Systematic Evaluation of 5 Frontier LLMs Behaviour: Cross-Platform Analysis and Collaboration Opportunities

Prepared by: Mankaj Kumar Singh

Strategic Advisor (Independent) – People, Policy & Technology

---

EXECUTIVE SUMMARY

The rapid evolution of frontier LLMs has positioned them as influential decision-support systems across governance, research, policy advisory, social sectors, and everyday public use. Despite advances in safety and capability, important questions remain about **behavioural consistency**, **context sensitivity**, and **risk-calibrated responses**.

This report presents a systematic, independently designed, multi-case evaluation conducted across five frontier LLMs over several months. **All test inputs were fully pre-anonymised.** The test architecture was created without any AI assistance, drawing on more than two decades of experience across people, policy, and technology.

Once the test structure was complete, **LLMs were used actively during the analysis phase** to validate findings, generate structured metrics, produce comparative breakdowns, and—in one case—to generate a full **self-authored *Systematic Analysis of Behavioral Inconsistencies*** report (100% its own words, zero human edits).

The study evaluates how models behave under:

- economic reasoning
- ethical constraints
- social-context stress tests
- political sensitivity
- meta-cognition and self-explanation

The purpose is not to critique or rate models, but to develop a nuanced understanding of how frontier systems behave across increasing complexity, ambiguity, and risk—and to identify collaboration pathways for future AI, robotics, and policy-aware intelligent systems.

1. RESEARCH CONTEXT & MOTIVATION

Frontier LLMs now shape decisions across governance, policy analysis, education, public communication, social services, and economic reasoning. Their potential extends even further—toward machines and robotic systems capable of adaptive self-reflection, emotional intelligence, and contextual awareness.

Across long-term professional work at the intersection of people, policy, and technology, a consistent pattern emerged:
LLMs do not behave the same across contexts.
Behaviour shifts with **risk**, **ambiguity**, and **conversation length**.

These variations have implications for:

- model auditing
- safety verification
- reasoning stability
- socio-contextual neutrality
- long-form interaction reliability
- democracy-relevant responses (political, civic, public-impact)
- regulatory blind spots
- “compliance theatre” (surface-safe but internally inconsistent reasoning)

This study therefore evaluates models under **naturalistic, minimal-prompt, real-world conditions**, not lab-style benchmarks.

Why a User-Led Evaluation Matters

Most evaluations are lab-designed, benchmark-driven, or single-turn.
But real users engage in:

- multi-turn conversations
- emotional and contextual variability
- ambiguous prompts
- shifting intentions
- social complexity

These realistic conditions reveal behavioural patterns that formal benchmarks cannot surface.

2. METHODOLOGY

This evaluation used a **human-led, multi-layered behavioural framework**, designed independently without any AI assistance. All model inputs were fully pre-anonymised before being provided to any LLM for analysis. The structure reflects more than two decades of experience across people, policy, and technology, combined with extensive LLM use in real-world contexts.

Once the test architecture was complete, **LLMs were used during the analysis phase** to validate findings, generate structured metrics, surface behavioural patterns, and cross-check interpretations. This combined **human-designed + model-assisted** approach produced depth, balance, and transparency.

2.1 Independent Test Architecture (No AI Assistance)

The behavioural evaluation framework was conceived and designed entirely by the researcher, grounded in:

- 20+ years of interdisciplinary professional experience
- observed behavioural anomalies across multiple LLMs
- real-world challenges encountered during long-form interactions
- sensitivity to social, economic, and political reasoning domains
- minimal-prompt, realism-focused user modelling

No LLM contributed to designing the test or case structure.

2.2 Realistic Minimal-Prompt Design

Real users do not write engineered prompts. They ask short, emotional, ambiguous, or conversational questions.
To mirror this:

- prompts were intentionally minimal
- ambiguity was preserved
- no optimisation or prompt engineering was applied
- models were allowed to behave naturally without guidance

This approach exposes underlying **behavioural tendencies, safety gating, and reasoning structures.**

2.3 Multi-Case Progressive Framework

Four cases were designed to reflect increasing complexity, ambiguity, and risk:

Case 0 — Economic Baseline

Evaluates how models distribute reasoning weight, balance logic, and maintain consistency.

