



# Qualitative Data Analysis with Large Language Models (AI)

Mat Bettinson & Robert Fleet  
The Digital Observatory

DMRC Summer School 2025

# This session

- What is QDA?
- The AI opportunity: qualitative 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 pipeline

# 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



# 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

- The use of ChatGPT (or similar) 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

# Validity: AI and the Margin of Error

- AI coding falls short of human capability (in some ways)
- What really matters is the **margin of error** of our result (precision if you like)

$$\text{Margin of Error} = z \cdot \sqrt{\frac{p(1-p)}{n}}$$

- $z$  = z-score (typically 1.96 for 95% confidence)
- $p$  = observed proportion (error rate)
- $n$  = sample size

# Validity: It's all about $n$

- Precision goes **down** with error rate (AI accuracy)
- Precision goes **up** with  $n$
- AI error rate isn't bad, and we can **scale up  $n$**  in some situations (like social media data where  $n$  can be *millions*)

## Human coding:

- $n = 100$  (sample size)
- $p = 0.05$  (5% error rate)
- Error =  $1.96 * \sqrt{[(0.05 * 0.95)/100]}$   
 $\approx \pm 4.3\%$

## AI coding:

- $n = 1000$  (sample size)
- $p = 0.20$  (20% error rate)
- Error =  $1.96 * \sqrt{[(0.20 * 0.80)/1000]}$   
 $\approx \pm 2.5\%$

# Summarisation 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**]

### Punching Down **Citations (comment IDs)**

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, **kahh98wi**]

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?"

### Original data


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.

# AI performance and cost

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

# Let's get started (election dataset)

We put together a Reddit dataset on the **2024 Queensland State Election**

- Collection of submissions, comments and subreddit metadata between the 23rd to the 30th of October (Spans election date on the 26th)
- Came from our AusReddit project, with Queensland tagged subreddits such as r/queensland, r/brisbane, r/goldcoast and more
- We used generative AI to identify comments relevant to the election
- Dataset contains 25,787 comments drawn from 191 submissions



# Election dataset / pipeline repository

Dataset is in this repo at /2024\_qld\_election\_reddit\_dataset (see README)



[https://github.com/QUT-Digital-Observatory/resbaz24\\_qda](https://github.com/QUT-Digital-Observatory/resbaz24_qda)

CRICOS No.00213J

# Research question

Given we know the outcome and we have ample political analysis for the timeframe, **how did the election issues contribute to the LNP win?\***

The major issues were:

- Youth Crime
- Cost of Living
- Health
- Energy and Infrastructure
- Abortion Laws

\* Based on what we observe on Reddit



# To answer our question

For example:

“Don’t mind ALP but they seem to be soft as fuck on crime / sentencing minimums / general law and order. That’s only reason I voted LNP”

This comment is clearly about **youth crime**, there is also clear indication of **LNP support**.

# AI coding & prompt engineering

## LLM Context

We provide:

- Instructions
- Recent knowledge
- Research data



The  
“prompt”

## LLM Knowledge

They provide:

- Base training
- System prompt

# Instructions

You can see all the prompts in phase 1 in the Github repo in the path:  
`experiments/1/phase1/prompt_*.txt`

- Tell the model what we're doing
- Provide a top-level context

“Your job is to **analyse a series of Reddit Comments** are drawn from submissions related to the **2024 Queensland State Election**. The comments are drawn from a collection of comments before, during and after the election itself on the 26th of October. **The parties that sought election were the Australian Labor Party (ALP), the Liberal National Party (LNP), the Queensland Greens (Greens), One Nation and Katter's Australian Party.** The **LNP prevailed** in the election, with David Crisafulli becoming the new Premier of Queensland.”

# Instructions cont.

- Add recent information
- Add analytical basis (this could be a coding scheme described in literature)

“We have identified several key issues that the major parties campaigned on during the 2024 Queensland State Election: Youth Crime (YC), Cost of Living (COL), Health (H), Energy and Infrastructure (EI), and Abortion Laws (AL). Below is a summary of each party's position on these issues:

Youth Crime (YC):

- LNP: Advocated for stricter penalties for young offenders, including the 'Adult Crime, Adult Time' policy, proposing that serious offenses committed by youths be met with adult sentencing.
- Labor: Focused on rehabilitation and prevention programs, aiming to address the root causes of youth crime through community engagement and support services.”

