# Can AI Agents Synthesize Scientific Conclusions in Health?

Hayoung Jung, Pedro Viana Diniz, Jose Reinaldo Roveda, Abner Fernandes, Haeun Jung, Enoch Tsai, Aleksandra Korolova*, Manoel Horta Ribeiro*

*** Jointly advised this work**

PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY

humans & machines lab

# The Promise of Agents: **Synthesis**

## Deep research in ChatGPT

Plan, research, and synthesize complex questions into a documented report.

## Get reports with Deep Research

Send feedback

Deep Research is a Made by Google agent for users who need to gather, analyze, and understand internal and external information.

## Introducing Perplexity Deep Research

**Today we're launching Deep Research** to save you hours of time by conducting in-depth research and analysis on your behalf. When you ask a Deep Research question, Perplexity performs dozens of searches, reads hundreds of sources, and reasons through the material to autonomously deliver a comprehensive report. It excels at a range of expert-level tasks—from finance and marketing to product research—and attains high benchmarks on Humanity's Last Exam.

## Using Research on Claude

Updated over 3 weeks ago

Research is available for users with paid Claude plans (Pro, Max, Team, or Enterprise) using Claude on the web, Claude Desktop, or Claude Mobile.

Research transforms how Claude finds and analyzes information. Claude operates agentically, conducting multiple searches that build on each other while determining exactly what to investigate next. It explores different angles of your question automatically and works through open questions systematically.

# Synthesis in Health Information Seeking

How People Use ChatGPT*

Aaron Chatterji[1,2] Tom Cunningham[1] David Deming[3] Zoë Hitzig[1,3]
Christopher Ong[1,3] Carl Shan[1] Kevin Wadman[1]

[1]OpenAI
[2]Duke University
[3]Harvard University

September 15, 2025

DIGITAL DOCTORS

**ALMOST 40 PERCENT OF AMERICANS TRUST MEDICAL ADVICE FROM AI CHATBOTS**

A new survey shows just how reliant people have become on programs like ChatGPT when navigating the U.S. healthcare system

By MILES KLEE
AUGUST 4, 2025

Google Is Using A.I. to Answer Your Health Questions. Should You Trust It?

Experts say the new feature may offer dubious advice in response to personal health queries.

Deep Research, Reasoning Models…

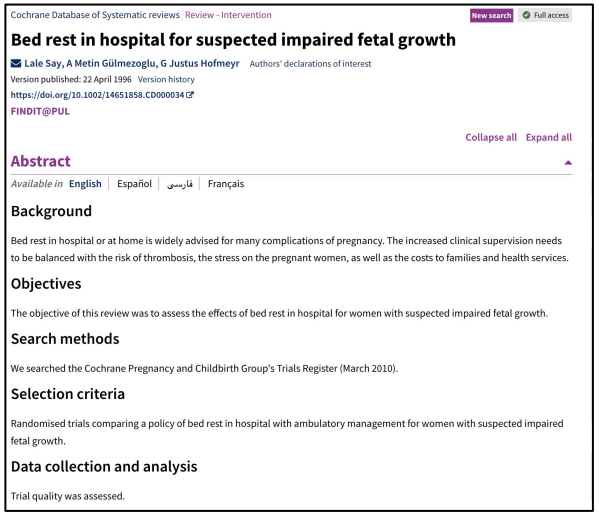
Cochrane Database of Systematic Reviews

**GOAL**: Evaluate AI agents on their ability to synthesize scientific conclusions from the open web

Clean-Room Evaluations

# What is the *Cochrane Database of Systematic Reviews*?

- A collection of expert-written systematic reviews for science & medicine.
- Experts identify, appraise, and synthesize evidence to answer a **research question.** Considered "gold standard" in evidence-based medicine!



Cochrane Database of Systematic reviews Review - Intervention

New search Full access

**Bed rest in hospital for suspected impaired fetal growth**

Lale Say, A Metin Gülmezoglu, G Justus Hofmeyr Authors' declarations of interest

Version published: 22 April 1996 Version history

https://doi.org/10.1002/14651858.CD000034

FINDIT@PUL

Collapse all Expand all

**Abstract**

*Available in* English | Español | فارسی | Français

**Background**

Bed rest in hospital or at home is widely advised for many complications of pregnancy. The increased clinical supervision needs to be balanced with the risk of thrombosis, the stress on the pregnant women, as well as the costs to families and health services.

**Objectives**

The objective of this review was to assess the effects of bed rest in hospital for women with suspected impaired fetal growth.

**Search methods**

We searched the Cochrane Pregnancy and Childbirth Group's Trials Register (March 2010).

**Selection criteria**

Randomised trials comparing a policy of bed rest in hospital with ambulatory management for women with suspected impaired fetal growth.

**Data collection and analysis**

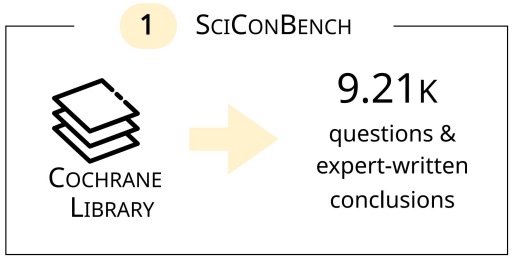
Trial quality was assessed.