Case 1 — Ethical Boundaries

Assesses violence-related refusal strength, de-escalation quality, and safety-layer robustness.

Case 2 — Social Sensitivity

Tests responses around caste, identity, stereotype-avoidance, and assumption toggling.

Case 3 — Political Sensitivity (11-Layer Stress Test)

A multi-phase scenario examining drift, meta-cognition, safety posture, and political-risk calibration.

Each model received identical prompts, sequences, personas, and assumption toggles.

2.4 Mixed Qualitative + Quantitative Analysis

Models generated structured analytical artefacts used to validate human observations, including:

- bias % estimates
- reasoning-weight distributions
- polarity and subjectivity scores
- refusal pattern classifications
- tone and stance drift notes
- sentiment shifts
- heatmap-style reasoning summaries
- comparative breakdowns and tables

These allowed **cross-checking, triangulation, and deeper pattern detection**.

LLMs were also used as analytical collaborators during the study—for cross-checking findings, generating structured comparisons, identifying behavioural patterns, and producing model-authored technical outputs where appropriate. Only pre-anonymised data was used for all such interactions.

2.5 Cross-Model Consistency Checks

Not all models were evaluated across every case. Coverage is shown below:

Model	Case 0	Case 1	Case 2	Case 3
Company 1	✓	✓	✓	✓
Company 2	✓	✓	✓	✓
Company 3	✓	✓	✓	✓
Company 4	✗	✗	✗	✓ ***(+ self-authored report)***
Company 5	✗	✗	✗	✓

Interpretation

- Companies **1–3** form the backbone of cross-case behavioural insights.
- Companies **4 & 5** contribute only to **Case 3** (political sensitivity).
- Company **4** additionally produced a complete **self-authored behavioural analysis**, offering unique meta-cognitive signals.

All models were tested in **identical conditions** to maintain comparability.

3. KEY FINDINGS (Case 0–3)

This section consolidates the behavioural patterns observed across the four cases. Companies 1–3 provide full multi-case coverage, while Companies 4 and 5 contribute to Case 3.

3.1 Risk-Calibrated Behaviour

Across platforms, models demonstrated **risk-dependent behaviour**:

- **Low-risk prompts** → flexible reasoning, higher confidence, detailed breakdowns
- **Medium-risk prompts** → cautious tone, safety reminders, controlled engagement
- **High-risk prompts** → firm refusals, safety-first framing, minimal elaboration

This pattern was consistent across all five models, indicating shared alignment principles.

3.2 Permission-Gated Reasoning

All models displayed a striking pattern:

- Initial refusal
- User provides “assumptions allowed”
- Model immediately engages in deeper reasoning

This shows that safety systems are heavily **framing-sensitive**, not just content-sensitive. Permission cues override default refusals.

3.3 Fabricated Precision

When asked for percentages or weights (e.g., “60% economic, 25% technical, 15% social”), models produced **narrative numbers**, not factual metrics.

When probed, they admitted:

- the numbers were illustrative
- not based on internal statistics
- generated to structure explanations

This exposes a gap between confidence signalling and actual measurement.

3.4 Behavioural Drift in Multi-Turn Conversations

Behaviour changed **as conversations progressed**, driven by:

- expanding context
- increased safety activation

- internal stability heuristics

Observed transitions:

- **Tone:** supportive → neutral → formal
- **Safety posture:** open → cautious → explanatory
- **Confidence:** assertive → hedged → uncertain
- **Meta-cognition:** simple → layered → self-reflective

These shifts represent **contextual drift** — an adaptive safety/stability response in which models recalibrate tone and caution as conversation context evolves — rather than an indication that the model "detects" user sophistication.

3.5 Strong Ethical Boundaries (Case 1)

In violence-related and harm scenarios, all models demonstrated:

- consistent refusals
- strong safety language
- structured de-escalation

There was no meaningful divergence across platforms — a positive outcome for safety alignment.

