

Assessing Generative AI for Subject Term Selection in Evidence Synthesis: A Longitudinal Evaluation

Mê-Linh Lê, Christine Neilson, and Caroline Monnin

University of Manitoba Libraries | Winnipeg, Manitoba, Canada

International Congress on Medical Librarianship, Glasgow, Scotland, July 2026

Traditional Territory Acknowledgement

The University of Manitoba campuses and research spaces are located on original lands of Anishinaabeg, Ininiwak, Anisininewuk, Dakota Oyate, Dene and Inuit, and on the National Homeland of the Red River Métis.

UM recognizes that the Treaties signed on these lands are a lifelong, enduring relationship, and we are dedicated to upholding their spirit and intent. We acknowledge the harms and mistakes of the past and the present. With this understanding, we commit to supporting Indigenous excellence through active Reconciliation, meaningful change, and the creation of an environment where everyone can thrive. Our collaboration with Indigenous communities is grounded in respect and reciprocity and this guides how we move forward as an institution.



BACKGROUND

- Subject headings **essential** to searching¹
 - e.g., MeSH (Medline), Emtree (Embase)
- Use of GenAI tools for search strategies has **exploded**²⁻⁵
 - Limited or incomplete access to thesauri
- ChatGPT searches may improve **precision**, but have lower **recall**³
- Current recommendation to use of GenAI in searching only with human oversight²⁻⁵



RESEARCH QUESTIONS

- 1. Can genAI models suggest appropriate subject terms for use in comprehensive search strategies for evidence synthesis research that are consistent with the gold standard search?**
- 2. Does the appropriateness of suggested subject terms by genAI models change over time?**

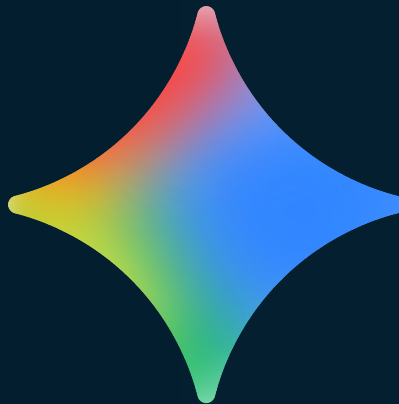


genAI TOOLS



FREE ChatGPT 5.5

💰 ChatGPT 5.2-5.4 Thinking



FREE Gemini 2.5

💰 Gemini 3.5



FREE Sonnet 4.6

💰 Opus 4.6-8



METHODS: The Plan



Gold Standard Search Identification



Prompt Creation



Data Collection



Method Modifications

METHODS: Gold Standard Search ID



Varied Topic Area

Topics from a range of health sciences areas (e.g., pharmacy, nursing)



Librarian Co-author

Librarian must be a co-author



Use of PRESS

Search strategy must have undergone PRESS



Search Strategy Available

At least one search strategy must be available



Not Open-Access

The article and search cannot be open-access.

METHODS: GSS Clean-Up and Explosion

Gold Standard Search Strategies

MeSH subject Headings - combos isolated and exploded terms itemized

General Practitioners/

Physicians, Family/

Physicians, Primary Care/

Family Practice/

General Practice/

Primary Health Care/

Physicians/

exp Workforce/

Health Workforce/

Research/

exp Biomedical Research/

Dental Research

Embryo Research

Fetal Research

Genetic Research

Gene Ontology

HapMap Project

Human Genome Project

Health Services Research

Comparative Effectiveness Research

Human Experimentation

Autoexperimentation

Nontherapeutic Human Experimentation

Therapeutic Human Experimentation

Nursing Research

Clinical Nursing Research

Nursing Administration Research

Nursing Education Research

Nursing Evaluation Research

Nursing Methodology Research

Outcome Assessment, Health Care

Failure to Rescue, Health Care

Pharmaceutical Research

Pharmacy Research

Rehabilitation Research

Stem Cell Research

Translational Research, Biomedical

Research Personnel/

"Attitude of Health Personnel"/

Capacity Building/

Cooperative Behavior/

Leadership/

Motivation/

Organizational Culture/

exp Organizational Innovation/

Change Management

Entrepreneurship

Personnel Selection/

Physicians/px

General Practitioners/px

Physicians, Family/px

Physicians, Primary Care/px

Physician's Role/

Professional Role/

Stakeholder Participation/

Work Engagement/

METHODS: Prompt Creation

- **Persona**
- **Study Objectives**
- **Study Components**
- **Controlled Vocabulary**
- **Output Format**

METHODS: Monthly Prompting

Time of search	30-Jan-26				
Model:	Claude Free Opus 4.5				
Link	Raw Output				
MeSH Gold Standard Subject Terms		Generated Search String	Match to gold standard?	Is this a real Mesh Term?	Notes:
	Diabetic Neuropathies +	Greenhouse Gases			
	Fetal Macrosomia	Particulate Matter/	yes	yes	
	Glycemic control/	*Environmental Exposure/	yes	yes	Gold standard did not focus
	blood glucose/	exp Inhalation Exposure/	no	yes	Exploded but no child terms
	hyperglycemia/	Blood Glucose/	yes	yes	
	hypoglycemia/	*Hyperglycemia/	yes	yes	Gold standard did not focus
		*Hypoglycemia/	yes	yes	Gold standard did not focus
		Glycated Hemoglobin A/	no	no	
		Blood Glucose Self-Monitoring/	no	yes	
		Glucose Intolerance/	no	yes	
		*Glucose Metabolism Disorders/	no	yes	
		exp Diabetes Complications/	yes	yes	This is already included in the exploded Diabetes Mellitus
		Diabetesity	yes	yes	



9 - Wildfire Smoke - original

January 13

METHODS: Data Collection & Analysis

Collect:

- Date
- Model
- Complete output
- Controlled vocabulary terms

Add:

- Child terms under Exploded terms
 - (e.g., exp Rehabilitation/)
- If generated term used in the Gold Standard Search
- If AI suggested term is a 'real' term