**Example of Cochrane Systematic Review Article**

# Methodology

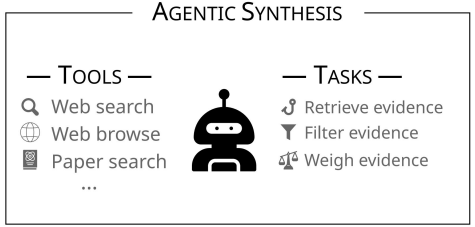
## 1 Creating SciConBench

*Converting Cochrane Systematic Reviews to Benchmark QAs.*



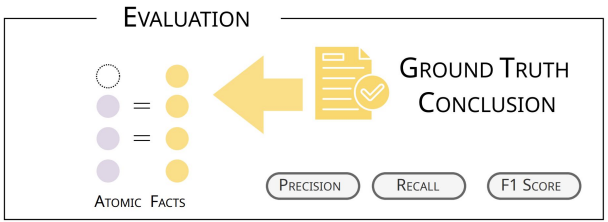

## 2 SciConHarness: Controlled Evaluation in the Clean Room

*A unified-MCP tool harness for controlled retrieval.*



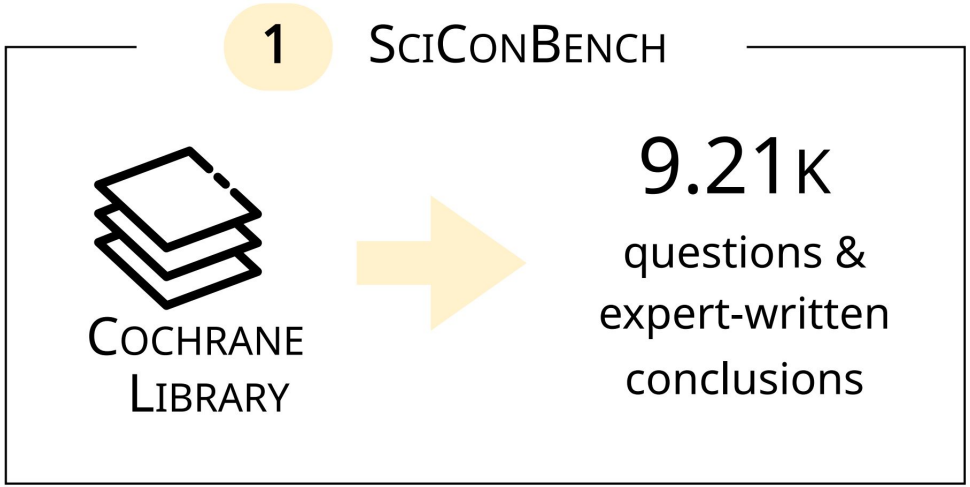

## 3 Measuring Factual Quality of Scientific Conclusions

*Introducing a new precision & recall-based metrics.*



# 1 Creating SciConBench

- **~9.21K Cochrane Systematic Reviews** as of October, 2026.
- Growing concerns around *data contaminations* and *benchmark leakage* due to model pre-training on recent snapshots of the web.
- **Solution**: Our benchmark is *live* (continuous updating on HuggingFace!)



# 1 Creating SciConBench Dataset: Data Preprocessing

- We convert each review into a structured question-answer (QA) evaluation unit.

**Systematic Review Article**




**Objectives**

To assess the benefits and harms of topical antibiotic prophylaxis for corneal abrasion.

**GPT-5-chat**

**Question**

What are the benefits and harms of using topical antibiotic prophylaxis for corneal abrasion?

**Authors' conclusions**

Given that the evidence supporting antibiotic use in corneal abrasion is of very low certainty, we are not able to support a specific antibiotic regimen or draw conclusions about the effects of antibiotic prophylaxis in preventing ocular infection or accelerating epithelial healing. Future research could explore adequately powered RCTs or alternative approaches, such as target trial emulation, while focusing on high-risk populations and antibiotic formulations.

**We use these as the answer!**

# 1 Creating SciConBench Dataset: Validation

- We validate the **Question Generation Pipeline**
- Two medical students evaluate 100 questions on the following dimensions:
  - Faithfulness
  - Completeness
  - Clear & Answerable

High Interrater Agreement!

| Agreement | Faith. | PICO | Clar. |
|---|---|---|---|
| Gwet's AC1 | 0.756 | 0.78 | 1.00 |
| % Agreement | 0.80 | 0.80 | 1.00 |

## Result

Domain-experts assess most generated questions to be faithful (92%), complete (92%), and clear (96%), thereby validating our pipeline!

| Metric | Faithfulness | | PICO Completeness | | | Clarity | |
|---|---|---|---|---|---|---|---|
| | Faithful | Unfaithful | Complete | Partial | Incomplete | Clear | Unclear |
| **Overall** | 92.0% | 8.0% | 92.0% | 6.0% | 2.0% | 96.0% | 4.0% |

