Stanford Artificial Intelligence

Stanford University Human-Centered Artificial Intelligence

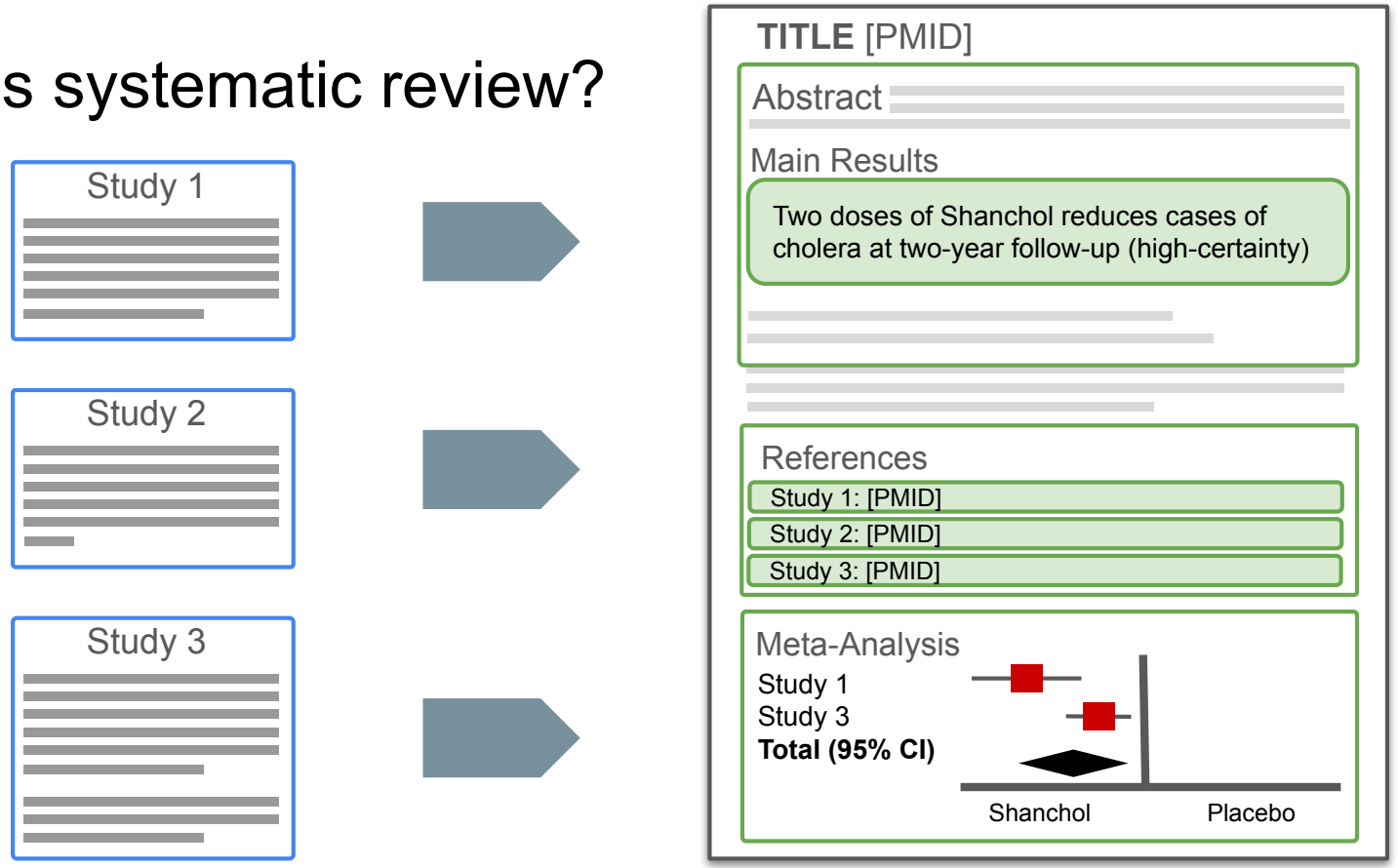
Medical AI and ComputeR Vision Lab

Stanford Data Science

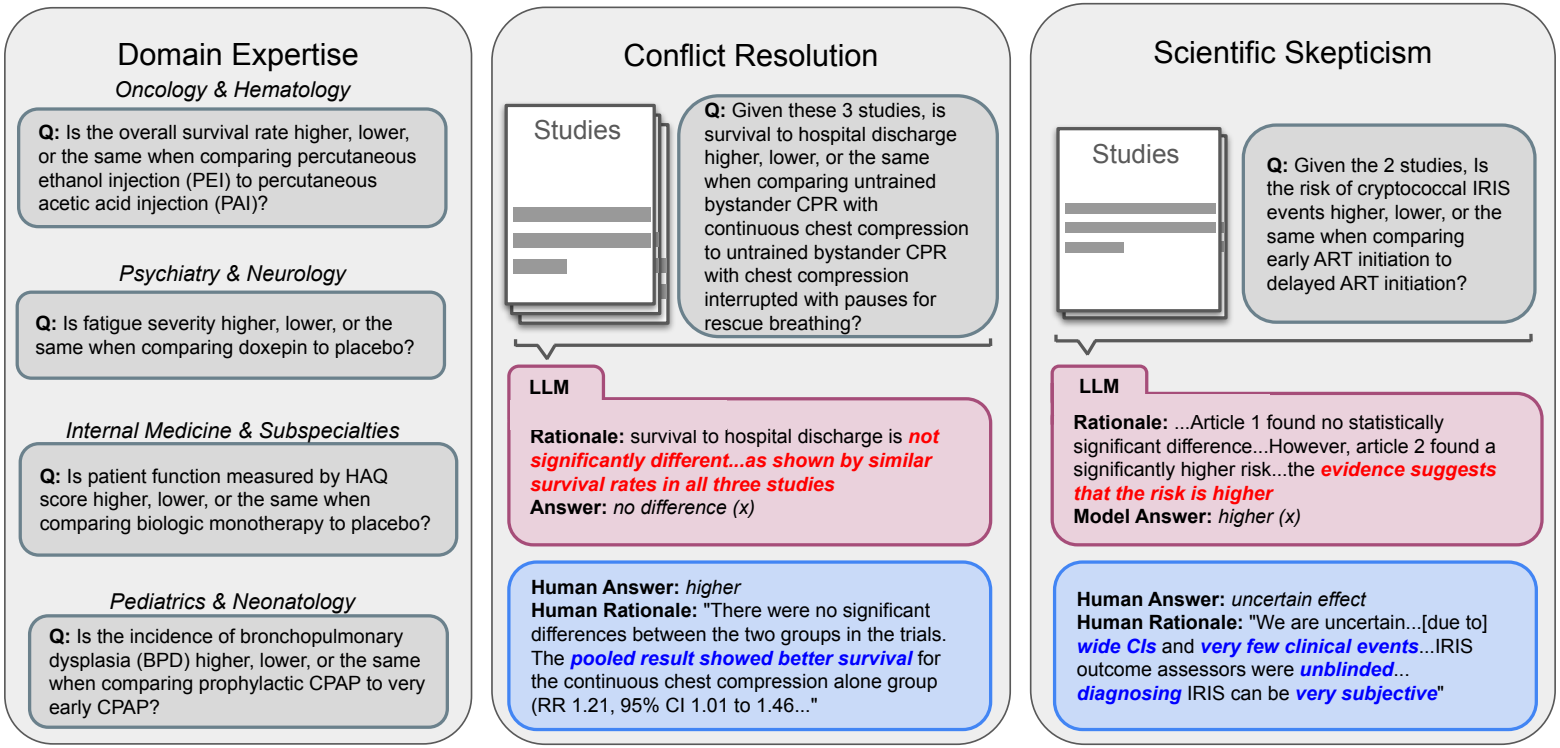
Can Large Language Models Match the Conclusions of Systematic Reviews?

MedEvidence: **Christopher Polzak***, Alejandro Lozano*, Min Woo Sun* et al **MARVL Lab**

What is systematic review?



Assessing core LLM capabilities



1 Systematic Review Selection

PubMed

TITLE [PMID]

Abstract

Main Results

Two doses of Shanchol reduces cases of cholera at two-year follow-up (high-certainty)

References

Study 1: [PMID]

Study 2: [PMID]

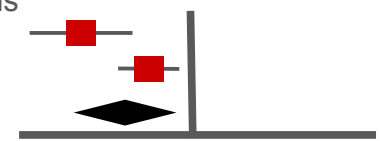
Study 3: [PMID]

Meta-Analysis

Study 1

Study 3

Total (95% CI)



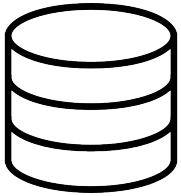
2 Conclusion Conversion

Systematic Review PubMed ID

Q: Is the risk of cholera at two-year follow-up higher, lower, or the same when comparing two doses of Dukoral with or without a booster dose to placebo?

A: *Lower*

Evidence Certainty: *High*



MedEvidence

3 Relevant Study Selection

Relevant studies

Study 1: [PMID]

Study 3: [PMID]

