Qualitative Data Analysis with Large Language Models (AI)

Dr Mat Bettinson
Greybeard Hacker @ The Digital Observatory

Resbaz 2024 – ACU
07/11/2024

This session

- What is QDA?
- The AI opportunity: qualitive analysis at scale
- Extraction, summarisation, coding
- Explainability and Errors
- About commercial AI models
- Baby steps: Using AI through programming (Queensland Election dataset)
- Concepts and Code AI coding pipelin

Qualitative data analysis

Qualitative Data Analysis Is:

- Exploring the *"how" and "why"* behind human experiences and behaviors
- Working with *non-numerical* data (text, images, audio, video)
- *Identifying patterns*, themes, and relationships in unstructured data
- *Interpreting meaning* from context and narrative
- Using methods like *coding* and *thematic* analysis

Qualitative Data Analysis Is Not:

- *Counting* frequencies or calculating statistics
- Testing hypotheses with *numerical data*
- Focusing primarily on *measurement or quantities*
- Seeking to prove *statistical significance*

The problem we’re solving

- Qualitative coding is:
  - Painstaking precise work
  - Hiring and training research assistant is expensive and time consuming
- Sometimes we need to explore data without prejudice to identify interesting patterns (Grounded Theory, cf. Glaser & Strauss 1967)
- Automation can boost research ambition for those without large grants

CRICOS No.00213J

A word on ethics

Do research assistants (students!) out of a job?!

- Hopefully not…
- Plenty of work in hand-coding gold standard examples, and checking AI-generated outputs
- In other words, more varied, ‘executive’ style work than repetitive work

That said…

- Some people are going to try and do things on the cheap with AI, we should do better

AI issues (nicked from Demystifying AI)

- The use of ChatGPT is **incompatible** with:
  - Ethics protocols for sensitive data, e.g. identifiable people.
  - Data governance requirements for anything other than public data (usually)

Presence of ✨ in popular software doesn't make it right.

- Wide variety of mitigating strategies:
  - Ensure only public data is used
  - Use an AI product through a cloud provider with:
    - Local deployment in Australia
    - No-train guarantees - Google GCP, Microsoft Azure have solutions
  - Use a locally deployed LLM
    - Fairly involved, far less capable

Qualitative operations

What sort of stuff can we do with AI?

Two basic themes:

1. Data exploration (not research outputs)
   - Extraction
   - Summarisation
2. Qualitative coding (research outputs)
   - Extractive open-ended classification
   - Controlled vocabulary coding (fixed list of things)

Qualitative operations cont.

What sort of stuff *can't* we do with AI?

1. Data exploration
    - Extraction
        - tell us HOW something was extracted
    - Summarisation
        - tell us precisely what contributed to the summary
2. Qualitative coding
    - Extractive (open) classification
        - tell us how this was determined
    - Controlled vocabulary coding
        - ditto, cannot be interrogated regarding mistakes

Dealing with AI (and errors)

- We need to explain our use of AI (in our methods)
- Some artefacts like summaries are *not research products* (because we can't reference/check them)
- For research outputs, we must check the output by:
    a. Ensuring we persists links to raw data (for reproducibility):
    b. Count the errors over a subset of human coded data
    c. Classify errors, e.g. identify patterns in errors
- Refine prompting strategies to get better results

Example of reproducibility

"The following is a series of reasons for opposing the comedy of controversial comedian Matt Rife (or believing that humour can be harmful generally), extracted from comments from the subreddit */r/abusiverelationships*.
Your task is to analyze the reasons and group them into similar themes."

Theme analysis for \r\abusiverelationships

Lazy and Unoriginal Humor

Comments criticizing Matt Rife's humor as lazy, unoriginal, or poorly delivered.

[kahw5g7, kagv93m, kah2tni, kah94k3, kahdq78, kajtw1o, kak9qgd]

Citations (comment IDs)

Punching Down

Comments highlighting that Matt Rife's jokes target vulnerable groups, particularly women and the working class.

[kahw5g7, kahtr01, kah94k3, kaj52wj, kajjy8r, kakjd67]

Normalizing Violence Against Women

Comments expressing concern that Matt Rife's jokes normalize or trivialize violence against women.

[kagqck8, kakew6x, kahh98w]

Original data

Yikes. Not even funny. Subpar content and bad delivery. If you're gonna be edgy, at least actually be funny. He's only 28 and he's telling lame boomer "jokes?"

Some of the very same men condoning Matt Rife's joke would suddenly be using male survivors as pawns if a woman made his same joke about a man - a lot of those men would then suddenly "care" and would not be viewing it as a joke any longer.

