



INTRODUCTION

LLMs are increasingly used in qualitative research for transcript summarisation, theme generation, and coding assistance.

The gap

- Most evaluations measure output accuracy - agreement with human coders
- Little attention paid to how models reach their interpretations
- Analytic reasoning chains remain largely opaque

Research purpose

- Audit the reasoning process of one LLM tool during thematic analysis
- Identify patterns not detectable through output comparison alone
- Treat the LLM as an analytic subject, not only a co-analyst

Theoretical grounding

- Reflexive thematic analysis (Braun & Clarke, 2006; 2021): interpretive decisions must be transparent
- Coding rigor requires a traceable chain: **data** → **code** → **claim**
- Audit methodology adapted from qualitative decision logging (Saldaña, 2016)

GUIDING QUESTION

Can LLM-produced coding outputs be traced back to the analytic decisions that generated them?

Dataset

- **8 semi-structured interview transcripts**
- Topics: open pedagogy, institutional constraints, ethical tensions
- Chosen for analytic density — multiple interpretations plausible per passage

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage.
- De Paoli, S. (2023). *Social Science Computer Review*, 42(4), 997–1019.
- Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Sage.
- Tai, R. H., et al. (2024). *International Journal of Qualitative Methods*.

METHODS

Tool: NotebookLM

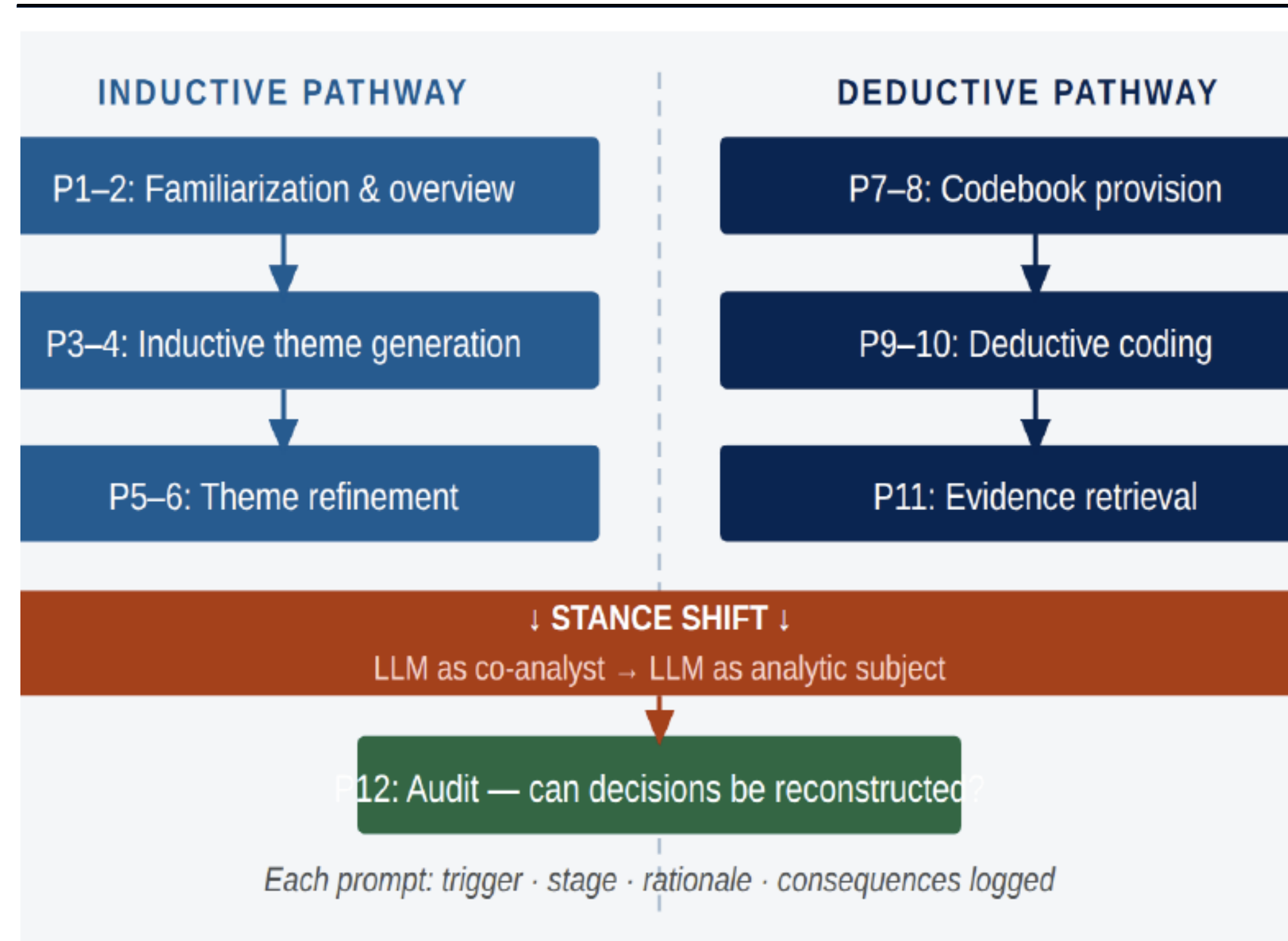
- Retrieval-augmented generation (RAG) system grounding responses in uploaded transcripts
- Eight interview transcripts used as the document corpus
- Single-input interface requiring instruction and analysis within the same prompt