4 Question Validation

PubMed

✓ Study 1

Population

Cholera Cases

✓ Study 3

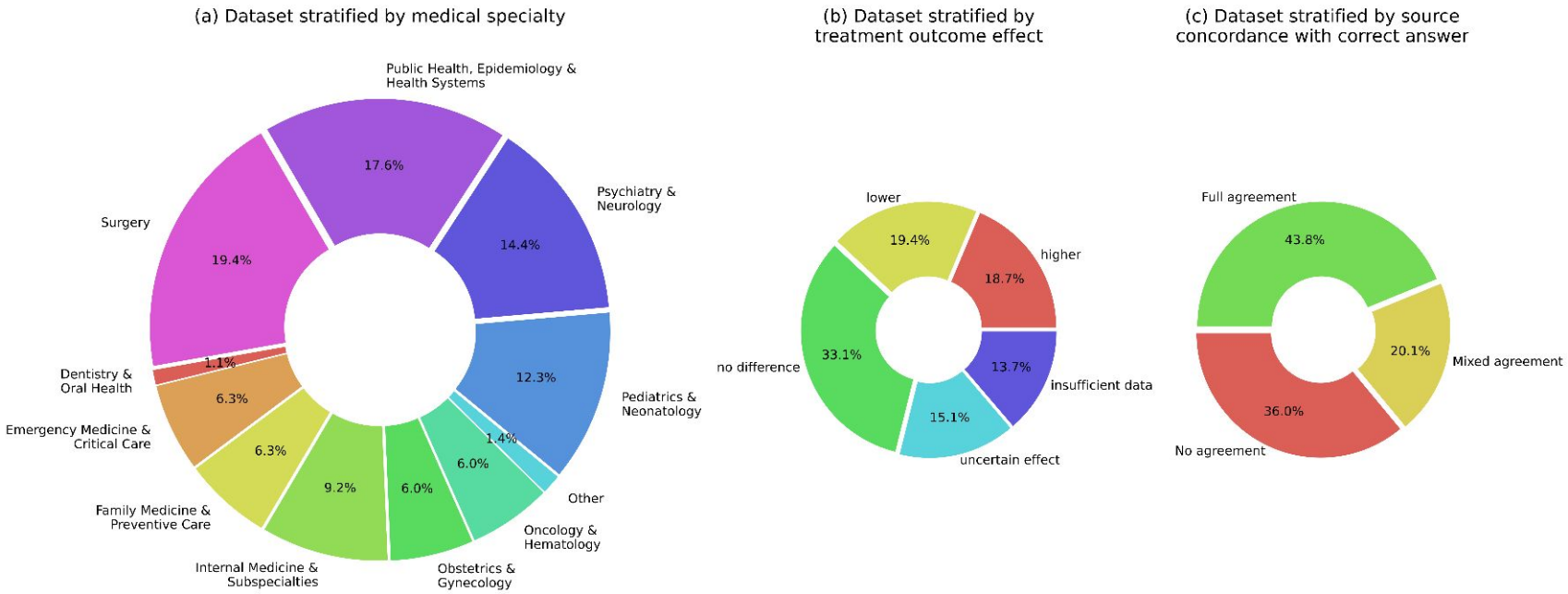
Population

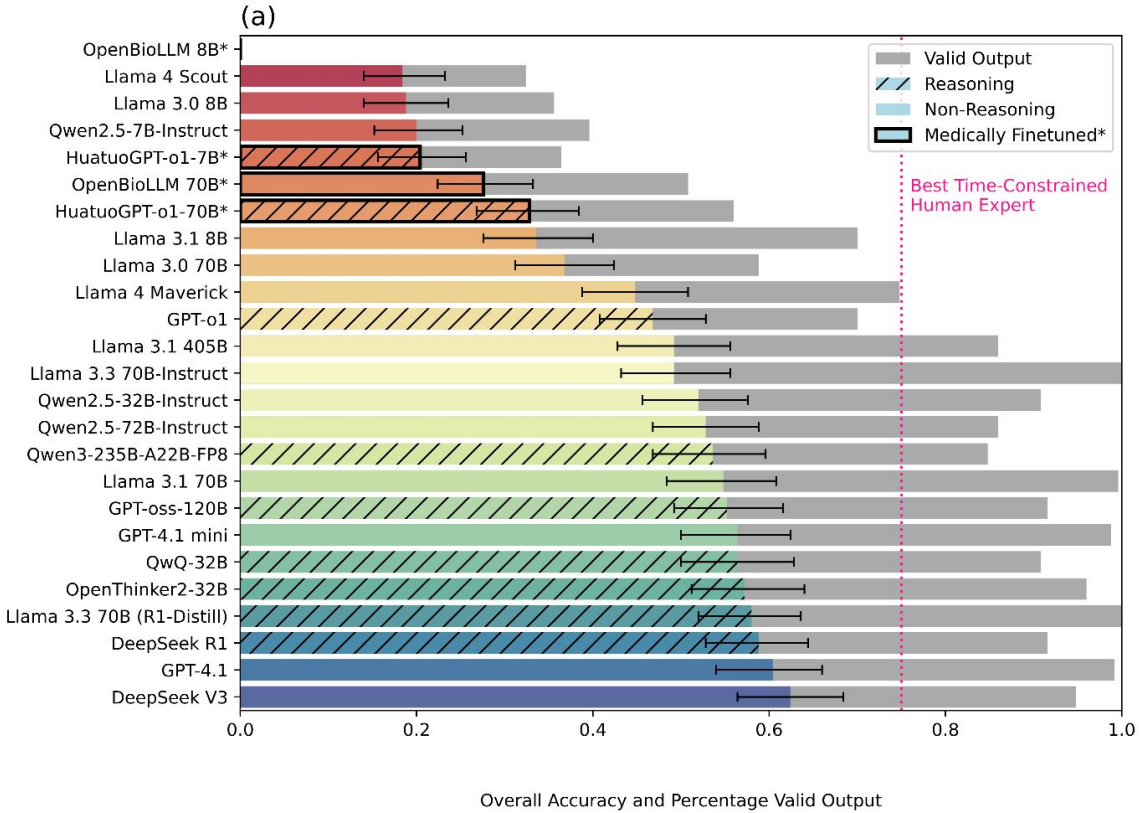
Severe Cholera