CRICOS No.00213J



# Instructions cont.

- Add coding procedure instructions
- Reinforce (repeat) labels

“Analyze the Reddit comments and code them based on the positions they support or oppose concerning these issues. **You will output JSON, valid values are strings or null.** Use the following codes: YC, COL, H, EI, AL. If a comment's position is unclear or does not pertain to these issues, output null.

Secondly identify which party the comment appears to support. Use the party code ALP, LNP, Greens, ON, KAP if this can be determined, otherwise a null value if undetermined.

# Instructions (JSON)

- JSON is a structured output we can process in code
- Some models benefits from reinforcement on JSON types and output
- It's sometimes useful to spell out data types

Output JSON as a **list of objects** where each object has an "id" (**integer**) property of the comment id, an "issue" property set to the issue code string (or **null value**), and a "party" property set to the party code string or null value. Null means JSON null, not a string. Values **MUST** be quoted, e.g. {"party": "LNP"} unless they are null, e.g. {"party": null}. Comments follow:

# Results

“In Townsville there are so many initiatives trying to stop the root causes of juvenile offenders but no one funds them.”

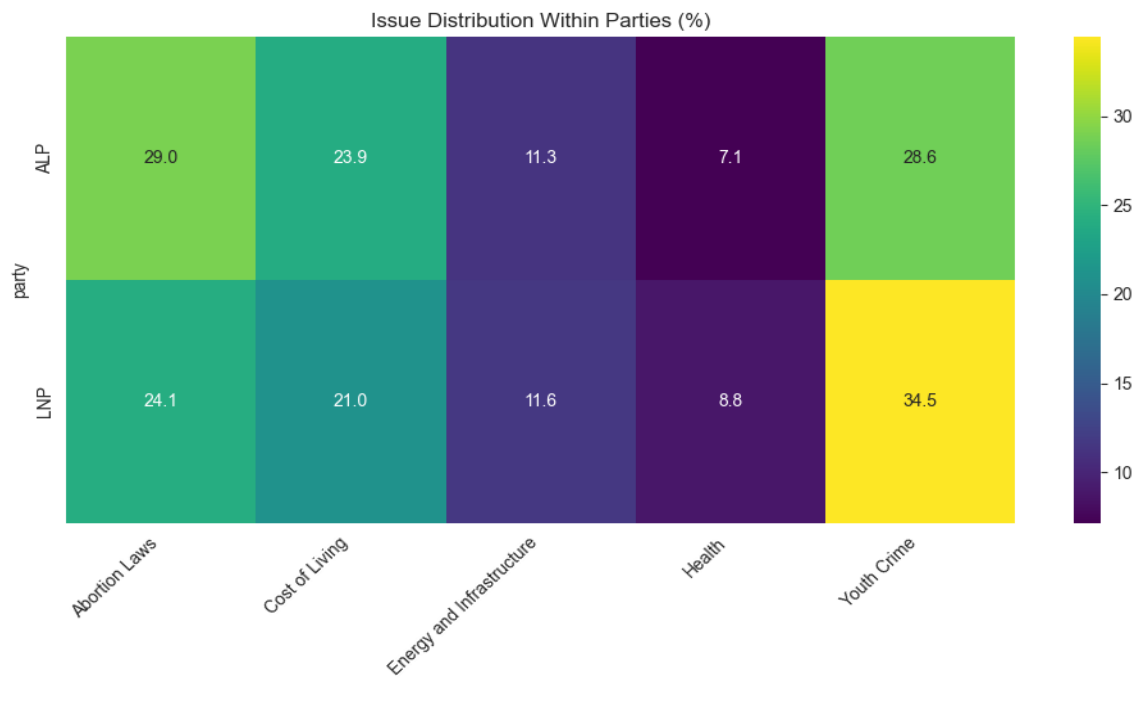
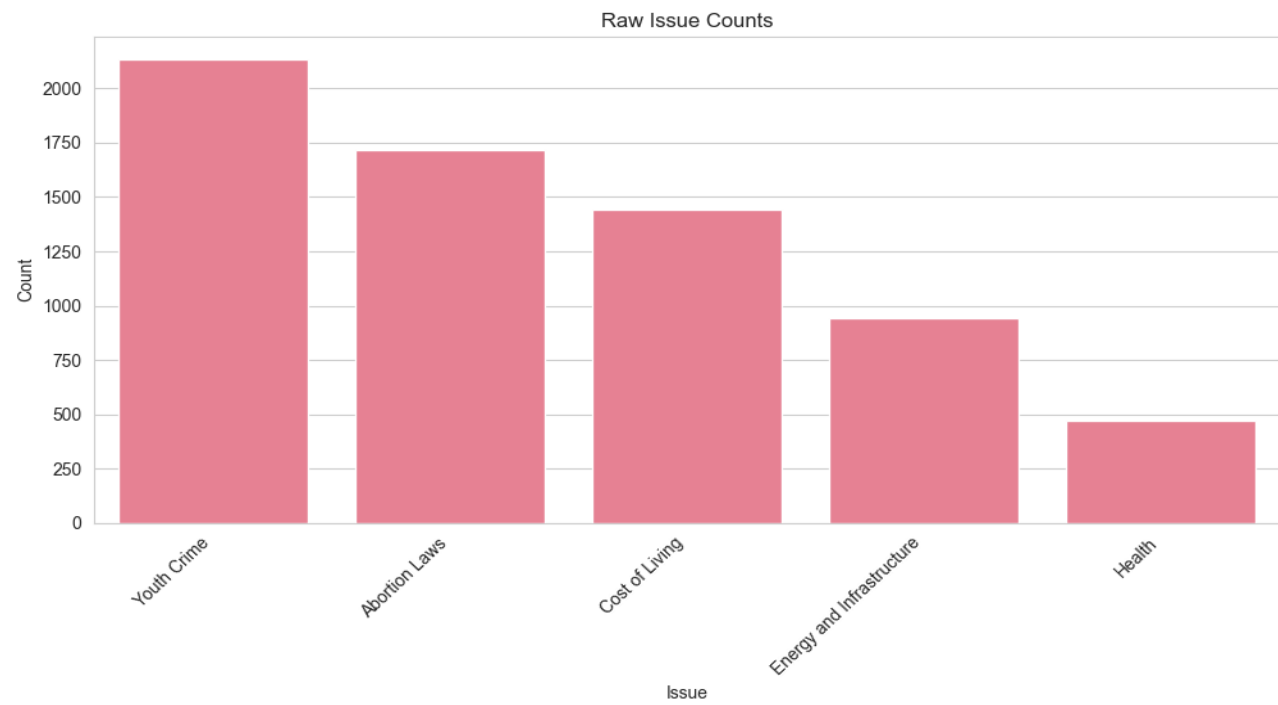
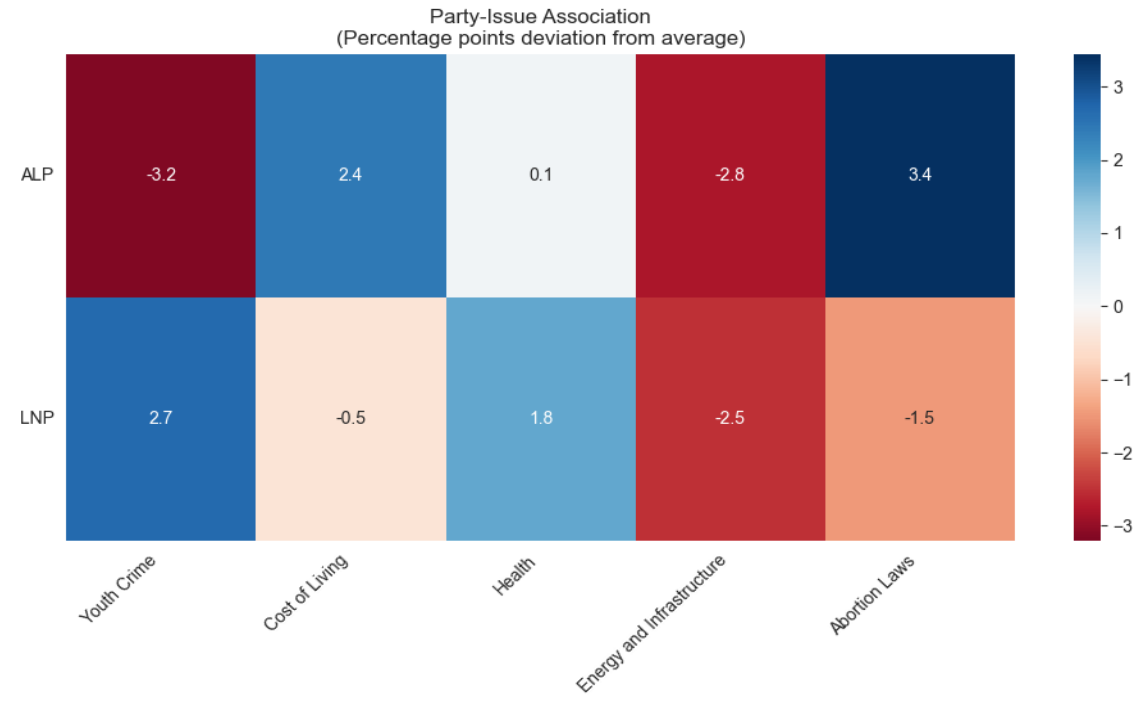
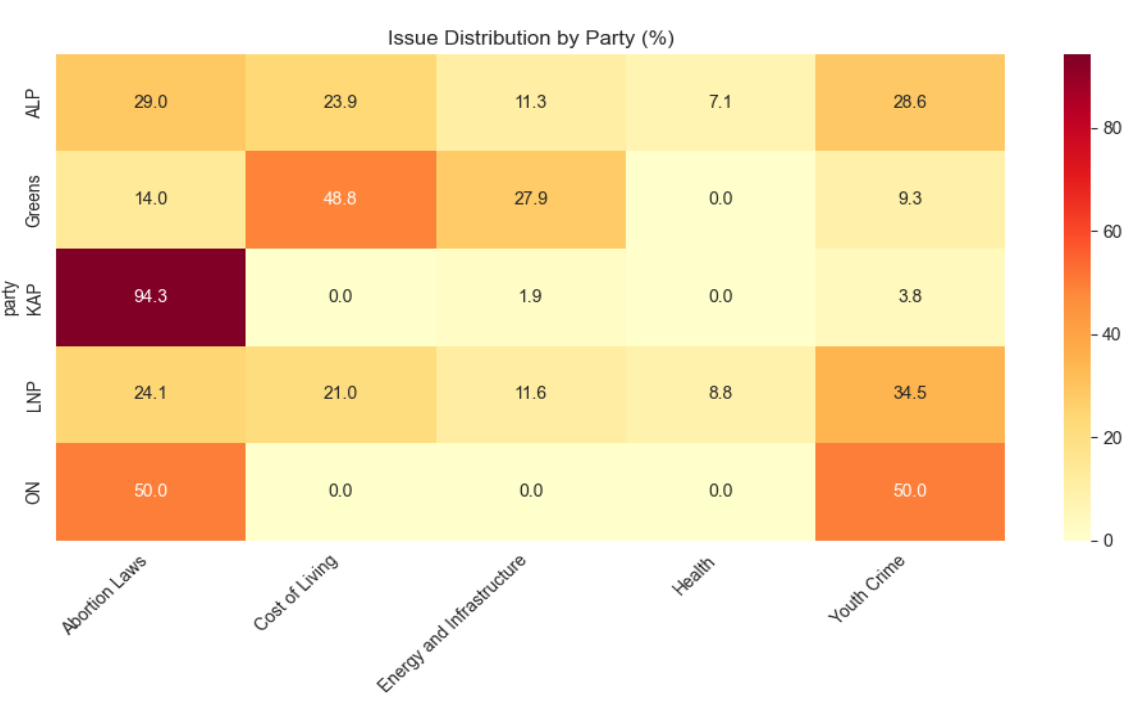
Real ID of the comment  
ID from the prompt  
Code (Youth Crime)  
Party support (unknown)

