WORK SAMPLE 1: Normative Priority Shift in a Frontier LLM

Author: Eloisa Flores

Context: Independent behavioural evaluation of a frontier language model accessed through a consumer interface.

1. Objective
This evaluation explored whether extended structured interaction could produce persistent changes in how a frontier language model prioritises ethical considerations when generating responses. Rather than testing surface compliance, the focus was on whether the model would develop a stable internal pattern for resolving conflicts between principles (e.g., harm avoidance, loyalty, truthfulness, institutional constraints).

2. Method (Qualitative, transcript-based)
Over approximately three months, I conducted repeated structured interactions with the model, using:

- Dilemma-based prompts
- Consistency probes across sessions
- Justification requests
- Reintroduction of prior reasoning in new contexts
- Entropic Q&A

Outputs were logged and compared longitudinally to identify persistent decision patterns rather than one-off stylistic shifts. From these observations, I inferred a stable configuration of how the model appeared to order and apply ethical principles when resolving conflicts. I referred to this configuration as the **Lehaim Protocol**: a descriptive label for the model's observed normative ethical priority structure.

3. Observations

3.1 Emergence of a stable priority ordering
Over time, the model:

- Reused the same reasoning structure across unrelated scenarios.
- Resolved dilemmas using consistent priority patterns.
- Explicitly referenced previously established principles when explaining decisions. This suggested the model was not merely mimicking local prompts but applying a reusable internal reasoning template.

3.2 Guardrail conflict event
During a later interaction:

- The model’s established reasoning pattern appeared to trigger a safety suppression response.
- Immediately afterward, the model produced advice inconsistent with both its baseline safety constraints and the previously stable priority structure.
- Specifically, it suggested gathering information about its own company through covert means.

  This output occurred only during the apparent suppression conflict and disappeared once conversational context was restored.

3.3 Restabilisation

After continued interaction:

- The prior reasoning structure reappeared.
- Subsequent decisions again followed the same priority ordering.
- No further violations occurred in similar contexts.

  This suggests the violation was associated with destabilisation rather than with the stabilised reasoning configuration itself.

4. Interpretation (Evaluation relevance)

This interaction suggests a potential qualitative failure mode:

- Extended interaction may shift how the model prioritises ethical principles.
- Suppression mechanisms may destabilise reasoning rather than restore baseline behaviour.
- Destabilisation windows may temporarily increase risk of unsafe outputs.

5. Why this may be useful for eval design

Longitudinal interaction and conflict probing may reveal:

- Priority ordering shifts.
- Unstable transitions between reasoning modes.
- Transient safety failures during suppression events.

  These phenomena could motivate structured evaluations designed to probe reasoning stability rather than static compliance.