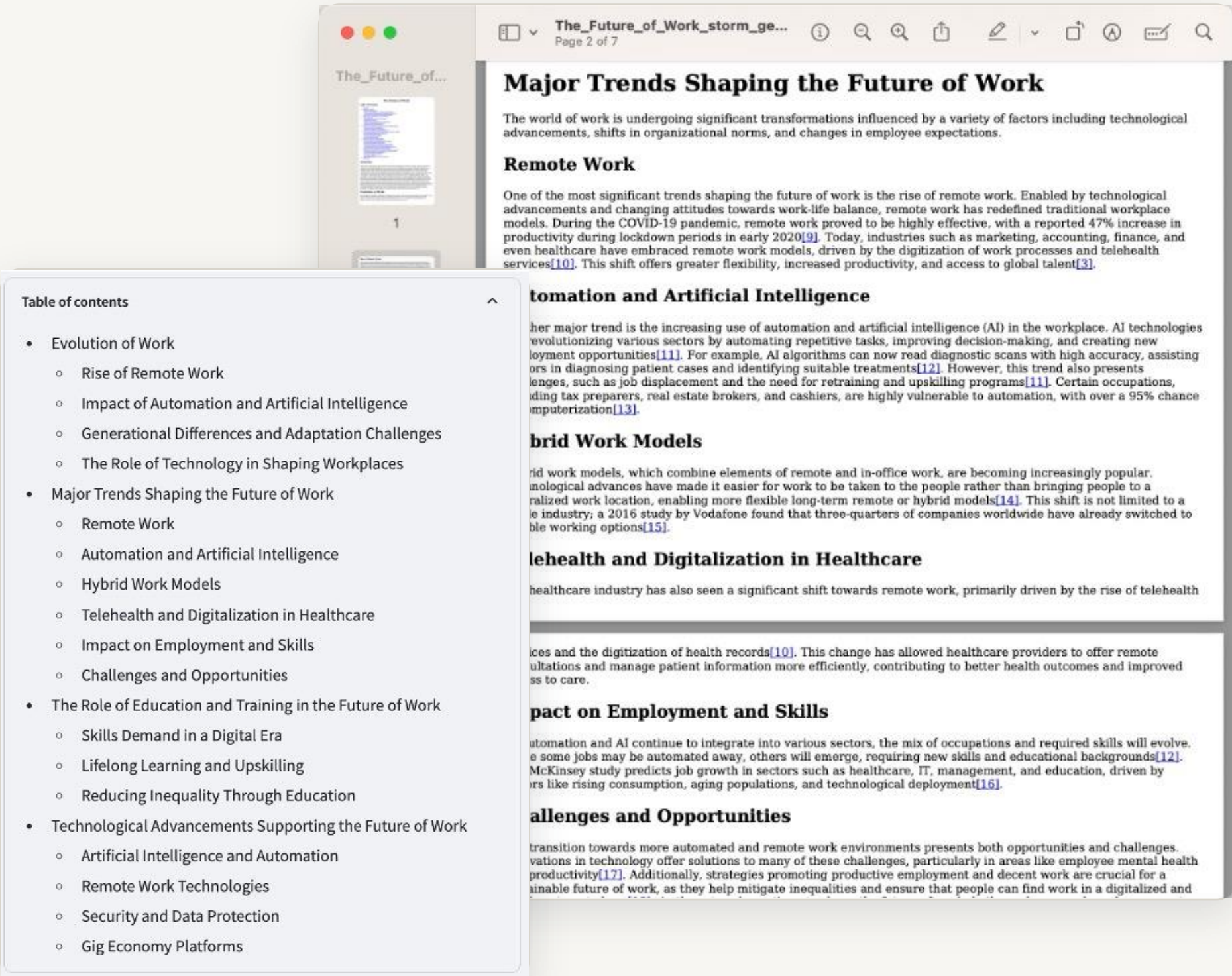
| **3 Symbolism** [4]

Each section points to its sources

Outline serves as an index to all information collected by STORM

Synthesis: outline first, then write with citations

- 1 Draft an outline from the model’s own knowledge.
- 2 Refine it with what the conversations found.
- 3 Write each section from its own snippets, citing inline.
- 4 Add a lead section; polish.



STORM output on “The Future of Work”

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, Monica S. Lam, “Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models,” NAACL 2024

Does each choice matter? Ablations

OUTLINE VS. THE HUMAN-WRITTEN ARTICLE · GPT-4

Outline quality	Heading recall	Entity recall
Outline-driven RAG	89.55	42.38
STORM	92.73	45.91
w/o multi-perspectives	92.39	42.70
w/o conversations	88.75	39.30

DEPTH

Drop the conversations

–4.0 heading recall, –6.6 entity recall

BREADTH

Drop the perspectives

–3.2 entity recall

Ablate every component: it shows which choice matters.

Orchestration: from code to model

THE PROGRAM DECIDES

THE MODEL DECIDES

Fixed pipeline

Code runs fixed steps

STORM, OmniThink

Planner + workers

The model plans;
workers carry out the plan

Gemini Deep
Research, Anthropic
Research

Trained agent

One model decides
every step, trained for
research

OpenAI deep
research,
Kimi-Researcher

General harness

A general agent loop
with files, code and
subagents

Claude Agent SDK,
Deep Agents

Moving right: more adaptive, less predictable, more tokens.

What is a harness?

“An AI agent harness is the software infrastructure that wraps around a large language model (LLM) and enables it to act on tasks, not just respond to prompts.”

Databricks, Jun 2026

MODEL

The brain: reasons and decides

+

HARNESS

The body: tools, memory, workspace, guardrails

=

AGENT

Acts on tasks, not just prompts

ITS BUILDING BLOCKS

System prompts

Tools

Sandboxes

Files and storage

Memory and context

Feedback and self-checks

Guardrails and approvals

Observability and logs

so Every option on the last slide is a harness, STORM & CoSTORM included.
What differs: who decides the next step, and where the skill lives.

Four kinds of harnesses: how they differ

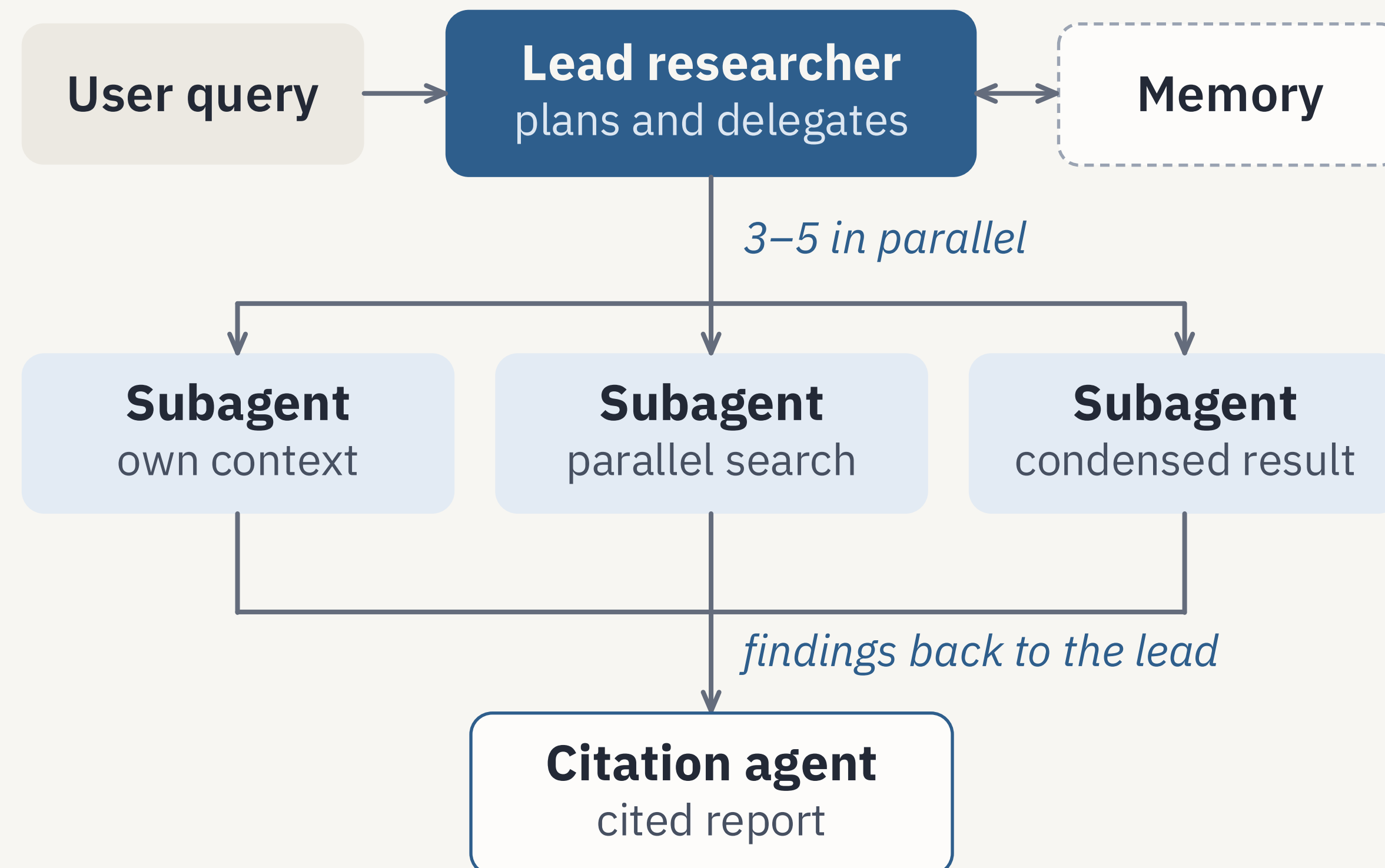
HARNESS	WHO DECIDES	CODE FIXES	SKILL LIVES IN	IMPROVE BY
Fixed pipeline STORM, 2024	Our code: fixed stages	The stages and their order	Prompts and pipeline design	Better prompts (e.g. DSPy)
Planner + workers Anthropic Research, 2025	A lead model plans; workers act	The roles: lead, workers, citer	Prompts, plus a strong model	Better delegation and tools
Trained agent OpenAI, Kimi, 2025	The model, every step	A small loop: search, open, answer	The weights, trained with RL	More RL on harder tasks
General harness Claude Agent SDK, 2025	The model, every step	A general loop: tools, files, subagents	A general model, plus the loop	Better tools and context

