# ContextEval: Evaluating LLM Agent Context Policies for ML Experiment Design

Hikaru Isayama  Adrian Apsay  Julia Jung  Raghavan Narasimhan  Mentor: Ryan Lingo

hisayama@ucsd.edu adapsay@ucsd.edu jmjung@ucsd.edu naraghavan@ucsd.edu ryan_lingo@honda-ri.com

Honda Research Institute US

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

## Motivation & Problem Statement

LLM agents are increasingly used for iterative ML experimentation, but context is typically treated as an implementation detail rather than a controlled variable. More context is not always beneficial — we therefore treat context visibility as a first-class experimental variable.

## Research Question

*"How does context visibility affect the efficiency and behavior of LLM agents performing iterative ML experimentation?"*

## Methodology & Experimental Setup

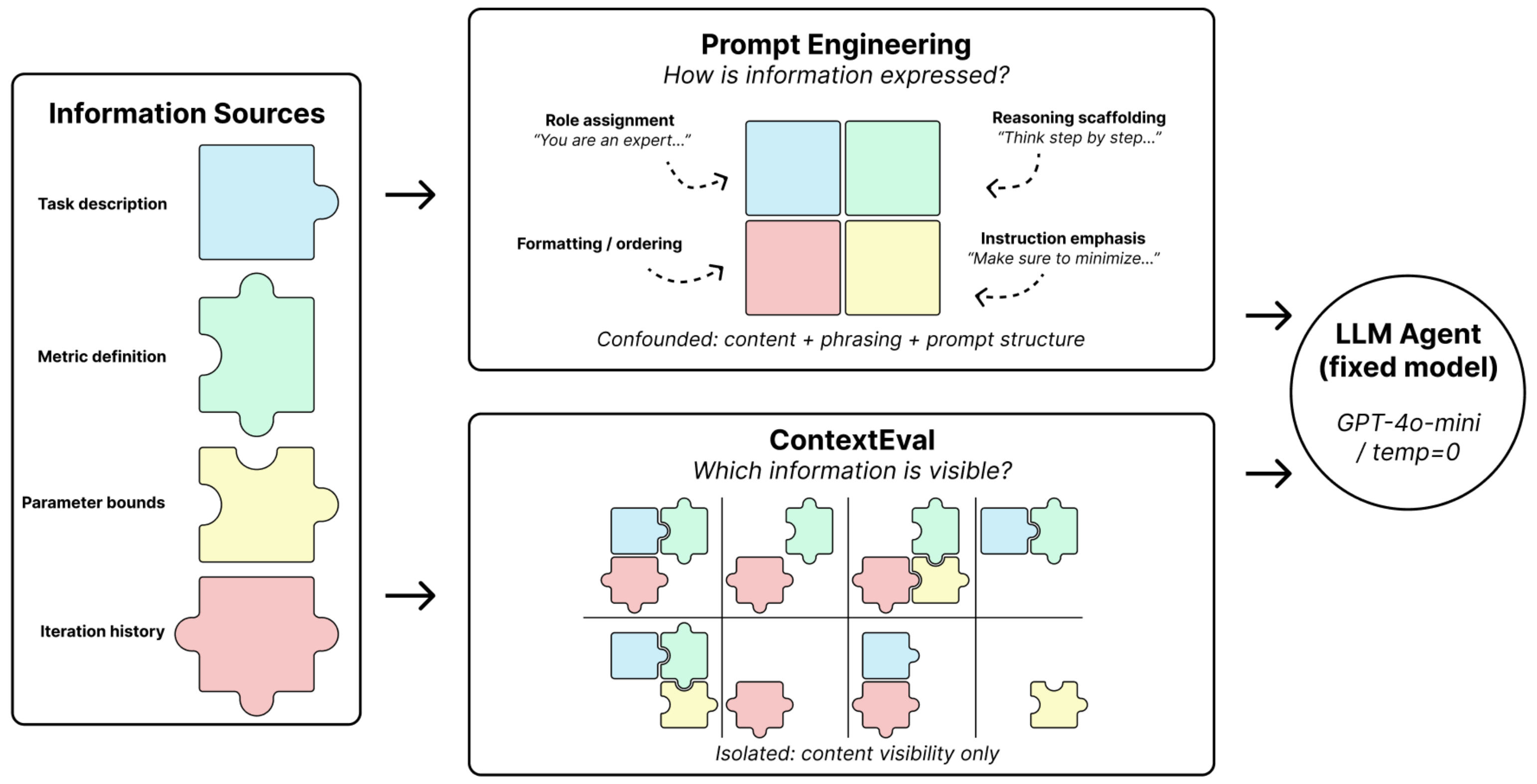


Figure 1. ContextEval isolates information visibility as the sole experimental variable, in contrast to prompt engineering approaches where content and phrasing are confounded.