Which AI?

We need an AI that will ideally:

- Work with long prompts (contexts)
- Be reliable, fast and cost-effective

Three ‘classes’ of AI are applicable:

1. Nano – Gemini Flash 8B, self-run models
   Good for simple things, cheap or free
2. Small – GPT 4o mini, Gemini Flash, Claude Haiku
   Sweet spot – Gemini and Claude 💚
3. Large – GPT 4, Gemini Pro, Claude Sonnet
   Unreliable (GPT4), severely rate limited (Claude), slow (all), can be costly

Consider the 2024 Qld Election dataset

- 25,787 comments
- ~ 1.2m input tokens, ~500k output tokens

Model	Input cost	Output cost	Total cost
Gemini Flash 8B	4.5c	7.5c	12c
Gemini Flash	9c	15c	24c
GPT 4o Mini	18c	30c	48c
Claude Haiku	$1.50	$2.50	$4.00
Gemini Pro	$1.50	$2.50	$4.00
GPT 4o	$3.00	$5.00	$7.00
Claude Sonnet*	$3.60	$7.50	$11.10

* Not practical given current rate limits

AI pipeline

- Enough with the boring slides already, LETS CODE

https://github.com/QUT-Digital-Observatory/resbaz24_qda

Error analysis

- Let’s examine apparent party support with the big election issues
- Does this seem right to you?

Party	Abortion Laws	Cost of Living	Enviro. & Infr.	Health	Youth Crime
ALP	196	652	333	106	328
LNP	1166	596	494	305	1110

Let’s drill in

> “12. Not looking forward to being a disability pensioner under an LNP State government. Less healthcare, longer wait times for urgent surgery, let alone what is considered elective. I'm so glad that I'm past menopause so my reproductive health won't be taken out of my hands by the LNP.”

Consider our prompt: “…identify which party the comment appears to support”

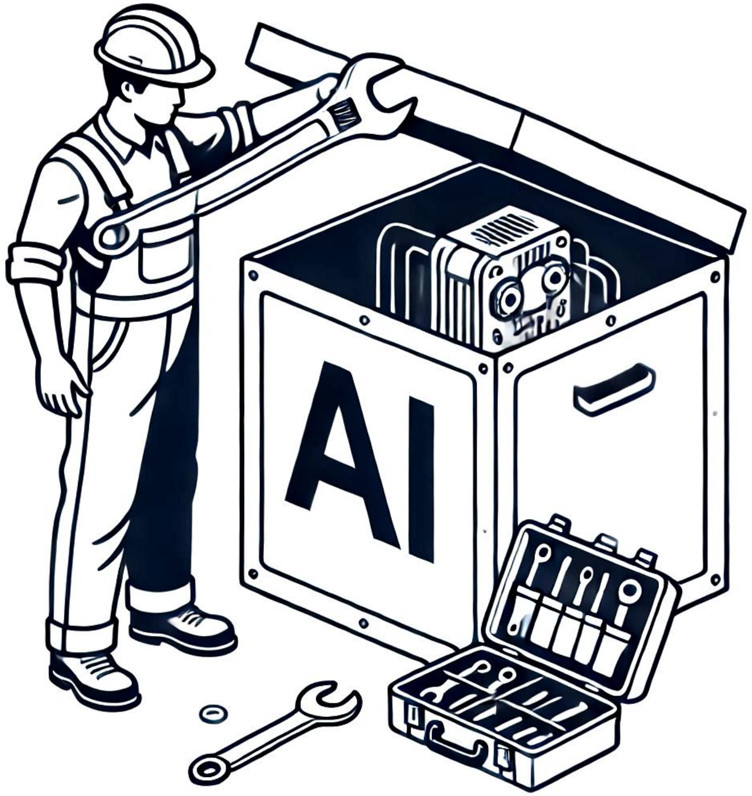
```
{
    "id": "ltwyh9n",
    "prompt_id": 12,
    "issue": "AL",
    "party": "LNP"
}
```

How to fix?

- Make a minimal prompt from prompt 62.txt
- Even better, get a range of LNP + Abortion Law examples
- Modify the prompt instructions
- Run directly in a chat window

Prompt tips:

- Increase the salience of the ‘party’ coding instructs.
- We could spell out the ALP+LNP duopoly as context
- We could elaborate specifically



That’s all folks!

- The Digital Observatory can help with:
  - QUT PhD and research projects
  - External projects with funding
- We have a lot of experience in AI processing of research data
- https://www.digitalobservatory.net.au

THANK YOU!

CRICOS No.00213J