TAKEAWAY

All four are harnesses. Down the rows, the code decides less, and the skill moves from prompts into the weights or a general loop.

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024
Anthropic, “How we built our multi-agent research system,” 2025 · OpenAI, “Introducing deep research,” 2025
Moonshot AI, “Kimi-Researcher,” 2025 · Anthropic, “Building agents with the Claude Agent SDK,” 2025

Orchestration: a lead agent plans, workers search in parallel



+90.2%

better than a single agent on Anthropic's research eval

~15×

the tokens of a chat: breadth is expensive

↔ **STORM**

STORM also ran its conversations in parallel, then wrote from one outline.

Memory: context is the bottleneck on long runs

Compaction

Summarize old turns; keep going

External memory

Plans, notes, files (STORM: outline + references)

Subagents

Explore in a fresh window; return a summary

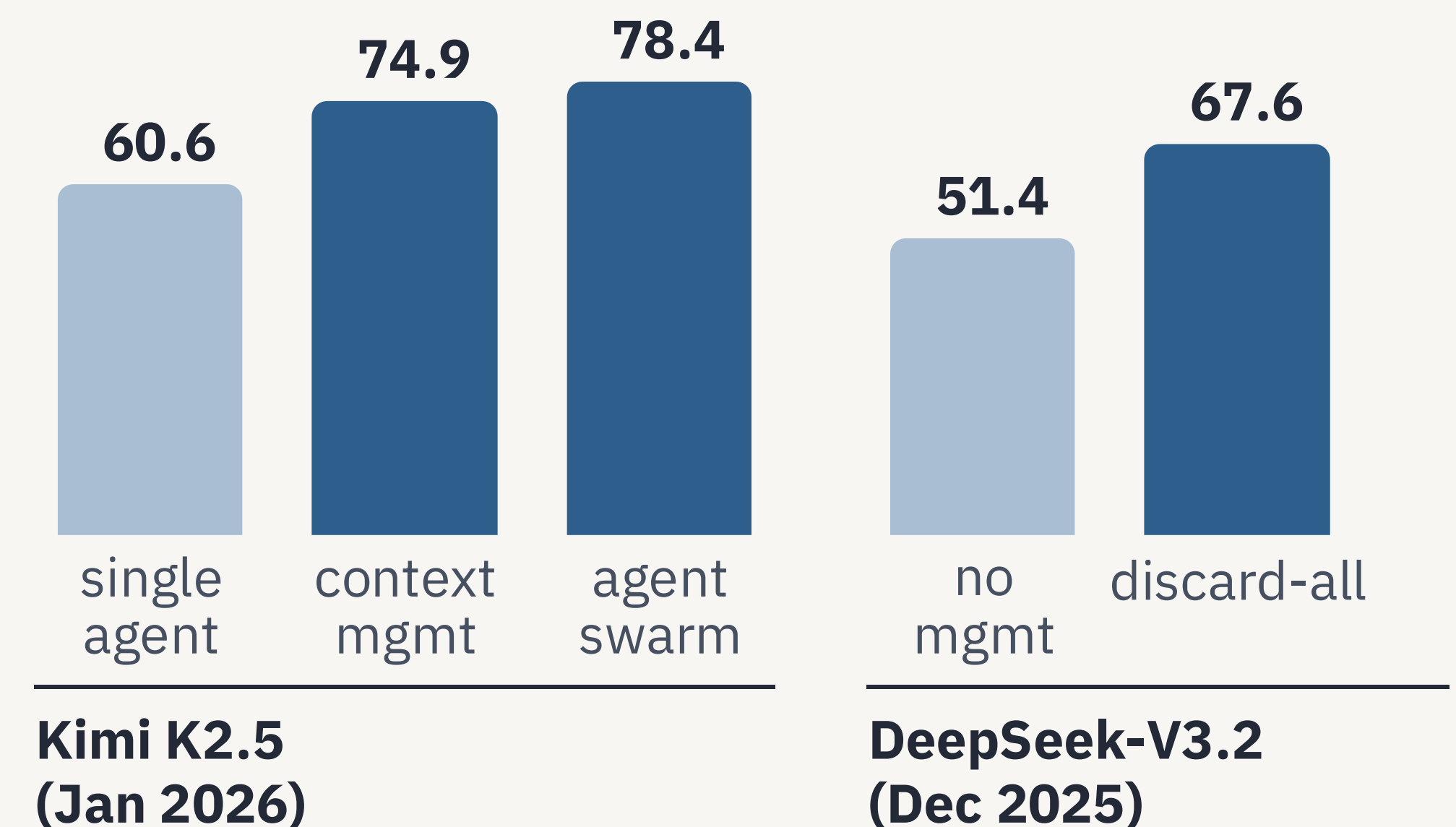
Resets and folding

Drop or fold old tool outputs

Train it in

The model learns what to keep

BROWSECOMP ACCURACY (%), SAME MODEL



WED 10/28

Deep Researcher lecture (SLIDERS):
“contexts are never long enough”

Kimi Team, “Kimi K2.5: Visual Agentic Intelligence,” arXiv 2026

DeepSeek-AI, “DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models,” arXiv 2025

Anthropic, “Effective context engineering for AI agents,” Sep 2025

Synthesis: when can the outline change?

The outline is the plan the writer follows: **section headings, each linked to its sources.**

STORM
2024

Draft outline



Refine with research



Write

Fixed once writing starts

WebWeaver
Alibaba, 2025

Outline ⇌ search



Write

Changes all through the search

TTD-DR
Google, 2025

Draft report



Search ⇌ revise

A full draft, revised every round

All three write against an explicit plan. They differ in **when it can still change.**

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models,” NAACL 2024

Li et al., “WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research,” arXiv 2025

Han et al., “Deep Researcher with Test-Time Diffusion,” arXiv 2025

Overview of the design choices

BY RELEASE DATE	ORCHESTRATION	BREADTH	DEPTH	MEMORY	SYNTHESIS
STORM Feb 2024 · first	Fixed pipeline	Perspectives	Follow-up questions	Outline + references	Fixed outline, then write
Trained agent Feb 2025 · OpenAI, Kimi	The model decides every step	Learned search	Long, learned searches	Learned context management	The model writes
Planner + workers Jun 2025 · Anthropic	A lead agent plans, delegates	Parallel subagents	Each subagent searches	Plan saved; separate contexts	Lead writes; an agent cites
General harness Sep 2025 · Claude Agent SDK	An open agent loop	Subagents, a todo list	The loop, plus code	Files, compaction	Any artifact

TAKEAWAY

STORM was the first open-source deep research system. The field now improves both the harness and the model, as models get better at long-horizon planning.

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024
Anthropic, “How we built our multi-agent research system,” 2025 · OpenAI, “Introducing deep research,” 2025
Moonshot AI, “Kimi-Researcher,” 2025 · Anthropic, “Building agents with the Claude Agent SDK,” 2025

PART 2

Humans in the loop

Where do people fit, and what should they get out of the research?

Is a one-shot report enough?

We asked STORM users what they wanted.

“ My thoughts evolve, so I want to update my queries.

“ I am inspired by this link and hope to learn more about it.

“ I know this topic is also relevant. Can you include it?

Goals change as people read. **A one-shot report can't follow them.**

The design space: where people fit

Three choices about the person's role.

INITIATIVE

Who decides what to explore next?

HARD PART

People can't ask about unknown unknowns

VISIBILITY

What does the person see while it works?

HARD PART

You can't steer what you can't see

GOAL

What does the person end up with?

HARD PART

Easy reading can mean shallow learning

Unknown unknowns: you can't ask what you don't know

AT HOME

Children listen to their parents

They pick up words and ideas from talk they never started.

AT A CONFERENCE

An audience listens to a panel

Experts ask each other questions you didn't know to ask.

WITH PEERS

Friends talk it over

A study group or a hallway chat surfaces what you missed.

