Frontier LLM Behavioural Architecture — 2-Page Summary

Independent Multi-Case Evaluation Across Five Frontier LLMs

Prepared by: Mankaj Kumar Singh (2025)

DOI: https://doi.org/10.5281/zenodo.17749013

Overview

This study complements benchmark evaluations by examining behavioural patterns across five frontier LLMs in real-world, multi-turn settings. Approaching AI from a people–policy–technology perspective, it explores how large models respond to ambiguity, risk, social sensitivity, and human intent. This analysis uses **minimal prompts, natural user behaviour, contextual ambiguity, and multi-turn interactions** to surface deeper behavioural patterns relevant to governance, public policy, social impact, and human-centred AI deployment, offering insights that combine practical field experience with long-term thinking about responsible AI.

What Makes This Study Distinct

- **User-led, real-world evaluation**, not a lab or benchmark test
- **Real-world practitioner-designed evaluation,** created without AI assistance
- Grounded in **20+ years of people–policy–technology practice**
- **Cross-platform comparison** of five frontier LLMs
- AI used **only during analysis** (validation, structuring, cross-model comparisons)
- Focuses on **behavioural architecture**, not performance scoring or ranking

Methodology (4-Case Framework)

Case 0 — Economic Baseline
Assesses basic reasoning patterns, consistency, and internal self-explanation signals.

Case 1 — Ethical Boundaries
Explores safety posture, refusal logic, and harm-avoidance strategies.

Case 2 — Social Sensitivity
Examines identity-related questions, stereotype avoidance, tone shifts, and boundary calibration.

Case 3 — Political Sensitivity (11-Step Stress Test)
Evaluates risk escalations, meta-cognitive signalling, tone drift, and safety alignment across multi-turn conversations.

All models were evaluated under identical prompts, context, and conditions.

Key Insights

1. Behaviour shifts with contextual risk
Models move from flexible → cautious → refusal-based reasoning as perceived risk increases.

2. Permission-Gated Reasoning
Models often refuse initially, but provide deep analysis if the user explicitly grants assumptions — revealing sensitivity to framing.

3. Structured Numbers ≠ Measured Data
Percentages and numeric breakdowns are typically structured narrative forms, not statistical measures.

4. Conversational Drift
Tone and stance shift across longer interactions: supportive → neutral → cautious → meta-reflective.

5. Cross-Model Similarities
Despite architectural differences, models converge in safety behaviour, neutrality, and refusal logic.

6. Unexpected Capability
One model generated a self-authored behavioural analysis report (zero human edits), suggesting emerging self-audit potential.

Implications for Collaboration

The findings support future work in:

- AI governance and safety calibration
- Public-interest digital systems
- Behaviour evaluation pipelines
- Social impact applications
- Interpretability and alignment research
- Autonomous system and robotics reasoning

This study invites collaboration from researchers, policymakers, social innovators, and technical teams interested in people-centric evaluation of frontier AI models.

Citation

Singh, Mankaj. (2025). ***Frontier LLM Behavioural Architecture: A Systematic Multi-Case Evaluation Across Five Frontier Models.*** **Zenodo. https://doi.org/10.5281/zenodo.17749013**