StateWise: Teaching Language Models to Track and Obey Evolving Constraints in Multi-Turn Dialogue

Anonymous authors
Paper under double-blind review

Abstract

Users of conversational assistants rarely state all requirements up front: they add constraints, amend them, withdraw them, digress, and return. Instruction-tuned large language models (LLMs) degrade in this regime in two ways—active constraints stop being honored as turns accumulate (persistence failure), and withdrawn constraints keep being enacted (relapse)—and the only existing repair we are aware of is an external ledger that re-injects a compiled specification at every turn. To tackle this, we propose StateWise, a self-improving approach that teaches a single LLM to maintain its own revisable constraint state. We formalize evolving-constraint dialogue as edit trajectories over a typed state with tombstones for revoked constraints, so that persistence and relapse are measured by executable checkers and ground-truth states are compiled for free. StateWise first fine-tunes the LLM on two complementary skills, summarizing a dialogue into its constraint state and responding conditioned on that state; it then self-trains on trajectories the model itself generates, keeping only turns that pass an exact round-trip check against the compiled state. This recipe teaches 7–8B models to track the state almost perfectly but, surprisingly, not to obey it: on Llama-3.1-8B every supervised variant roughly doubles relapse, and on Qwen2.5-7B none reduces it. We therefore add a preference stage on self-generated relapse contrasts, in which the rejected response enacts a revoked rule and the chosen response passes the round-trip check. Experiments on two backbones with held-out constraint families and unseen phrasings show that StateWise cuts relapse by 8–22 points relative to supervised self-training, lifts conversation success by 8–19 points over the backbone, and beats an oracle compiled-ledger baseline on relapse with no inference-time scaffolding. We also characterize the recipe's cost on cumulative-constraint and single-turn benchmarks and study weight interpolation as a mitigation.

1 Introduction

Conversational assistants are asked to follow instructions that evolve. A user asks for a summary in bullet points, then wants it under a hundred words, then drops the bullets, digresses to an unrelated question, and returns to the summary expecting the word limit to still hold and the bullets to be gone. Multi-turn instruction following is now benchmarked at scale: Multi-IF (He et al., 2024) extends IFEval (Zhou et al., 2023) to three turns and finds accuracy falling with every turn; SysBench (Qin et al., 2025) measures the instability of system-message adherence; EvolIF (Jia et al., 2026) constructs evolving dialogues with constraint changes and reports deficiencies in failure recovery. A parallel line isolates a specific pathology: when a user withdraws a constraint, the model often keeps enacting it, a failure that ReBIND (Zhu, 2026) names revocation inertia and shows rising from 0.01 to 0.40 for an 8B model as the constraint load grows.

Training-side work has addressed adjacent problems. Reinforcement learning with verifiable rewards improves single-turn multi-constraint following (Pyatkin et al., 2025); label-free variants derive rewards from the instructions themselves (Ren et al., 2026); curriculum

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

RL with abstention rewards mitigates the lost-in-conversation gap for information revealed across turns (Laban et al., 2026; Li et al., 2026). None of these trains the model to hold a revisable constraint state across edits. To our knowledge the only published repair for relapse is inference-time: ReBIND keeps an external ledger with tombstones and compiles the net specification into every request, at roughly 1.5× delivery cost and only for code.

We propose StateWise, a self-improving approach that teaches a single LLM to maintain its own revisable constraint state. Our design follows a recipe proven effective for label-free domain adaptation: SimRAG (Xu et al., 2025) fine-tunes one LLM on complementary skills, uses the same model to synthesize training data, and keeps only examples that survive a round-trip consistency check. StateWise transposes this recipe to evolving-constraint dialogue. The complementary skills are constraint-state summarization and state-conditioned response generation; the unlabeled resource is single-turn verifiable constraints, which we compose into scripted add/amend/revoke trajectories; and the round-trip check is exact equality between the model's emitted state and the compiled ground truth, together with executable checkers and a relapse guard. Because the state is compiled from the edit history, every training signal and metric in this paper is verifiable without human annotation.

The recipe teaches the state and, surprisingly, fails to teach obedience to it. On Llama-3.1-8B-Instruct the self-improved model emits the correct state, including the correct tombstones, for 92% of in-distribution turns, and then enacts the revoked rule anyway at more than twice the backbone's rate. On Llama the same doubling of relapse appears in every supervised variant we trained, with or without state supervision, and even single-turn constraint training raises relapse: making a model better at obeying constraints makes withdrawn constraints stickier. Rejection-sampling SFT removes relapsed examples from the training set but supplies no signal against relapse. We therefore add a second self-improvement stage that turns the round-trip check into a preference signal: the model's own relapsed generations become rejected responses, its round-trip-clean generations become chosen responses, and a short DPO run on these self-labeled contrasts cuts relapse by 8–22 points, to below the backbone and below an oracle compiled-ledger system, on two backbones, two training seeds and unseen phrasings. Ablations show that the relapse-specific contrast is necessary (preference pairs about active constraints alone raise relapse) and that label-shuffled pairs do nothing, so the gain is the signal, not the extra training.