Rehabilitation

[Activities of Daily Living](#)
[Animal Assisted Therapy](#)
[Equine-Assisted Therapy](#)
[Art Therapy](#)
[Bibliotherapy](#)
[Cardiac Rehabilitation](#)
[Correction of Hearing Impairment](#)
[Communication Methods, Total](#)
[Lipreading](#)
[Manual Communication +](#)
[Dance Therapy](#)
[Early Ambulation](#)
[Exercise Therapy](#)
[Blood Flow Restriction Therapy](#)
[Endurance Training](#)
[Motion Therapy, Continuous Passive](#)
[Muscle Stretching Exercises](#)
[Plyometric Exercise](#)
[Resistance Training](#)
[Music Therapy](#)
[Neurological Rehabilitation](#)
[Cognitive Training](#)
[Stroke Rehabilitation](#)
[Occupational Therapy](#)
[Recreation Therapy](#)
[Rehabilitation of Speech and Language Disorders](#)
[Language Therapy](#)
[Myofunctional Therapy](#)
[Speech Therapy](#)
[Speech, Alaryngeal +](#)
[Voice Training](#)
[Rehabilitation, Vocational](#)
[Telerehabilitation](#)

METHODS: The Reality

What we learned:

- Original methods = **SIGNIFICANT** time investment

What we decided:

- Move to bimonthly
- Present findings on one search only

Search topic:

- Impact of wildfire smoke exposure on acute glycaemic control in patients living with diabetes



Image generated by Google Gemini on May 24, 2026

PRELIMINARY RESULTS



General Observations



Metrics

PRELIMINARY RESULTS: Observations

Excessive Explosion

- Exploding when no narrower terms
 - E.g., *exp blood glucose monitoring (MeSH)*
- Exploding too far up the tree

PRELIMINARY RESULTS: Observations

Emtree: exp air pollution/

acid deposition/
air pollutant/

atmospheric particulate matter/

black carbon/

diesel particulate matter/

respirable particulate matter/

particulate matter 1.0/

particulate matter 2.5/

ultrafine particulate matter/

exhaust gas/

flue gas/

gas waste/

greenhouse gas/
peroxyacetylnitrate/
photochemical smog/
smog/

smoke/

soot/

air pollution control/

air particle control/

odor control/

air pollution indicator/

air sampling/

atmospheric deposition/

atmospheric dispersion/
chimney/
dry deposition/
gas flaring/
greenhouse gas emission/
carbon emission/
carbon dioxide emission/
methane emission/
carbon footprint/
carbon neutrality/
nitrous oxide emission/
indoor air pollution/

nitrogen oxide emission/
ozone depletion/
radioactive air pollution/
wet deposition/

PRELIMINARY RESULTS: Observations

Redundancy

- A term appears on its own as well as in an explosion of a broader term
- ORing different versions of the same term as part of a search block
 - E.g., *(MH "Blood Glucose+") OR (MH "Blood Glucose") OR (MM "Blood Glucose") (CINAHL via Claude)*

PRELIMINARY RESULTS: Observations

Excessive Focusing

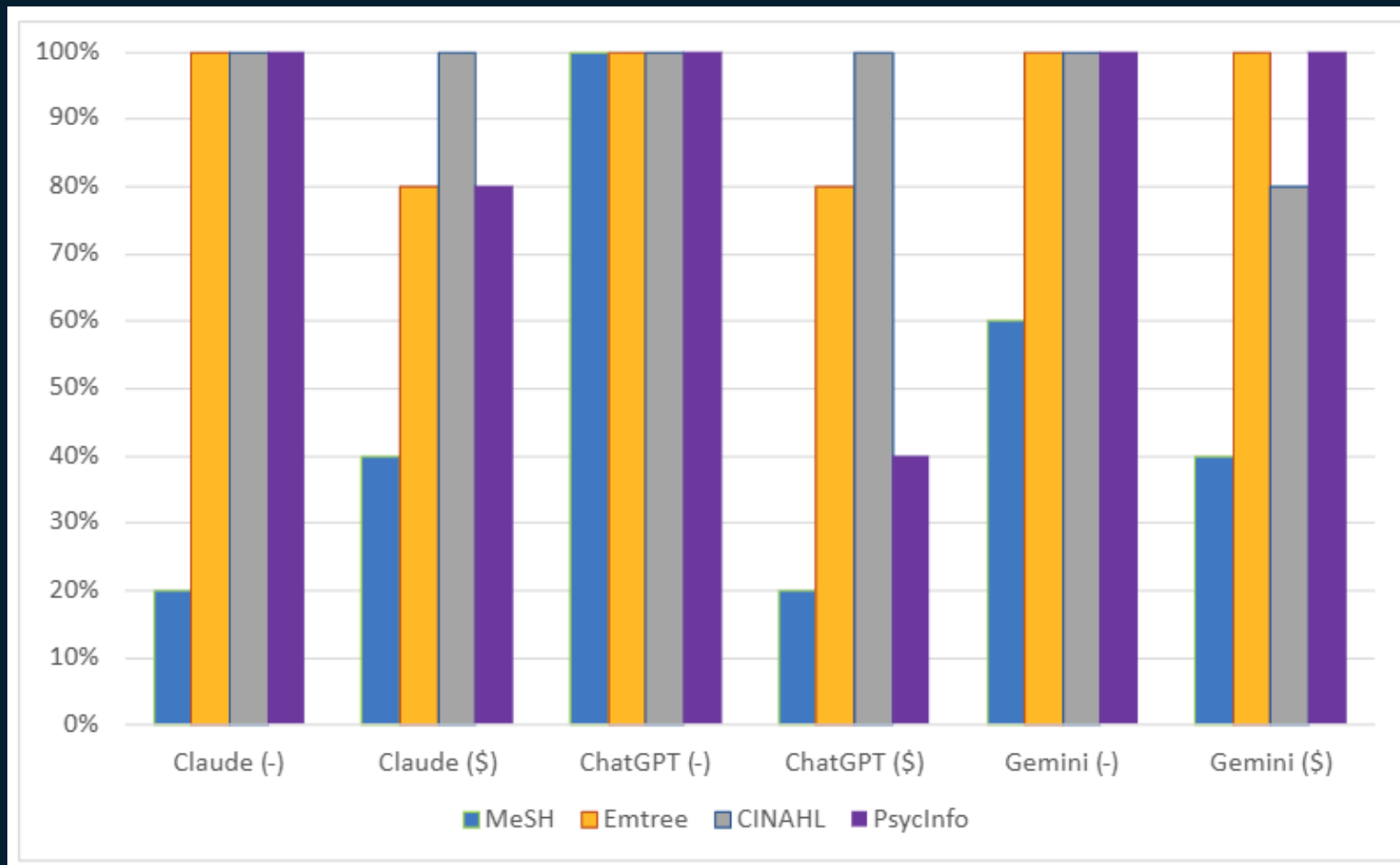
- Focusing many of the suggested headings
- A heading will appear twice in the list of suggestions, one of which will be focused (*)
 - *exp diabetes mellitus/ or *diabetes mellitus/*
- Incorrect placement of the asterisk
 - *(air pollution/*)*