We fix the model, environment, and prompt template — varying only which information axes are visible to the agent: 16 context policies across 4 benchmarks and 3 initialization strata (576 runs, GPT-4o-mini, $T = 10$, 3 seeds).

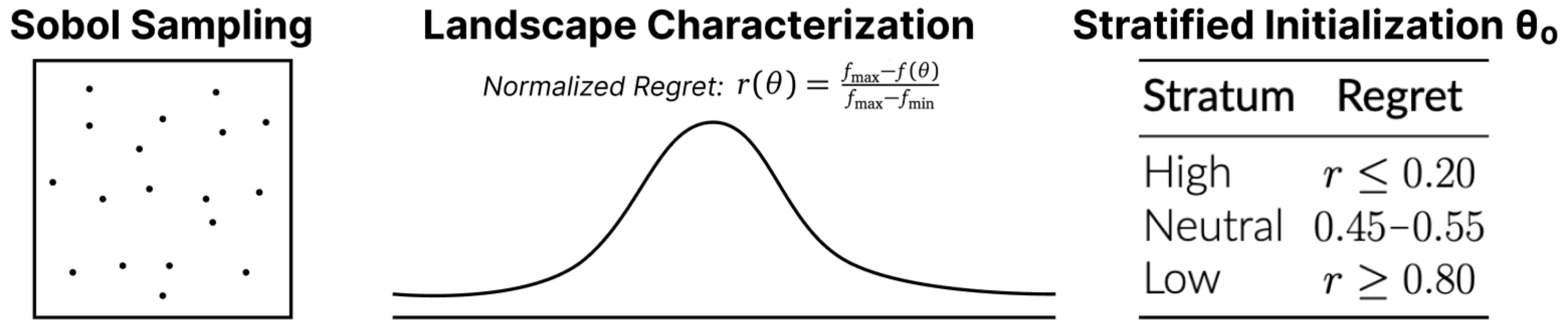
Context Policy Axes

| Axis | Exposed |
|---|---|
| `show_task` | Task context |
| `show_metric` | Metric def. |
| `show_bounds` | Bounds $\Omega$ |
| `feedback_depth` | History $d \in \{1, 5\}$ |

Benchmarks

| Name | Task | Metric |
|---|---|---|
| NOMAD | Materials reg. | RMSLE ↓ |
| Jigsaw | Text classif. | AUC ↑ |
| Forest | Tabular classif. | Acc. ↑ |
| Housing | Housing reg. | RMSE ↓ |

Stratified Initialization via Sobol Sampling



Sobol Sampling
Landscape Characterization
Normalized Regret: $r(\theta) = \frac{f_{\max} - f(\theta)}{f_{\max} - f_{\min}}$
Stratified Initialization θ₀

| Stratum | Regret |
|---|---|
| High | $r \leq 0.20$ |
| Neutral | $0.45–0.55$ |
| Low | $r \geq 0.80$ |

Figure 2. 256 quasi-random Sobol configurations characterize the objective landscape. Normalized regret $r(\theta) \in [0, 1]$ computed via min-max scaling for cross-benchmark comparison.

## Results & Analysis

Feedback depth is the dominant context factor across all benchmarks — and its effect depends critically on where optimization begins.

| Benchmark | fd | task | metric | bounds |
|---|---|---|---|---|
| NOMAD | +0.019 | −0.018 | +0.004 | −0.029 |
| Jigsaw | +0.263 | −0.018 | −0.015 | −0.023 |
| Housing | +0.103 | −0.037 | +0.032 | +0.002 |
| Forest | +0.019 | −0.014 | −0.002 | +0.010 |

Table 1. Marginal effect of each context axis on final regret. Positive = higher regret (worse). Feedback depth is the dominant factor; task descriptions consistently help.