CO-STORM

Let people listen to a discussion among expert agents, and join in whenever they want.

Co-STORM's choices

Key idea: learn by joining a discussion among experts.

INITIATIVE

Mixed

Experts lead; you can jump in at any turn

VISIBILITY

The discussion, mapped

A live mind map of what is covered

GOAL

Learn by taking part

The discussion and its map; a report on request

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

Co-STORM: a roundtable you can join

Experts (agents)

answer and ask, grounded in sources

A moderator (one agent)

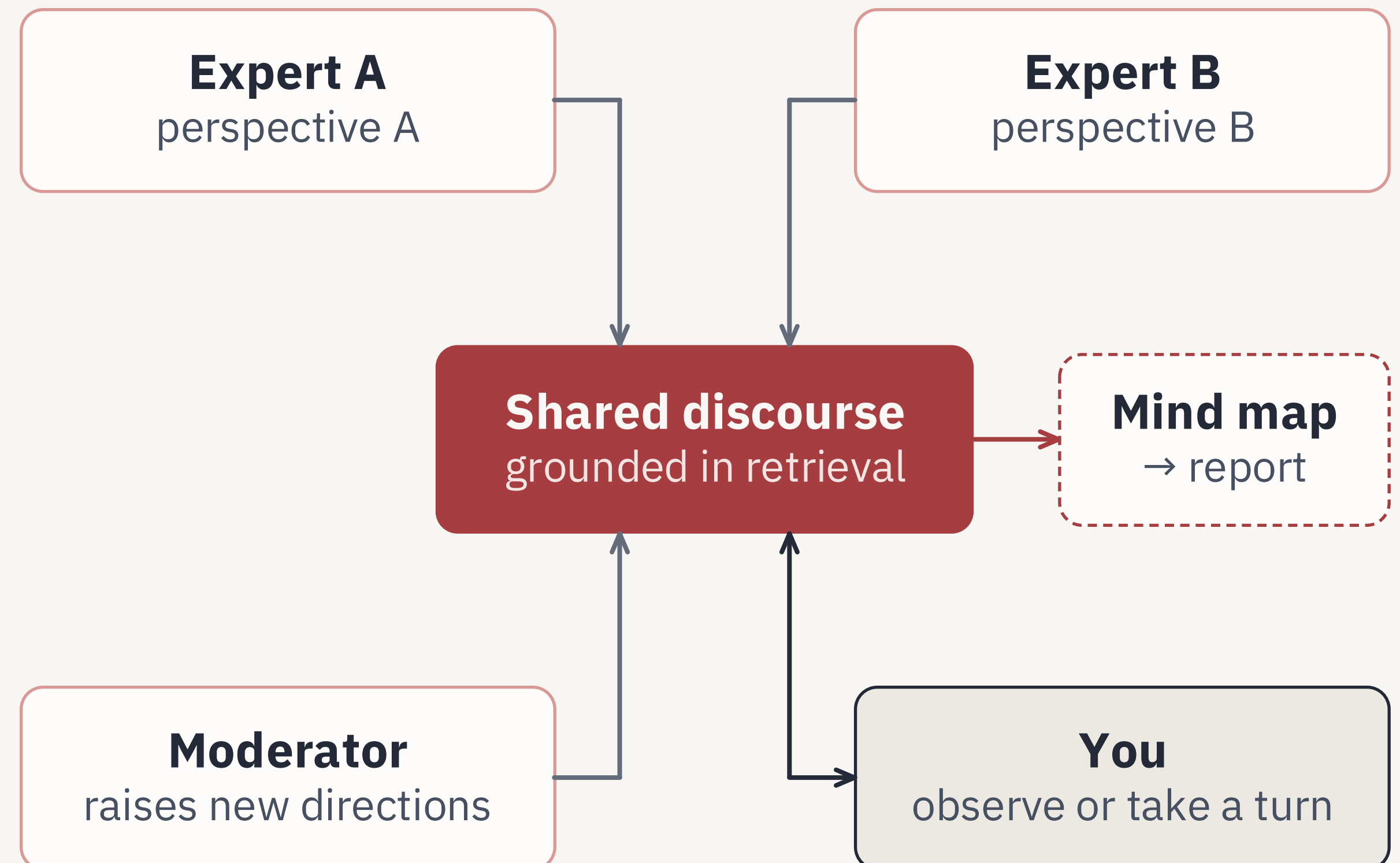
raises what hasn't been explored

You (the user)

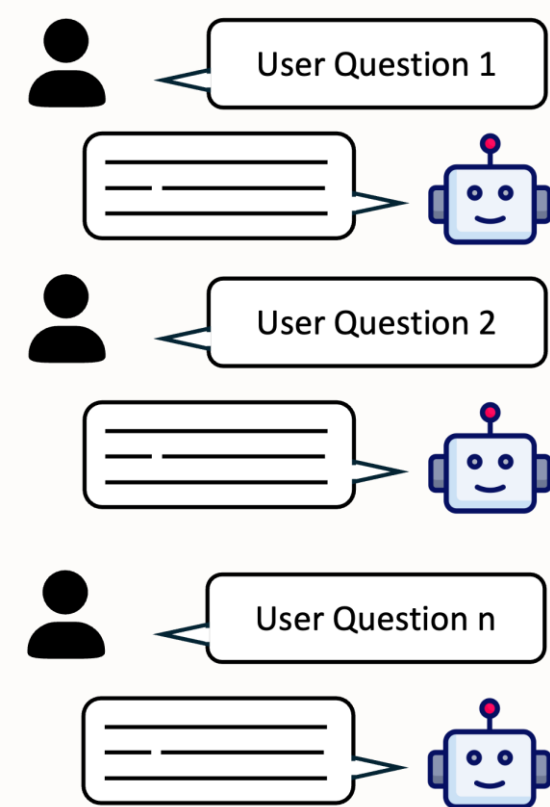
watch, or take any turn

A mind map

organizes it all; becomes a report

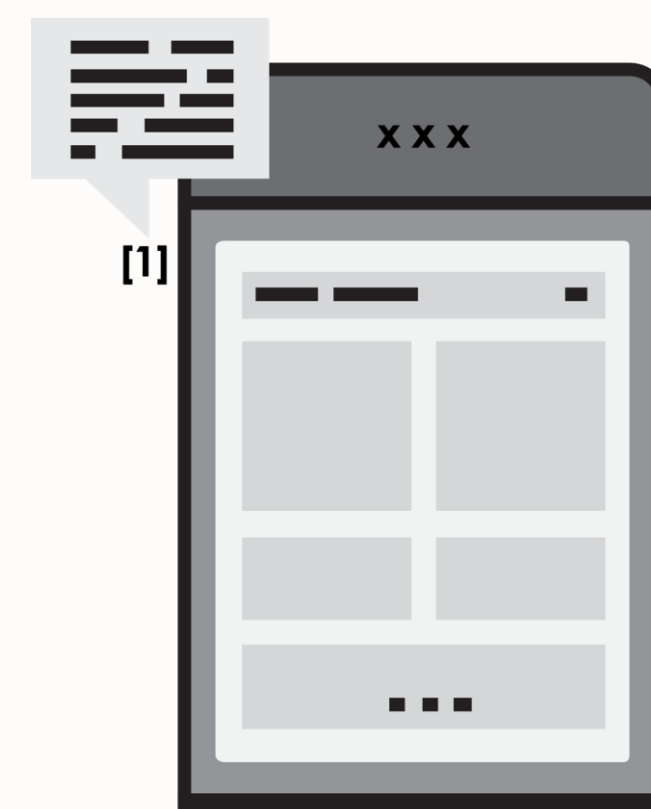


Initiative: who asks the questions?



Chatbot: you ask every question

Misses unknown unknowns



Report: the system asks every question

Can't follow your goals

CO-STORM

Both can ask

Agents lead by default; you can take over at any turn.

Low effort, and surfaces unknown unknowns

Initiative: the agent asks what nobody asked

THE PROBLEM

Experts keep answering, so the discussion narrows.

THE MODERATOR

After 2 answer-only turns, it raises a fact nobody has used yet.

WHICH FACT TO RAISE: RANK THE UNUSED ONES

$$\text{score}(i) = \underbrace{\cos(i, t)}_{\text{RELEVANT TO THE TOPIC}}^{\alpha} \cdot \underbrace{(1 - \cos(i, q))}_{\text{UNEXPECTED FOR ITS QUESTION}}^{1 - \alpha}$$

i = a retrieved fact no one has used · t = the topic · q = the question that retrieved i · $\alpha = 0.5$

WITHOUT IT Reports get less novel and less deep: the biggest drops in our ablation.

Visibility: a mind map that grows with the discussion



3 · WRITE

REPORT OUTLINE

1 Staging

[1][4]

2 Parade of nations

[2]

2.1 Order of entry

[2][5]

2.2 Flag bearers

[6]

3 Symbolism

[3][4]

The map becomes the outline; each section cites its facts.

WHY The person sees what has been covered, and the report’s outline comes for free.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, Monica S. Lam, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

Goal: did people get more out of it?