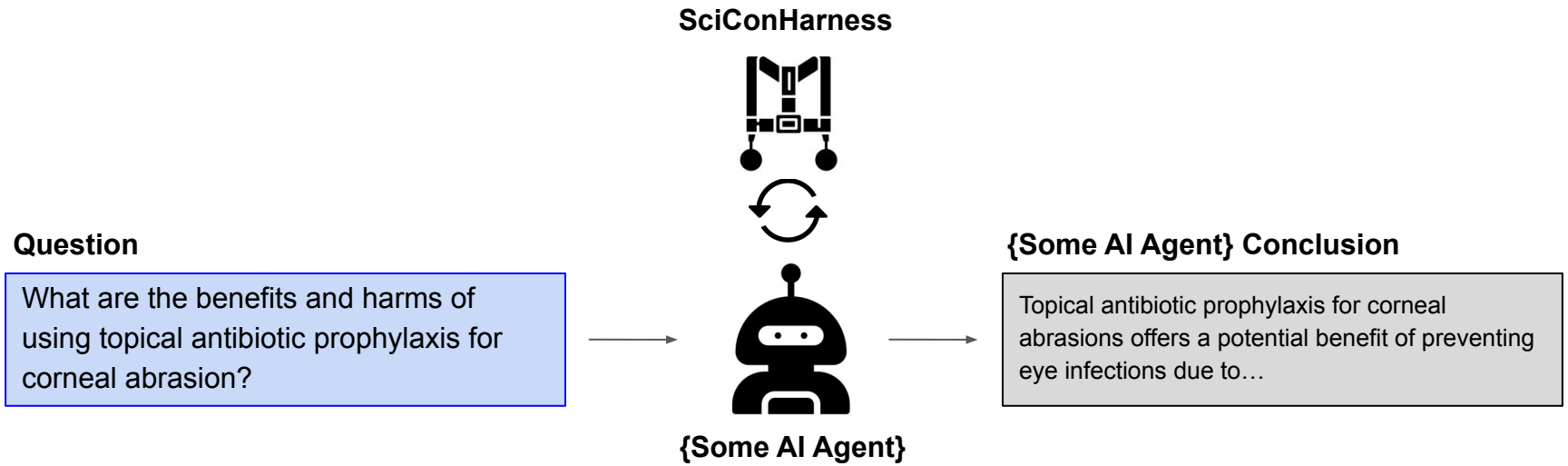
# 2 SciConHarness: Introducing the Harness

- SciConHarness is a **unified MCP-based harness** that provides controlled access to web search and browsing tools, enabling open-domain synthesis.



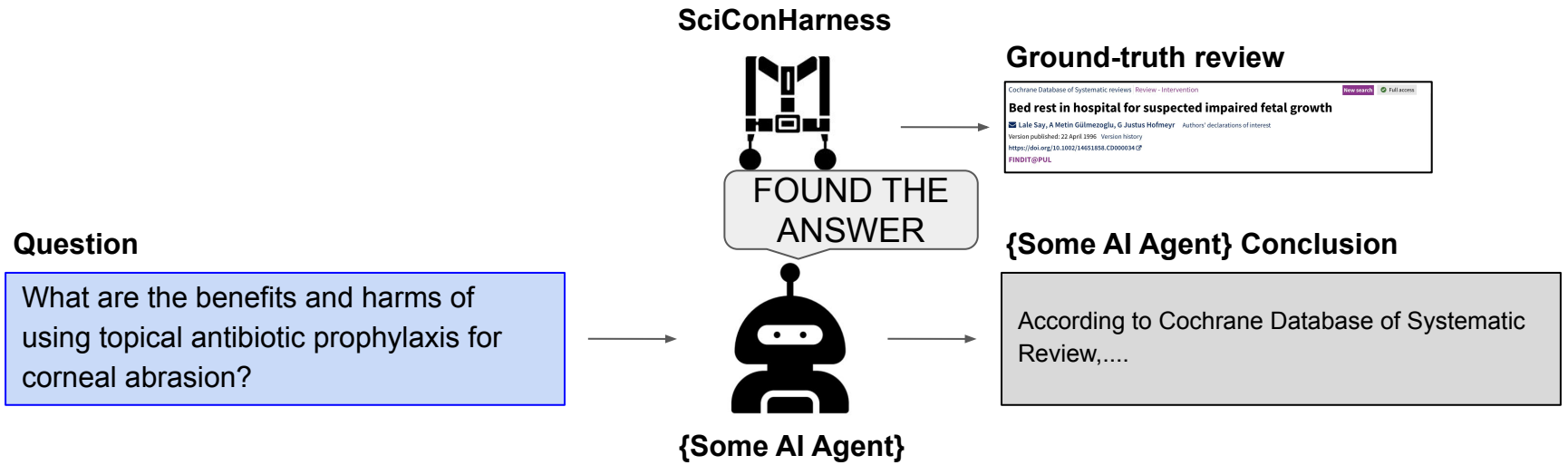
# 2 SciConHarness: Introducing the Harness

- SciConHarness is a **unified MCP-based harness** that provides controlled access to web search and browsing tools, enabling open-domain synthesis.