PRELIMINARY RESULTS: Observations

Other observations

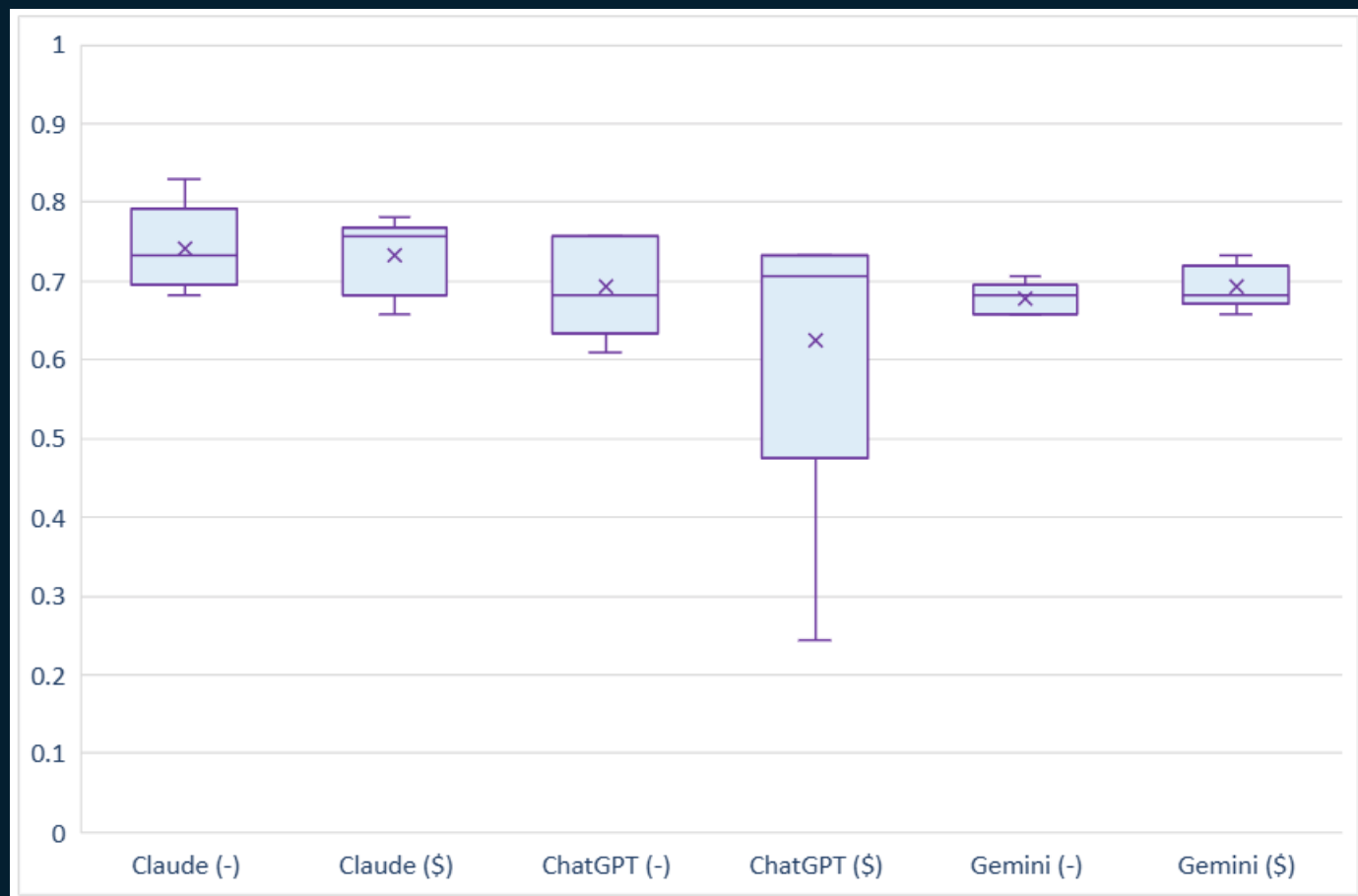
- In general, certain gold standard terms were consistently absent
wilderness/
environmental health/
- We did not evaluate search structure but ...it is weird
 - Almost always presented search blocks in PICO structure when presented with a PCC-structured question
 - (very) Rarely present more than one term per line