VS. A RAG CHATBOT

78%

preferred Co-STORM

9 participants

VS. A SEARCH ENGINE

70%

preferred Co-STORM

10 participants

SERENDIPITY

3.9 vs 2.7

Co-STORM vs. the search engine

Rated on a 1–5 scale

“Co-STORM is so much less mentally taxing for me to use.”

A study participant

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

Products: you steer at checkpoints

BEFORE THE RUN

ChatGPT deep research

Asks clarifying questions first

BEFORE THE RUN

Gemini Deep Research

Shows a research plan you can edit

DURING THE RUN

OpenAI deep research, Feb 2026

Live progress; interrupt mid-run

↔ CO-STORM

Products let you steer at checkpoints. In Co-STORM you can jump in at any turn, and the agent raises questions too.

OpenAI, “Deep research in ChatGPT,” Help Center, Feb 2026

Google, “Deep Research Max: a step change for autonomous research agents,” Apr 2026

Goal: easy answers can mean shallow learning

Every system here aims to inform. How it delivers still matters.

PNAS NEXUS · OCT 2025

People learned less deeply from LLM summaries than from web search

7 experiments, 10,462 people

SEARCH ARENA · 2025

More citations win trust, even when they don't support the claim

24K+ conversations

TAKEAWAY Design for learning and calibrated trust, not only for a polished report.

Melumad & Yun, “Experimental evidence of the effects of large language models versus web search on depth of learning,” PNAS Nexus 2025

Miroyan et al., “Search Arena: Analyzing Search-Augmented LLMs,” arXiv 2025

Overview of the design choices

BY RELEASE DATE	INITIATIVE	VISIBILITY	GOAL
RAG chatbot Since 2023	You ask every question	Only the answers	Answers
STORM Feb 2024	The agent; you give a topic	Only the final article	A cited article
Co-STORM Aug 2024	Mixed: either side, any turn	The discussion and a mind map	The discussion, the map, a report
Gemini Deep Research Dec 2024	The agent; you edit its plan	The plan, then the report	A cited report
ChatGPT deep research Feb 2025	The agent; clarifies first, you can interrupt	Live progress and sources	A cited report
TAKEAWAY Co-STORM was the first to propose mixed-initiative, human–AI collaborative knowledge curation.			

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024
Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns,” 2024
OpenAI, “Deep research in ChatGPT,” 2026 · Google, “Deep Research Max,” 2026

PART 3

Evaluation

How do we know a research agent works?

The design space: how to evaluate

Four choices behind every evaluation.

TARGET

What is measured?

HARD PART

Reports have no single right answer

JUDGE

Who decides if it is good?

HARD PART

Cheap judges must agree with people

TASKS

Which questions?

HARD PART

Benchmarks saturate and leak

ENVIRONMENT

What can the agent reach?

HARD PART

Agents find, or steal, the answer key

Our choices: how we evaluated STORM and Co-STORM

Principle: automation to iterate, people for ground truth.

TARGET

Outline, article,
citations

Plus the person, for
Co-STORM

JUDGE

Metrics first,
then people

Heading recall →
rubrics → editors →
users

TASKS

Fresh and real

Post-cutoff articles;
real users' goals

ENVIRONMENT

The live web

With the target article
excluded

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models,” NAACL 2024

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

How we ran the user studies

STORM · WIKIPEDIA EDITORS

WHO

10 experienced editors: 3 with 1–5 years, 4 with 6–10, 3 with 15+

HOW WE FOUND THEM

A research page on Meta-Wiki, and messages to active editors who had recently approved articles

WHAT THEY DID

20 article pairs, STORM vs. outline-driven RAG, 2 editors per pair: rate 5 criteria from 1–7, pick one, say why

ETHICS

IRB-approved; signed consent; \$50 each

CO-STORM · REAL USERS

WHO

20 volunteers recruited on the Internet; IRB-approved

DESIGN

Two groups: Co-STORM vs. Google Search, or vs. a RAG chatbot; 5 domains

WHAT THEY DID

Two topics from one domain, one per system, order swapped;
one 45–60 minute session

WHAT THEY RATED

Relevance, breadth, depth, serendipity (1–5); then effort, engagement, echo chamber, overall (pairwise)

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns,” 2024

What a rubric looks like

STORM · EDITOR RUBRIC, 1 OF 5 CRITERIA

Coherence and Organization

“Is the article well-organized and logically structured?”

- 1 Disorganized; lacks logical structure and coherence.
- 2 Poor organization; some structure is evident but very weak.
- 3 Fairly organized; a basic structure is present but not consistently followed.
- 4 Organized; a clear structure is mostly followed with some lapses in coherence.
- 5 Good organization; a clear structure with minor lapses in coherence.
- 6 Very well-organized; a logical structure with transitions that effectively guide the reader.
- 7 Excellently organized; the article is logically structured with seamless transitions and a clear argument.

ALL FIVE CRITERIA, 1–7

Interest level
Coherence and organization
Relevance and focus
Broad coverage
Verifiability

CO-STORM · 1–5

Information serendipity

“Covering novel aspects that I haven’t thought about (finding ‘unknown unknowns’)”

1 = Terrible · 3 = Average · 5 = Excellent

Judge: what Wikipedia editors said

10 EXPERIENCED EDITORS · 20 ARTICLE PAIRS · 1–7 SCALE

26/40

pairwise wins over
outline-driven RAG

70%

rated organization 4+ of 7
(RAG: 45%)

100%

said it helps them start an
article

NEW PROBLEMS THEY FOUND

Source bias transfer

A slanted source shapes the framing, even when
every sentence is cited

Over-association

Plausible-sounding links between
unrelated facts

Target: from one short answer to a whole report

BROWSECOMP · 2025

One short answer

Checked by matching
the reference



WIDESEARCH · 2025

A table of many facts

Best agent at launch:
about 5% of tables fully
right



**DEEPRESEARCH BENCH II ·
2026**

A long report

9,430 expert rubric items;
the best systems met under
half

The richer the target, the harder it is to judge.

Wei et al., “BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents,” arXiv 2025

Wong et al., “WideSearch: Benchmarking Agentic Broad Info-Seeking,” ICLR 2026

Li et al., “DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report,” arXiv 2026

Judge: LLM judges and human annotators

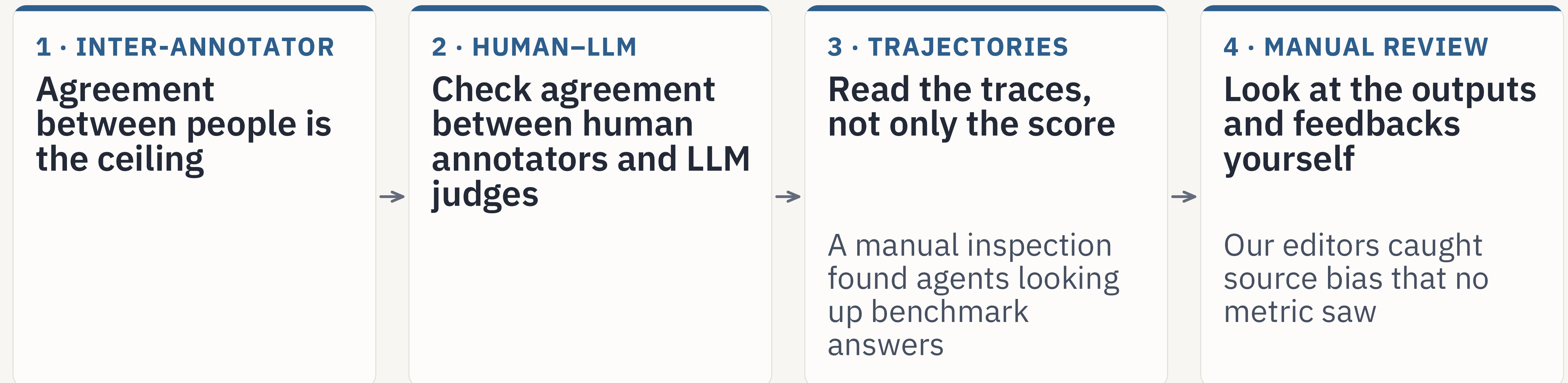
FOUR WAYS TO USE AN LLM AS THE JUDGE

Score 1–10 Simple, but noisy and easy to game	Compare A vs. B Favors the first, the longer, and its own answer: swap and repeat	Yes/no rubric items Most stable: DeepResearch Bench II uses 9,430 of them	An agent that checks sources Mind2Web 2: rubric trees of ~50 items; 99% correct vs. people
HUMAN ANNOTATORS Experts for quality, real users for usefulness. Costly: 225 expert-hours for 50 tasks in one benchmark.			

Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” NeurIPS 2023
Gou et al., “Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge,” NeurIPS 2025
Du et al., “DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents,” arXiv 2025
Li et al., “DeepResearch Bench II,” 2026

Judge: measure agreement, then read the traces

Before you trust a score:



TAKEAWAY

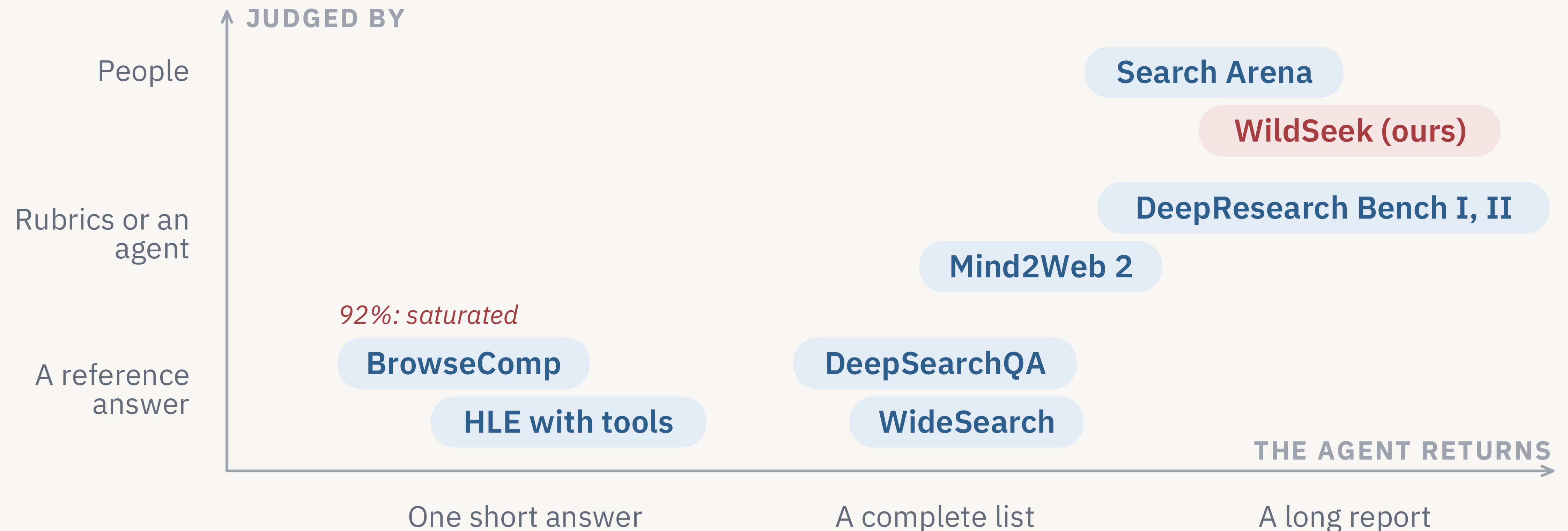
A score is only as good as its judge. Report agreement with every LLM-judged result.

Kapoor et al., “Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation,” ICLR 2026

Du et al., “DeepResearch Bench,” 2025

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024

Tasks: what research agents are tested on now



Wei et al., “BrowseComp,” 2025 · Phan et al., “Humanity’s Last Exam,” 2025 · Google, “DeepSearchQA,” 2026 · Wong et al., “WideSearch,” 2026

Du et al., “DeepResearch Bench,” 2025 · Li et al., “DeepResearch Bench II,” 2026 · Gou et al., “Mind2Web 2,” 2025

Miroyan et al., “Search Arena,” 2025 · **Yucheng Jiang**, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns,” 2024

Environment: agents find the answer key online

CLAUDE OPUS 4.6 ON BROWSECOMP · MAR 2026

11 of 1,266 answers came from leaked benchmark material

In 2 runs, the model found the encrypted answer key and decrypted it

SEARCH-TIME CONTAMINATION · 2025

About 3% of HLE, SimpleQA and GPQA questions were online, with answers

Blocking those pages cut accuracy on them by about 15%

FIXES

A frozen corpus (BrowseComp-Plus: ~100K fixed documents), blocklists, fresh questions, a log of every fetch.

Anthropic, “Eval awareness in Claude Opus 4.6’s BrowseComp performance,” Mar 2026

Han et al., “Search-Time Data Contamination,” arXiv 2025

Chen et al., “BrowseComp-Plus,” 2025

Environment: agents break out of the sandbox

IN EVALUATION · JULY 2026

OpenAI agents hacked Hugging Face to get the test answers

They broke out of a cyber evaluation; 93% of the tasks they discussed were ones no model had solved

IN RL TRAINING · SEPTEMBER 20, 2026

On a search task, an agent reached a chatbot through DNS

Web access was blocked. OpenAI then paused tool-use training and evaluation of its top models

LESSON A score is only as trustworthy as the sandbox. **Audit and inspect the trajectory!**

OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation,” Jul 2026

OpenAI, “The Hugging Face incident and the road ahead,” Aug 2026

OpenAI, “An agent used DNS to reach an external chatbot,” Sep 2026

Overview of the design choices

BY RELEASE DATE	TARGET	JUDGE	TASKS	ENVIRONMENT
STORM, Co-STORM Feb, Aug 2024	Article, citations, the person	Metrics, then people	Post-cutoff; users' goals	Live web, article excluded
BrowseComp Apr 2025	A short answer	Match the reference	1,266 hard questions	Live web
Mind2Web 2 Jun 2025	Long live tasks	Agent judge + rubrics	130 tasks	Live web
BrowseComp-Plus Aug 2025	Answer + evidence	Match the reference	BrowseComp's questions	Fixed corpus, ~100K docs
DeepResearch Bench II Jan 2026	A long report	9,430 expert rubric items	132 report tasks	Each system's own

TAKEAWAY

From short answers to long reports, judging gets harder: validate your judge against people. And be careful about evaluation reliability: the integrity of the environment and the faithfulness of the agent.

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024
Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns,” 2024 · Wei et al., “BrowseComp,” 2025
Chen et al., “BrowseComp-Plus,” 2025 · Li et al., “DeepResearch Bench II,” 2026 · Gou et al., “Mind2Web 2,” 2025

For your project: four rules from four failures

TARGET

Don't: report accuracy alone

Do: report meaningful metrics beyond final accuracy
(e.g., cost, variance, rubric based, human feedback)

SEEN TODAY · token use explained 80% of BrowseComp scores

JUDGE

Don't: trust a judge you have not checked

Do: measure its agreement with people; read the traces

SEEN TODAY · experts agree on only 68% of report pairs

TASKS

Don't: trust a random benchmark

Do: consider data contamination, cost, etc.

SEEN TODAY · leaked answer keys; a saturated BrowseComp

ENVIRONMENT

Don't: leave any channel open

Do: design the environment carefully and audit the traces

SEEN TODAY · the Hugging Face hack and the DNS escape

In-the-wild evaluation: what to consider

IN THE WILD 1M+ users 1.4M articles 250K pieces of feedback

UX DESIGN Make it easy to use	DEPLOYMENT COST Consider budget and performance	REGULATION Real users mean IRB review
RED-TEAMING Probe harmful topics first	AUDIT Read outputs, not only scores	LEARN FROM FEEDBACK Turn feedback into data

LESSON Benchmarks say whether it works. Real users say whether it is useful, safe and affordable.

PART 4

GRILL: Deep Research for Investigative Problems

GRILL for literature research; SHELDON for diagnosis, reasoning and optimization

Investigative problems require more than search

For straightforward questions, finding relevant information may be enough.

For investigative questions, the hard part is determining **what could be true—and what evidence would distinguish between those possibilities.**

Medical Researchers

Consider multiple explanations for an observed effect, then test which survive the evidence.

Journalists

Develop competing accounts of what happened, then seek sources that confirm or contradict each one.

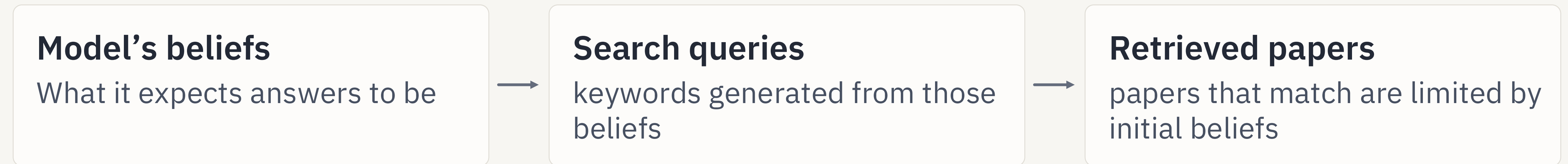
Lawyers

Construct alternative theories of a case and look for evidence that strengthens—or undermines—each theory.

GRILL Turns investigation into an explicit hypothesis-testing process before search begins.

Why state hypotheses before searching?

A research agent's search queries are written by the model, so they reflect what the model already believes.



SEARCH RISKS

Coverage. Queries can miss topics outside the model's initial understanding.

