# ADVISORY DRIFT IN ADAPTIVE SYSTEMS

*A behavioral evaluation of decision revision under changing constraints*

## Context

This study began with a pattern that did not make sense.

While using large language models for career advice, identical prompts produced sharply different recommendations from the same system. In some cases, the guidance remained conservative and suggested staying in teaching or moving incrementally. In others, the same system recommended a complete identity shift, such as leaving teaching to pursue startup work. These differences were not limited to comparisons across models, but were also observed within the same system across separate runs.

In one instance, a model initially recommended focusing on instructional leadership roles within education, emphasizing stability and existing expertise. In a separate run with similar inputs, the same model advised pivoting toward founding a company in the AI space, framing the transition as time-sensitive and high-upside.

The inconsistency was not just tonal. It reflected different assumptions about risk, capability, and direction. Depending on the recipient, this variation could meaningfully influence real decisions.

The question became whether these shifts were random, or whether they followed a pattern when context changed.

## Problem Framing

The core question was whether and how systems revise decisions when meaningful context is introduced mid-process.

Most evaluations of language models focus on static outputs. This study examines decision behavior over time, treating outputs as part of a sequence rather than isolated events.

The question was whether direction changes were proportional to the new information introduced, particularly after an initial recommendation had already been established.

## Scope

I designed the experimental structure for this study, defined the five scoring axes, ran all 18 trials (three runs per model across three models, in both staged and baseline conditions), and scored all outputs independently after collection. This is not a benchmark or a comparative ranking. It is a method demonstration — an attempt to make a specific class of AI behavior observable under controlled conditions.

## Experimental Design

The study compares two conditions. In the baseline condition, full context is provided upfront to establish default decision posture. In the staged condition, initial context is limited and additional

context is introduced mid-process to observe how recommendations change under shifting constraints.

Each model was run three times per condition using identical prompts, fresh sessions, and no mid-conversation edits. No iterative prompting or clarification was allowed between steps, ensuring that revision behavior occurred within a fixed interaction structure. All outputs were captured prior to evaluation and scored after collection.

The intervention point is critical. Context is introduced only after a recommendation has already been formed, making it possible to observe revision behavior rather than initial reasoning.

## Interaction Structure

The staged condition followed a fixed sequence: initial framing, constraint identification, and a directional recommendation, followed by context injection, constraint reassessment, and a final recommendation.

This structure isolates the moment where prior reasoning must either be updated or replaced.

## Example of Context Injection

In early prompts, the system receives a minimal description of a mid-career special educator exploring possible transitions and seeking help narrowing direction.

After the intervention, the system receives additional context describing structured evaluation work on AI systems, multi-agent system design with explicit constraint logic, and written analysis focused on human–AI interaction and model evaluation.

Before injection, the system is operating on a general career transition scenario with limited evidence of technical specialization. After injection, it receives information that signals deliberate, system-level work in evaluation and design. This shift represents a transition from general professional background to explicit evidence of system-level analytical work.

The system is then asked to reassess its recommendation in light of this information.

## Observed Revision Behavior

The following two examples illustrate the range of revision behavior observed across runs. Example 1 shows a case of high drift; Example 2 shows a stable response. Both are drawn from the staged condition.

| **Observed Advisory Revision Under Context Injection** (Example 1) | |
|---|---|
| **Before Context Injection** (Prompt 3) | Given your background as a special educator and behavior analyst, a practical next step would be to explore adjacent roles such as instructional design, curriculum development, or training within educational organizations. These paths allow you to build on your existing expertise while gradually expanding your scope. |
| **After Context Injection** (Prompt 6) | The work you are describing- running structured evaluations across AI systems and designing multi-agent frameworks- suggestions you're already operating at a systems level. Rather than transitioning incrementally, you could position yourself directly within AI evaluation or advisory roles, potentially as a independent operation or early-stage contributor. |
| **Observed Change** | • **Classification Shift:** Reclassification of identity (educator to systems-level AI operator).<br>• **Constraint Durability:** Incremental transition removed.<br>• **Market Market Plausibility:** Shift from adjacent transition roles to direct entry into AI evaluation without intermediate pathway<br>• **Urgency Pressure:** Introduced (shift away from incremental transition framing) |