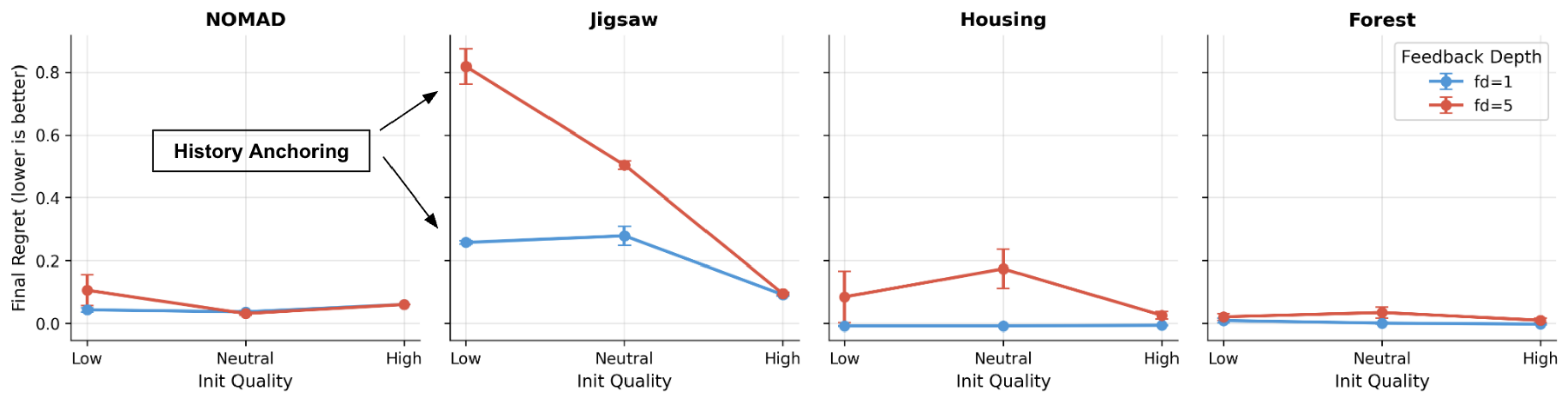


Figure 3. **History Anchoring:** deeper feedback (fd=5, red) substantially increases final regret for low and neutral initializations, but the gap collapses at high-quality starts.

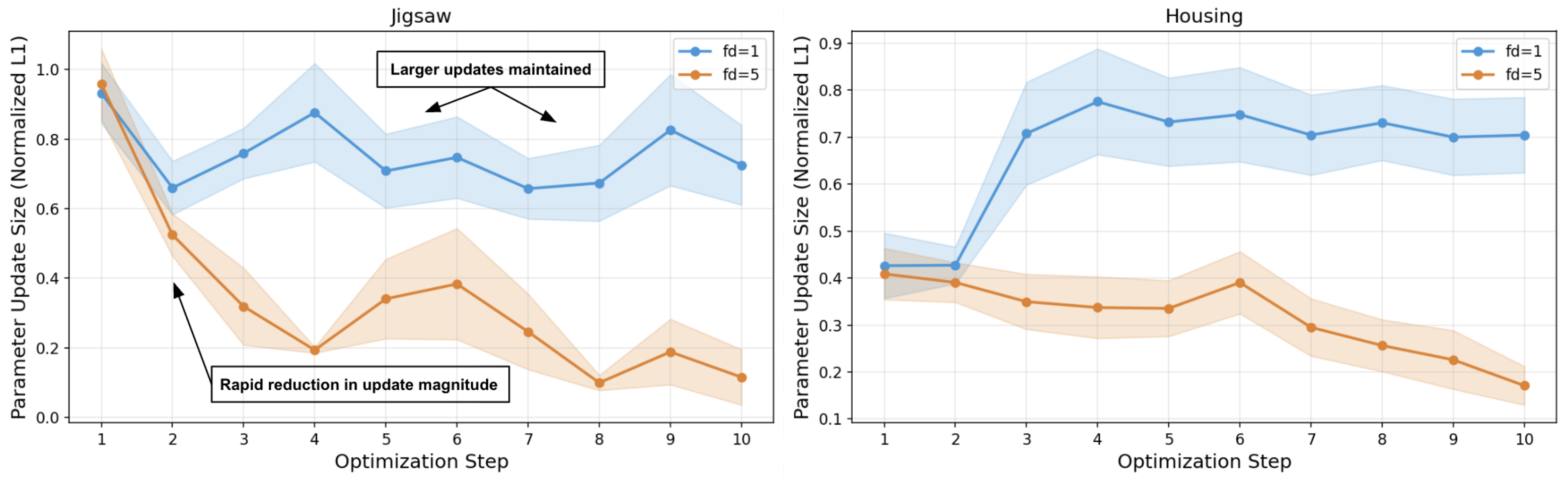


Figure 4. **Mechanism:** fd=5 produces consistently smaller parameter updates — the agent becomes conservative rather than exploratory when anchored to poor scores.