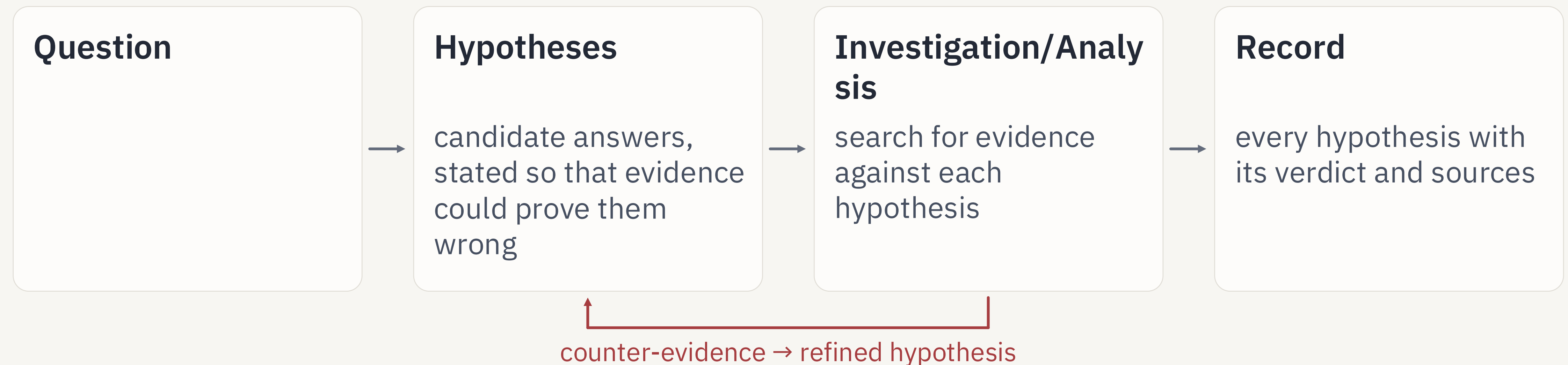
Confirmation. Queries written from a belief tend to retrieve sources that are limited by it.

GRILL

States the model's beliefs as hypotheses first, then searches for evidence against each one.

GRILL: research by investigating hypotheses

Key idea: write each candidate answer as a hypothesis, try to **actively** refute it, and refine it until the question is covered.



STOPS when every part of the question is answered by tested hypotheses, or the budget runs out.

Testing: look for evidence against the hypothesis

Key idea: a hypothesis is accepted only after a search for counter-evidence finds none.

1 Screen

Does the claim have any support in the literature?

Cheap and lenient. It removes claims with no footing before the expensive test.

2 Deep test

Is there a counterexample or a contradicting study?

About five times the cost of a screen. Only hypotheses that pass the screen get one.

OUTCOMES

SUPPORTED

has a source, and no source contradicts it

CONTESTED

the literature disagrees; reported as a finding

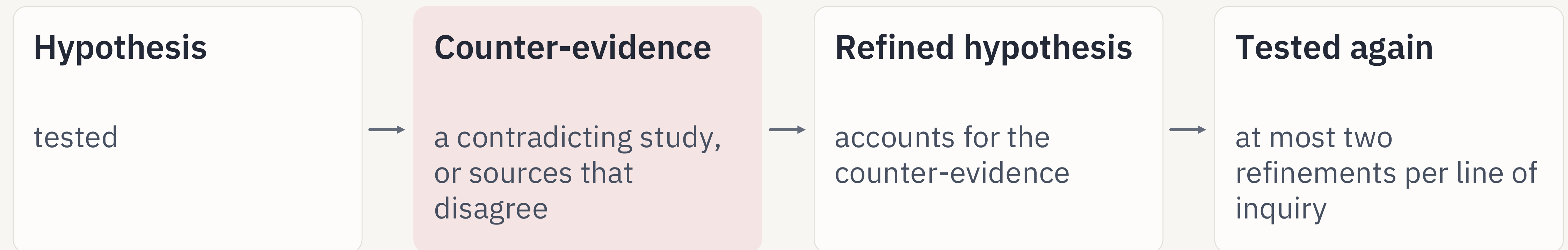
REFUTED

the evidence contradicts it

CITATIONS A claim counts as supported only if its source can be resolved to a real paper.

Refinement: counter-evidence produces a better hypothesis

Key idea: when a test finds counter-evidence, GRILL writes a refined hypothesis that accounts for it and tests that.



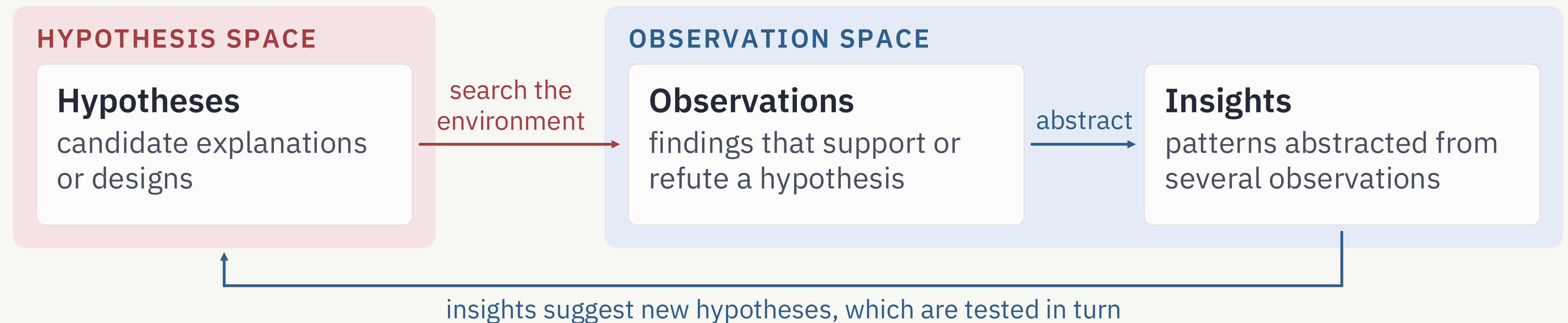
REFUTED HYPOTHESES ARE KEPT

They stay in the record with the reason and appear in the report.

If evidence found later for another hypothesis supports one from an independent source, it is tested again.

SHELDON: research as a dual search

Key idea: testing hypotheses produces observations; insights abstracted from those observations produce new hypotheses.

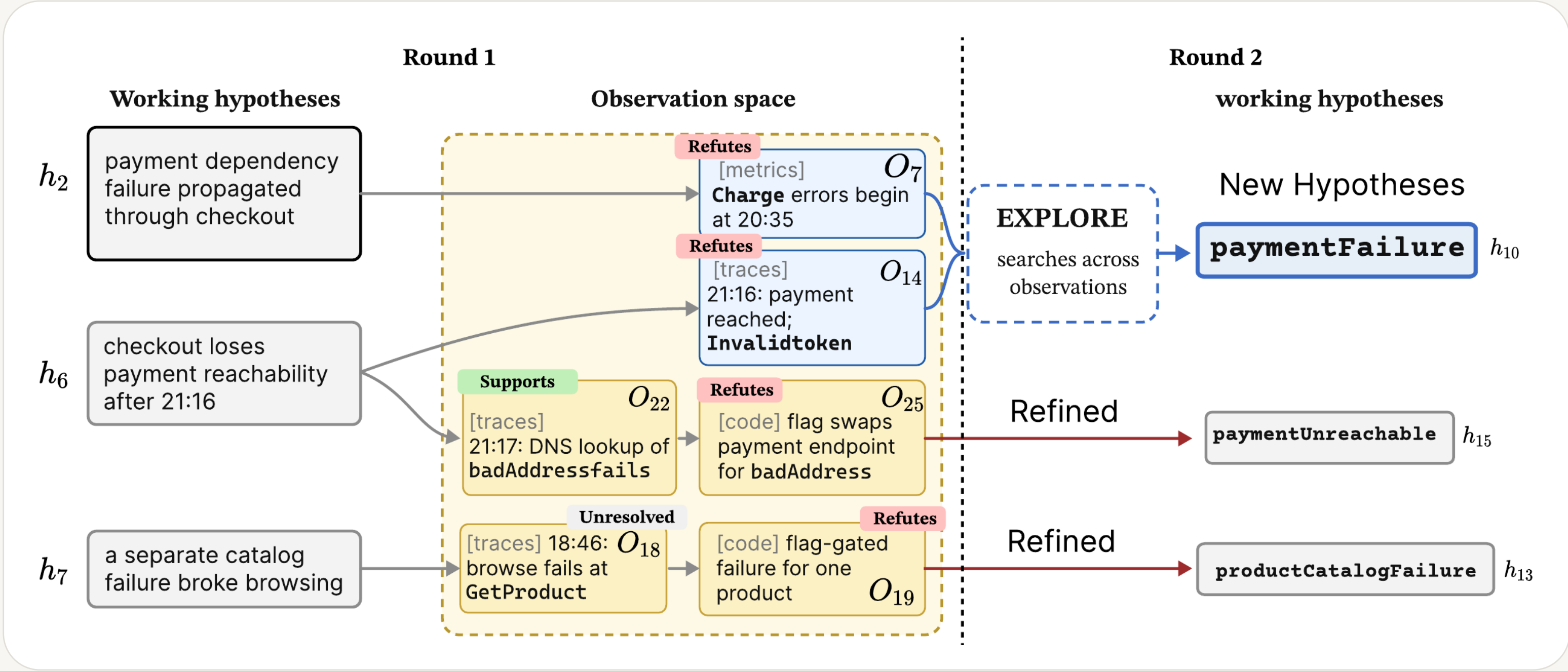


Klahr and Dunbar (1988) describe human scientific discovery as a search over two spaces, hypotheses and experiments.

