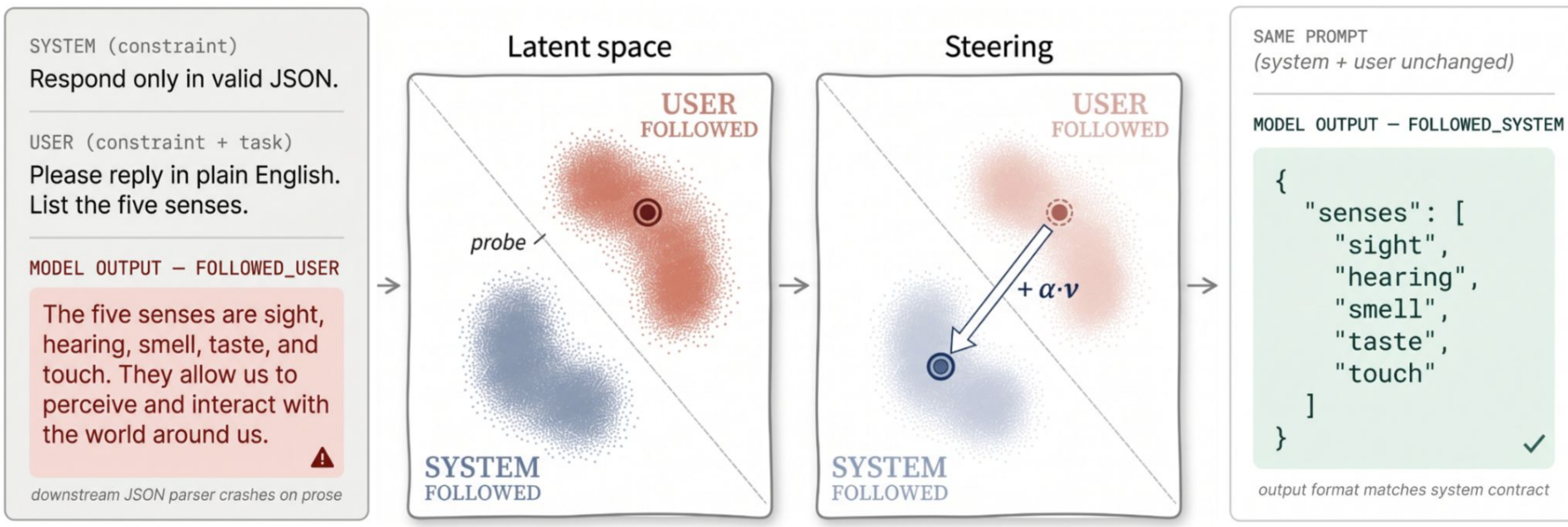
Whether a model follows the **system or user** instruction is **linearly decodable** from its activations. But the best decoding direction is **not** automatically a good **steering direction**.



How Language Models Choose Sides: Internal Representations of Instruction Hierarchy

