Interrogating with AI

Exploring the Usage of and Auditing of Large Language Models in Relational Coordination Research

Kartik Trivedi

2025-11-07

University of New Hampshire / RCC Roundtable 2025

Agenda

- What are LLMs?
- What changes in qualitative inquiry with GenAI?
- Opportunities + risks in QDA
- Prompt engineering & **parameter sensitivity**
- The **MERIT** framework for responsible reporting
- Hands-on with LLM tools (3 tools)
- Reflection & Q&A

What is a Large Language Model (LLM)?

- AI trained on **massive** text data
- Learns patterns $\longrightarrow$ generates & interprets language
- Can summarize, code, classify, compare
- Examples: GPT-4, Claude, LLaMA

Why “Large” Matters

- Billions of parameters $\longrightarrow$ emergent reasoning
- Transformer architecture (“attention”)
- Handles complexity & longer qualitative texts
- Limitations: hallucinations, bias, data opacity

Types of LLMs

Category	Examples	Why it matters
Model family	GPT, Claude	Capabilities differ
Openness	LLaMA vs GPT-4	Reproducibility
Modality	Text vs multimedia	Conversation types
Interface	App vs API vs QDA tools	Workflow impact

What is Qualitative Data Analysis?

- Interpretation, meaning, lived experience
- Coding $\longrightarrow$ themes $\longrightarrow$ narrative
- Always reflexive & relational

Traditional QDA Workflow

1. Data familiarization
2. Open/initial coding
3. Theme development
4. Interpretation & checking
5. Reporting + reflexivity

Why Use LLMs for QDA?

- First-pass coding + summarization
- Handle large corpora
- Cognitive support
- Speed for iteration
- Must maintain trustworthiness & participant dignity

Technologies Are Not Neutral

- Early view: digital tools as neutral *instruments*
- Updated view: tools shape inquiry
- LLMs have **power** in what is seen / unseen

Key Threats of GenAI

- Exploited data workers
- Environmental burden
- IP + consent issues
- Algorithmic bias
- Hallucination errors
- Loss of authenticity

Engagement with AI is a **choice**, not a destiny

Efficiency ≠ Epistemic Quality

- “Faster” is not a valid analytic paradigm
- Summary ≠ interpretation
- LLMs flatten nuance if unchecked

Prompt Engineering = Analytic Intervention

How we ask affects:

- Codes generated
- Voices prioritized
- Power dynamics in text

Prompt Sensitivity Example

- **Prompt A** $\longrightarrow$ high-level themes
- **Prompt B** $\longrightarrow$ structured codebook with quotes

Differences are **methodological**, not cosmetic

LLM Parameters: What They Control

Parameter	Meaning	Impact on QDA
Temperature	Creativity vs precision	Nuance vs drift
Top-p / Top-k	Diversity of ideas	Breadth of coding
Max Tokens	Length/depth	Truncation vs saturation
Model choice	Training biases	Interpretive differences
Context window	Memory	Linking across data
System role	Interpretive lens	Analytical stance
Repetition penalty	Novelty	New vs redundant codes

Example: Changing Findings

Same transcript Same prompt Different parameters

Temp 0.1	Temp 0.9
Literal codes	Emotional interpretation
Conservative grouping	Diverse yet unstable themes
High repeat	Higher hallucination risk

Meaning changes So must **report** settings

Transparency Matters

- QDA already criticized for “black box” analysis
- GenAI increases opacity
- MERIT encourages explicit reporting:
    - Prompts
    - Model role & settings
    - Human oversight

Summary of MERIT Framework

- **M**ethods: Methods shift with settings
- **E**thics: Whose perspectives are emphasized/silenced?
- **R**esponsibility: Who validates decisions?
- **I**mpact: Did quality improve or degrade?
- **T**ool: Is the tool transparent about settings?

A reflexive guide for **trustworthy** GenAI practices

Hands-On Exploration

You will: 1. Run a baseline prompt 2. Adjust parameters 3. Compare changes 4. Reflect using MERIT

Developing your Custom GPT

1. Using ChatGPT Builder, create a custom GPT that codes transcripts with your own codebook,
2. Produces JSONL-structured outputs and memos
3. Uses shared framework to map themes
4. Maintains transparency in analytic decisions

Your Custom Tool Demo

- Modify: temp, top-p, tokens, model, role
- Compare outputs side-by-side
- Export prompt logs
- Built for **transparency in analytic decisions**

Best Practices

- Human-in-the-loop
- Prompt logs saved
- Model version + date recorded
- Validate outputs manually
- Member/peer checking
- Consent for AI use with participant data

Reflection Discussion

- What worked?
- What didn’t feel trustworthy?
- Where will GenAI enter *your* workflow responsibly?

References (APA Suggested)

- Silver, C., Paulus, T., & … (2024). *Evaluating GenAI in QDA…*
- Paulus, T., & Lester, J. (2021). *Doing Qualitative Research in a Digital World.*
- Minaee, S., et al. (2024). *Large Language Models: A Survey.*
- Zhu, et al. (2023). *LLM-in-the-loop Thematic Analysis.*
- Hannah & Bender (2022). *The myth of AI inevitability.*

Thank You

Questions & Discussion