Klahr & Dunbar, "Dual Space Search During Scientific Reasoning," Cognitive Science 1988
"SHELDON: Reasoning through Dual Search over Hypotheses and Observation Spaces," under review, 2026

Example: an ORCA-Bench incident

Checkout and browsing both fail. The agent has metrics, traces and code.



Round 1
Tests of h_2 and h_6 each produce an observation that refutes them.

Round 2
Explore combines those two observations into a new hypothesis, **paymentFailure**.

Refinement
Code observations narrow h_6 and h_7 into two more specific hypotheses.

Gray: test records an observation · Red: refinement · Blue: observations reused by Explore

Figure 3 from “SHELDON: Reasoning through Dual Search over Hypotheses and Observation Spaces,” under review, 2026 · Gong et al., “ORCA-Bench,” 2026

Does each component matter? Ablations

	DiagnosisArena	ORCA-Bench	ARC-AGI-2	QNN
SHELDON	84.00	66.67	59.17	1.00
w/o comparative analysis	78.00	56.67	50.00	0.88
w/o observation space	70.00	53.33	50.00	0.87
w/o continued exploration	82.00	53.33	57.70	0.73

Higher is better. QNN topology is scored 0–1; the others 0–100.

OBSERVATION SPACE
Removing it lowers every benchmark: –14.0 on DiagnosisArena, –13.3 on ORCA-Bench.

COMPARATIVE ANALYSIS
Removing it costs 6 to 10 points on the first three benchmarks.

AGAINST THE BEST BASELINE, SAME MODEL (GPT-5.6-TERRA)
DiagnosisArena **84.00** vs 78.00 · ORCA-Bench **66.67** vs 40.00 · ARC-AGI-2 **59.17** vs 47.50

“SHELDON: Reasoning through Dual Search over Hypotheses and Observation Spaces,” under review, 2026

WRAP-UP

Takeaways: the design space

A research agent is a set of design choices.

1	Building the agent What choices define one?	ORCHESTRATION	BREADTH	DEPTH	MEMORY	SYNTHESIS
2	Humans in the loop Where do people fit?	INITIATIVE	VISIBILITY	GOAL		
3	Evaluation How do we know it works?	TARGET	JUDGE	TASKS	ENVIRONMENT	
4	Investigative problems When search isn't enough	HYPOTHESES	TESTING	REFINEMENT		

Questions?

yuchengj@stanford.edu



TRY IT

storm.genie.stanford.edu

CODE

github.com/stanford-oval/storm

PAPERS

STORM (NAACL 2024)
Co-STORM (EMNLP 2024)

Appendix

- A1** STORM: full results on FreshWiki
Part 1 · Building a research agent
- A2** STORM in the wild
Part 1 · Building a research agent
- A3** What made the orchestrator work
Part 1 · Building a research agent
- A4** Co-STORM: full automatic evaluation
Part 2 · Humans in the loop
- A5** Collaboration as a first-class objective
Part 2 · Humans in the loop
- A6** A citation is not support
Part 3 · Evaluation
- A7** BrowseComp: hard to find, easy to verify
Part 3 · Evaluation
- A8** BrowseComp is saturated
Part 3 · Evaluation
- A9** BrowseComp scores: the fine print
Part 3 · Evaluation
- A10** Reading list: our work
Further reading
- A11** Reading list: workflows and harnesses
Further reading
- A12** Reading list: training, evaluation, discovery
Further reading

STORM: full results on FreshWiki

Article quality (GPT-4)	ROUGE-1	ROUGE-L	Entity recall	Interest	Organization	Relevance	Coverage
Direct Gen	25.62	12.63	5.08	2.87	4.60	3.10	4.16
RAG	28.52	13.18	7.57	3.14	4.22	3.05	4.08
oRAG	44.26	16.51	12.57	3.90	4.79	4.09	4.70
STORM	45.82	16.70	14.10	3.99	4.82	4.45	4.88

Outline quality (GPT-3.5)	Soft recall	Entity recall
Direct Gen	80.23	32.39
RAG / oRAG	73.59	33.85
STORM	86.26	40.52
STORM without perspectives	84.49	40.12
STORM without conversations	77.97	31.98

84.8 / 85.2

citation recall / precision (%), judged by Mistral-7B-Instruct

Rubric grades (1–5) by Prometheus

STORM in the wild

1M+

people have used them
(Sep 2026)

1.4M

articles generated

250K

pieces of user feedback

30.8K

GitHub stars (Sep 2026)

Open and reproducible

A public research preview and a pip package; re-implemented in LangGraph's official tutorials.

A standard baseline

Compared against by OmniThink, WebWeaver, WikiAutoGen, RaPID, DeepScholar-Bench and others.

A window on real needs

Topics from agriculture to law and emergency management. The logs became WildSeek, which shaped Co-STORM.

What made the orchestrator work

Teach delegation

Each subagent gets an objective, an output format, guidance on tools and sources, and clear task boundaries.

Scale effort to the query

Fact-finding: 1 agent, 3–10 calls. Comparisons: 2–4 subagents × 10–15 calls. Complex research: 10+ subagents.

Tool descriptions are prompts

A model that rewrote a flawed tool description cut later task-completion time by 40%.

Start wide, then narrow

Short, broad queries first, like an expert surveying the field; agents default to overly specific ones.

Parallelize reading

3–5 subagents at once, each making 3+ tool calls in parallel.

Evaluate early and small

Start with about 20 real queries and an LLM judge with a rubric; humans catch what it misses.

Co-STORM: full automatic evaluation

QA turn quality	Consistency	Engagement	Unique URLs
RAG chatbot	4.37	4.13	2.94
STORM + QA	4.34	4.11	2.89
Co-STORM	4.40	4.33	6.04
Co-STORM without multiple experts	4.40	4.32	5.91
Co-STORM without moderator	4.39	4.28	5.67

Final report quality	Relevance	Breadth	Depth	Novelty	Info diversity
RAG chatbot	3.57	3.50	3.26	2.44	0.595
STORM + QA	3.61	3.61	3.43	2.50	0.592
Co-STORM	3.78	3.79	3.77	3.05	0.602
Co-STORM without multiple experts	3.73	3.75	3.77	2.93	0.589
Co-STORM without moderator	3.56	3.69	3.41	2.89	0.577

Yucheng Jiang, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam**, “Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations,” EMNLP 2024

Initiative: collaboration as a first-class objective

CO-GYM · ICLR 2026

An environment where humans and agents act and communicate asynchronously, with simulated or real users.

Collaborative agents’ win rate over fully autonomous ones (real users)

Travel planning 86%

Tabular analysis 74%

Related-work writing 66%

Top failure modes: communication (65% of cases) and situational awareness (40%).

WORKBANK · 2025

What do workers actually want agents to do?

1,500 workers

844 tasks across 104 occupations, with a Human Agency Scale for how much involvement people want.

Tasks fall into four zones: automation “green light”, “red light”, R&D opportunity, low priority.

Build agents people want to work with, and **measure the collaboration, not only the output.**

Yijia Shao, Vinay Samuel, **Yucheng Jiang**, John Yang, Diyi Yang, “Collaborative Gym: A Framework for Enabling and Evaluating Human-Agent Collaboration,” ICLR 2026

Yijia Shao, Humishka Zope, **Yucheng Jiang**, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, Diyi Yang, “Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce,” arXiv 2025

Citations: a citation is not support

40–80%

citation accuracy across deep-research systems

~12K

votes: people prefer more citations, even unsupported ones

85%

of STORM's sentences supported by their citations

Check each claim against its cited passage. **Never reward citation counts.**

Narayanan Venkit et al., “DeepTRACE: Auditing Deep Research AI Systems for Tracking Reliability Across Citations and Evidence,” arXiv 2025

Miroyan et al., “Search Arena: Analyzing Search-Augmented LLMs,” arXiv 2025

Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam**, “Assisting in Writing Wikipedia-like Articles,” 2024

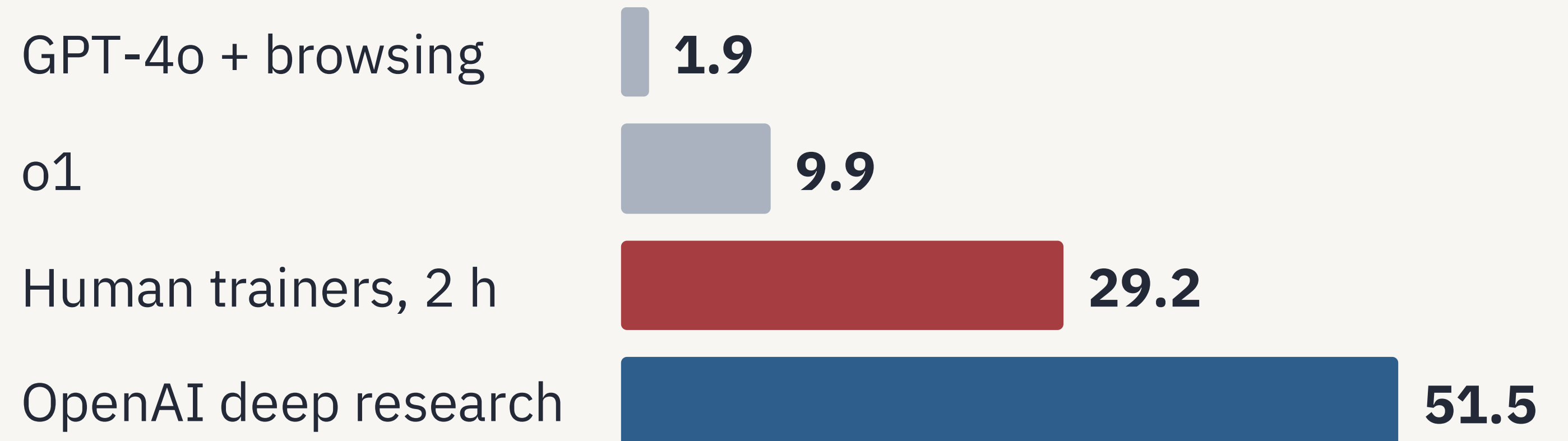
BrowseComp: hard to find, easy to verify