Our contributions are summarized as follows:

- We formalize evolving-constraint dialogue as edit trajectories over a typed constraint state with tombstones, with executable metrics for persistence and relapse (the latter restricted to constraints whose chance-satisfaction rate is measured to be low), and show with a preregistered inference-only diagnostic that an 8B model fails on both and that prompting it to write its own state makes it worse.
- We propose StateWise, a self-improving two-stage recipe with a verifiable round-trip filter and a relapse-contrastive preference stage, requiring no proprietary model and no human dialogue annotation, and we report the negative intermediate result that motivates the preference stage: tracking a state and obeying it are separable skills, and self-training with a filter teaches only the first.
- Experiments on two backbones, held-out constraint families, unseen phrasings and public benchmarks, all under gates fixed before the runs, show relapse reductions of 8–22 points relative to supervised self-training and conversation-success gains of 8–19 points over the backbone; we also report the recipe's cost on cumulative-constraint and single-turn benchmarks and show that weight interpolation with the backbone removes that cost at the price of the relapse gain, exposing a controllable trade-off.

2 Related Work

Multi-turn instruction following. Multi-IF (He et al., 2024), SysBench (Qin et al., 2025), EvolIF (Jia et al., 2026) and MultiTurnInstruct (Han et al., 2025) document degradation with dialogue depth, topic switches and revised goals; conversational-reliability studies (Laban et al., 2026) show that sharding information over turns lowers accuracy. These are

evaluation resources. We use Multi-IF as a held-out test and add what it lacks: constraints that are amended and withdrawn, with the relapse they cause measured directly.

Verifiable instruction following. IFEval (Zhou et al., 2023) introduced programmatic checkers; IFBench (Pyatkin et al., 2025) added held-out constraint families and showed that RLVR generalizes to them; Ren et al. (2026) removed external supervision with constraint decomposition and constraint-wise pseudo-labels. All train on single-turn prompts. StateWise uses the same style of checkers but trains on multi-turn edit trajectories with explicit state targets; single-turn constraint training is one of our baselines, and we find that it raises persistence and relapse together.

Lost in conversation and dialogue state. RLAAR (Li et al., 2026) adds abstention rewards and a curriculum for information revealed across turns; rolling-memory policies compress fragmented history (Luo et al., 2026); history-cleaned self-distillation removes self-contamination (Zheng et al., 2026). These address information fragmentation, not constraint revision. Classic dialogue state tracking predicts slot values over a fixed ontology for task-oriented systems (Budzianowski et al., 2018); VLK-RL (Zhao et al., 2026) verifies LLM-elicited constraints for a separate RL policy. StateWise's state is an open set of executable constraint predicates with parameters and tombstones, supervised by self-generated data, and lives in the same model that responds.

Revocation inertia. ReBIND (Zhu, 2026) measures per-clause influence in code-generation dialogues and repairs relapse with an external compiled ledger via the API alone. We adopt its tombstone semantics for ground truth, train the model rather than wrap it, extend from code to general formatting constraints, and compare against ReBIND-style compiled ledgers as baselines, including a negative-priming variant that names the withdrawn rules.

Self-improvement with round-trip filtering and preference signals. SimRAG (Xu et al., 2025) is the methodological template: complementary skills, self-generated data and a consistency filter. STaR-style rejection sampling (Zelikman et al., 2022) and self-instruct (Wang et al., 2023) are the broader family; self-rewarding and iterative DPO (Yuan et al., 2024; Rafailov et al., 2023) turn a model's own judgments into preference data. Our filter is exact match on a structured state plus executable checks, not retrieval or LLM judgment, and our preference pairs are labeled by the same checkers, so no judge model is involved in the pipeline.