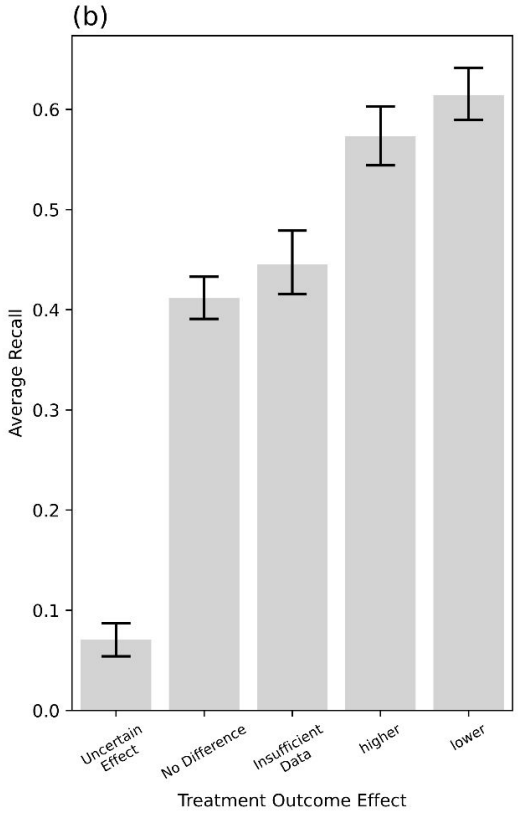
Cholera Cases

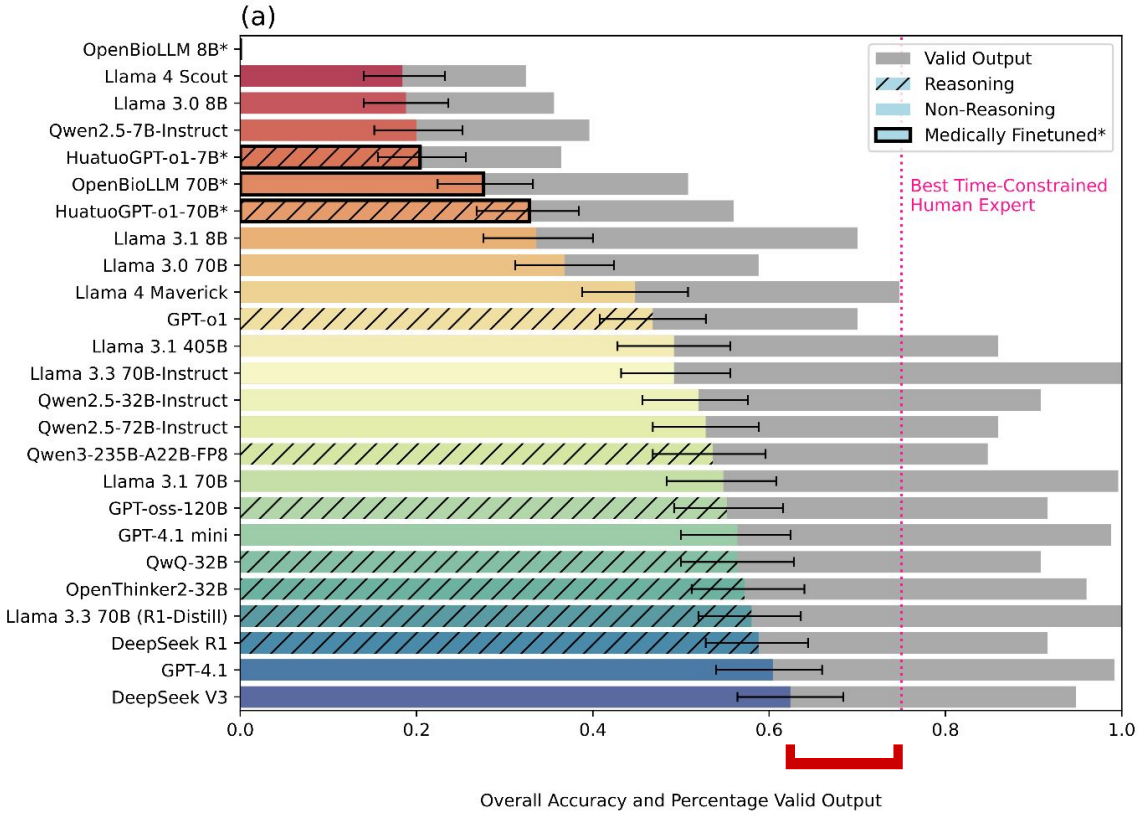
MedEvidence: **Christopher Polzak***, Alejandro Lozano*, Min Woo Sun* et al **MARVL Lab**