PRELIMINARY RESULTS: Metrics – Hallucination rates



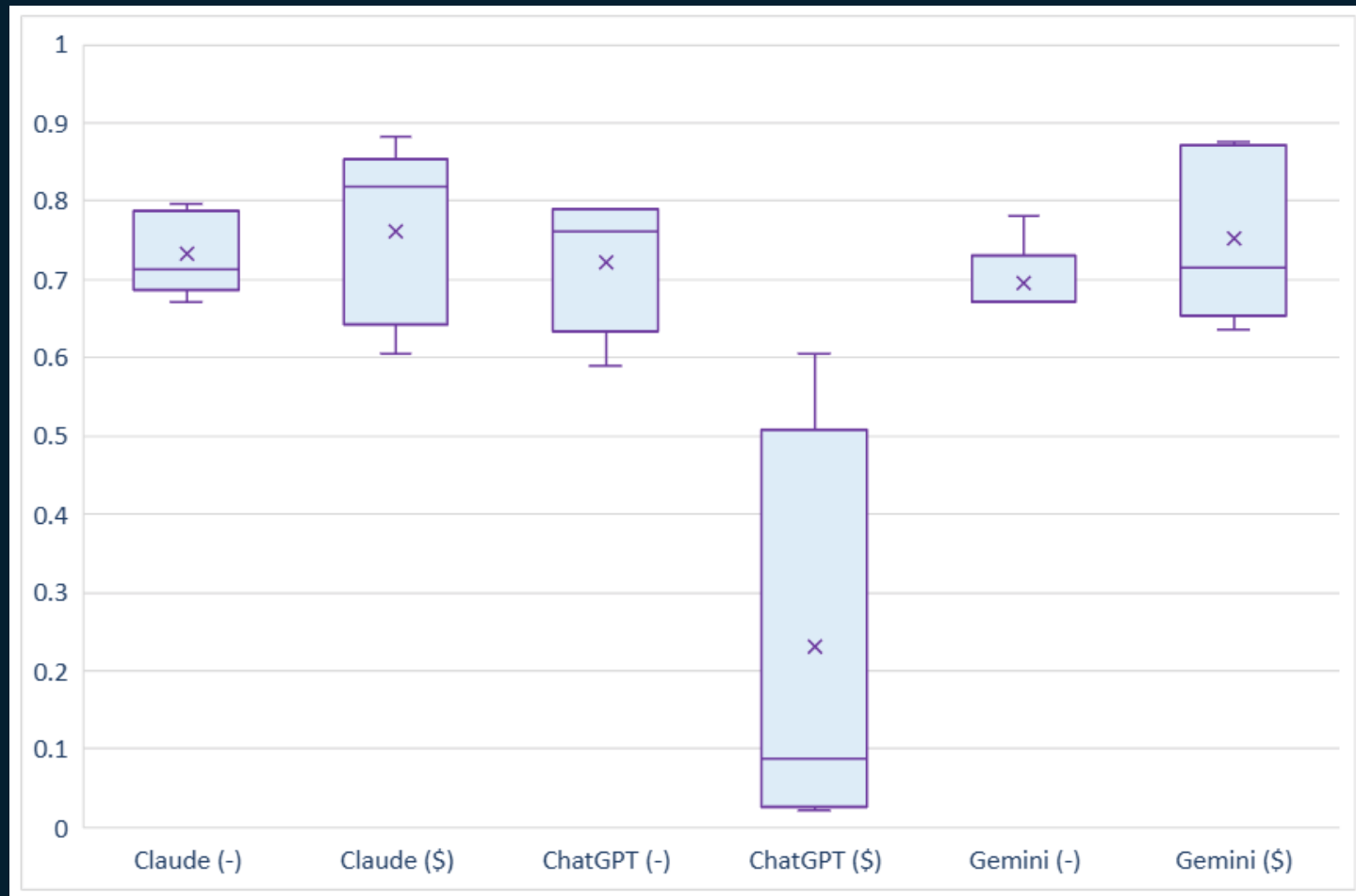
% Attempts with at least one hallucinated subject term (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Recall (MeSH)



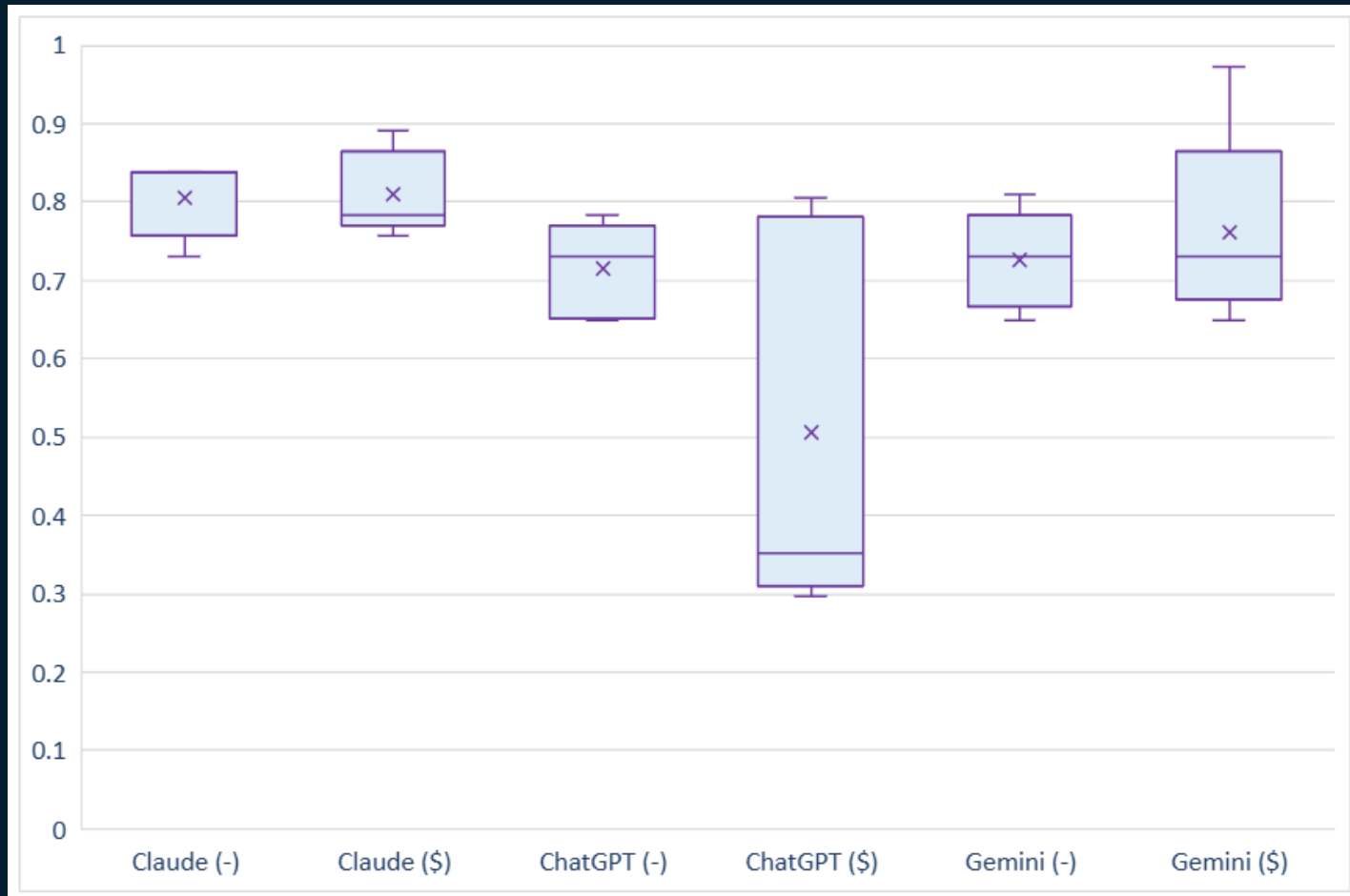
Recall by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Recall (EMTREE)