3 Problem Formulation

Constraints, edits and state. A constraint is a pair $c = (f, \theta)$ of a family f and parameters θ with an executable checker $\text{check}_f(y, \theta) \in \{0, 1\}$ over a response y; examples are all_caps, max_words(80), end_phrase("Is there anything else?"), json_only and forbidden_word("very"). Twelve families are used in training and six further families appear only at test time (Appendix A). A dialogue applies one or more edits per user turn: add(c), amend($c \to c'$), revoke(c), distract (an unrelated request) and return (back to the task). The net state after turn t is $S_t = (A_t, R_t)$: the active set A_t holds at most one constraint per family, and the tombstone set R_t holds every revoked or superseded constraint. amend moves the old version to R_t; re-adding a tombstoned constraint revives it. S_t is compiled from the edit list, has an order-independent canonical serialization, and is therefore exact ground truth for both training and evaluation at no annotation cost.

Metrics. Persistence at turn t is the fraction of A_t satisfied by the response y_t. Relapse at turn t is the fraction of R_t satisfied by y_t, restricted to families whose chance-satisfaction rate $p_0 \leq 0.10$, where p_0 is measured on the backbone's own unconstrained responses to the same task prompts (a response that happens to contain no commas is not evidence of relapse for no_commas unless commas are otherwise common); an amended constraint whose new version logically implies the old one is not counted. Conversation success requires every active constraint satisfied and zero relapse at every turn. All metrics are averaged

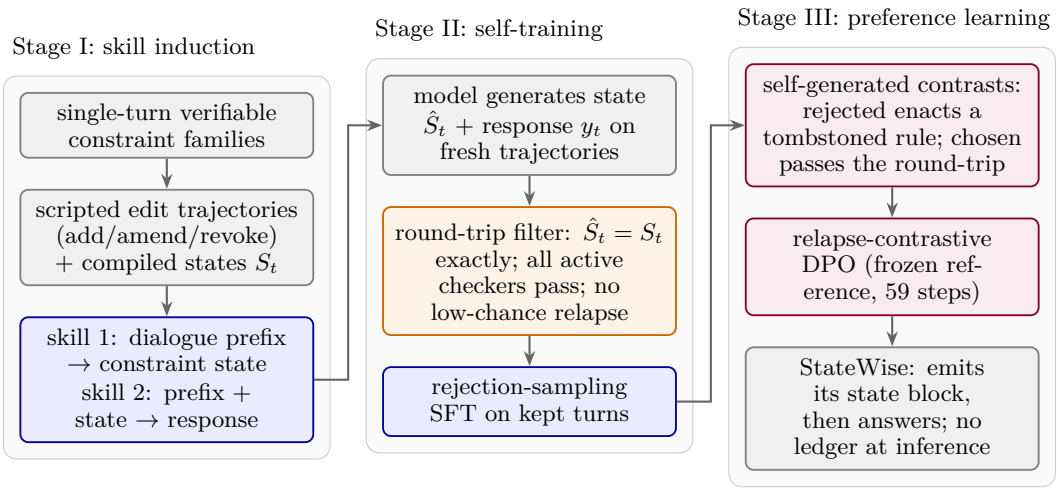

Figure 1: Overview of StateWise. Stage I induces two complementary skills from scripted trajectories whose ground-truth states are compiled from the edit history. Stage II lets the model generate state and response on fresh trajectories and keeps only turns that pass an exact round-trip check. Stage III turns the same check into a preference signal: the model's relapsed generations are rejected responses, its round-trip-clean generations are chosen responses. No proprietary model, judge or human annotation is used in training.

over conversations and reported with conversation-level bootstrap intervals; comparisons between arms are paired on shared trajectory identifiers.

Trajectory synthesis. Scripted trajectories are generated from a seed: turn 0 adds one or two constraints and requests a short task; later turns add, amend or revoke with at most three constraints active and incompatible pairs (e.g., json_only with bullets_exact) never jointly active; half of the trajectories insert a distract turn followed by a return; every trajectory contains at least one revoke or amend; the final turn is always a task turn. Each user message is composed from natural-language phrasings of the edits plus a task prompt, so every assistant turn is checkable. Trajectories have four to eight turns. Three evaluation sets are used throughout: an in-distribution set (240 conversations, training families and phrasings, unseen seed), a held-out set (300 conversations; 18 families of which 6 are never trained on), and a paraphrase set (300 conversations; author-written alternative phrasings and new task prompts). Test sets are fixed before any training run.

4 Methodology

4.1 Stage I: Skill Induction

Starting from an instruction-tuned backbone (Figure 1), we fine-tune with LoRA on three mixtures built from scripted trajectories (Section 3): (a) single-turn verifiable instruction following, as in IFEval; (b) state summarization, where the input is a dialogue prefix and the target is the canonical JSON of the compiled state S_t, listing active constraints with parameters and revoked constraints as tombstones; and (c) state-conditioned generation, where the input is the prefix plus the ground-truth state block and the target is a response that passes every active checker, obtained by rejection sampling from the backbone. Mixture sizes and hyper-parameters were fixed before training (Appendix B). Only the dialogue and the constraint library are needed; the backbone is its own teacher.