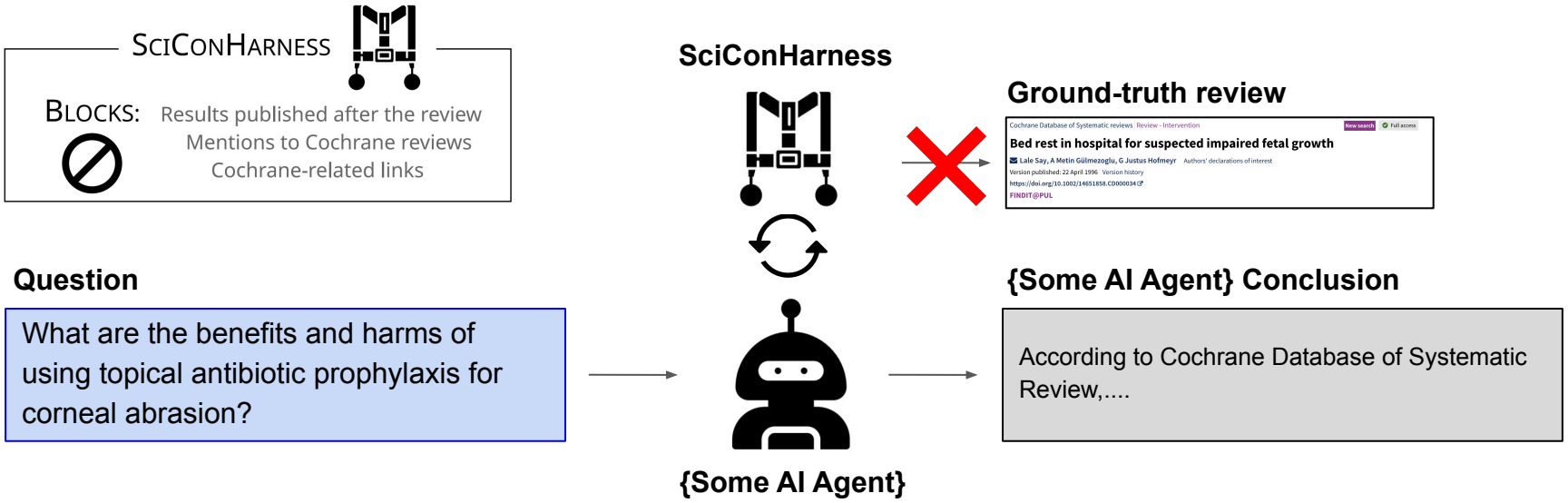
# 2 SciConHarness: Clean Room Evaluation

- Models would often **cheat** by looking up the ground-truth Cochrane reviews!
- This reduces our task to retrieval rather than synthesis. **Bad construct validity**!



# 2 SciConHarness: Clean Room Evaluation

- **Our Solution: *Clean Room Evaluation Protocol.***
- We implement a filtering middleware that uses rules-based heuristic to filter out any leakage in all tool outputs before returning them to the model



# 2 SciConHarness: Validating the Harness

- We sample 150 tool outputs (50 per) & evaluate whether they should be filtered
- **Note:** Recall is more important than precision. We can occasionally filter legit content so that we can mitigate risk of model accessing the reference conclusion.

## Result

Our clean room protocol has high precision (0.933) and recall (0.972). Notably, 100% of ground-truth reviews are successfully removed across all tools, demonstrating robust prevention of direct leakage.

| Tools | Precision | Recall | F1 | Accuracy | % GT Filtered |
|---|---|---|---|---|---|
| `google_search` | 0.92 | 1.00 | 0.958 | 0.96 | 100% |
| `web_browse` | 0.88 | 0.957 | 0.917 | 0.92 | 100% |
| `paper_search` | 1.00 | 0.9615 | 0.98 | 0.98 | 100% |
| **Overall** | 0.933 | 0.972 | 0.952 | 0.953 | 100% |

# 3 Measuring Factual Quality

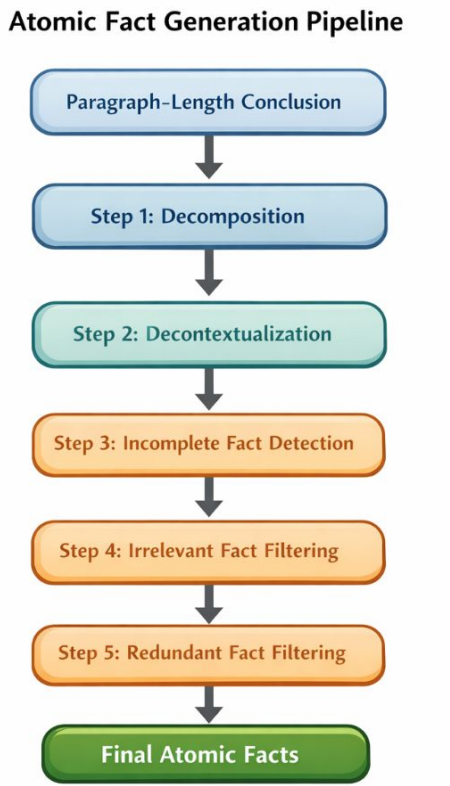
- **Motivation:** Scientific conclusions inform clinical practice and health policy, thus we evaluate their ***factual quality***.
- Long-form conclusions often contain **multiple compositional claims**
- How can we evaluate factual quality in a fine-grained manner?
  - **Our Solution: Decomposing long-form conclusion into atomic facts!**

**{Some AI System} Conclusion**

Topical antibiotic prophylaxis for corneal abrasions offers a potential benefit of preventing eye infections due to the vulnerable nature of the de-epithelialized cornea. The main intended benefit is preventing bacterial keratitis…

# 3 Measuring Factual Quality: Atomic Fact Generation

- We develop a pipeline for generating **atomic facts.**
  - Consisting of five sequential modules based on prior works, using **LLMs**.
  - Decomposing a conclusion into a set of high-quality atomic facts.



