McMaster University

Family Medicine

Making informed choices about GenAI in health research

Watch a recording of this session

Stephanie Verkoeyen

March 14th, 2026

Making informed choices about GenAI in health research

Stephanie Verkoeyen

Special Advisor, Generative AI

Where is AI showing up in research?

- Idea generation
- Literature screening & synthesis
- Qualitative analysis
- Data analysis & coding
- Writing & summarization

Using GenAI well depends on understanding what it is

LLMs predict the next token that is most statistically likely, given the previous tokens

What if you want to do complex math or data sorting?

"What are the main themes of these 500 survey responses?"

vs.

"Write a Python script using K-Means clustering to segment these responses into 4 distinct, non-overlapping audience personas."

What kinds of tasks do LLMs do well?

- Predicting and generating language based on patterns
- Synthesizing and summarizing text
- Identifying recurring patterns in language
- Reframing or re-expressing ideas
- Generating plausible continuations or alternatives

Research depends on judgement

- What to include
- What to exclude
- What matters
- What something means

When pattern generation meets judgement

Where does judgement now sit?

What is being prioritized?

What becomes harder to see or justify?

LLM use in research is rarely a one-shot substitution

It is often an interactive loop where:

- prompts encode assumptions
- outputs suggest directions
- researchers refine prompts
- those refinements change what is surfaced
- and the researcher’s own sense of what is worth asking may shift

Judgement becomes entangled with prompt iteration

Case 1: Screening for systematic reviews

Purpose: selection of relevant articles, define the evidence base

Traditionally:

- multiple reviewers
- iterative calibration of inclusion/ exclusion criteria
- judgement through comparison & discussion

With LLMs:

- criteria are translated into prompts
- inclusion decisions are mediated through model outputs

ChatGPT achieved up to ~82% accuracy in screening articles

- performance shifts with prompt design and corpus characteristics
- high recall (finds most relevant articles)
- low precision (includes lots of irrelevant ones)

Syriani, E., David, I., & Kumar, G. (2024). Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages*, *80*, 101287.

AI reshapes what enters the knowledge base

Inclusion defines what counts as evidence, what gets synthesized, what shapes conclusions

Epistemic risk is also:

- undocumented calibration
- hidden prompt tuning
- gradual normalization of certain inclusion logics

Case 2: Thematic analysis

Purpose: Identifying patterns and themes in qualitative data, interpreting meaning

ChatGPT achieved 80% agreement with human coding

Proposed uses:

- Generate themes and subthemes
- Extend or complement analysis

Suggests AI performs well and can support thematic analysis

"*the quality of the outcomes produced by LLMs does not currently rival the work of qualitative researchers or their role in generating knowledge*" (p. 203)

LLMs miss nuance, overgeneralize, produce coherent but weakly grounded interpretations

Suggests AI is limited in supporting meaningful interpretation

Castellanos, A., Jiang, H., Gomes, P., Vander Meer, D., & Castillo, A. (2025). Large language models for thematic summarization in qualitative health care research: comparative analysis of model and human performance. *JMIR AI*, *4*, e64447.

Vikan, M., Aryan, R., Kannelønning, M. S., Riegler, M. A., & Danielsen, S. O. (2026). Reflecting on LLM support in reflexive thematic analysis: An exploratory study. *Qualitative Health Research*, *36*(2-3), 191-205.

Judgement is handled differently across studies

Early/stabilized (deductive)

- Codes are defined upfront
- Categories are clarified early
- Interpretation becomes:
  - structured, shared, more consistent

Ongoing/evolving (inductive)

- Interpretation develops through engagement with data
- Themes are refined iteratively
- Meaning remains:
  - open, situated, reflexive

AI reshapes how knowledge is constructed

Not only hallucination, bias, overgeneralization, hidden assumptions

Epistemic risk as:

- **Path dependency**: early prompts and outputs can shape later questions, categories, and interpretations
- **Narrowing**: some lines of thought feel more salient, coherent, or actionable than others
- **False reflexivity**: because the process is interactive, it can feel thoughtful and dialogic
- **Undocumented co-construction**: if prompt evolution is not tracked, the research path becomes harder to explain, justify, or reproduce

So, what's being gained?

AI doesn't just make research faster; it changes what becomes visible

Scale: ability to process large datasets quickly

Pattern detection: identifying recurring structures across data

Generativity: surfacing alternative framings or connections

Exploratory value: surfacing possibilities

Contrastive value: challenging researcher assumptions

Augmentative value: extending analytic perspectives

These gains reshape judgement

Gains and risks are produced by the same shift

Scale →less deep engagement with individual data points

Pattern detection →may miss contextual nuance

Generativity →may introduce ungrounded interpretations

Making informed choices about AI in research

When considering AI use, ask:

- How does the task align with how LLMs work?
- What kind of judgement does it require? (e.g., selection, interpretation, framing)
- What does AI make easier or more visible?
- What might it obscure or reshape?
- What safeguards are needed?

Where in your research would AI feel most useful? Why?

What would you be gaining in that use case? What risk/cost might you be introducing?