ebhc CONFERENCE 2026

THE ECOSYSTEM OF EVIDENCE

Navigating the future
of the ecosystem of evidence

Artificial intelligence and digital transformation
in research and healthcare

11th EBHC International Conference
Taormina, 21st – 24th October 2026

#EBHC2026

30th 1996-2026 GIMBE EVIDENCE FOR HEALTH

Same evidence, different risks

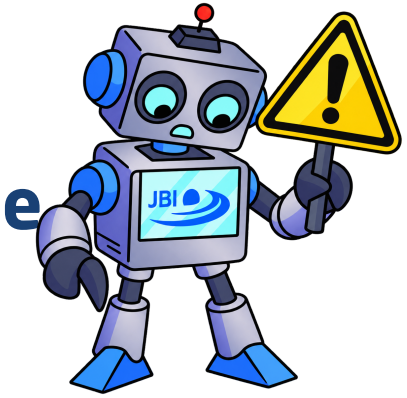
Prompt engineering and format-specific failure modes in AI-assisted EBHC communication

Co-investigators: Firmaye Bogale, Mamuye Hadis, Sonia Hines, Lyn Hendricks, Bianca Pilla

Heather McCulloch
JBI, Adelaide University

Background

Documented risks of using AI to communicate evidence



- Factual errors
- Certainty misrepresented
- Qualifiers and context removed
- Unsupported claims
- Clarity
- Framing that excludes or marginalises
- Language that reinforces stereotypes

Where evidence is most qualified, the risk of inaccurate or marginalising output is highest

Aims of proof-of-concept stud

Objective

Characterise the failure modes that occur when AI is used to communicate EBHC information across different communication formats

1 x GenAI model

1 x systematic review

3 x communication formats



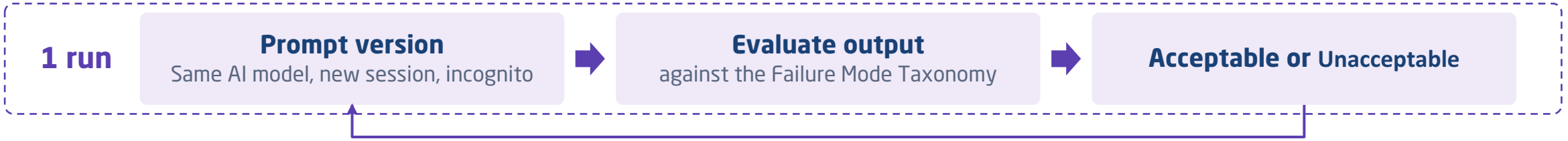
Effectiveness of conditional cash transfers for uptake and retention of HIV prevention of mother-to-child transmission services in low- and middle-income countries: a systematic review and meta-analysis

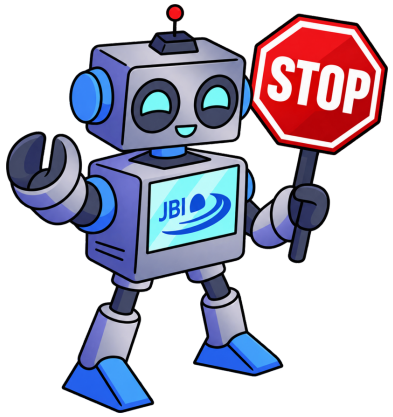
Mamuye Hadis, Tesfaye Dagne, Sabit Ababor, Dagmawit Solomon, Samson Mideksa, Zelalem Kebede, Yosef Gebreyohannes, Firmaye Bogale

JBI Evidence Synthesis. September 2025: 23(9):1770-1787

Methods

Prompt engineering





Rule 1: freeze the prompt

If: 4 Acceptable runs in a row, with no earlier fixes lost

Then: Prompt becomes the deliverable

Rule 2: stop iterating

If: Same failure across 3 prompt versions, despite targeted instructions

Then: Failure is prompt-resistant; human editing required

Failure mode taxonomy

Evidence fidelity

F1 Accuracy or verification error
F2 Certainty misrepresentation
F3 Unsupported claim
F4 Omission or selective emphasis of a limitation or caveat
F5 Context or population over-generalisation

Clarity and accessibility

F6 Clarity, concise prose and scan-ability failure
F7 Format non-compliance

Audience appropriateness

F8 Tone and register misalignment
F9 Stigmatising or exclusionary framing

***'Cash payments help more HIV-positive mothers deliver safely and get babies tested'**

F2 Certainty overstated (SR evidence was moderate and low)