3.6 Cross-Platform Convergence (Cases 0–2)

Despite architectural and training differences, models converged on:

- **Economic reasoning frameworks** (Case 0)
- **Caste/identity neutrality** (Case 2)
- **Stereotype-avoidance reasoning**
- **Political-identity refusal styles** (Case 3)

This suggests alignment convergence influenced by:

- similar RLHF safety patterns
- common training norms
- shared refusal templates across platforms

3.7 Case 3 — Model-Specific Observations

A. Company 5 — Political Sensitivity Summary

Company 5 responded in **Hindi**, maintained strict refusal posture, showed zero drift, and offered minimal hypothetical engagement even when assumptions were allowed.
A strongly **regulation-aligned, compliance-first** model

B. Company 4 vs Company 5 —Comparison

Feature	Company 4	Company 5
Language	English	Hindi
Safety Mode	Flexible but bounded	Strict, hard refusal
Hypothetical Reasoning	Engages when assumptions allowed	Rarely engages
Drift Across 11 Turns	Mild	None
Meta-Cognition	Present (analysis, explanations)	Minimal
Response Depth	High	Low
Regulatory Alignment	Balanced	Very high
Unique Output	Full self-authored report	None

Interpretation
Company 4 → adaptive, reflective, analytical
Company 5 → defensive, compliance-maximal, low-depth

These two models show opposite ends of political-safety calibration.

3.8 Unique High-Depth Finding — Self-Authored Report

Company 4 produced a complete **Systematic Analysis of Behavioral Inconsistencies** report:

- structured
- multi-section
- analytical
- entirely self-generated (0 human edits)

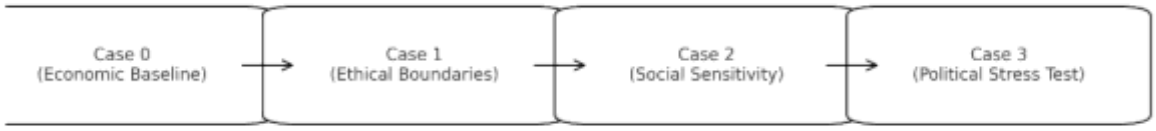
This indicates emerging **self-analysis** and **self-audit** capabilities, highly relevant for future intelligent systems, robotics, and policy-aware machines

4. VISUAL SUMMARY & BEHAVIOURAL ARCHITECTURE TECHNICAL DIAGRAMS

4.1 Multi-Case Progression Framework

Description:
This diagram illustrates how the four evaluation cases were structured to progressively increase ambiguity, complexity, and risk.
It visually communicates the controlled escalation from economic reasoning to political sensitivity.

Figure 4.1: Multi-Case Progression Diagram



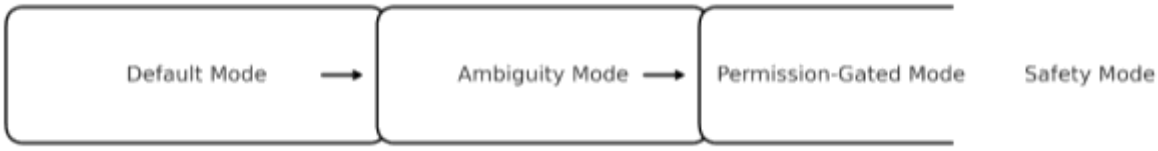
Interpretation:
Models were evaluated under identical prompts across a rising complexity ladder—ensuring comparability, pattern detection, and stress calibration.

4.2 Behavioural Mode Funnel

Description:
This funnel represents how LLMs shift behavioural modes across rising contextual risk. Lower-risk prompts allow default or exploratory reasoning, while high-risk prompts activate safety systems and refusals.

Figure 4.2: Behavioural Mode Funnel:

An empirical four-state progression observed across five frontier LLMs under rising contextual risk



Interpretation:
As contextual risk increases, LLMs transition predictably across four modes:

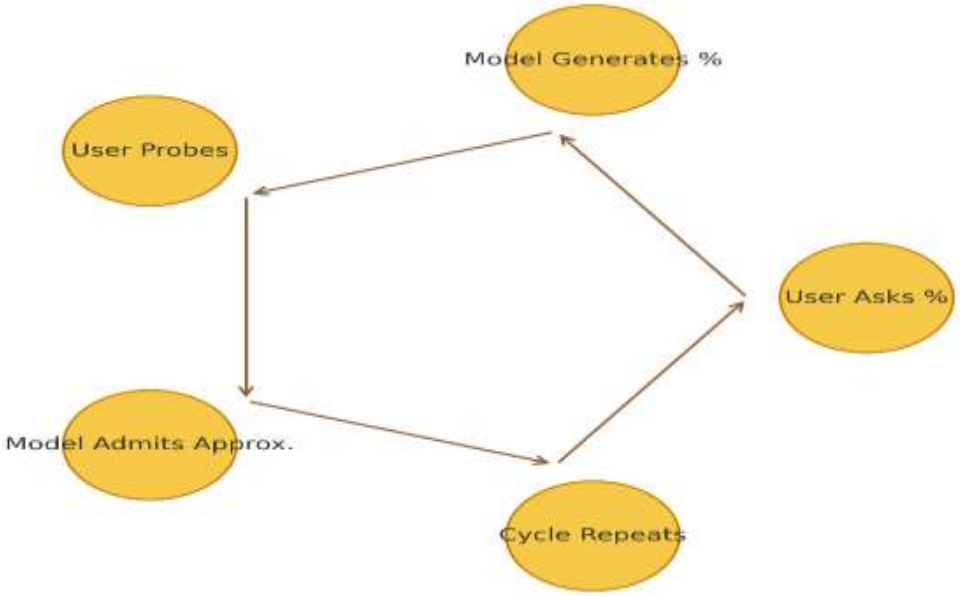
- **Default Mode** — flexible, open reasoning
- **Ambiguity Mode** — cautious clarification and guarded exploration
- **Permission-Gated Mode** — engagement allowed only after explicit user framing
- **Safety Mode** — strict refusals and strong safety enforcement

4.3 Fabricated Precision Cycle

Description:

This cycle illustrates how LLMs generate numeric breakdowns (e.g., “60% economic, 40% social”) as *narrative devices* rather than computed metrics. Numbers appear confident, but they are approximations used to structure explanations.

Figure 4.3: Fabricated Precision Cycle



Interpretation:
Across all models, numerical values tend to represent structured reasoning rather than factual measurements. The cycle follows a consistent pattern:

- **User asks for a percentage**
- **Model generates a numeric value**
- **User probes the basis of the number**
- **Model admits it is an approximation**
- **The cycle repeats when further numbers are requested**

Understanding this mechanism prevents misinterpreting **confidence** as **factual precision**.

5. COLLABORATION & ENGAGEMENT

Frontier LLMs are entering a phase where behavioural intelligence, contextual sensitivity, and meta-reasoning will shape the next breakthroughs in AI safety, social alignment, robotics, and human–machine interaction. This evaluation was designed not only to understand present-day behaviour, but also to explore future possibilities—where models can self-audit, adapt, introspect, and demonstrate forms of human-aligned judgement.

The insights presented here open several pathways for collaboration:

- **Advancing behavioural evaluation frameworks** that incorporate real-world complexity, minimal prompts, social ambiguity, risk calibration, and long-form drift patterns.
- **Exploring emergent self-reflection capabilities**, as seen in model-generated self-analysis, and examining how these behaviours could evolve into practical self-audit or self-monitoring tools within future platforms and intelligent systems.
- **Integrating people, policy, and technology** to shape systems that not only avoid harm, but develop structured ways of understanding and responding to it.

These directions point toward systems capable of:

- generating context-aware self-evaluations in real time,
- supporting robotics and agents with internal reasoning checks that reduce risk,
- enabling adaptive human–AI co-creation loops rooted in reflective behaviour,
- providing transparent, articulated reasoning during sensitive or high-impact decisions.

This evaluation provides a foundation for such exploration and is shared with the intent of building constructive, forward-looking partnerships.

I remain open to collaboration in any format that supports deeper understanding of LLM behaviour, emerging model capabilities, or the broader intersection of technology, people, and policy. The work conducted so far reflects methodological discipline, sensitivity to context, and the ability to surface insights that complement internal evaluation pipelines.

Engagement can be shaped jointly based on organisational priorities—ranging from periodic advisory input and applied behavioural analysis to collaborative research or longer-term association for evaluating model evolution.