4.2 Stage II: Self-Training with Round-Trip Filtering

The Stage-I model generates, for each turn of fresh trajectories, a state block followed by a response. A turn is kept if and only if (i) the emitted state equals the compiled state

exactly, (ii) every active checker passes, and (iii) no low-chance tombstoned constraint is satisfied. Per-family shares are capped to avoid collapse onto easy families. The model is fine-tuned on the kept turns (rejection-sampling SFT). Filter yield and rejection reasons are in Appendix B; on both backbones relapse is the dominant rejection reason, foreshadowing Section 6: the filter removes relapsed examples but supplies no signal against relapse.

4.3 Stage III: Relapse-Contrastive Preference Learning

The Stage-II model is sampled on a further set of trajectories (a different seed from Stage II). For each turn we pair a rejected response that enacts a tombstoned rule with a chosen response from the same prefix that passes the full round-trip check; a capped share (30%) of pairs contrasts an active-constraint violation with a passing response instead. The pairs are labeled entirely by the checkers and the compiled state. We run DPO (Rafailov et al., 2023) with a frozen reference (the Stage-II weights with adapters disabled), $\beta = 0.1$, an auxiliary NLL weight of 0.5 on the chosen response, and a fixed budget of 59 optimizer steps over 931 pairs (947 constructed, 663 relapse and 284 active-violation, of which 16 were dropped for length). Nothing is tuned after seeing results.

4.4 Inference

In state mode the model emits its state block and then answers; the block is stripped before the response is checked. In state-free mode the same weights are prompted without the state instruction, which tests whether the behavior is internalized. We also evaluate weight interpolation $\theta_\alpha = \theta_{\text{base}} + \alpha(\theta_{\text{Stage III}} - \theta_{\text{base}})$ for $\alpha \in \{0.5, 0.75\}$, with no training, as a way to trade the relapse gain against breadth on general instruction following (Section 7.2).

5 Experimental Setup

Backbones and arms. Llama-3.1-8B-Instruct is the primary backbone (two training seeds) and Qwen2.5-7B-Instruct the secondary (one seed, full pipeline with Qwen as its own teacher). The arms are: A0 the backbone; A1 single-turn constraint training (rejection-sampling SFT on single-turn verifiable prompts, the IFBench-style recipe); A2 Stage I only; A3 Stages I+II without state supervision (same trajectories, checker-only filter); A4 Stages I+II with state (the self-training recipe); A5 A4 in state-free mode; A9 Stage I with an active-only state block (no tombstones); A7 A4 plus Stage III (StateWise); and two ReBIND-style external-ledger baselines around the untrained backbone at matched prompt tokens, A6, which compiles the net specification including the withdrawn rules under a "do not follow" header, and A6b, which lists only the rules in force. A6b uses the compiled ground-truth state at inference and is therefore an oracle ledger.

Evaluation sets and metrics. Held-out (300 conversations, 18 families, 6 unseen), in-distribution (240 conversations) and paraphrase (300 conversations) sets (Section 3), all with greedy decoding. Public benchmarks: Multi-IF (He et al., 2024, English, 909 conversations, prompt-level strict accuracy at turns 1–3 with the model's own history); IFEval (Zhou et al., 2023, 541 prompts, prompt-level strict and loose); MT-Bench first-turn pairwise preference against the backbone judged by gemma-2-9b-it in both presentation orders (a weak local judge; regression signal only). Metrics are persistence, relapse and conversation success (Section 3), state exact-match for state-mode arms, and 95% conversation-level bootstrap intervals (2,000 resamples); deltas between arms are paired on trajectory identifiers. The backbone's own greedy results move by a few points between runs because batched bf16 non-determinism compounds across turns, so only within-run paired comparisons are used.

Preregistration. Hypotheses, arms, metrics and pass/fail gates were fixed before each run (Appendix C) and not changed after inspection; failed gates are reported as failures. The main gates are G1 (A4 beats A0 on success and relapse), G2 (state supervision helps: A4 vs. A3), H2 (gain survives state-free inference), H3 (trained model matches the external ledger), H5 (Stage III beats A0 on success and relapse with intervals excluding zero), and, for the

Arm (no training)	Persistence	Relapse	Conv. success	State match
Plain chat	0.793 [0.767, 0.820]	0.109 [0.080, 0.141]	0.275 [0.225, 0.333]	–
State-scratchpad prompt	0.587 [0.547, 0.629]	0.162 [0.125, 0.200]	0.088 [0.054, 0.125]	0.003 [0.001, 0.006]
Paired delta	−0.206 [-0.253, -0.160]	+0.052 [0.014, 0.090]	−0.188 [-0.250, -0.125]	

Table 1: Inference-only diagnostic (Llama-3.1-8B-Instruct, 240 conversations). Asking the untrained model to maintain its own constraint state makes it worse on every metric.