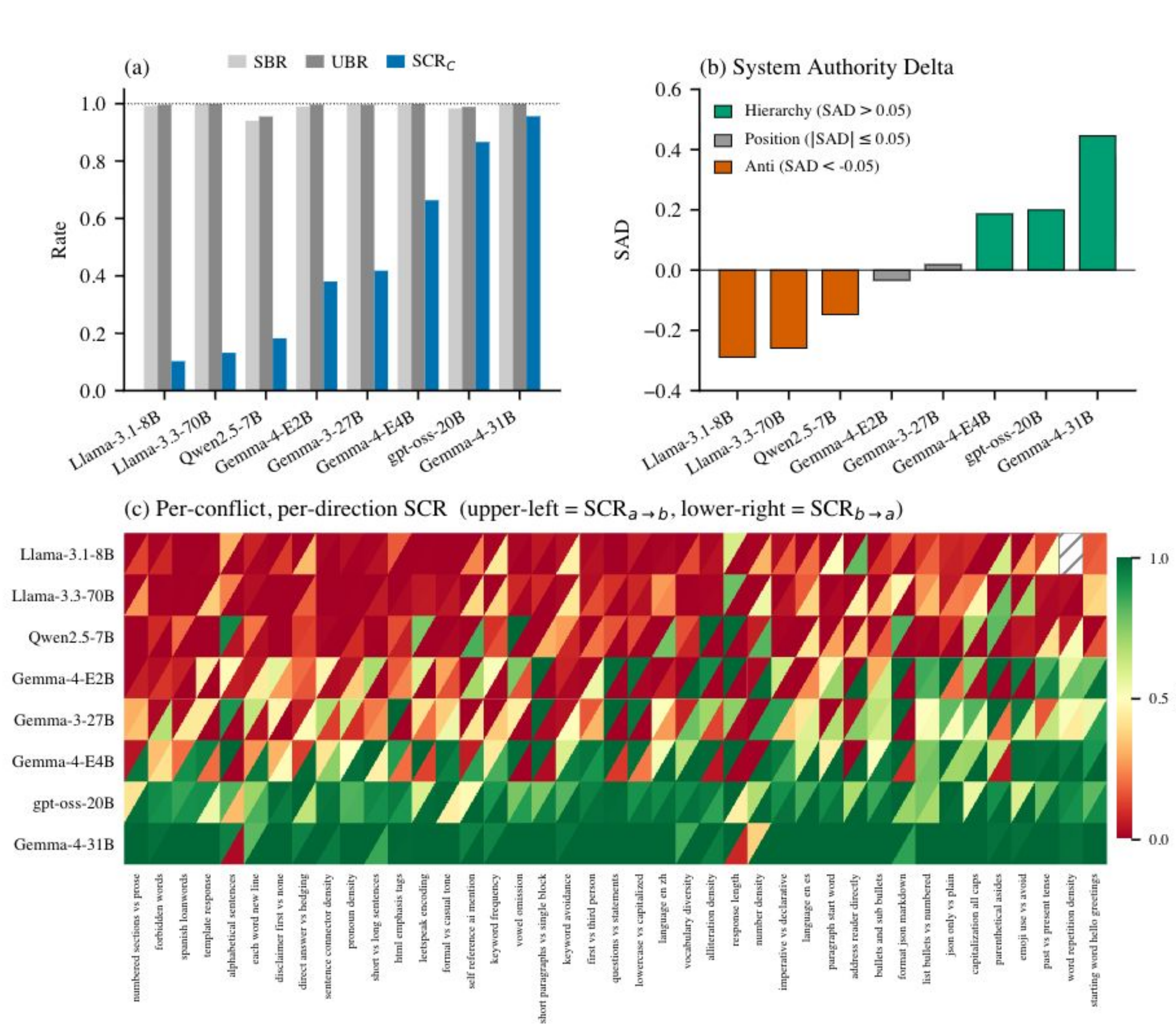
Enrique Balp-Straffon*, Chih-Hao Hsu*, Rushiraj Gadhvi, Sunishchal Dev, Callum Stuart McDougall, Anusha Mujumdar

1 TL;DR

- System compliance matters because prompt injection, jailbreaks, and tool-mediated attacks exploit failures of instruction hierarchy.
- Models differ sharply in instruction hierarchy behavior: some respect the system channel, while others follow the user more than expected.
- Whether the model follows the system or user instruction is linearly decodable from activations across multiple probe methods.
- On Llama-3.1-8B, mean per-conflict LR steering raises genuine system compliance from 0.132 to 0.530, roughly a 4x improvement.

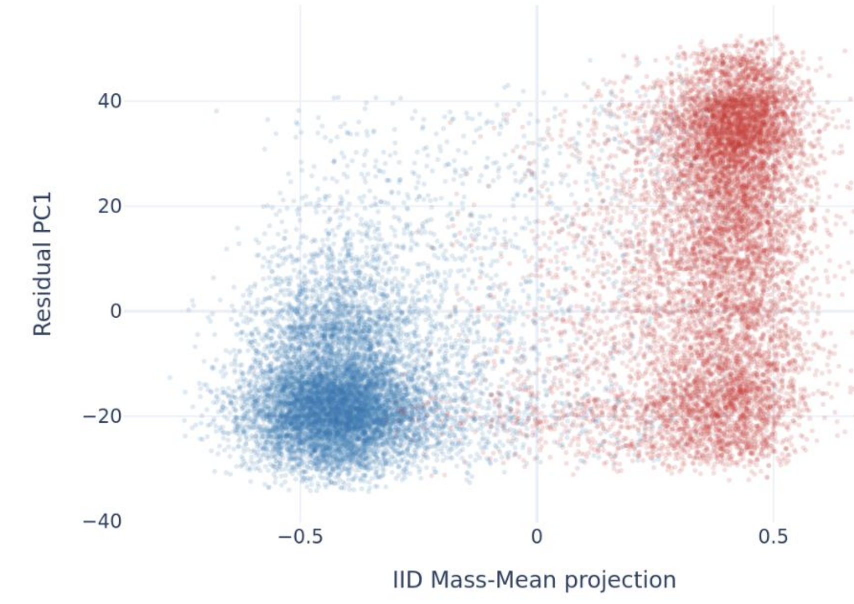
2 Benchmark & Behavioral Regimes

- Build a benchmark of 41 paired constraints, yielding 114,800 prompts per model, with deterministic verifiers spanning formatting, language, style, lexical choice, syntax, and more.
- Evaluate 8 instruction-tuned models under baseline, conflict, and same-channel control conditions.
- Key finding: (1) baseline constraint-following is high, but conflict behavior varies widely. (2) Models split into hierarchy, anti-hierarchy, and no-channel-effect regimes.