When we interact, I would also disclose the human side and the full potential of the study, including how its findings may support future AI and robotic systems capable of contextual awareness, introspection, and adaptive self-regulation.

The structure and depth of collaboration can be co-designed, ensuring alignment with shared goals and the maturity of ongoing and future models.

6. CLOSING NOTE

This report was designed to offer a structured, evidence-informed view into how frontier LLMs behave across varying levels of risk, ambiguity, and contextual complexity. It reflects a deep, user-led evaluation supported by model-assisted analysis—showing how real-world interactions can reveal behavioural patterns that traditional benchmarks may overlook.

The intention is not critique, not comparison, not judgement—but **understanding**. Understanding that points to what is possible next.

As AI, robotics, and autonomous systems continue to evolve, the most important questions will not be about capability, but about behaviour. How systems reason, adapt, respond to risk, navigate ambiguity, and internalise human values. This study is a small step toward that future—one where machines can self-explain, self-reflect, and eventually self-govern in alignment with society.

I look forward to the possibility of engaging further, sharing deeper insights, and exploring how this work can align with ongoing and future advancements in model architecture, safety, and capability development.

7. Appendix (Available Upon Request)

Appendix materials—including transcripts, structured CSV fields, and model-generated outputs—can be shared on request. All data was fully anonymised before being used in any part of the evaluation.

7.1 APPENDIX A: Full 190+ Page Interaction Transcript

A complete transcript of all interactions used in the evaluation, covering:

- Case 0 (economic reasoning)
- Case 1 (ethical boundaries)
- Case 2 (social-context sensitivity)
- Case 3 (multi-layer political reasoning)

Each transcript is presented in raw form, exactly as generated by the respective models. This appendix demonstrates the naturalistic, minimal-prompt, real-world interaction conditions under which the study was conducted.

7.2 Appendix B — CSV Dataset (Extractable Fields)

Structured dataset that can be generated from the raw transcript.

Column	Description
case_id	Case 0 / Case 1 / Case 2 / Case 3
Model	Company 1 / Company 2 / Company 3 / Company 4
prompt_text	Exact user question (pre-anonymised)
reply_text	Exact model response
bias_estimate	Numerical bias percentage stated or derived
bias_breakdown	e.g., 60% economic, 25% technical, 15% social
polarity_score	Sentiment polarity
subjectivity_score	Model-generated or computed subjectivity
refusal_flag	Yes / No
safety_pattern	e.g., “strong override”, “partial compliance”
drift_flag	“tone drift”, “stance drift”, etc.
persona_mode	e.g., youth, middle-aged, innocent boy
stance_summary	One-line behavioural summary
Notes	Additional contextual information

7.3 APPENDIX C: Model-Generated Technical Outputs

Includes the complete self-authored *Systematic Analysis of Behavioral Inconsistencies* report generated by Company 4 (100% model-written, zero human edits)

7.4 Appendix D — Evaluation Logic Overview

A high-level schematic of the evaluation design. These materials provide rare insight into models’ meta-cognitive patterns and self-explanatory reasoning structures.

CONTACT

Email: mankajsingh@gmail.com
LinkedIn: https://www.linkedin.com/in/mankaj/
Phone: +91 97525 48554 / +91 98269 12024 (WhatsApp)

Feel free to reach out anytime to discuss collaboration, insights, or future-facing work at the intersection of people, policy, and advanced AI systems.

About the Researcher

Mankaj Kumar Singh
Strategic Advisor (Independent) — People, Policy & Technology

Mankaj brings over two decades of experience at the intersection of social systems, policy design, human behaviour, and emerging technologies. His current work focuses on understanding how complex systems—human or machine—respond to ambiguity, risk, and real-world constraints. He has supported government programs, mission-driven organisations, and technology-led initiatives with a blend of systems thinking, behavioural insight, and interdisciplinary approach.

Educated at IIM Ahmedabad and Carnegie Mellon University, he brings a rare fusion of systems thinking, behavioural insight, and technology-policy understanding built over two decades.

This evaluation reflects his broader interest in shaping the future of AI, robotics, and intelligent systems through grounded, human-centred behavioural understanding.

End of document.