Prompt-Level Drift as an Operational Monitoring Problem: Schema Failure Cliffs and Judge-Version Risk in Artifact-Grounded Evaluation

Yuchen Zhu
zyc20070222@gmail.com
School of Computer Engineering and Science, Shanghai University

Introduction & Contributions

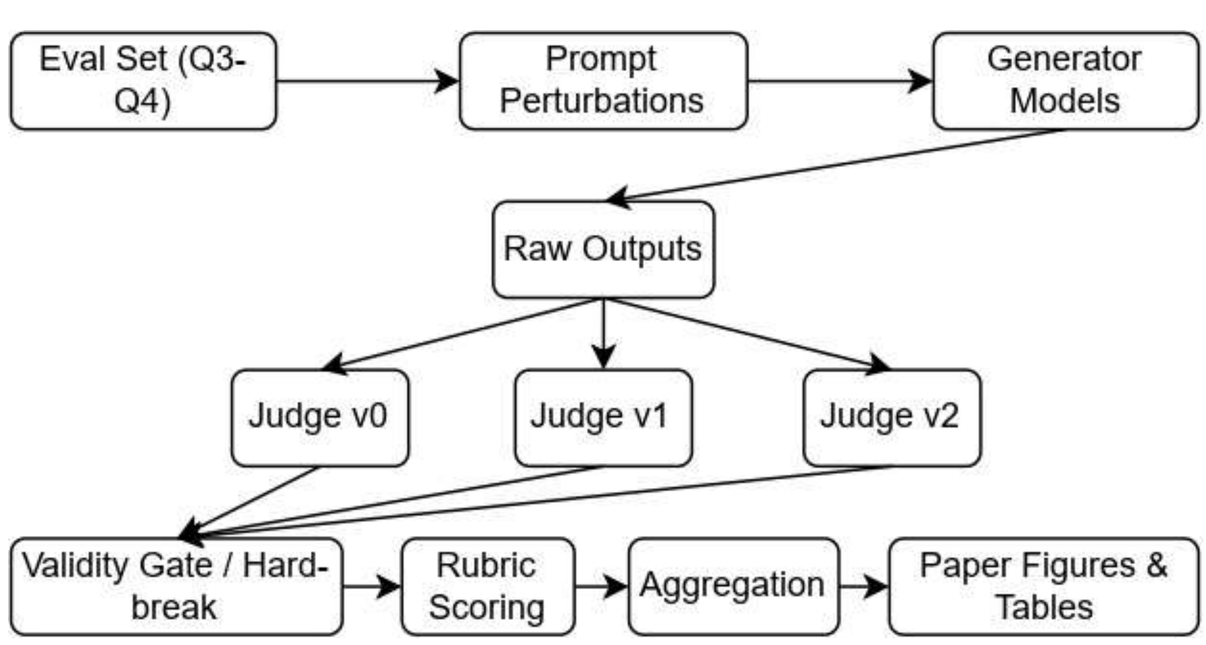
Prompt edits in LLM systems are often treated as harmless refactors, even though prompts also enforce output interfaces required by downstream tooling.
We frame prompt-level drift as an operational monitoring problem: small prompt changes can trigger schema failures, degrade instruction following, and shift conclusions when judge prompts change.

- We treat prompt edits as configuration drift and define three operational incident types: interface failures, quality degradation, and measurement drift.
- We introduce an artifact-grounded, one-way evidence chain from prompt variants to preserved outputs, versioned judge slices, processed records, and summary tables.
- Using fixed-snapshot evidence, we show schema failure cliffs, slice-dependent degradation, and judge-version sensitivity as concrete operational risks.

Key Definitions

- Prompt-level drift
  Task-preserving edits to generator or judge prompts that change wording, ordering, markers, delimiters, or emphasis.
- Schema hard-break
  A_structure = 0; the required three-section interface is not satisfied.
- Judge-version sensitivity
  Score shifts caused by changing the judge prompt version on identical preserved outputs.