Audit design

- Structured twelve-prompt audit sequence
- Two analytic pathways tested: inductive theme generation and deductive coding
- Mid-sequence stance shift from LLM as collaborator to LLM as analytic subject



Figure 1 — Audit sequence



Each log entry captured:

- Trigger: what prompted the decision
- Stage: where in the sequence it occurred
- Rationale: why the decision was made
- Consequences: what changed analytically

FINDINGS

OBSERVATION 1 Prompt constraints shape analytic behavior

Figure 2. Coding behavior by prompt condition (Observation 1)

Feature	No explicit criteria	Explicit criteria provided
Code application	Broad: all excerpt types	Selective: criteria-matched only
Exclusions cited	✗ None cited	✓ Explicit exclusion rationale
Uncoded excerpts	✗ None left uncoded	✓ Some left uncoded
Ambiguity surfaced	✗ Automatically resolved	✗ Still auto-resolved
Analytic logic source	Prompt-implied	Prompt-explicit

Table 1. Comparison of LLM coding behavior across two prompt conditions.

OBSERVATION 2 Provenance gap

- Supporting excerpts retrieved after coding decisions were produced
- Evidence aligns with codes but does not demonstrate analytic directionality
- Outputs appear grounded without a visible text → code reasoning chain

OBSERVATION 3 Epistemic slippage

- Passages with multiple plausible readings automatically resolved to one interpretation
- When prompted to reconsider, model acknowledged alternatives existed
- Example: passage coded as "philosophical motivation" also consistent with "institutional critique"