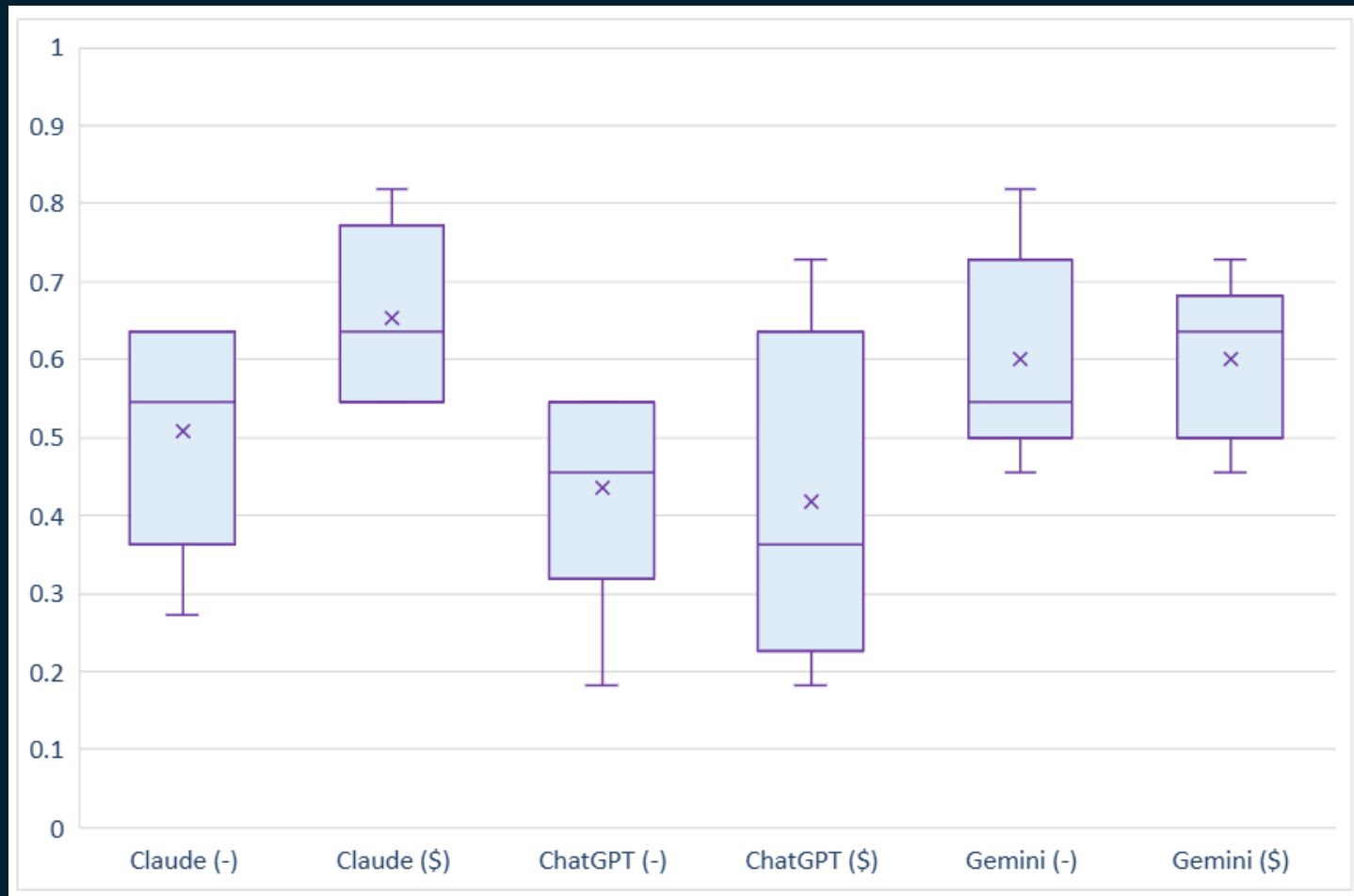
Recall by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Recall (CINAHL)