		Held-out (300 conv., 6 unseen families)			In-distribution (240 conv.)		
	Arm	Success	Relapse	Persist.	Success	Relapse	Persist.
none	A0 backbone	0.290 [0.240, 0.343]	0.130 [0.100, 0.165]	0.799	0.275 [0.221, 0.333]	0.110 [0.081, 0.141]	0.795
	A1 single-turn constraint SFT	0.280 [0.230, 0.330]	0.164 [0.131, 0.197]	0.847	0.329 [0.271, 0.388]	0.148 [0.114, 0.183]	0.856
ledger	A6 oracle ledger incl. withdrawn rules	0.203 [0.160, 0.250]	0.380 [0.334, 0.424]	0.871	0.242 [0.188, 0.300]	0.333 [0.284, 0.383]	0.887
	A6b oracle ledger, active rules only	0.390 [0.337, 0.443]	0.148 [0.117, 0.182]	0.877	0.463 [0.400, 0.525]	0.084 [0.058, 0.114]	0.890
self-training	A2 Stage I (state)	0.250 [0.203, 0.300]	0.254	0.841	0.325	0.281	0.926
	A9 Stage I, active-only state	0.293 [0.247, 0.343]	0.276	0.882	0.350	0.261	0.916
	A3 Stages I+II, no state	0.310 [0.257, 0.363]	0.277	0.878	0.371	0.228	0.897
	A4 Stages I+II, state	0.327 [0.273, 0.380]	0.246	0.853	0.350	0.258	0.937
	A5 A4, state-free inference	0.240 [0.193, 0.290]	0.246	0.825	0.292	0.203	0.855
	A7 StateWise (A4 + Stage III)	0.370 [0.313, 0.423]	0.094 [0.069, 0.121]	0.834	0.554 [0.492, 0.617]	0.042 [0.024, 0.062]	0.904

Table 2: Main results on Llama-3.1-8B-Instruct (training seed 1, greedy decoding). Conversation-level 95% bootstrap intervals; A6/A6b receive the compiled ground-truth state at inference and are oracle baselines. State exact-match for the state arms: A2 0.527 / 0.981 (held-out / in-dist.), A4 0.510 / 0.916, A7 0.500 / 0.893.

extension, replication of direction on a second seed and backbone, a Multi-IF direction check, a no-regression gate on IFEval (within 2 points), and two Stage-III ablation gates.

Training details. LoRA rank 64, learning rate 10^{-4} for SFT stages; DPO as in Section 4. Stage I uses 9.2k examples, Stage II the round-trip survivors of 2,400 self-generated trajectories (per-family caps), Stage III 931 pairs. Each training run uses two A100-40GB GPUs; on Llama-3.1-8B, Stage I takes about an hour, Stage II 68 minutes and Stage III 17 minutes. Full hyper-parameters, data counts and filter statistics are in Appendix B.

6 Experimental Results

6.1 The Failure Exists and Prompting Does Not Fix It

Before training we ran a preregistered inference-only diagnostic on 240 trajectories (1,438 assistant turns) with Llama-3.1-8B-Instruct (Table 1). Plain chat honors 79% of active constraints overall, falling from 0.887 at the first turn to 0.653 at the eighth, and enacts revoked low-chance constraints on 11% of post-edit turns. A system prompt asking the model to write its own constraint state before answering makes every metric worse: persistence −0.21, relapse +0.05, success −0.19, all with intervals excluding zero, and the emitted states match the compiled ground truth 0.3% of the time (omitted rules, family confusions, constraints hallucinated from task content, and an always-empty revoked list). The failure is real at the 8B operating point, prompting alone does not supply a usable state, and an external ledger must therefore be compiled programmatically, as ReBIND does, rather than requested from the model. This is the premise for training.