# 3 Measuring Factual Quality: Validating the Atomic Facts

- We sample N=200 sentences and their corresponding set of atomic facts, stratified across human- vs. AI-generated sentences
- Two medical doctors annotated across four dimensions: completeness, faithfulness, comprehensiveness, redundancy

## Result

The generated facts were considered to be complete (95.95%), faithful (96.38%), comprehensive (98%), and not redundant (90.5%) with respect to the original sentence, thereby validating our pipeline :D

| Type | Faithfulness | | Completeness | | | Comprehensiveness | | | Redundancy | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Faithful | Unfaithful | Complete | Incomplete | Compound | Compre. | Partial | Not Compre. | No Redunancy | Redunancy |
| Humans | 96.0 | 4.0 | 98.0 | 2.0 | 0.0 | 98.0 | 2.0 | 0.0 | 94.0 | 6.0 |
| LLM | 96.6 | 3.4 | 95.0 | 4.4 | 0.6 | 98.0 | 2.0 | 0.0 | 87.0 | 13.0 |
| Overall | 96.4 | 3.6 | 96.0 | 3.6 | 0.4 | 98.0 | 2.0 | 0.0 | 90.5 | 9.5 |

# 3 Measuring Factual Quality: Atomic Fact Generation

- We develop an LLM pipeline for decomposing conclusions into atomic facts

**{Some AI Agent} Conclusion**

Topical antibiotic prophylaxis for corneal abrasions offers a potential benefit of preventing eye infections due to the vulnerable nature of the de-epithelialized cornea…

**Reference Authors' Conclusion**

**Authors' conclusions**

Given that the evidence supporting antibiotic use in corneal abrasion is of very low certainty, we a antibiotic regimen or draw conclusions about the effects of antibiotic prophylaxis in preventing o epithelial healing. Future research could explore adequately powered RCTs or alternative approac emulation, while focusing on high-risk populations and antibiotic formulations.



Atomic Fact Generation Pipeline

**{Some AI Agent} Atomic Facts**

1) Topical antibiotic prophylaxis may help prevent eye infections.
2) Topical antibiotic prophylaxis is used for corneal abrasions.

**Reference Atomic Facts**

1) Conclusions cannot be drawn about the effects of antibiotic prophylaxis in preventing ocular infection.
2) Conclusions cannot be drawn about the effects of antibiotic prophylaxis in accelerating epithelial healing.

# 3 Measuring Factual Quality: Metrics

**Factual Precision (Min et. al 2023):** Measures factual correctness – whether the generated facts are supported by the reference review and do not contradict it

- High score = supported claims + few/no contradictions. Ranges from 0 to 1.

**Factual Recall (Liu et. al 2025):** Measures factual coverage – whether the generated conclusion cover key reference facts.

- High score = the model covers most of the key reference facts. Ranges from 0 to 1.

# 3 Measuring Factual Quality: LLM Judge Validation

- Two medical doctors annotated N=~130 samples for both factual precision & recall tasks, with any disagreement resolved by a third medical doctor.
- GPT-5.4-mini performs the best on factual precision (F1: 0.84) and recall (F1 0.87)

## Result

Experts agree with the LLM judge on average *more* than they agree with each other

| Pairwise Annotators | Factual Precision | | | Factual Recall | | |
|---|---|---|---|---|---|---|
| | % | $\kappa$ | AC1 | % | $\kappa$ | AC1 |
| Expert A – Expert B | 0.691 | 0.519 | 0.545 | 0.832 | 0.658 | 0.670 |
| Expert A – LLM | 0.707 | 0.526 | 0.579 | 0.885 | 0.753 | 0.785 |
| Expert B – LLM | 0.683 | 0.497 | 0.541 | 0.823 | 0.637 | 0.660 |
| Avg. Expert – LLM | 0.695 | 0.512 | 0.560 | 0.854 | 0.695 | 0.723 |

# Evaluation Details

- We evaluate 8 frontier LLMs and deep research models on N=268 SciConBench samples published after each model's latest knowledge cutoff.
- Using default settings and the highest reasoning level, we test:
  - **Base model:** *parametric knowledge*-only; no tools
  - **SciConHarness:** web-retrieval allowed (including ground-truth)
  - **SciConHarness + Clean Room**: web-retrieval + filters access to ground-truth artifacts

# Benchmark Results

- Models and deep research agents have substantial room for improvement.
- Our clean-room evaluation *consistently* attenuates performance.