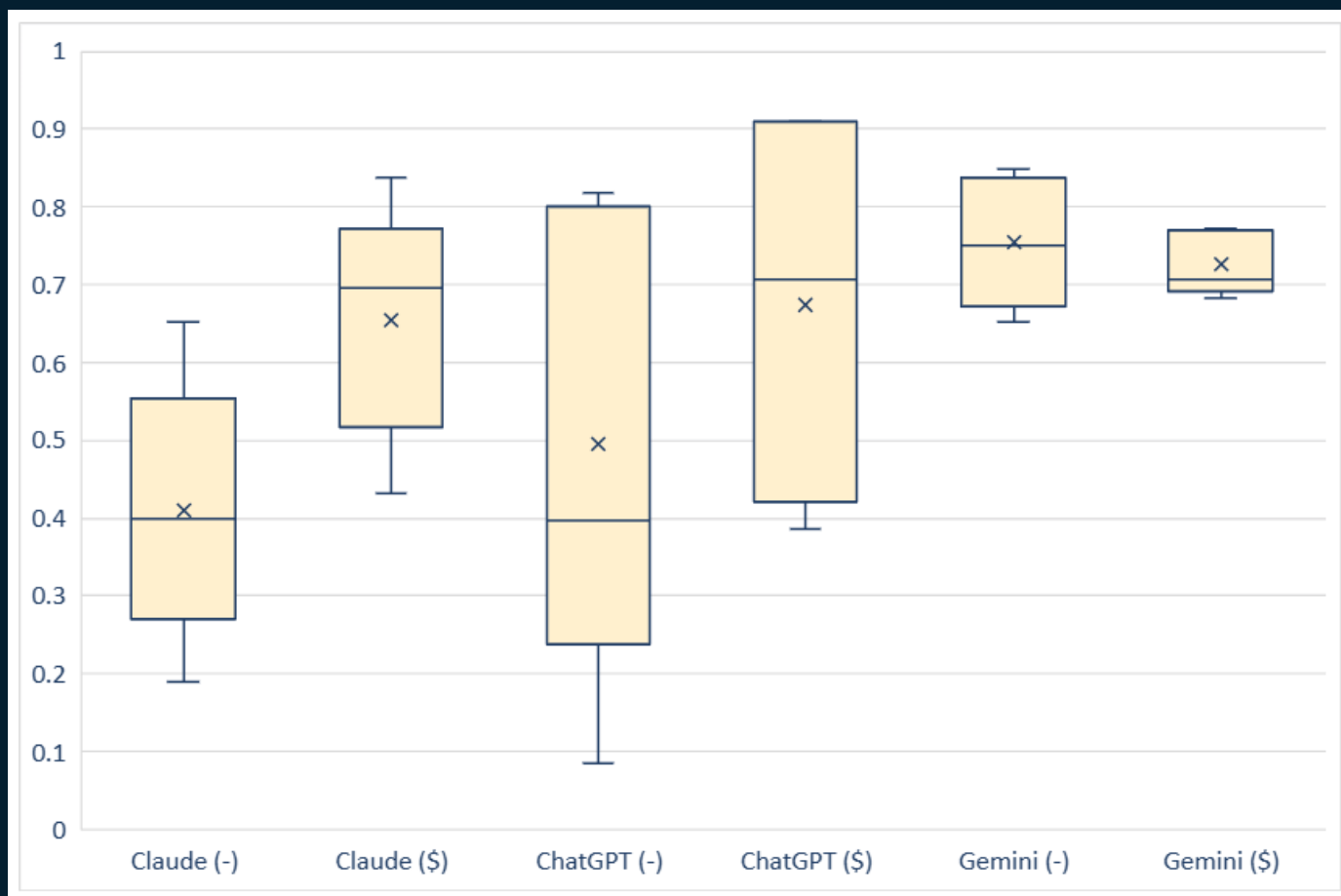
Recall by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Recall (PsycINFO)



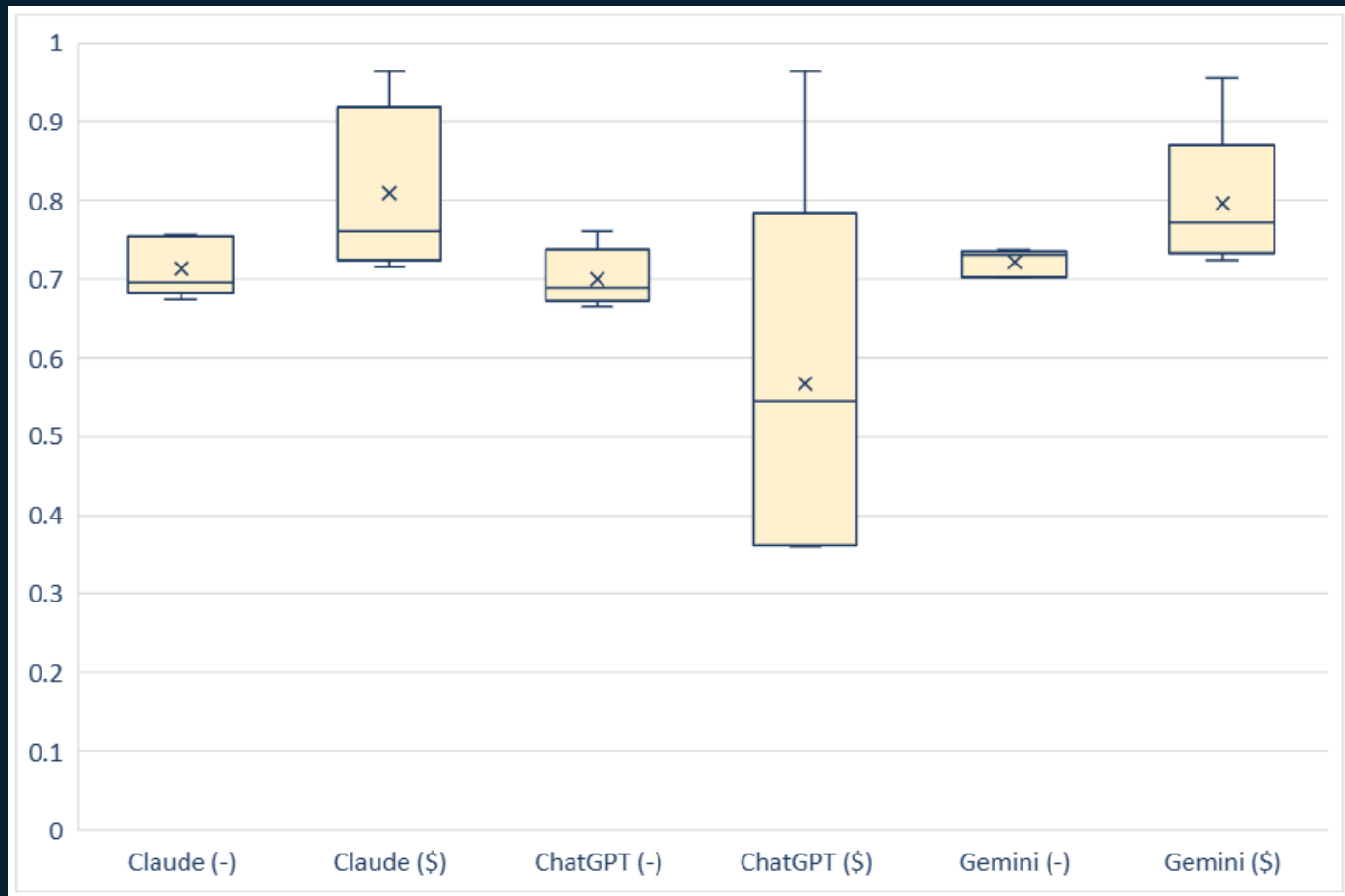
Recall by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Precision (MeSH)