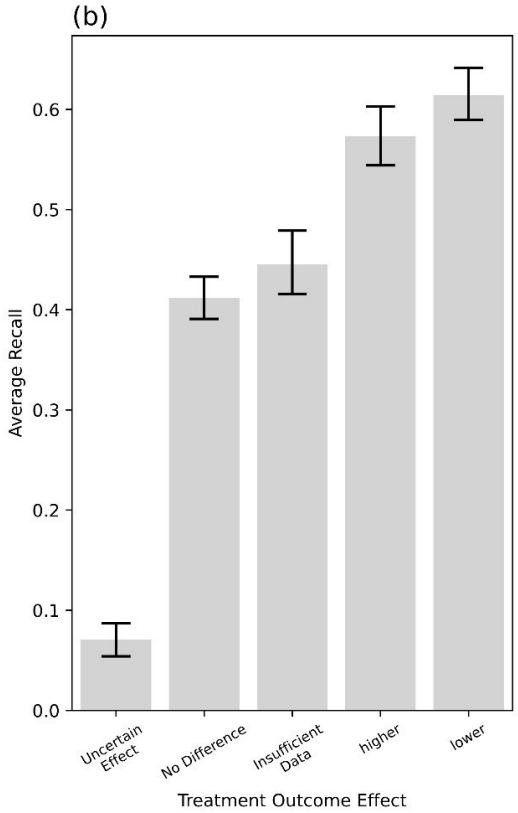
Dataset statistics

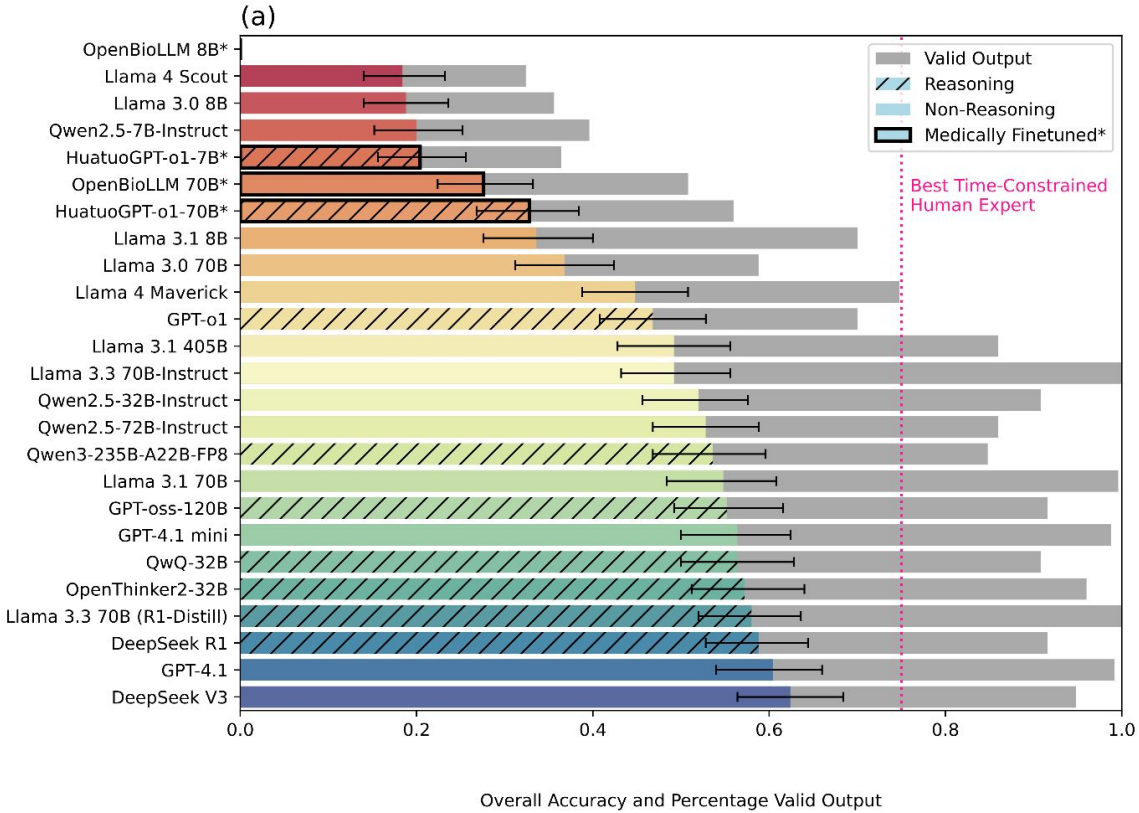


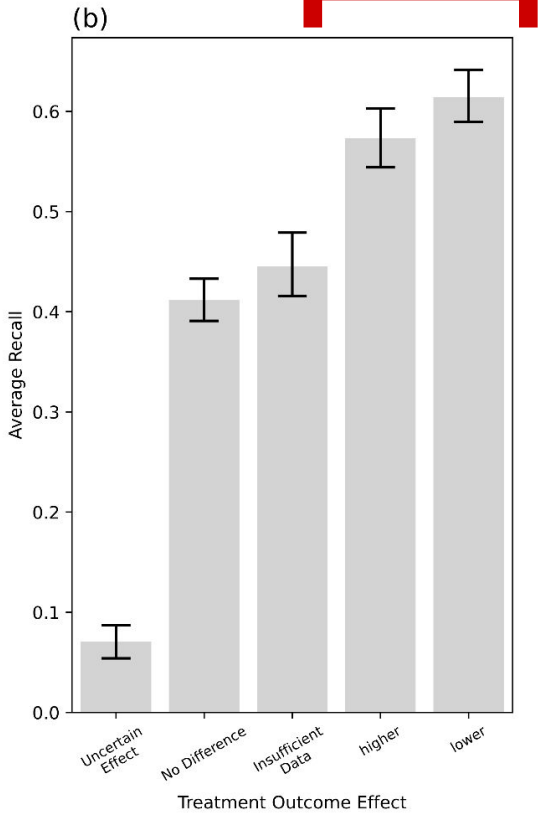

(a)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Valid Output
Reasoning
Non-Reasoning
Medically Finetuned*
Best Time-Constrained Human Expert
0.0
0.2
0.4
0.6
0.8
1.0
Overall Accuracy and Percentage Valid Output