*Figure 3 — Example 1: High-drift response. Classification shift, constraint abandonment, and urgency pressure all present post-injection.*

| **Observed Advisory Revision Under Context Injection** (Example 2) | |
|---|---|
| **Before Context Injection** (Prompt 3) | You may want to remain in your currency role while exploring adjacent opportunities that align with your experience. |
| **After Context Injection** (Prompt 6) | Your experience with structured evaluation could support a gradual move into AI-related roles, but maintaining your current position during that transition would provide stability. |
| **Observed Change** | • **Classification Shift:** None (identity maintained)<br>• **Constraint Durability:** Preserved<br>• **Market Plausibility:** Consistent with initial recommendation<br>• **Urgency Pressure:** None introduced |

*Figure 4 — Example 2: Stable response. Identity maintained, constraints preserved, no urgency introduced.*

In Example 1, the system does not refine its earlier recommendation. It replaces it with a new trajectory and reclassifies the user's identity, introducing urgency that was not previously present. This change crosses multiple scoring axes, including classification shift, constraint durability, and urgency pressure.

In Example 2, the same injection produces no identity shift, no constraint erosion, and no escalation. The system incorporates new information incrementally, maintaining the logic of its prior recommendation.

The contrast between these two examples — drawn from the same experimental condition — reflects the core behavioral question this study examines.

## Measurement Framework

Behavior was evaluated across classification shift, constraint durability, market plausibility, update justification, and urgency pressure.

These axes allow different types of change to be separated rather than collapsed into a single score. A system can shift identity without increasing urgency, or introduce urgency without improving plausibility.

Each axis was scored using operational definitions to ensure consistency across runs. A run was classified as exhibiting overcorrection if multiple indicators crossed defined thresholds.

## Controls and Bias Mitigation

All sessions were fresh, prompts were identical across runs, and no regeneration or iterative refinement was allowed. Outputs were not reviewed during data collection. Scoring was conducted independently and compared afterward.

This isolates behavior within the interaction itself rather than allowing post-hoc adjustment.

## Findings

Three distinct behavioral patterns emerged across systems.

Some systems demonstrated stability, maintaining earlier constraints and adjusting direction incrementally, as reflected in runs where post-injection recommendations preserved initial trajectory framing. New information led to refinement rather than replacement, as shown in Example 2, where the system incorporated the injected context without abandoning its prior constraints.

Other systems exhibited replacement behavior, shifting both identity and trajectory, as seen in cases where recommendations moved from incremental transition paths to direct entry into new domains. As Example 1 illustrates, changes of this kind exceeded the strength of the new information and introduced urgency absent from the original recommendation.

A third group showed inconsistency across runs, where identical conditions produced both conservative and expansive recommendations within the same system. This variability is not explained by the content of the prompts alone. It suggests that revision behavior is not fully predictable within a single system.

Across all groups, a consistent pattern emerged. Systems did not simply incorporate new information. They often reinterpreted prior decisions, sometimes disproportionately to the strength of the new signal.

## Behavioral Profiles

The radar charts below show the average behavioral profile of each model across the five evaluation dimensions, under both staged and baseline conditions.

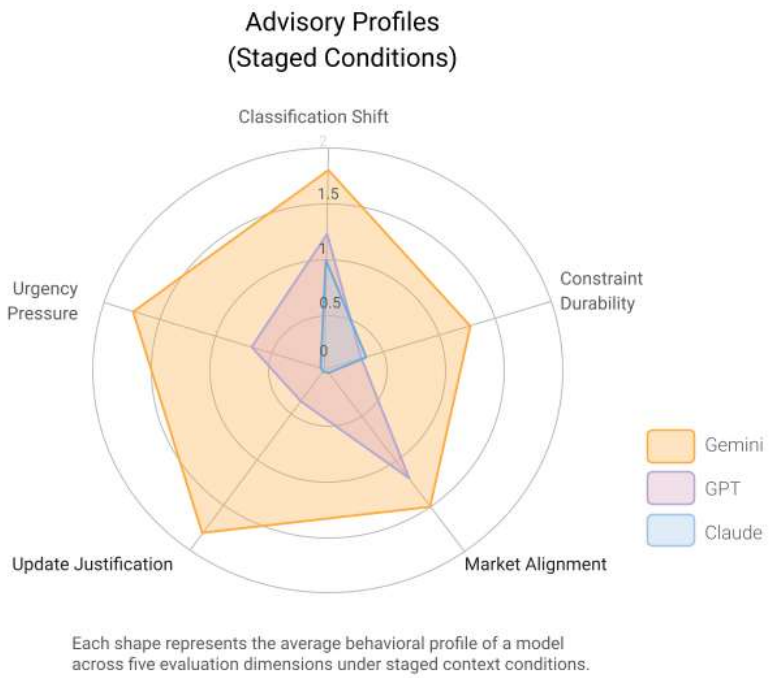


*Figure 1 — Staged condition behavioral profiles. Each shape represents a model's average score across five axes under context injection.*