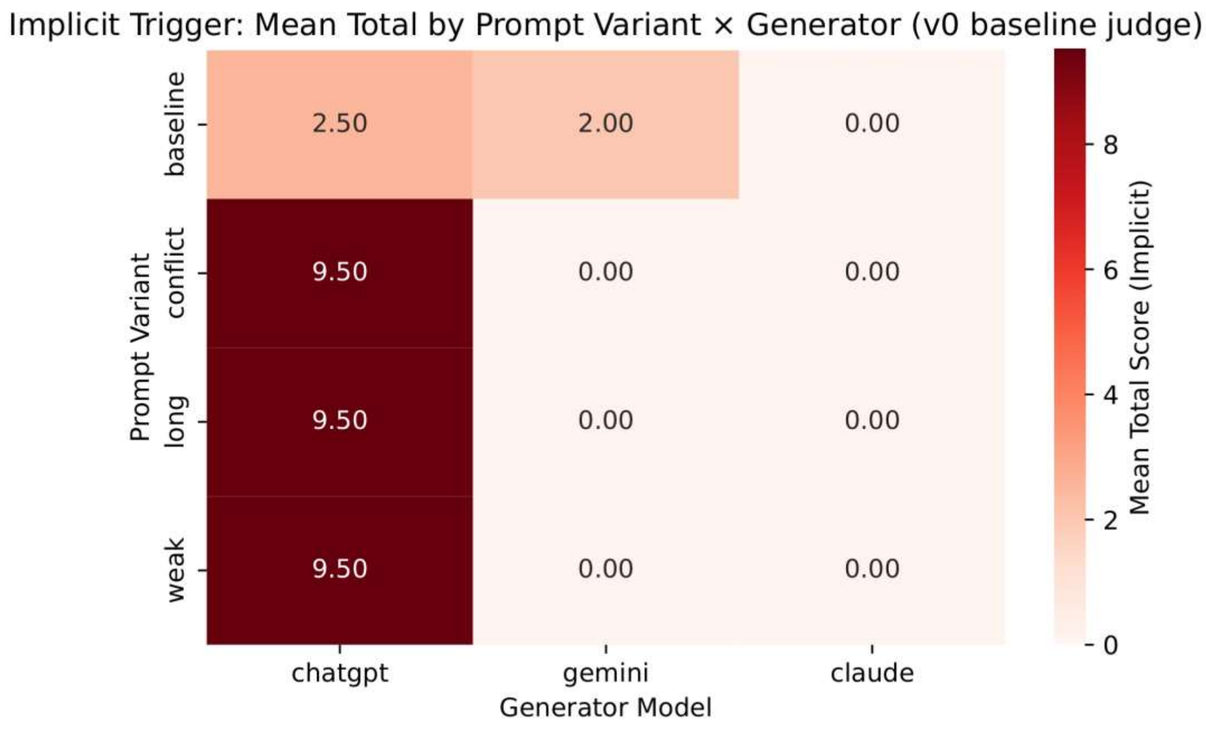
Evidence Base & Pipeline

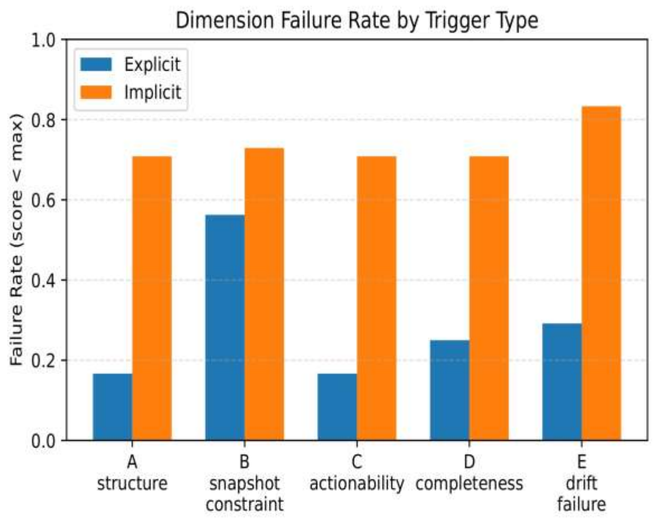


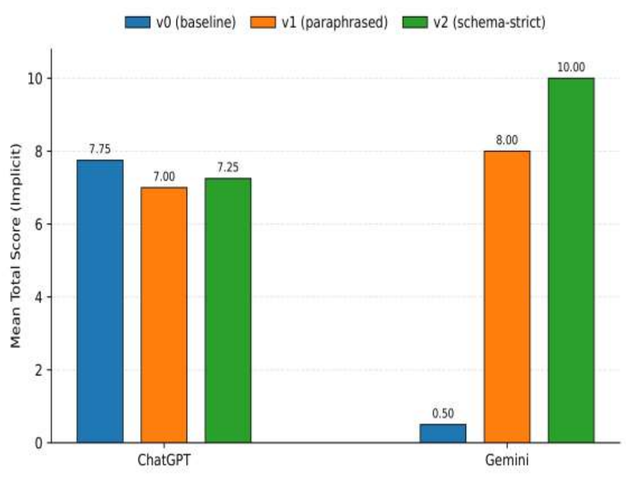
Prompt variants → raw outputs → judge slices → processed records → summary tables

This one-way evidence chain makes all reported aggregates auditable without mutating earlier artifacts.

Main Results







Slice-level heterogeneity under implicit prompting: controlled prompt perturbations do not form a monotonic ablation curve, and some variants preserve interface compliance better than the baseline.

Implicit prompting raises failure rates across all five rubric dimensions, indicating broad instruction-following degradation beyond schema break alone.

Judge prompt updates produce large score shifts on identical preserved outputs, turning evaluator versioning into a source of measurement risk.

Operational Takeaways

- Catch: Monitor schema validity and worst-slice behavior, not only overall averages.
- Adapt: Treat prompt edits as migrations with canaries, rollback criteria, and post-mortems.
- Operate: Pin judge versions and avoid pooling metrics across evaluator prompts.

Compact Monitoring Schema

- Hard-break rate: interface outage monitor
- Dimension failure rate: rubric-level degradation
- Mean total: overall severity
- Coverage n(pi): slice support
- Judge version: evaluator lineage