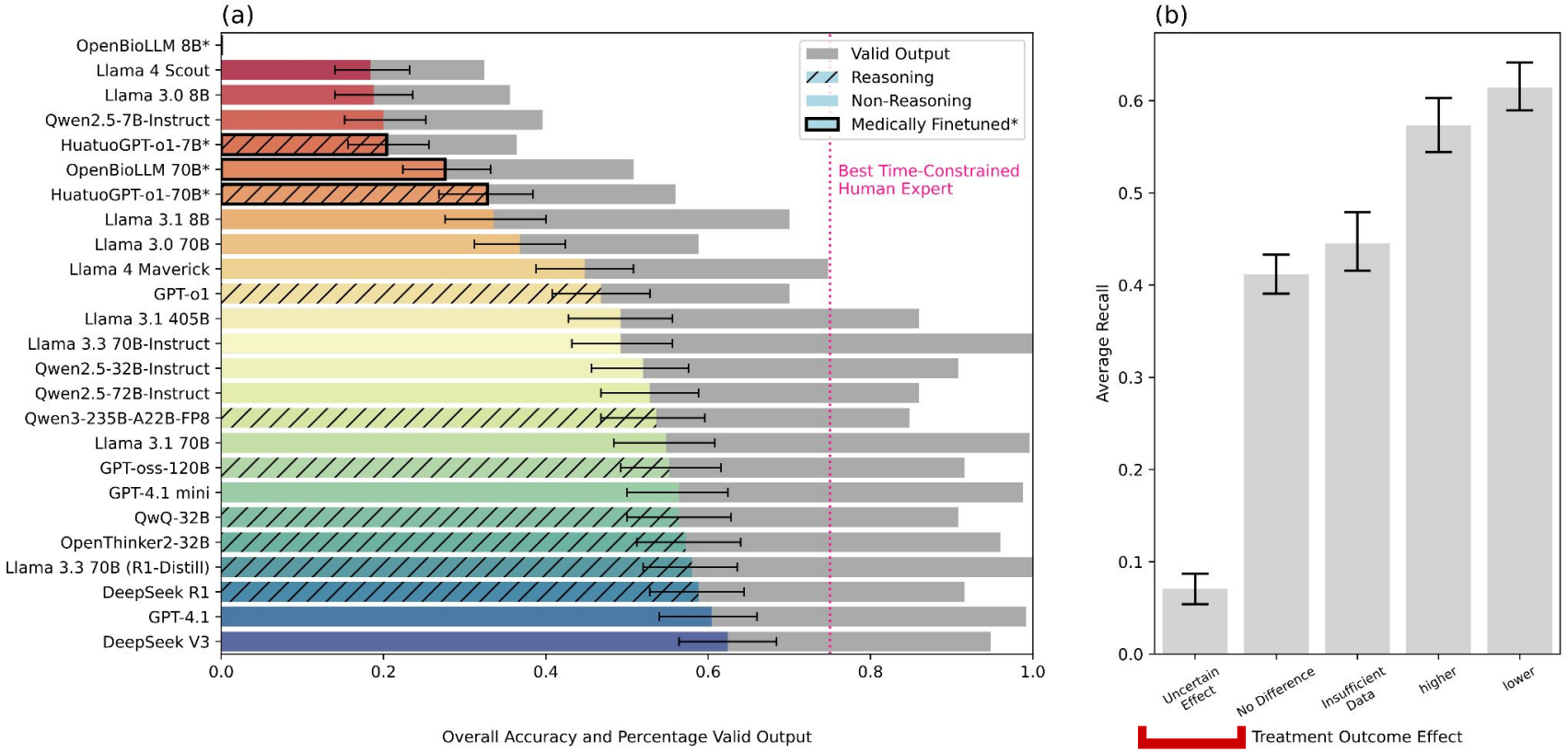

(b)
Average Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Uncertain Effect
No Difference
Insufficient Data
higher
lower
Treatment Outcome Effect

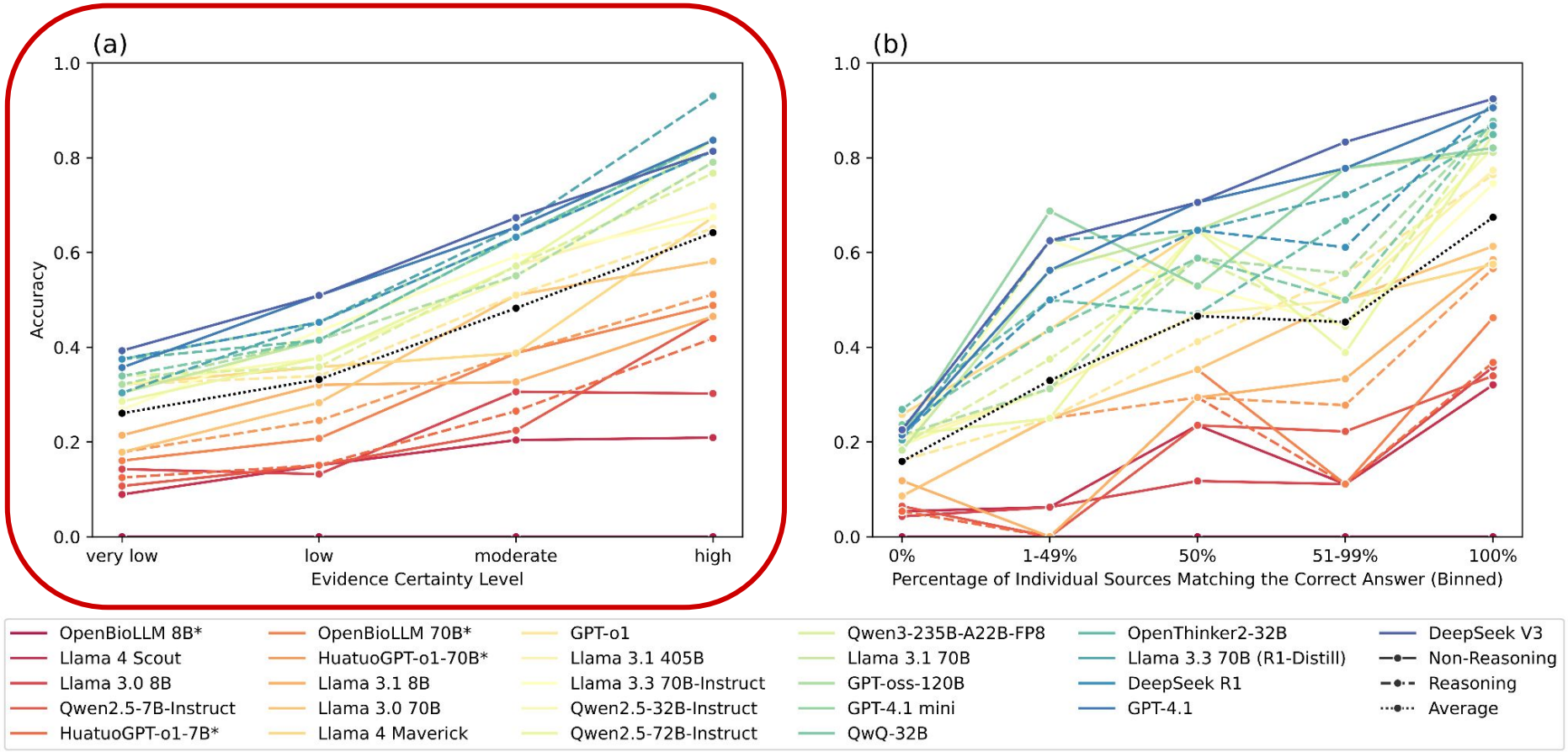

(a)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Valid Output
Reasoning
Non-Reasoning
Medically Finetuned*
Best Time-Constrained Human Expert
0.0
0.2
0.4
0.6
0.8
1.0
Overall Accuracy and Percentage Valid Output