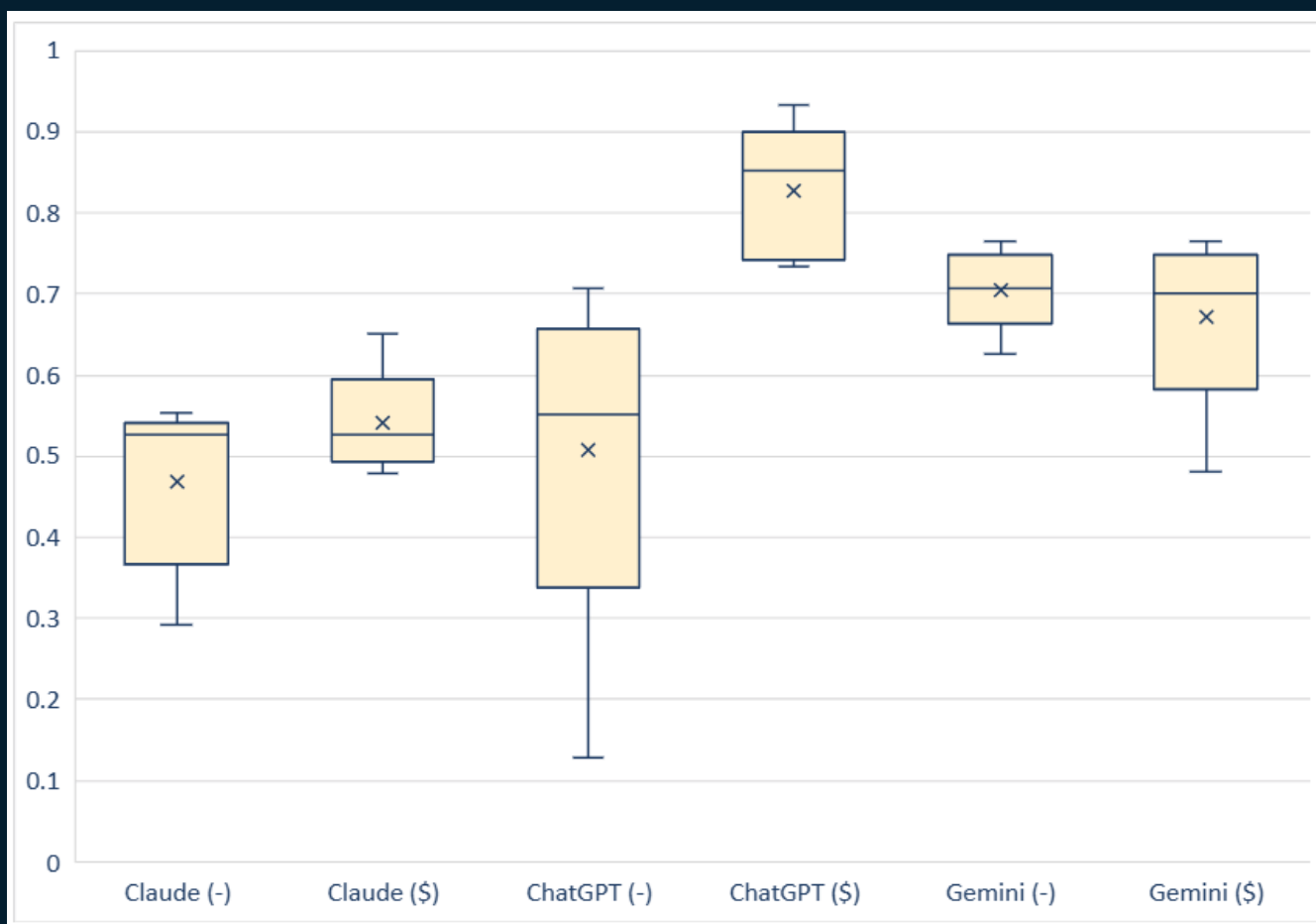
Precision by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Precision (EMTREE)