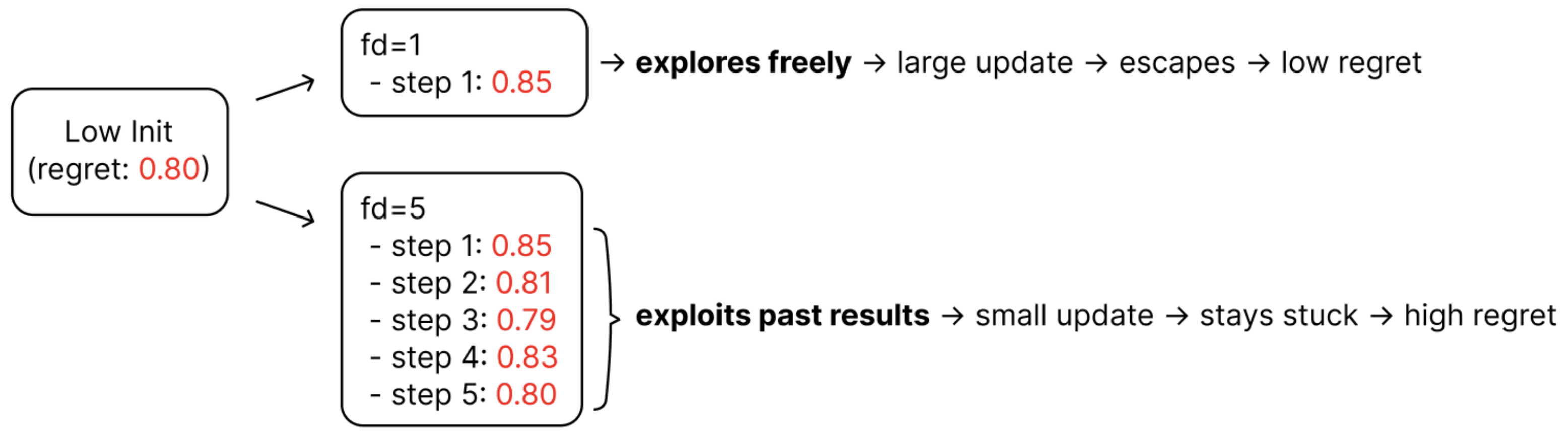
When exposed to a long history of poor scores, the agent shrinks its parameter updates and stops exploring — a history of poor outcomes is **worse than no history at all**.

### History Anchoring

**Outcome.** Jigsaw fd=5 Low init reaches regret **0.818** vs **0.258** for fd=1 — a 3× penalty for deeper history from a poor start.

**Conditional effect.** The gap collapses at High init — fd=5 vs fd=1 difference is just **+0.003** on Jigsaw High init. Deeper history only hurts when the history is poor.

**Mechanism.** fd=5 produces consistently smaller parameter updates throughout the trajectory — the agent becomes conservative rather than exploratory when anchored to poor scores.



Low Init (regret: 0.80)

fd=1
- step 1: 0.85

→ **explores freely** → large update → escapes → low regret

fd=5
- step 1: 0.85
- step 2: 0.81
- step 3: 0.79
- step 4: 0.83
- step 5: 0.80

**exploits past results** → small update → stays stuck → high regret

Figure 5. Given the same low-quality initialization, fd=1 explores freely ($\Delta\theta$ large) and recovers to regret **0.258**, while fd=5 exploits its history of poor scores ($\Delta\theta$ small) and remains stuck at **0.818**. When history is poor, exploitation becomes a trap.