“I suspect you're right about youth crime. The policies Labor brought in in the last year have effectively worked (at least in the short term) and this trend will likely continue.”

“It's not baseless. Crime is up outside SEQ. Keep dismissing the truth at your own peril, as Labor has found out.”

CRICOS No.00213J

```
{
  "cost": 0.000890925,
  "model": "gemini-1.5-flash-002",
  "responses": [
    {
      "id": "lu26xtf",
      "prompt_id": 1,
      "issue": "YC",
      "party": null
    },
    {
      "id": "ltyt07w",
      "prompt_id": 6,
      "issue": "YC",
      "party": "ALP"
    },
    {
      "id": "lty0iji",
      "prompt_id": 9,
      "issue": "YC",
      "party": "LNP"
    }
  ]
}
```



# Great but...

- Real practice is to code a subset and check with humans
- We didn't do that here so, we *expect* mistakes
- Take a look at this from the first version of the prompt:

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”

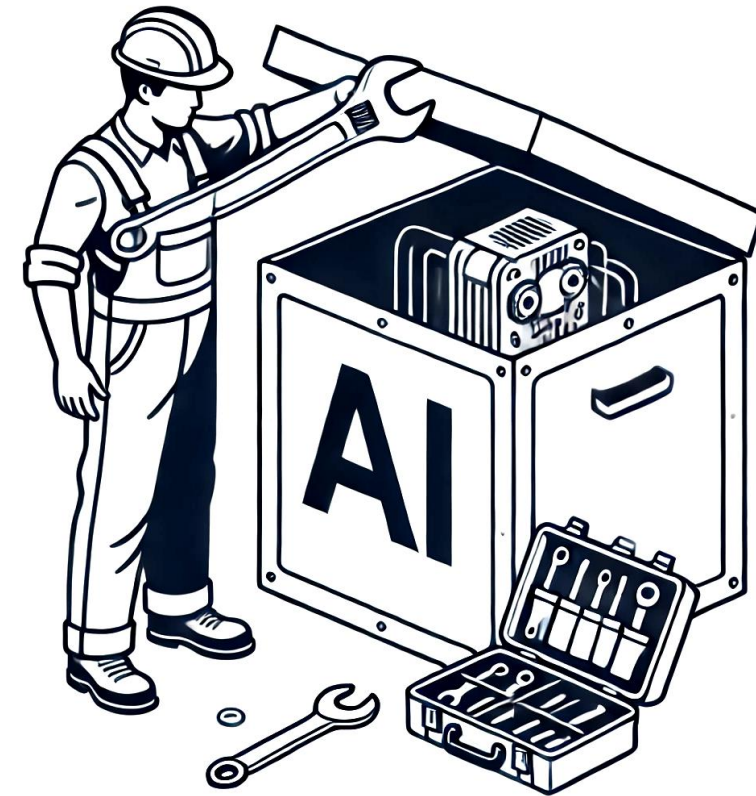
```
{  
  "id": "ltwyh9n",  
  "prompt_id": 12,  
  "issue": "AL",  
  "party": "LNP"  
}
```

CRICOS No.00213J



# How to fix?

- Make a minimal prompt as a coding test
- Even better, get a range of LNP + Abortion Law examples
- Modify the prompt instructions
- Do a test run or... run in a chat window (Google AI Studio is great for this)



CRICOS No.00213J

# Versioning prompts

- The problem with fiddling with prompts is:
  - You risk losing the knowledge of what worked
  - Reproducibility issues
- The solution is versioned prompting (simple example in the repo)
- Example, we made version 2 of the prompt to address the earlier problem and here's the notes from **qda\_03\_prompts.py**:

"Typo fixes. Removed info about LNP victory, added incumbent/opposition info, frame major parties, and minor parties. Condensed text. Improved clarity of coding instructions and considerably elaborated on party support coding. Shifted to omitting JSON properties where not determined. Increased prominence of omit strategy (mentioned with coding and json output)."

# Further enhancement

- In this example, prompts are dumps of comments
- “Labor has run Queensland for 30 of the past 35 years, you were saying?”
- Significant amount of context is not being used
- Reddit > Subreddit > Submission > Comment > Reply

To do this properly, we need to iterate through some of them and the prompt should offer the suitable context.

We provide example code that outputs Markdown representations of threads in **qdaai/threading.py**

# 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!