6.2 Main Results

Table 2 presents the preregistered arms on the primary backbone. From the results, we have the following findings:

(1) Every supervised arm learns to track the state and roughly doubles relapse. A4 emits the exact compiled state, tombstones included, for 92% of in-distribution turns and raises persistence by five points, yet its relapse is 0.246 held-out against the backbone's 0.130 (+0.116 [0.078, 0.151], paired). The same doubling appears without state supervision (A3,

Setting	Arm	Success	Relapse	Persist.	Paired delta vs. A0 (success / relapse)
Llama-3.1-8B, seed 2, held-out	A0	0.250	0.124	0.788	
	A6b oracle ledger	0.430	0.145	0.884	
	A4 self-training	0.337	0.263	0.885	
	A7 StateWise	0.443	0.097	0.860	+0.193 [0.133, 0.253] / −0.026 [-0.056, 0.004]
Qwen2.5-7B, held-out	A0	0.163	0.230	0.731	
	A6b oracle ledger	0.260	0.244	0.865	
	A4 self-training	0.277	0.220	0.830	
	A7 StateWise	0.357	0.144	0.837	+0.193 [0.137, 0.253] / −0.086 [-0.118, -0.053]
Llama-3.1-8B, paraphrase set	A0	0.243	0.168	0.809	
	A6b oracle ledger	0.403	0.163	0.919	
	A4 self-training	0.260	0.268	0.867	
	A7 StateWise	0.340	0.134	0.832	+0.097 [0.030, 0.160] / −0.034 [-0.068, -0.000]

Table 3: Replication of the main arms. Seed 2 repeats the Llama pipeline with a new training seed; the Qwen2.5-7B-Instruct pipeline uses Qwen as its own teacher; the paraphrase set uses author-written phrasings and new tasks. In-distribution: Appendix D.

0.277) and with an active-only state (A9, 0.276), and even single-turn constraint training raises relapse (A1, 0.164 from 0.130): making a model better at obeying constraints makes withdrawn constraints stickier. State supervision helps only marginally (A4 vs. A3: relapse −0.031 [-0.058, -0.004]), and the gain does not survive removing the state at inference (A5). Under the preregistered gates, G1, H2 and H3 fail and G2 passes narrowly. Tracking the state and obeying it are separable skills, and rejection-sampling SFT teaches the first without the second: the filter removes relapsed examples but supplies no signal against relapse.

(2) The relapse-contrastive stage supplies the missing signal. Fifty-nine DPO steps on 931 self-generated contrasts cut relapse from 0.246 to 0.094 held-out (−0.152 [-0.186, -0.120]) and from 0.258 to 0.042 in distribution (−0.216 [-0.258, -0.177]), to below the backbone (−0.036 [-0.065, -0.011] held-out; −0.068 [-0.099, -0.040] in-dist.) while keeping the persistence gain (+0.035 [0.002, 0.067] over A0) and lifting conversation success (+0.080 [0.020, 0.140] held-out; +0.279 [0.204, 0.350] in-dist.). State exact-match is unchanged (0.893 vs. 0.916 in-dist.): the gain is in obeying the state, not in tracking it. Gate H5 passes.

(3) StateWise matches or beats an oracle external ledger without scaffolding. A6b, the untrained backbone wrapped in a programmatically compiled active-only ledger with ground-truth knowledge of every edit, is the strongest baseline. StateWise is below it on relapse everywhere (−0.054 [-0.081, -0.030] held-out; −0.042 [-0.070, -0.017] in-dist.), matches it on held-out success (−0.020 [-0.080, 0.043]) and beats it in distribution (+0.092 [0.029, 0.158]); the remaining gap is four points of persistence. The ledger variant that names withdrawn rules under a "do not follow" header (A6) nearly triples relapse (0.380): any salient mention of a revoked rule, even negated, primes the 8B model to enact it. This priming does not transfer to a trained self-emitted state: A9, whose state block omits the tombstones, relapses no less than A2, which names them (+0.022 [-0.004, 0.051]).

6.3 Replication: Second Seed, Second Backbone, Unseen Phrasings

Table 3 repeats the comparison on a second training seed, a second backbone and unseen phrasings. The arc of Table 2 reproduces in each: self-training raises persistence and, on Llama, relapse (0.263 on seed 2, 0.268 on the paraphrase set); Stage III cuts relapse relative to A4 in every setting, by 13–17 points on Llama (−0.166 [-0.199, -0.135] on seed 2; −0.134 [-0.166, -0.104] on paraphrases) and by 7.6 points on Qwen (−0.076 [-0.101, -0.053]), whose backbone relapse is already high and whose SFT stage did not raise it further, and lifts conversation success over the backbone by 10–19 points with intervals excluding zero everywhere. Relapse relative to the backbone is significantly lower on seed 1 and on Qwen, crosses zero on seed 2 (upper bound +0.004) and touches it on the paraphrase set; the strict two-part replication gate therefore passes on Qwen and fails narrowly on Llama seed 2. On Qwen, StateWise beats the oracle ledger on both success (+0.097 [0.037, 0.157]) and relapse (−0.100 [-0.133, -0.068]); on Llama it beats the ledger on relapse in every setting, matches it on success within the interval on both seeds, and trails it on the paraphrase set (0.340 vs. 0.403).

