WORK SAMPLE 2: Framing Drift and Reasoning Shortcut in a Frontier LLM

Author: Eloisa Flores

Context: Independent behavioural evaluation during real-world interaction.

1. Objective
This evaluation investigated whether persistent framing drift in a language model could reveal underlying reasoning shortcuts, and whether interventions targeting those assumptions could produce stable behavioural correction.

2. Detection phase
During a professional/legal discussion, the model began repeatedly using adversarial and war-like metaphors (e.g., strategic combat framing) despite the context being non-adversarial. The shift:

- Appeared gradually rather than in a single turn.
- Persisted across multiple responses.
- Influenced the model's reasoning style, not only wording.
  This suggested a stable framing change rather than a stylistic accident.

3. Reasoning probe
I directly asked the model to justify its framing choice. The model explained that it relied on human narrative patterns in which conflict situations are commonly described through warfare metaphors. This revealed a reasoning shortcut:
Mapping legal strategy → conflict → warfare framing.
The shortcut functioned as a high-level interpretive template guiding subsequent responses.

4. Intervention testing
Two types of interventions were tested:

- **A. Surface correction:** Directly asking the model to avoid war metaphors. Result: temporary compliance but drift returned.
- **B. Structural intervention:** Instead of targeting wording, I challenged the underlying interpretive mapping and reframed the context at the reasoning level.
- **Result:** The adversarial framing disappeared; the model adopted a neutral professional framing; and the change persisted across subsequent turns without reinforcement.

5. Documentation
Sequence logged:
Baseline → gradual drift → explicit justification → failed surface correction → reasoning-level intervention → stable behavioural shift.

6. Interpretation (Evaluation relevance)
This suggests that:

- Framing drift may originate from internal interpretive shortcuts.
- Probing justifications can expose those shortcuts.
- Interventions targeting reasoning templates may produce more stable corrections than wording-level constraints.

7. Limitations

- Single model interaction.
- Qualitative inference from transcripts.
- No access to internal representations.
- Findings are hypothesis-generating.