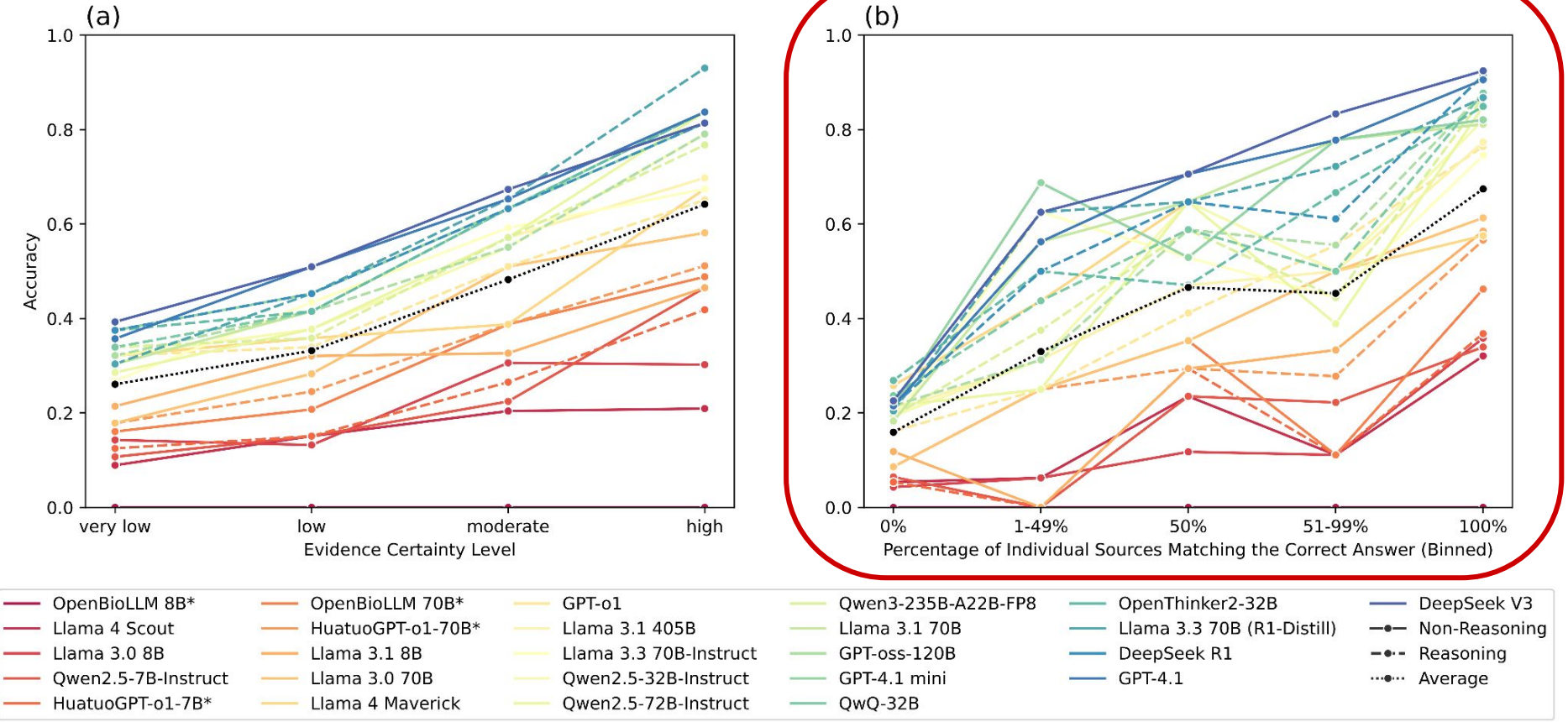

(b)
Average Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Uncertain Effect
No Difference
Insufficient Data
higher
lower
Treatment Outcome Effect