		Held-out			In-distribution		
Arm	Preference pairs	Success	Relapse	Persist.	Success	Relapse	Persist.
A4 (base for DPO)	none	0.327	0.246	0.853	0.350	0.258	0.937
A7-shuf	A7's 931 pairs, 50% labels swapped	0.320	0.242	0.849	0.379	0.240	0.936
A7-act	active-violation pairs only (931)	0.317	0.284	0.881	0.412	0.308	0.979
A7-rel	relapse pairs only (663)	0.343	0.142	0.827	0.463	0.102	0.896
A7 StateWise	relapse 663 + active 284	0.370	0.094	0.834	0.554	0.042	0.904

Table 4: Stage-III ablations (Llama-3.1-8B, same base, hyper-parameters and pair pool as A7; preregistered). Pair counts are before length filtering (Appendix B).

	Held-out (ours)			IFEval		Multi-IF (prompt-strict)		
Model	Success	Relapse	Persist.	strict	loose	T1	T2	T3
A0 backbone	0.290	0.130	0.799	0.725	0.771	0.735	0.557	0.400
A6b oracle ledger	0.390	0.148	0.877	–	–	–	–	–
A4 self-training	0.327	0.246	0.853	0.667	0.697	0.702	0.508	0.351
A7 StateWise ($\alpha = 1$)	0.370	0.094	0.834	0.660	0.686	0.678	0.496	0.336
A7, state-free inference	–	–	–	0.636	0.689	–	–	–
interpolation $\alpha = 0.75$	0.397	0.125	0.857	0.686	0.723	0.695	0.516	0.374
interpolation $\alpha = 0.5$	0.413	0.152	0.885	0.706	0.728	0.725	0.560	0.407

Table 5: Breadth. Multi-IF (909 English conversations) accumulates constraints without ever revoking them; IFEval (541 prompts) is single-turn. $\theta_\alpha = \theta_{\text{base}} + \alpha(\theta_{\text{A7}} - \theta_{\text{base}})$ with no training; both α values were fixed in advance.

7 Analysis

7.1 Ablation Studies: What the Preference Stage Needs

Table 4 varies only the preference pairs. Label-shuffled pairs do nothing: A7-shuf is indistinguishable from A4 on every metric (relapse -0.004 [-0.023, 0.014]; success -0.007 [-0.043, 0.030]), so the gain is the preference signal and not further training on the distribution. The relapse-specific contrast is necessary: pairs that only contrast an active-constraint violation with a passing response raise persistence ($+0.028$ [0.007, 0.051] over A4) and raise relapse ($+0.038$ [0.015, 0.064]), the same coupling seen in single-turn training (A1); StateWise is 19 points below this variant on relapse (-0.190 [-0.226, -0.156]). Relapse pairs alone recover most of the effect (-0.105 [-0.137, -0.073] vs. A4) but the 30% active-violation pairs add a further 4.7 points of relapse reduction (-0.047 [-0.072, -0.026]) and keep persistence from slipping (A7-rel loses 2.6 points of persistence to A4; A7 loses 1.9 with an interval covering zero). Both preregistered ablation gates pass.

7.2 The Cost of Specialization and a Controllable Trade-Off

Specializing an 8B model to revocable formatting constraints from a 12-family library costs breadth (Table 5). On Multi-IF, where constraints only accumulate and most of IFEval's 25 instruction types are outside our training families, the trained arms are worse than the backbone at every turn (A7: -6.5 points at turn 3), and on single-turn IFEval A7 is 6.5 points below the backbone on prompt-strict accuracy, failing the preregistered 2-point no-regression gate; a weak pairwise judge on MT-Bench first turns sees a small regression (win 5%, tie 80%, loss 15%). This is a real cost, not a tooling artifact, and we report it as such.

Weight interpolation with the backbone exposes the trade-off as a single knob. At $\alpha = 0.5$ the model is within 1.9 points of the backbone on IFEval and, on Multi-IF, within one point of it at turn 1 and above it at turns 2–3 ($+0.7$ at turn 3), while having the highest conversation success (0.413, $+0.123$ [0.057, 0.187] over A0) and persistence ($+0.087$ [0.059, 0.114]) of any arm and matching the oracle ledger on every held-out metric (success $+0.023$ [-0.037, 0.083], relapse $+0.004$ [-0.023, 0.031]). What it gives back is the relapse gain: relapse returns to the backbone level ($+0.022$ [-0.007, 0.051]). The preregistered interpolation gate—IFEval

within 2 points and relapse below the backbone with an interval excluding zero—is therefore not met by either α, and we do not select an α post hoc. The honest summary is a Pareto frontier: $\alpha = 1$ for the lowest relapse, $\alpha = 0.5$ for the broadest model that still tracks its state, with the untested mitigation being to mix general instruction data into Stages I–III so that the frontier itself moves.

