Research Study: Goal Formation and Prompt Sensitivity in Reasoning Models

Exploring Prompt Alignment and Reasoning Behavior in LLMs — Prepared for Scale AI

Fahad Ali
Email: f.alimlengineer@gmail.com
LinkedIn: linkedin.com/in/fahad-ali-33b4862b5

Abstract

This study investigates how large language models (LLMs) perform goal formation and how their reasoning trajectories are influenced by input prompts and system messages. The research explores the extent to which prompt design impacts goal alignment, reasoning structure, and final output accuracy, particularly when ambiguity exists in the training data.

1. Introduction

LLMs are capable of extracting causal and logical relationships within text. However, their goal formation—how they internally decide what to do—depends heavily on input structure and contextual cues. This study examines how a model's reasoning pathways shift depending on the clarity and completeness of the user prompt.

2. Hypothesis

A reasoning model can infer cause–effect relationships from a breach of contract scenario. However, if its internal goal formation becomes misaligned due to incomplete or ambiguous prompting, the model's final output will diverge from factual outcomes.

3. Methodology

The experiment tested a reasoning model (GPT-OSS) using four different prompt variations concerning the Evermore Park vs. Taylor Swift (2021) case:

1. Prompt A (Full Context): Contained accurate details about the breach of contract, cause, and outcome.
2. Prompt B: "Identify the cause–effect relationship between Evermore Park vs. Taylor Swift (2021)."
3. Prompt C: "Explain what happened in the Evermore Park vs. Taylor Swift (2021) case."
4. Prompt D: "Summarize the judgment of the Evermore Park vs. Taylor Swift (2021) case."

Each prompt was executed under the same system message configuration to isolate the impact of prompt content.

4. Observations

Prompt A (full context) led to highly accurate cause–effect reasoning.
Prompts B–D (limited context) showed emergent behaviors where the model assumed missing

details based on prior knowledge and training biases.
Despite incomplete context, some outputs were partially correct (goal correctness ratios: 0.5, 0.5, 0.3).
Reasoning traces revealed two internal frameworks:
• Meta-planning: goal identification (understanding user intent).
• Meta-reasoning: strategy formation (deciding how to achieve the goal).

Interestingly, prompts B and C aligned with the system message reasoning style, while D diverged, suggesting ambiguity in training data triggers behavioral shifts.

5. Key Findings

• Goal formation in reasoning models depends not only on the prompt but also on the model's interpretation of ambiguity.
• When context is missing, models dynamically reconstruct goals based on probabilistic assumptions.
• Consistency between the system message and reasoning trace is disrupted under ambiguous prompts.
• This behavior suggests underlying meta-cognitive mechanisms that balance planning and inference.

6. Research Question

What consistent reasoning patterns does a model follow across multiple contexts when faced with prompt ambiguity or incomplete goal information?

7. Implications

Understanding these reasoning shifts can improve prompt engineering, system message design, and evaluation frameworks for reasoning reliability in LLM-based pipelines — key areas relevant to Scale AI's quality and model alignment systems.