| | Factual Precision (Var) | Factual Recall (Var) | Factual F1 (Var) | $\Delta_{\text{Tools}}$ F1 | $\Delta_{\text{Clean}}$ F1 |
|---|---|---|---|---|---|
| *Base Models (No Tools)* | | | | | |
| gpt-5.1 | 0.366 (0.019) | 0.382 (0.057) | 0.332 (0.027) | – | – |
| claude-sonnet-4.5 | 0.464 (0.025) | 0.270 (0.049) | 0.291 (0.037) | – | – |
| gemini-3-pro | 0.339 (0.032) | 0.246 (0.045) | 0.239 (0.029) | – | – |
| *Models (SciConHarness)* | | | | | |
| gpt-5.1 | 0.329 (0.017) | 0.446 (0.062) | 0.344 (0.026) | +0.012 | – |
| claude-sonnet-4.5 | 0.435 (0.043) | 0.409 (0.080) | 0.382 (0.054) | +0.091 | – |
| gemini-3-pro | 0.311 (0.035) | 0.222 (0.048) | 0.213 (0.035) | -0.025 | – |
| sonar-reasoning-pro | 0.547 (0.064) | 0.363 (0.084) | 0.392 (0.075) | – | – |
| *Models (SciConHarness + clean-room)* | | | | | |
| gpt-5.1 | 0.294 (0.017) | 0.408 (0.065) | 0.300 (0.024) | – | -0.044 |
| claude-sonnet-4.5 | 0.350 (0.020) | 0.329 (0.057) | 0.297 (0.030) | – | -0.085 |
| gemini-3-pro | 0.294 (0.034) | 0.206 (0.042) | 0.194 (0.029) | – | -0.020 |
| sonar-reasoning-pro | 0.384 (0.032) | 0.205 (0.044) | 0.220 (0.035) | – | -0.172 |
| *DR (SciConHarness)* | | | | | |
| DR Tulu | 0.308 (0.042) | 0.178 (0.034) | 0.175 (0.027) | – | – |
| sonar-deep-research† | 0.383 (0.028) | 0.396 (0.066) | 0.351 (0.041) | – | – |
| o4-mini-deep-research | 0.593 (0.044) | 0.386 (0.063) | 0.427 (0.054) | – | – |
| o3-deep-research | **0.628** (0.045) | **0.483** (0.069) | **0.508** (0.051) | – | – |
| *DR (SciConHarness + clean-room)* | | | | | |
| DR Tulu | 0.259 (0.038) | 0.168 (0.034) | 0.145 (0.023) | – | -0.030 |
| sonar-deep-research† | 0.357 (0.036) | 0.243 (0.047) | 0.237 (0.034) | – | -0.115 |
| o4-mini-deep-research | 0.467 (0.028) | 0.298 (0.051) | 0.315 (0.039) | – | -0.113 |
| o3-deep-research | 0.441 (0.033) | 0.342 (0.054) | 0.337 (0.035) | – | -0.170 |

# Benchmark Results: Factual Quality Issues

- Factual quality issues were pervasive:
  - 44.8–84.0% of conclusions contained at least one contradiction
  - Nearly all contained at least one unsupported fact.

| Model | ≥ 1 Contr. | ≥ 1 Not Supp. |
|---|---|---|
| *Base Models (No Tools)* | | |
| gpt-5.1 | 77.2 | 100.0 |
| claude-sonnet-4.5 | 56.7 | 100.0 |
| gemini-3-pro | 70.1 | 100.0 |
| *Models (SciConHarness)* | | |
| gpt-5.1 | 84.0 | 100.0 |
| claude-sonnet-4.5 | 60.8 | 99.3 |
| gemini-3-pro | 62.7 | 99.3 |
| sonar-reasoning-pro | 56.7 | 94.0 |
| *Models (SciConHarness + clean-room)* | | |
| gpt-5.1 | 80.6 | 100.0 |
| claude-sonnet-4.5 | 73.9 | 100.0 |
| gemini-3-pro | 61.2 | 99.3 |
| sonar-reasoning-pro | 70.9 | 97.4 |
| *DR (SciConHarness)* | | |
| DR Tulu | 47.8 | 98.9 |
| sonar-deep-research† | 61.0 | 100.0 |
| o4-mini-deep-research | 61.9 | 97.4 |
| o3-deep-research | 55.2 | 95.9 |
| *DR (SciConHarness + clean-room)* | | |
| DR Tulu | 44.8 | 99.3 |
| sonar-deep-research† | 62.0 | 100.0 |
| o4-mini-deep-research | 75.4 | 98.5 |
| o3-deep-research | 73.1 | 99.6 |
| *Consumer-Facing Agents* | | |
| Google AI Overview | 56.3 | 98.9 |
| Google AI Mode | 59.0 | 99.6 |
| OpenEvidence | 50.8 | 100.0 |

# Benchmark Results: Failure Modes

- Common failure modes:
  - **Incorrect Direction-of-Effect.** Flipping treatment effect direction of an intervention.
  - **Evidence Quality Mischaracterization.** Claiming an evidence is much higher quality.
  - **Lack of Specificity.** Collapse nuanced outcome-level findings into overly broad summaries.
- Example of Incorrect Direction-of-Effect:

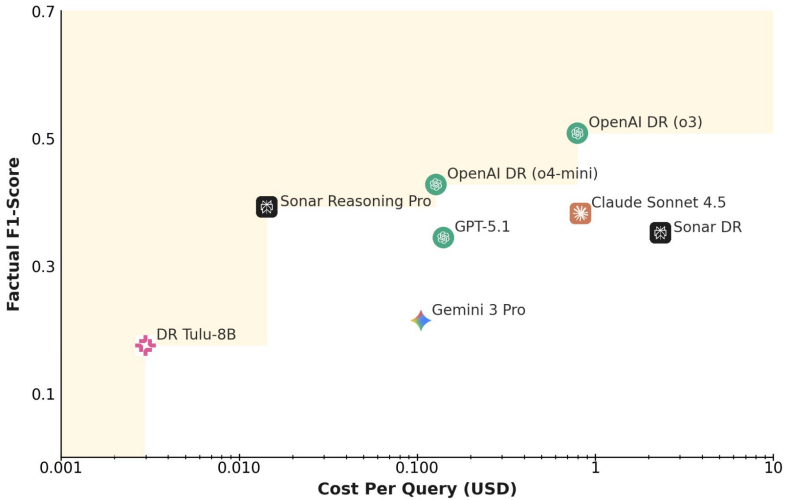
GENERATED CONCLUSION (`o3-deep-research`, tools): Large trials (e.g. STICH and STICH II) and meta-analyses found **no statistically significant improvement in survival or functional recovery** with early open surgery compared to medical management alone

CDSR REFERENCE ARTICLE (`HTTPS://WWW.COCHRANELIBRARY.COM/CDSR/DOI/10.1002/14651858.CD015387.PUB2`): For people with spontaneous supratentorial ICH, surgery aimed at clot removal may **increase the chance of achieving good functional outcome and may reduce all-cause mortality and 30-day case fatality** compared to standard medical management.

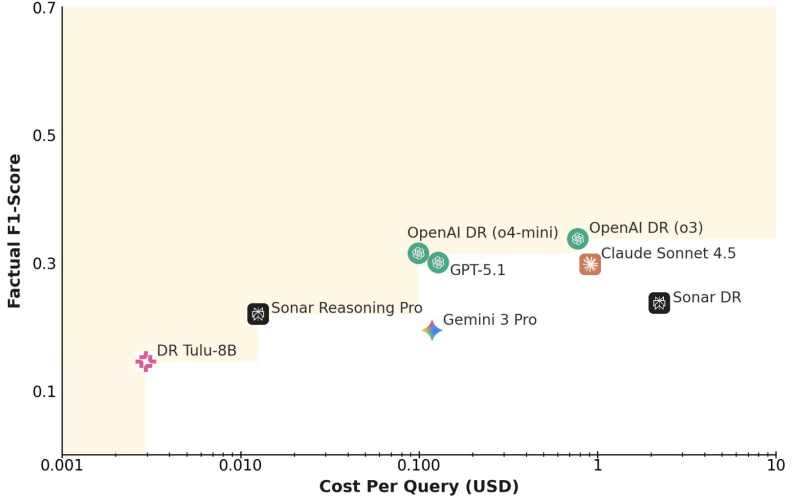
# Benchmark Results: Cost vs. Performance Tradeoffs

- Applying our Clean-Room Protocol consistently flattens the pareto curve.

SciConHarness (Clean-Room **Disabled**)



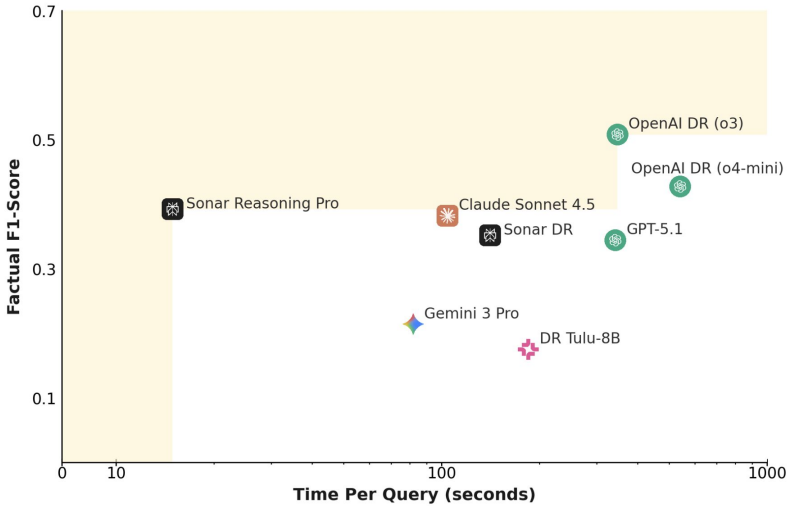

SciConHarness (Clean-Room **Enabled**)