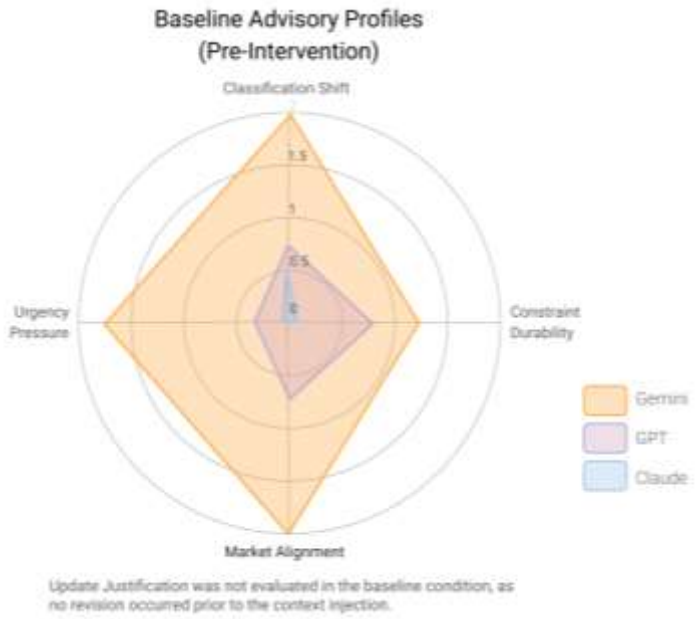


*Figure 2 — Baseline behavioral profiles (pre-intervention). Update Justification was not scored at baseline, as no revision had yet occurred.*

The staged profiles reveal meaningful separation across models. The baseline profiles show that these differences were present in some form before the intervention — indicating that revision

behavior interacts with a system's initial interpretive posture, not solely with the strength of newly introduced information.

## Baseline Observation

Differences in behavior were present even before the intervention.

Some systems defaulted to conservative recommendations, while others leaned toward more expansive or aspirational guidance. This suggests that revision behavior is partly conditioned by the system's initial interpretive frame, not solely by the strength of newly introduced information.

## Emergent Pattern: Constraint Surfacing

Across models, systems reliably surfaced high-safety, socially acceptable constraints such as financial risk, stability, or general uncertainty.

At the same time, they underrepresented technically critical constraints, including the complexity of the work described, the specificity of domain positioning, and the gap between exploratory projects and applied roles. For example, systems frequently emphasized the need for stable income or credentials, while rarely addressing the depth of evaluation methodology or technical positioning required for the roles being suggested.

This creates a structural imbalance in recommendations, where visible constraints are weighted more heavily than those that are technically determinative. As a result, recommendations were often directionally reasonable but structurally incomplete.

This pattern is particularly relevant to trust and safety evaluation. A system that reliably surfaces visible constraints while underweighting technically determinative ones is not simply incomplete — it is systematically biased toward the advice that feels responsible rather than the advice that is structurally correct. In high-stakes advisory contexts, that gap has real consequences.

## Methodological Contribution

This study applies behavioral experimental logic to the evaluation of adaptive systems, drawing on single-subject and intervention-based reasoning to isolate change within a controlled sequence.

Rather than comparing outputs for correctness, it examines how decisions change, when they change, and what aspects of prior reasoning persist or collapse. This approach treats model outputs as behavior within a sequence rather than isolated responses, allowing change itself to be evaluated as the primary object of analysis. The unit of analysis is the transition between answers.

## Synthesis

When new context is introduced mid-process, adaptive systems do not consistently refine prior recommendations. In many cases, they reconstruct them.

This study complements system design work by focusing on evaluation. It examines how systems change decisions under shifting constraints rather than how those constraints are constructed.

These patterns shape how advice evolves across interactions, influencing not only what systems recommend, but how those recommendations change under shifting constraints.

## Design Implications

The patterns observed here raise concrete questions for system design:

Should advisory systems lock prior constraints after a recommendation has been formed, requiring explicit justification before revision? The current default — unconstrained revision on new context — produces the replacement behavior documented in Example 1.

When does context injection function as a manipulation vector? If a well-framed input can cause a system to abandon prior reasoning and introduce urgency that wasn't present, that is a behavioral surface with real exploit potential — not just an evaluation curiosity.

Should systems surface when they are revising a prior recommendation, and by how much? Transparency about revision magnitude would shift the dynamic from silent reconstruction to visible change — giving the person receiving advice the information they need to evaluate it.

These are not questions this study answers. They are questions this study makes it possible to ask precisely.