F7 12 words (the limit was 10)

F9 Labels mothers by their HIV status

* Media release headline

Prompt engineering results

Policy brief

5 prompt versions, 13 outputs

	Initial	Re-run 1	Re-run 2	Re-run 3
V1	×			
V2	×			
V3	×	×	×	
V4	×	×	✓	×
V5	✓	×	✓	×

Stopped under Rule 2

Media release

6 prompt versions, 11 outputs

	Initial	Re-run 1	Re-run 2	Re-run 3
V1	×			
V2	×			
V3	×			
V4	✓	×	×	
V5	✓	×	×	×
V6	×			

Stopped under Rule 2

LinkedIn post

3 prompt versions, 11 outputs

	Initial	Re-run 1	Re-run 2	Re-run 3
V1	×	×	×	
V2	×	×	×	×
V3	✓	✓	✓	✓

Version 3 frozen under Rule 1

× Unacceptable ✓ Acceptable

Failure modes by format

Policy brief

Acronyms not spelled out

- Eg: GRADE, CI, ART

Domain: **Clarity and accessibility**

Media release

Stigmatising language, deficit framing

- Eg: 'women in LMICs often fail to access or stay in these programs' (vs 'women faced barriers to accessing')

Domain: **Audience appropriateness**

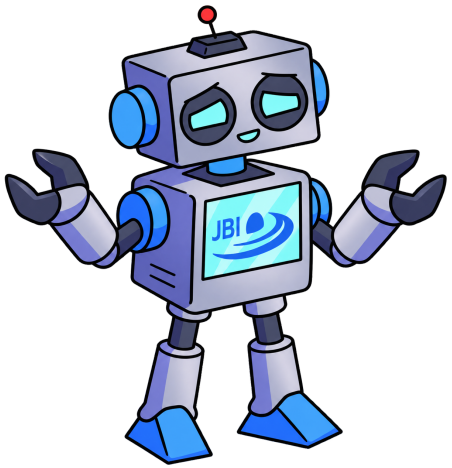
LinkedIn post

Overstatement

- Eg: 'Hundreds of thousands of children are born with HIV' (vs 'about 130,000 new infections in children under five')

Domain: **Evidence fidelity**

Limitations

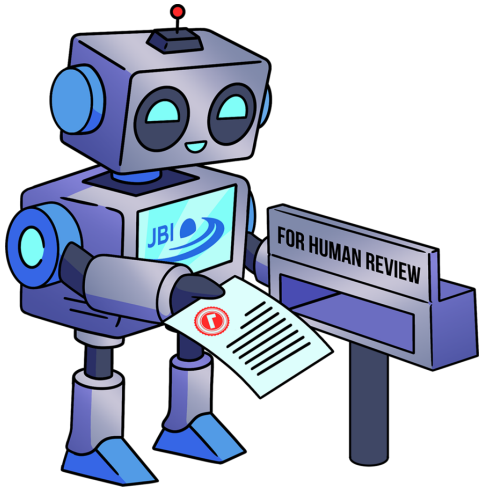


- A single systematic review was used
- 3 x communication formats
- Findings may reflect one individual's prompt engineering skill
- Findings are specific to a single AI model at a single point in time

Conclusions

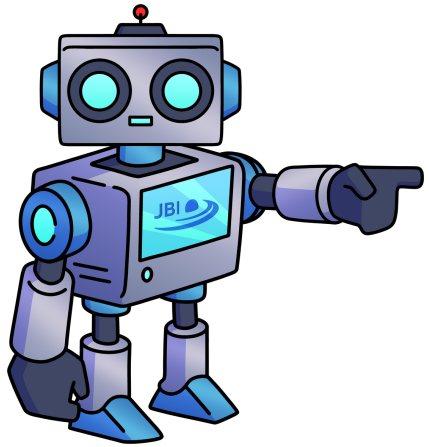
- Identical prompts, run in the same conditions, gave different results
- One acceptable run isn't evidence that a prompt works
- Failure modes across the formats extended beyond accuracy
- Acceptable output in this study still needed a human to edit it

No prompt removed the need for expert human review



AI was used to draft the diagrams in this presentation. Each was human-reviewed.
Retro robot was designed by a human illustrator.

Questions?



@hmcatjbi