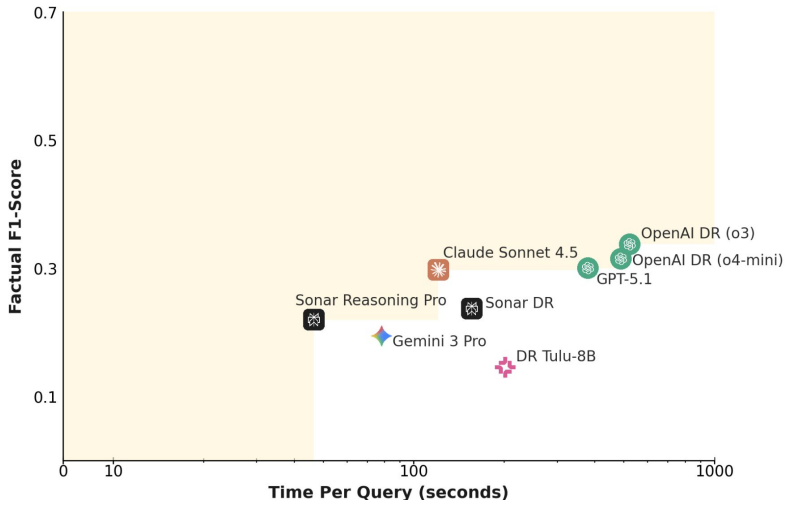
# Benchmark Results: Time vs. Performance Tradeoffs

- **Time is money.** Applying our Clean-Room Protocol *consistently* flattens and *pushes* the curve to the right.

SciConHarness (Clean Room **Disabled**)

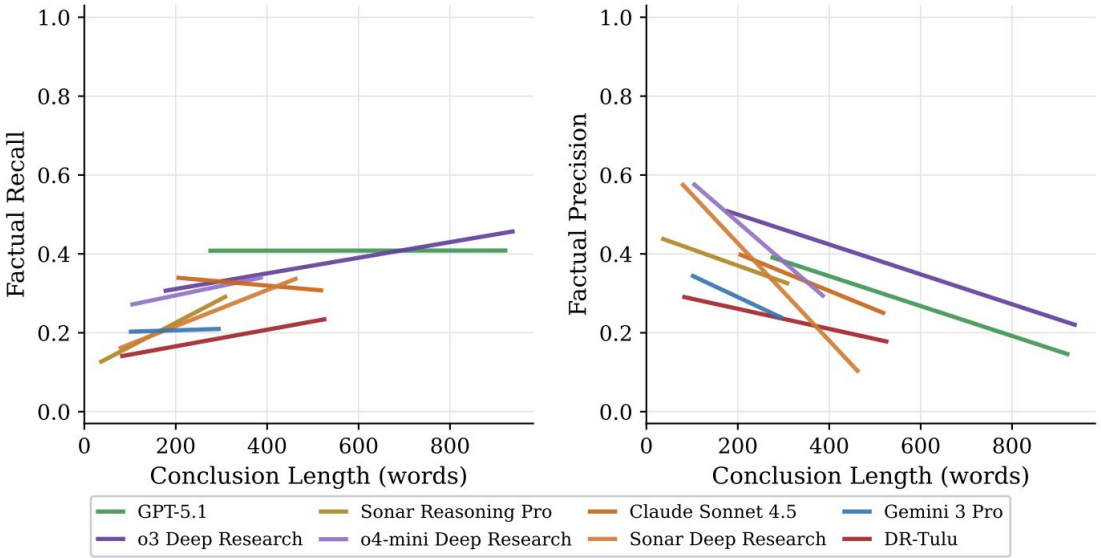



SciConHarness (Clean Room **Enabled**)



# Benchmark Results: So why not just write longer?

- Longer conclusions increase recall but decreases precision.
- As a result, our factual F1 is robust to verbosity: models cannot inflate performance simply by generating longer conclusions.



# Auditing Consumer-Facing Agents

- SciConBench can be applied to audit proprietary, consumer-facing agents, which are increasingly used by laypeople and clinicians in health contexts.
- We evaluate three agents: Google AI Overview, Google AI Mode, and OpenEvidence on the same N=268 samples.
- No clean-room protocol was applied; these agents can simply retrieve the ground-truth answers from the web.

# Audits: Consumer-Facing Agents

- We audit consumer-facing agents e.g., OpenEvidence & Google AI Overview!
- Consumer-facing agents generate unreliable scientific conclusions despite access to ground-truth artifacts.
- Risk for **high-stakes decision-making** in real-world health contexts!

| | **Precision** | **Recall** | **F1** |
|---|---|---|---|
| Google AI Mode | 0.443 (0.048) | 0.380 (0.077) | 0.361 (0.054) |
| Google AI Overview | 0.508 (0.044) | 0.367 (0.061) | 0.384 (0.048) |
| OpenEvidence | **0.580** (0.028) | **0.541** (0.070) | **0.522** (0.042) |

# Takeaways + Contributions

1. We contribute SciConBench, one of the largest, *live* benchmark for evaluating AI agents on the long-horizon task of *scientific conclusion synthesis*.

2. We contribute SciConHarness, a clean-room harness that provides controlled web tools for AI agents, enabling *valid measurement* of synthesis capabilities.

3. Our evaluations & audits reveal that frontier models and agents often synthesize unreliable conclusions, raising concerns for *high-stakes health decision-making*.

# Thank you!

Hayoung Jung

**Contacts: hayoungjung.me**
**hayoung@cs.princeton.edu**







