Evaluating Large Language Models (LLMs) & Infrastructure for Scalable Qualitative Annotation

Junran Lei | HASS Digital Research Hub

Email: Junran.Lei@anu.edu.au

Australian National University

Evaluating Large Language Models (LLMs) and Infrastructure for Scalable Qualitative Annotation

- Australian Research Council (ARC) Project: Public Interest Advocacy
- Analysing advocacy statements in news media
- Collaborators: Darren Halpin (ANU), Anne Sofie Cornelius Nielsen (ANU), Max Groemping (Griffith), Mat Bettinson (ANU)
- Work in Progress – Methodology Demo

The Challenge: Scaling Qualitative Annotation:

- Manual annotation doesn't scale with large datasets
- Multiple coders create consistency challenges
- LLMs accelerate work but involve trade-offs
- Must balance accuracy, cost, and data control
- Ethics requirements constrain infrastructure choices

Our 5-Step Workflow:

1. Ground-truth consensus dataset (3 researchers annotate subset)
2. Design prompts from codebook
3. Test across models and infrastructure types
4. Evaluate systematically: accuracy, cost, time, privacy, governance
5. Scale to full dataset with best-fit solution

Current status: 1 complete annotation; expect +10–20% accuracy with full consensus & prompt improvement

Four Infrastructure Types

- Local (DGX/Ollama): Easy setup, slow speed, full control, free
- Nectar A100 (80GB): Moderate setup, medium speed, data sovereignty
- NCI HPC (1-8×A100): Complex setup, fast, handles large models (235B+)
- APIs (OpenAI/Google/Bedrock/Openrouter): Instant setup, fastest, minimal control
- Key constraint: Ethics often requires institutional computing

Performance Trade-offs (250 records):

Platform	Time	Cost	Current Accuracy*
Local DGX (Qwen2.5 72B)	33 min	Free	46%
Nectar/NCI Single (Qwen2.5 72B)	19 min	2-3 / 18 SU	46%
NCI 8-GPU (Qwen3 235B)	23 min	222 SU	56%
Gemini 2.5 Flash	5 min	$0.09	54%
GPT-5	67 min	$2.51	63%

* **Expected improvement** to 56-83% with consensus annotation and prompt tuning.

Note: 60-70% accuracy matches human coder agreement

- **Scaling to 30k records:** Nectar ~40h, NCI 8-GPU ~46h, APIs 10-134h

Data Governance Matters:

Why research & institutional computing:

Data governance often decides the platform

Commercial APIs can change terms, reuse data, or retire models

Reproducibility requires stable infrastructure

Scaling large models (Qwen3 235B example):

OpenRouter: 8 min, $0.24 per 250 → ~$29 (30k), ~15h

NCI 8-GPU: 23 min, 222 SU per 250 → ~26,640 SU (30k), ~46h

Trade-off: speed and accessibility vs control and security

Key Takeaways:

- No universal best: Only best fit for your constraints
- Decision depends on: Data sensitivity, budget, technical capacity, scale
- Every platform trades: Cost, time, skill, privacy, and control
- Testing now prepares for future sensitive research
- We can use the same process for other models and platforms

Expected accuracy with consensus & prompt tuning

Model	Expected Accuracy	30k Cost	30k Time
GPT-5	73-83%	$300	~134h
Grok-4	71-81%	$10	~105h
Claude 4.5 Sonnet	70-80%	$72	~15h
Gemini 2.5 Flash	64-74%	$12	~10h
Qwen3 235B (NCI)	66-76%	26,640 SU	~46h

Note: Expected 60-70% range matches human annotator agreement