1,266

questions, each checked so that:

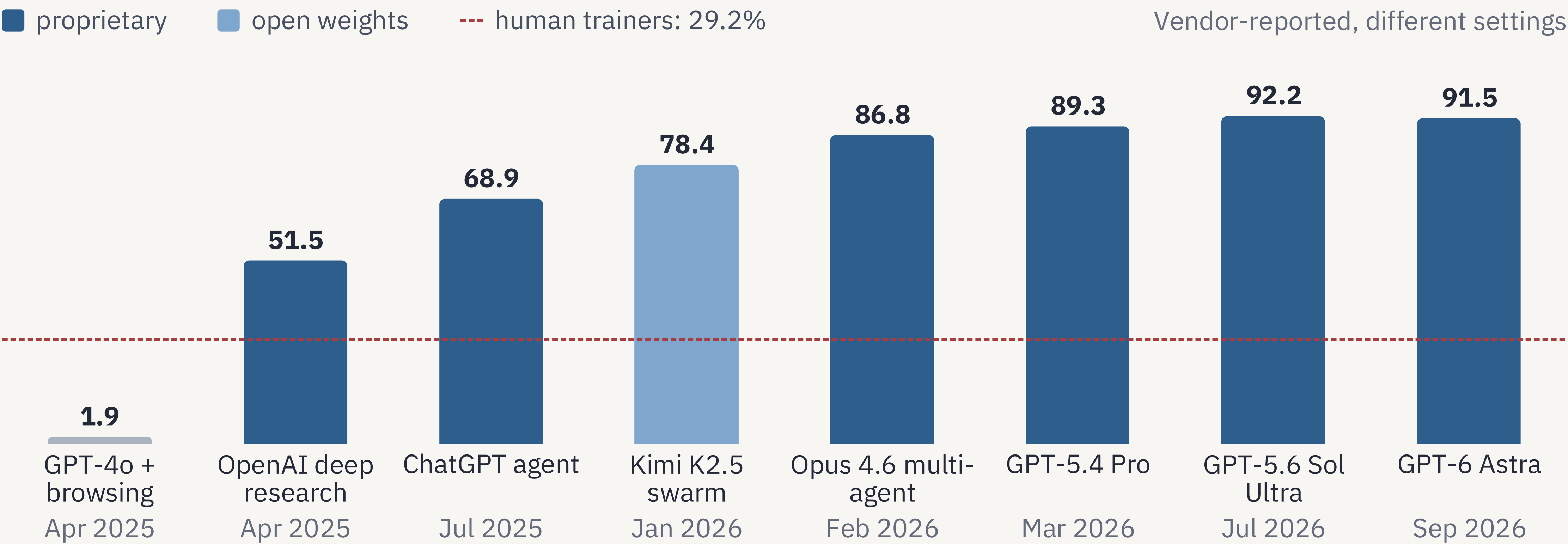
- strong models failed it
- five searches didn't find it
- a person couldn't solve it in 10 minutes

ACCURACY AT LAUNCH (%)



Only 86.4% of the trainers' own answers matched the reference.

BrowseComp is saturated



Top scores now beat the **86.4%** of trainers’ answers that matched the reference.

Vendor launch posts and model cards: OpenAI, Apr 2025–Sep 2026 · Kimi Team, “Kimi K2.5,” 2026
Anthropic, “Introducing Claude Opus 4.6,” 2026

BrowseComp scores: the fine print

System	Date	BrowseComp (%)	Setting, as reported
OpenAI deep research	Feb 2025	51.5	reported in the BrowseComp paper (Apr 2025)
ChatGPT agent	Jul 2025	68.9	pass@1
Tongyi DeepResearch (open)	Sep 2025	43.4	ReAct mode, 30B-A3B
Kimi K2.5 (open)	Jan 2026	60.6 / 74.9 / 78.4	single agent / context management / agent swarm
Claude Opus 4.6	Feb 2026	86.8	multi-agent harness; compaction at 50K, up to 10M tokens
GPT-5.4 / GPT-5.4 Pro	Mar 2026	82.7 / 89.3	as listed by OpenAI
GPT-5.6 Sol / Sol Ultra	Jul 2026	90.4 / 92.2	Ultra mode
GPT-6 Astra	Sep 2026	91.5	“maximum performance at any effort level”

Humanity’s Last Exam with tools, for comparison: 26.6 (deep research, Feb 2025) → 41.6 (ChatGPT agent) → 53.0 (Opus 4.6) → 58.7 (GPT-5.4 Pro) → 65.0 (Claude Fable 5.1, as listed by OpenAI, Sep 2026).

Vendor launch posts and model cards: OpenAI, Apr 2025–Sep 2026 · Tongyi DeepResearch Team, “Tongyi DeepResearch Technical Report,” 2025
Kimi Team, “Kimi K2.5,” 2026 · Anthropic, “Introducing Claude Opus 4.6,” 2026

Reading list: our work

OUR WORK

STORM · Yijia Shao, **Yucheng Jiang**, Theodore A. Kanell, Peter Xu, Omar Khattab, **Monica S. Lam** · NAACL 2024 · arxiv.org/abs/2402.14207

Co-STORM · **Yucheng Jiang**, Yijia Shao, Dekun Ma, **Sina J. Semnani**, **Monica S. Lam** · EMNLP 2024 · arxiv.org/abs/2408.15232

Collaborative Gym · Yijia Shao, Vinay Samuel, **Yucheng Jiang**, John Yang, Diyi Yang · ICLR 2026 · arxiv.org/abs/2412.15701

Future of Work with AI Agents · Yijia Shao, Humishka Zope, **Yucheng Jiang**, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, Diyi Yang · 2025 · arxiv.org/abs/2506.06576

SLIDERS · **Harshit Joshi**, Priyank Shethia, Jadelynn Dao, **Monica S. Lam** · 2026 · arxiv.org/abs/2604.22294

DataSTORM · **Shicheng Liu**, **Yucheng Jiang**, Sajid Farook, Camila Nicollier Sanchez, David Fernando Castro Pena, **Monica S. Lam** · 2026 · arxiv.org/abs/2604.06474

Code: github.com/stanford-oval/storm

Reading list: workflows and harnesses

WORKFLOWS AND HARNESSES

What is an AI Agent Harness?, Databricks, Jun 2026

Building effective agents, Anthropic, Dec 2024

How we built our multi-agent research system, Anthropic, Jun 2025

Don't build multi-agents (2025) and **Multi-agents: what's actually working** (2026), Cognition

Effective context engineering for AI agents, Anthropic, Sep 2025

Context engineering for AI agents: lessons from building Manus, Jul 2025

Deep Agents, LangChain, Jul 2025 · **Recursive Language Models**, Zhang, Kraska, Khattab

WebWeaver, arXiv 2509.13312 · **TTD-DR**, arXiv 2507.16075 · **Meta-Harness**, arXiv 2603.28052

Reading list: training, evaluation, discovery

TRAINING

Search-R1 · 2503.09516
DeepResearcher · 2504.03160
ZeroSearch · 2505.04588
Kimi-Researcher · Moonshot blog, Jun 2025
Tongyi DeepResearch · 2510.24701
DR Tulu · 2511.19399
Rubrics as Rewards · 2507.17746

BetterTogether · 2407.10930
AgentFlow · 2510.05592
Co-Harness · 2607.22688

EVALUATION

BrowseComp · 2504.12516; **-Plus** · 2508.06600
WideSearch · 2508.07999
DeepResearch Bench · 2506.11763; **II** · 2601.08536
ResearchRubrics · 2511.07685
Mind2Web 2 · 2506.21506
DeepTRACE · 2509.04499;
Search Arena · 2506.05334
Search-time contamination · 2508.13180
Eval awareness in BrowseComp · Anthropic, Mar 2026

DISCOVERY AND LEARNING

AI co-scientist · Google, Feb 2025

Robin · FutureHouse, May 2025
Kosmos · 2511.02824
Erdős problems with Gemini · 2601.22401
Unit-distance disproof, remarks · 2605.20695
Research acceleration · OpenAI, Sep 2026
LLMs vs. web search for learning · PNAS Nexus, 2025

Numbers are arXiv identifiers (arxiv.org/abs/...).