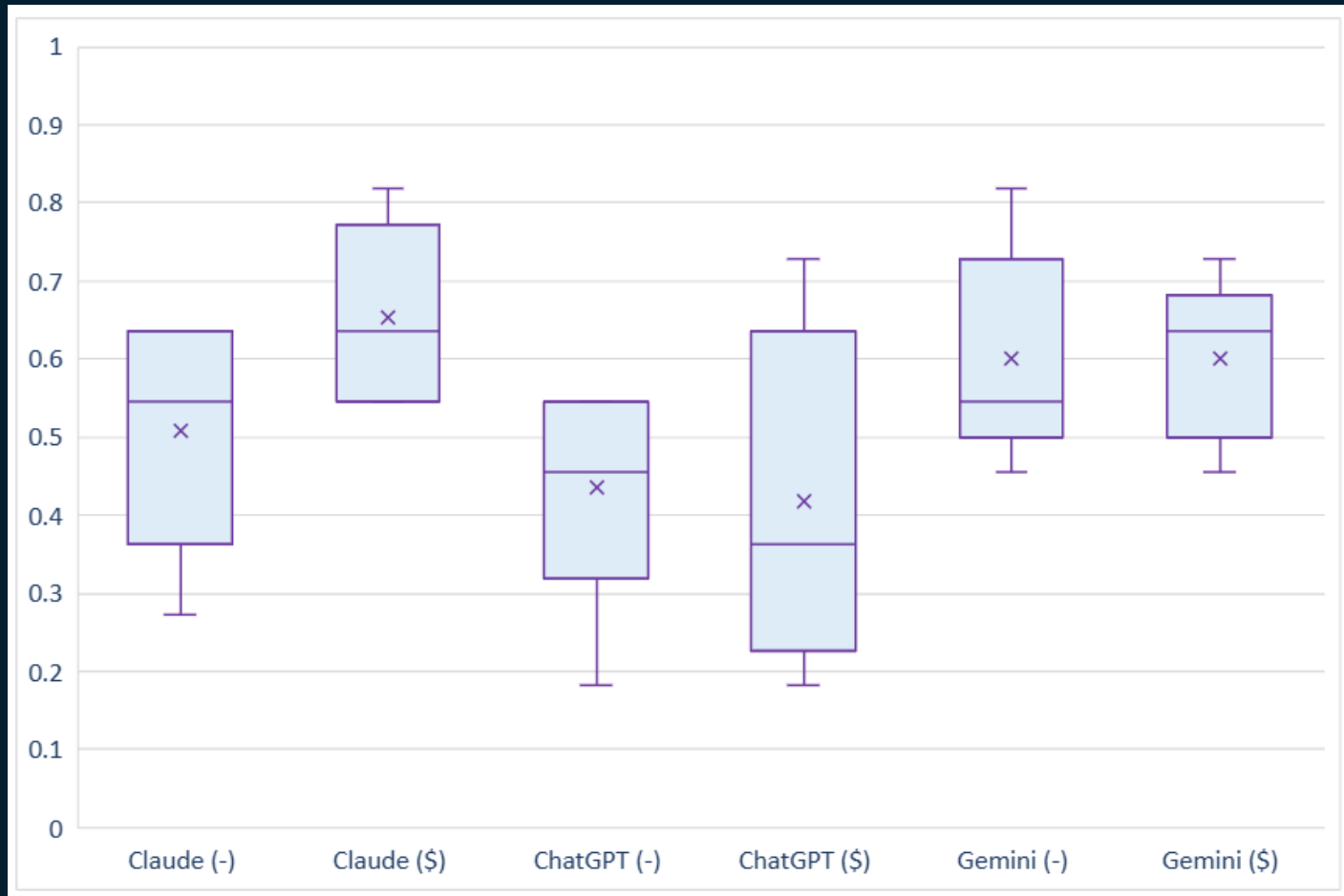
Precision by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Precision (CINAHL)



Precision by Model (June 30, 2026) / $N=5$

PRELIMINARY RESULTS: Metrics – Precision (PsycINFO)



Precision by Model (June 30, 2026) / $N=5$

SUMMARY

- **Hallucinations present for every controlled vocabulary**
- **Precision is a major concern**
- **Recall is a concern**

WHAT'S NEXT...

- The work continues ... stay tuned!

REFERENCES

1. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies [last updated March 2025]. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). Cochrane Handbook for Systematic Reviews of Interventions version 6.5.1 Cochrane, 2025. Available from cochrane.org/handbook.
2. Alaniz L, Vu C, Pfaff MJ. The utility of artificial intelligence for systematic reviews and boolean query formulation and translation. *Plastic and Reconstructive Surgery–Global Open*. 2023 Oct 1;11(10):e5339.
3. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* 2023 Jul 19 (pp. 1426-1436).
4. Yu F, Kincaide H, Carlson RB. An Empirical Study Evaluating ChatGPT's Performance in Generating Search Strategies for Systematic Reviews. *Proceedings of the Association for Information Science and Technology*. 2024;61(1):423-434. doi:[10.1002/pa2.1039](https://doi.org/10.1002/pa2.1039)
5. Clark J, Barton B, Albarqouni L, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*. 2025;16(4):601-619. doi:[10.1017/rsm.2025.16](https://doi.org/10.1017/rsm.2025.16)

Credits / Acknowledgements

- Mê-Linh Lê was funded by the University of Manitoba Librarians, Archivists and Librarian Interns Education/Travel Endowment Fund
- Presentation template by [SlidesCarnival](#)
- Photographs by [Startupstockphotos](#)

Contributor statements

- ML: Conceptualization, Investigation, Methodology, Writing – Original Draft, Writing – Review & Editing
- CM: Investigation, Methodology, Writing – Original Draft, Writing – Review & Editing
- CN: Investigation, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing

THANKS!

Any questions?

Caroline Monnin | caroline.monnin@umanitoba.ca

Mê-Linh Lê | me-linh.le@umanitoba.ca

Christine Neilson | christine.neilson@umanitoba.ca