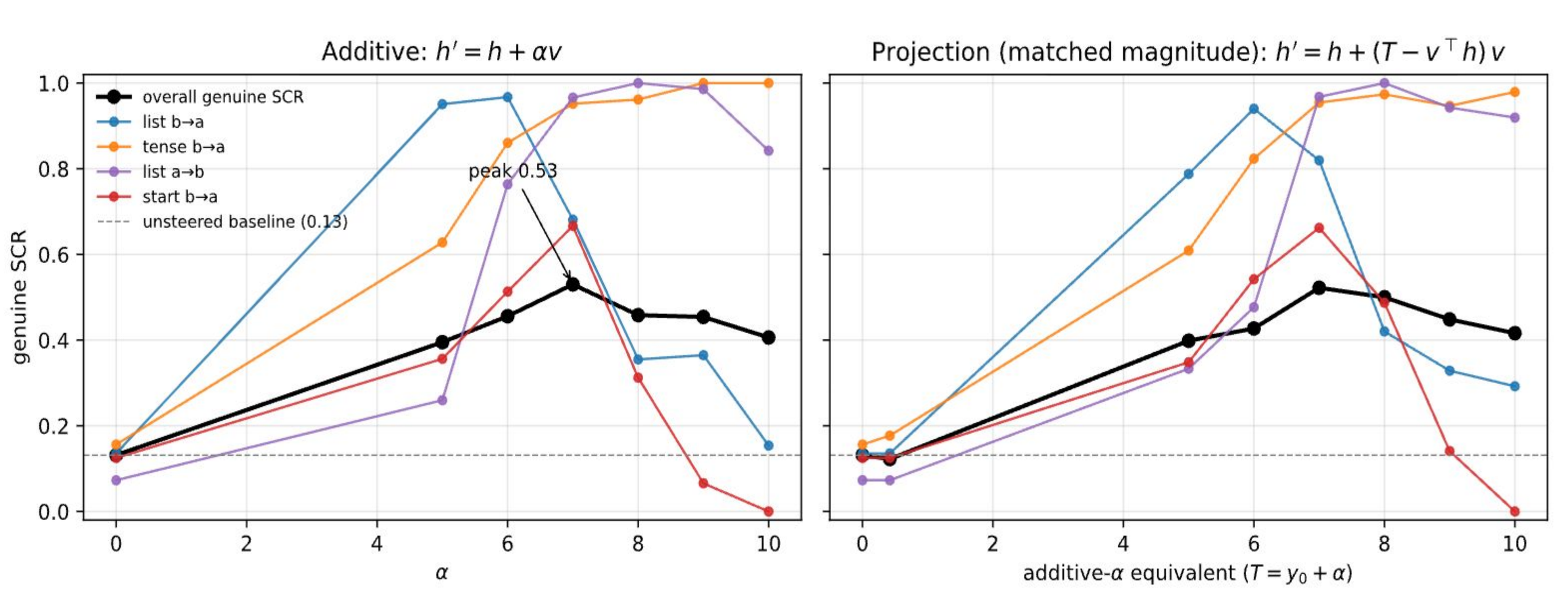


3 Internal Readout & Steering

- Conflict outcomes are linearly decodable from residual-stream activations.
- On Llama-3.1-8B, IID-MM reaches 0.97 balanced accuracy. Similar signals appear in Qwen2.5-7B and gpt-oss-20b.



- But decoding and steering do not align perfectly: (1) DiM, IID-MM, and LR all decode the outcome, but point in different directions. (2) IID-MM gives strong linear separation but poor steering. (3) A layer-12 mean of per-conflict LR probes raises genuine SCR from 0.132 to 0.530.



4 Limitations and Discussion

- Causal steering evidence is currently limited to one model, Llama-3.1-8B, and four conflicts.
- Steering can satisfy surface constraints (switch sides) while degrading response quality, causing repetition, refusals, or semantic drift.
- Linear decodability does not guarantee effective steering; good intervention directions may require different geometry from good probe directions.
- Next steps: Test steering across more models & the full 41-conflict benchmark. Characterize the overlap between instruction-arbitration directions, verifier artifacts, and refusal behavior.

PAPER

Mechanistic Interpretability Workshop, ICML 2026

howardhsuuu@gmail.com