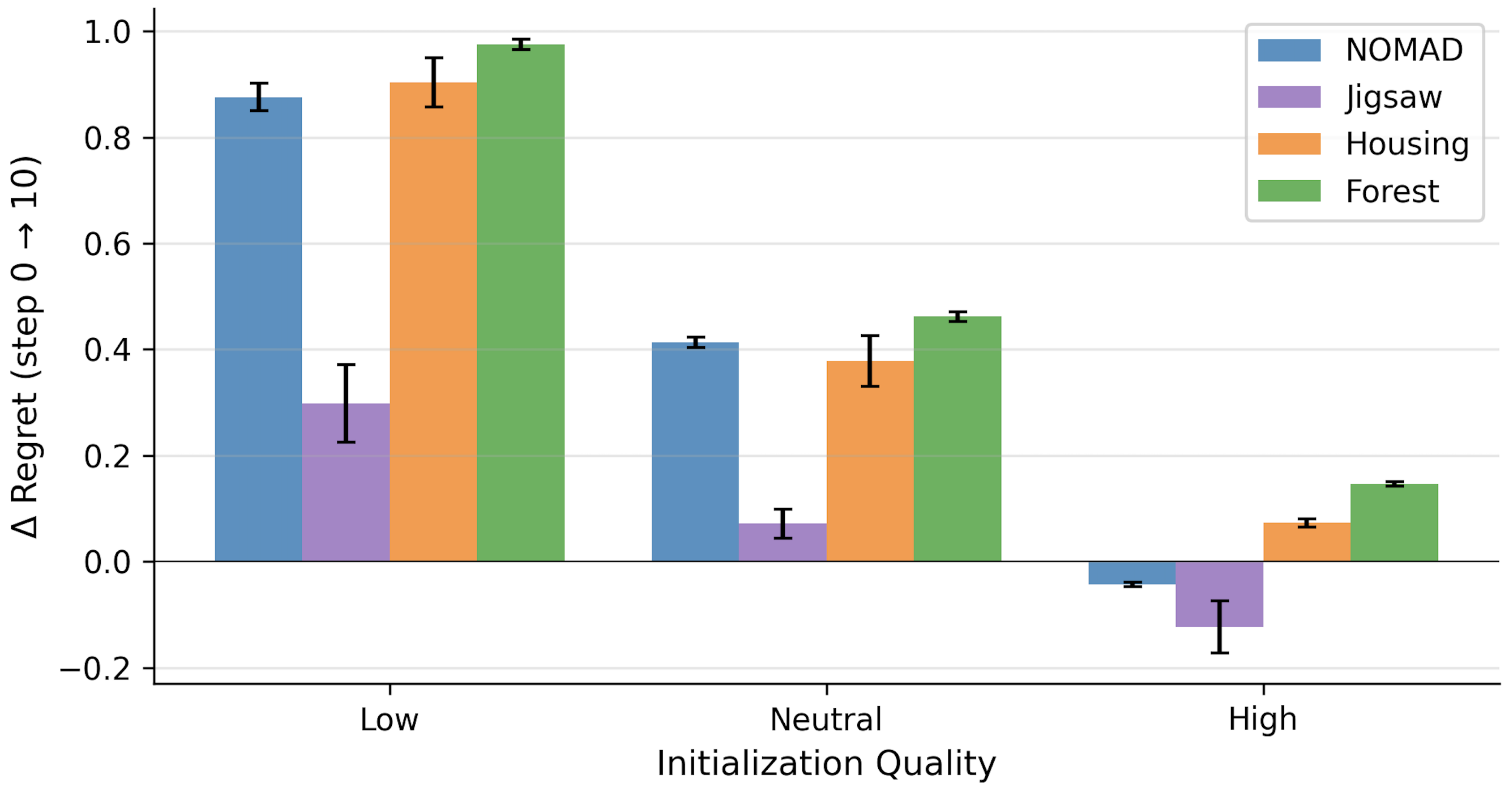


Figure 6. **Heuristic Effect:** the LLM optimizer rapidly repairs poor configurations but provides diminishing returns — and occasionally degrades — near strong ones.

The agent functions as a **heuristic** rather than a systematic optimizer — effective at escaping poor configurations but limited in its ability to refine strong ones.

### Corrective Heuristic

**Recovery.** Low-init runs improve by up to **0.98** regret units within 10 steps across benchmarks.

**Diminishing returns.** High-init runs show near-zero gains — initialization explains up to **49.7%** of variance (Jigsaw $\eta^2 = 0.497$).

## Limitations & Future Work

**Limitations.** Single model (GPT-4o-mini), two feedback depth values only ($d \in \{1, 5\}$), and tabular benchmarks only — generalization to larger models and deep learning pipelines remains untested.

**Future Work.** Evaluate reasoning-specialized models, adaptive context policies, and extend the action space to model selection.

Github: https://github.com/juliamsjung/context-eval