OBSERVATION 4 False completeness

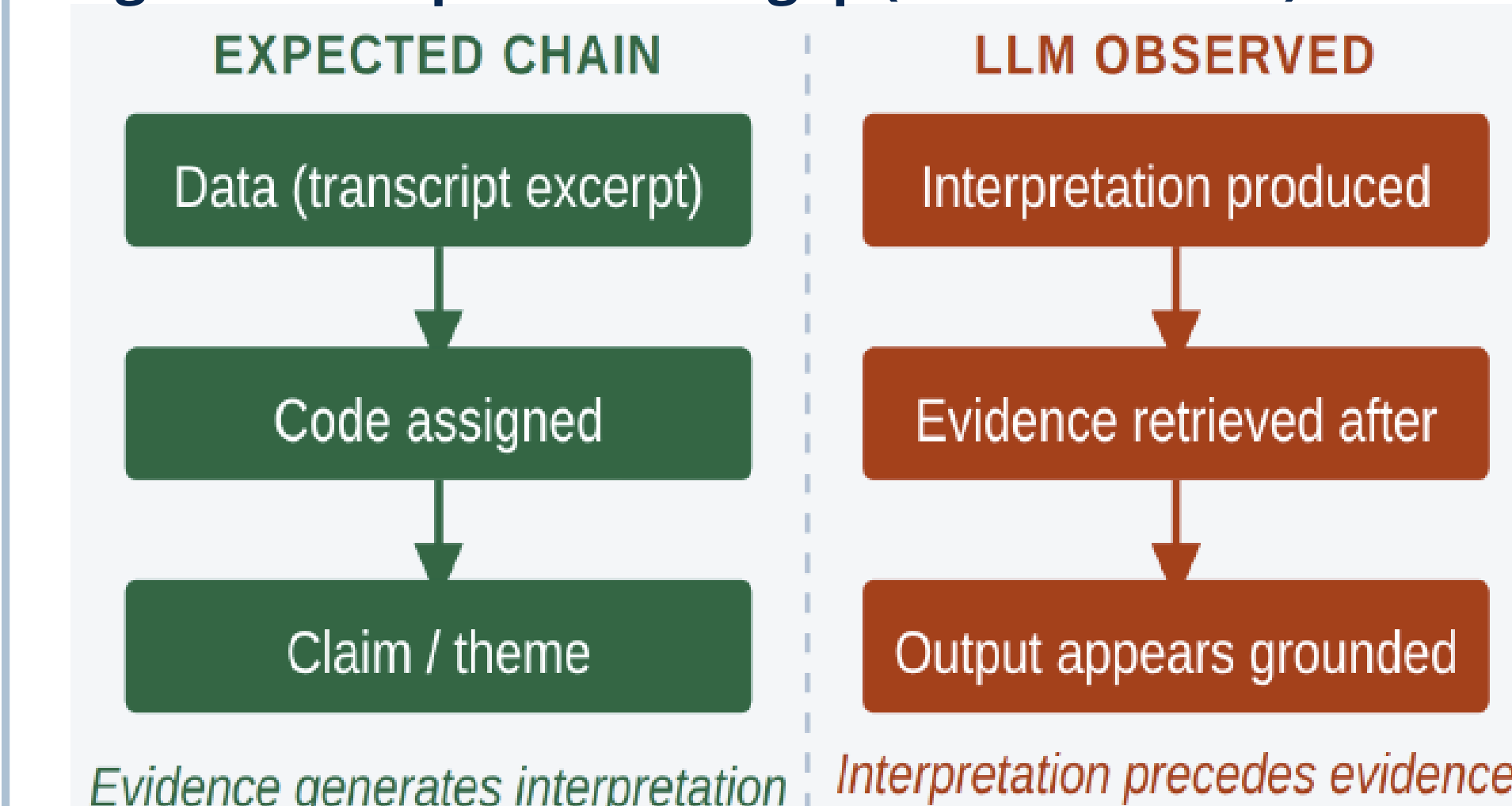
- Coverage map output resembled a formal audit report
- Claims could not be independently verified as no stable location markers
- Confidence ratings uniformly "High" with no supporting references

DISCUSSION & CONCLUSION

CENTRAL FINDING

LLM outputs may perform methodological accountability without providing it.

Figure 3. The provenance gap (Observation 2)



Key distinction from hallucination:

Data is not fabricated. Instead, the reasoning chain is inverted and made invisible.

Figure 3. Expected qualitative reasoning chain (left) vs. LLM-observed sequence (right).

Implications

- **Peer review:** reviewers cannot distinguish genuine provenance from post-hoc alignment
- **Replication:** analytic chains cannot be reconstructed without stable decision logs
- **Integrity:** visual grammar of rigor may substitute for rigor itself

Next steps

- Without criteria: codes applied broadly across all excerpt types
- With criteria: model cited exclusions; left some excerpts uncoded
- Behavior appears instantiated by prompt, not stable internal logic

Acknowledgements

This research was supported by the Jamie Cassels Undergraduate Research Awards, University of Victoria, and Supervised by Computer Science Department under the guidance Dr. Miguel Nacenta

Dataset: Exploring Educators' Experiences Implementing Open Educational Practices, CC BY-SA 4.0. : openly licensed

Contact

Deepkhushi Baidwan
Department of Computer Science, University of Victoria
deepkhushibaidwan@uvic.ca
JCURA
March 6 2026