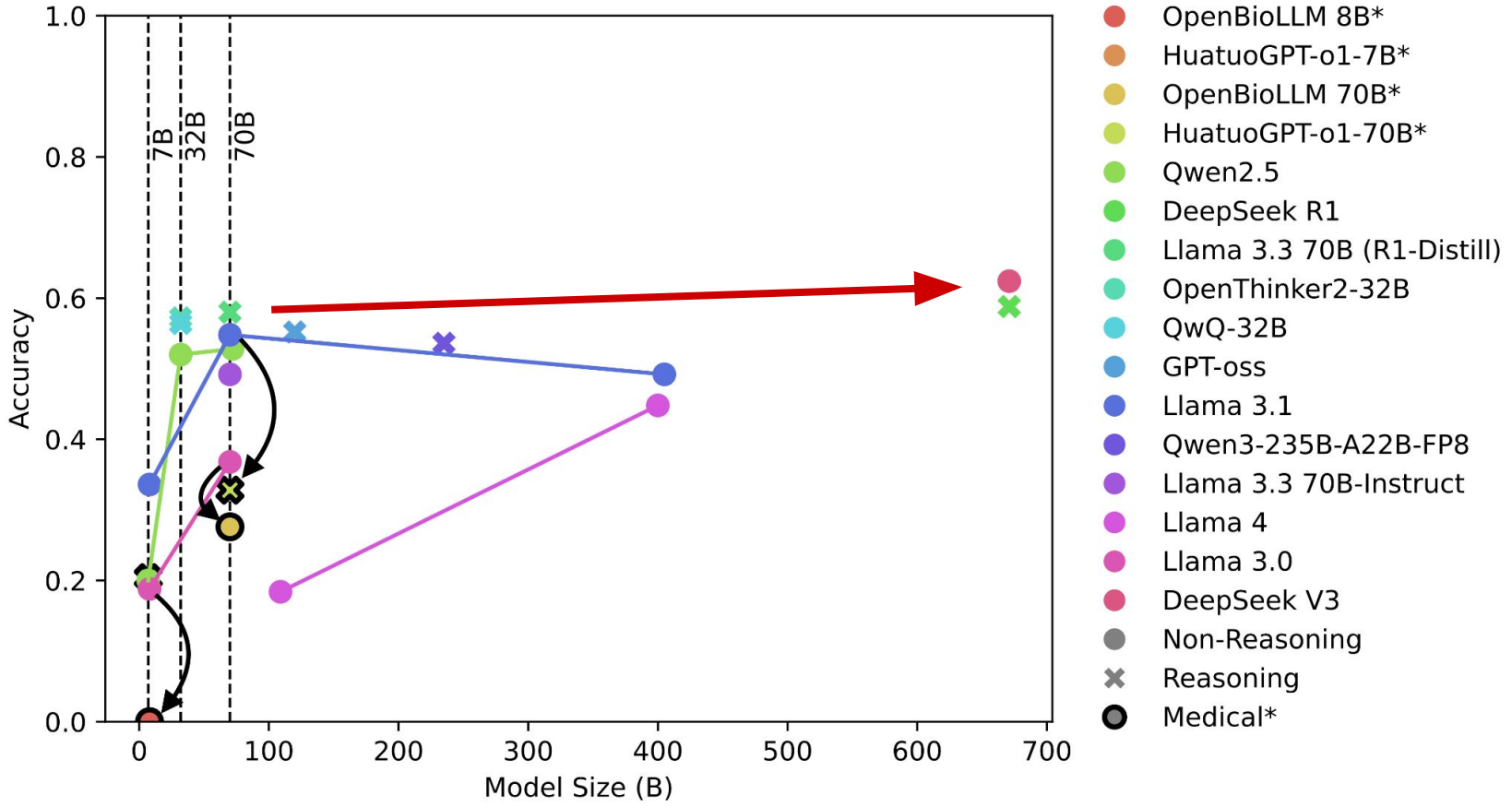

(a)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Valid Output
Reasoning
Non-Reasoning
Medically Finetuned*
Best Time-Constrained Human Expert
0.0
0.2
0.4
0.6
0.8
1.0
Overall Accuracy and Percentage Valid Output


(b)
Average Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Uncertain Effect
No Difference
Insufficient Data
higher
lower
Treatment Outcome Effect


(a)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Valid Output
Reasoning
Non-Reasoning
Medically Finetuned*
Best Time-Constrained Human Expert
0.0
0.2
0.4
0.6
0.8
1.0
Overall Accuracy and Percentage Valid Output
(b)
Average Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Uncertain Effect
No Difference
Insufficient Data
higher
lower
Treatment Outcome Effect


(a)
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
very low
low
moderate
high
Evidence Certainty Level
(b)
1.0
0.8
0.6
0.4
0.2
0.0
0%
1-49%
50%
51-99%
100%
Percentage of Individual Sources Matching the Correct Answer (Binned)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Non-Reasoning
Reasoning
Average


(a)
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
very low
low
moderate
high
Evidence Certainty Level
(b)
0%
1-49%
50%
51-99%
100%
Percentage of Individual Sources Matching the Correct Answer (Binned)
OpenBioLLM 8B*
Llama 4 Scout
Llama 3.0 8B
Qwen2.5-7B-Instruct
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Llama 3.1 8B
Llama 3.0 70B
Llama 4 Maverick
GPT-o1
Llama 3.1 405B
Llama 3.3 70B-Instruct
Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct
Qwen3-235B-A22B-FP8
Llama 3.1 70B
GPT-oss-120B
GPT-4.1 mini
QwQ-32B
OpenThinker2-32B
Llama 3.3 70B (R1-Distill)
DeepSeek R1
GPT-4.1
DeepSeek V3
Non-Reasoning
Reasoning
Average


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
7B
32B
70B
0
100
200
300
400
500
600
700
Model Size (B)
OpenBioLLM 8B*
HuatuoGPT-o1-7B*
OpenBioLLM 70B*
HuatuoGPT-o1-70B*
Qwen2.5
DeepSeek R1
Llama 3.3 70B (R1-Distill)
OpenThinker2-32B
QwQ-32B
GPT-oss
Llama 3.1
Qwen3-235B-A22B-FP8
Llama 3.3 70B-Instruct
Llama 4
Llama 3.0
DeepSeek V3
Non-Reasoning
Reasoning
Medical*

CAN LARGE LANGUAGE MODELS MATCH THE CONCLUSIONS OF SYSTEMATIC REVIEWS?

Christopher Polzak*[1] **Alejandro Lozano*[1]** **Min Woo Sun*[1]**

James Burgess[1] **Yuhui Zhang[1]** **Kevin Wu[1]** **Chia-Chun Chiang[2]**

Jeffrey J. Nirschl[1] **Serena Yeung-Levy[1]**

[1]Stanford University **[2]Mayo Clinic**

https://zy-f.github.io/website-med-evidence/