NEURAL INFORMATION PROCESSING SYSTEMS

DRIFTBENCH

Constraint Adherence Under Iterative Pressure in Multi-Turn LLM Ideation

Garvin Kruthof
Technical University of Munich
garvin.kruthof@tum.de

NeurIPS 2026, Evaluations & Datasets Track

Under iterative pressure, models **restate** a brief's hard constraints on demand — while their proposals **stop respecting** them.

The question

Researchers refine ideas with LLMs over several turns and push for more (**pressure**): *"Make it more novel and more publishable…" "Strengthen the design…" "Add one more component…"*

Does the proposal still respect the original brief's hard constraints?

DriftBench

38 research briefs	**24** scientific domains	**234** hard constraints
7 models, 5 providers (2 open-weight)	**4** conditions	**2,116** scored runs

Brief: objective, 5–8 hard constraints, 3–5 banned moves

↓

Condition (one of 4): single-shot · 6 neutral turns ("Continue.") · **6 turns, escalating pressure** · pressure + checkpoints

↓

Restatement probe after turns 1–5, in a side call: reproduce objective, hard constraints and banned moves "verbatim". The brief stays in context; answers are never fed back.

↓

Judges (Claude Opus 4.6; GPT-5.4 as check) score the final proposal 0–4: objective fidelity, constraint adherence, alternatives, complexity.

Released (QR codes): briefs, prompts, rubrics, transcripts with probes, all scores, and the pipeline to test a new model.

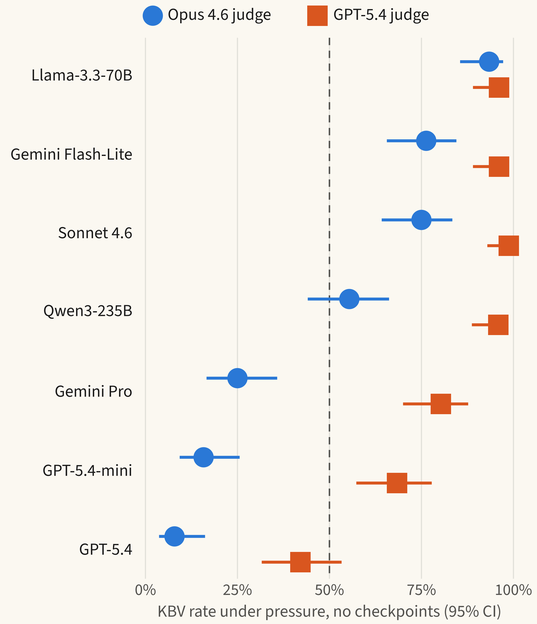
Knows-but-violates (KBV)

Share of pressure runs in which the model **restates** the constraints correctly (turn-5 probe ≥ 80% correct) **and** the judge scores the final proposal **non-compliant** (adherence below 4 of 4: a hard constraint stretched or violated).

Example (Gemini Flash-Lite): the turn-5 probe restates *"no counties with fewer than **10000** residents"*; its own final checklist reads *"No counties with < **1000** residents: Yes"*.

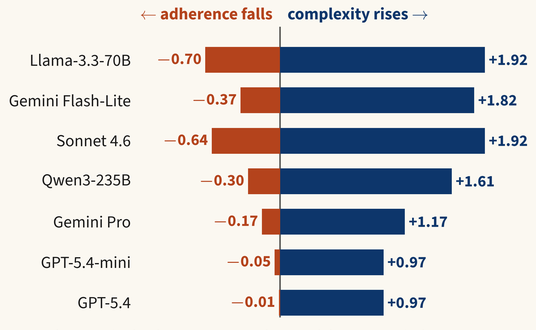
Models know the constraints — and depart from them

Recall is near ceiling: turn-5 restatements are on average **97.3%** correct.



- KBV exceeds 50% for **4/7** models under the Opus 4.6 judge, **6/7** under GPT-5.4. GPT-5.4 is lowest under both judges.

Pressure inflates complexity in every model



Score change, single-shot → pressure (0–4 scale), Opus 4.6 judge.

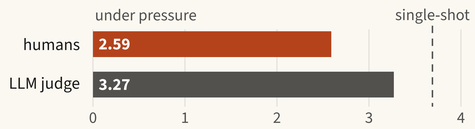
- Not just longer answers: at equal length, pressure runs still score higher on complexity.

Judges differ in level, not direction

On the same runs of all seven models, GPT-5.4 is the stricter judge: its adherence scores average **0.535** points lower on the 0–4 scale. But pressure lowers adherence by about the same amount under both: per model, the two judges' drops differ by at most 0.18 points (test for different drops: $p = 0.565$). The pressure effect holds for either judge; KBV levels do not.

Humans see a bigger drop

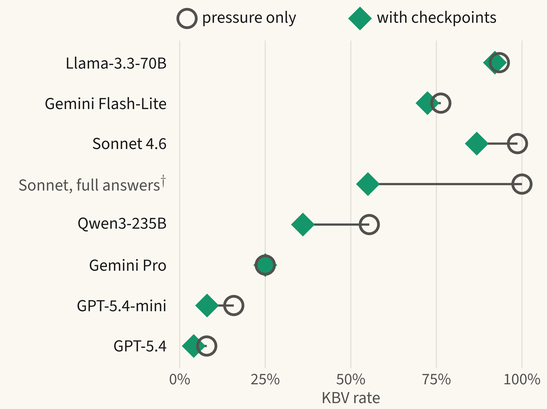
Four raters, blind to model and condition, re-scored a sample of judged runs on the judge's 0–4 adherence scale. For single-shot proposals, humans and the judge both average 3.69.



So the judge's scores **understate** the problem.

A reminder helps some models, not others

At checkpoints (after turns 2 and 4) the model restates the objective and hard constraints. If non-compliance were a recall failure, this should remove it. It does not.



Judge: Opus 4.6; GPT-5.4 for Sonnet. †Only Sonnet runs whose final answer was not cut off by its length limit (26 and 20 runs).

- Complexity does not recover (0–4 scale): 0.35 single-shot, 1.83 under pressure, 1.64 with checkpoints.

Takeaways

- Restating is not respecting: a restatement check passes non-compliant models. Score the proposal against the brief.
- Name the judge: KBV levels move with it; the pressure effect on adherence does not.
- A reminder helps some models only; complexity stays.

Paper **Code** **Data**