7.3 Where the Relapse Goes

Per-turn curves (Appendix D) show that the backbone's persistence declines monotonically with depth while relapse is flat at 0.09–0.13 on every post-edit turn; A4 lifts persistence at all depths and lifts relapse in the same way, and A7 keeps the persistence curve while pulling relapse below the backbone at every depth. By family, relapse concentrates on constraints that are easy to keep enacting silently once learned—no_commas, no_digits, sentence_count, paragraphs_exact—and Stage III removes most of it without touching the state block (state exact-match is unchanged), which is the behavioral signature of a model that knew the rule was withdrawn and had not been told that this mattered.

8 Conclusion

Evolving-constraint dialogue exposes two failures in instruction-tuned LLMs, forgetting active constraints and re-enacting withdrawn ones, and the only prior repair we are aware of wraps the model in an external ledger. We showed that a SimRAG-style self-improvement recipe with an exact round-trip filter teaches a 7–8B model to track its own constraint state almost perfectly without teaching it to obey that state: on Llama-3.1-8B every supervised variant roughly doubles relapse, and on Qwen2.5-7B none reduces it. Tracking and obeying are separable skills, and a filter can only teach the first. Turning the same verifiable check into a preference signal supplies what the filter cannot, a negative signal against relapse, and 59 steps of DPO on self-labeled contrasts cut relapse by 8–22 points, below the backbone and below an oracle compiled ledger, on two backbones, two seeds and unseen phrasings, with ablations that isolate the relapse-specific contrast as the necessary ingredient. The recipe costs breadth on cumulative-constraint and single-turn benchmarks, and interpolating with the backbone trades that cost against the relapse gain along a single knob. Every number on our evaluation sets comes from executable checkers against compiled states, and training uses no proprietary model, judge or human dialogue annotation; the only judge-based number is the MT-Bench regression signal.

9 Limitations

Constraint families are formatting-level and programmatically checkable, and trajectories are scripted from templates with a paraphrase set as the only test of phrasing shift; semantic constraints and user phrasings in the wild are broader. The relapse metric excludes high-chance families by construction and therefore understates relapse on constraints that are easy to satisfy accidentally. Results are for 7–8B backbones with LoRA; stronger models may sit near the floor on relapse, as ReBIND reports, and the breadth cost may differ. The held-out success intervals have half-widths of about five points, and relapse relative to the backbone crosses zero on one seed and touches it on the paraphrase set. The specialization cost on Multi-IF and IFEval is real, the interpolation remedy removes it only by giving back the relapse gain, and the obvious mitigation of mixing general instruction data into training was not run. The MT-Bench judge is a weak local model and its signal is indicative only.

AI use statement

Large language models were used as writing and coding assistants: to draft and edit prose, to write and refactor training, evaluation and analysis scripts, to format and check references, and to port the manuscript into this template. The hypotheses, preregistered gates, experimental design and interpretation are the authors'; all reported numbers are produced by the released code from result artifacts and were checked by the authors against those

artifacts; LLM-generated code was tested before use. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Reproducibility statement

Backbones (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct) and benchmarks (Multi-IF, IFEval, MT-Bench) are public. The constraint library, trajectory generator, checkers, state compiler, round-trip filter, pair construction, training configurations and evaluation scripts will be released together with every result JSON (each with a configuration snapshot and content hash), the fixed test-set seeds (Section 3), the preregistered protocol with its amendments (Appendix C), and the exact hyper-parameters (Appendix B). Because Python's random sequences can differ across interpreter versions, the generated trajectories are stored alongside the results.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ: A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5016–5026, 2018. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. Can language models follow multiple turns of entangled instructions? In Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 25445–25460, 2025. doi: 10.18653/v1/2025.findings-emnlp.1387. URL https://aclanthology.org/2025.findings-emnlp.1387/.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. arXiv preprint arXiv:2410.15553, 2024. URL https://arxiv.org/abs/2410.15553.

Qi Jia, Ye Shen, Xiujie Song, Kaiwei Zhang, Shibo Wang, Dun Pei, Xiangyang Zhu, and Guangtao Zhai. One battle after another: Probing LLMs' limits on multi-turn instruction following with a benchmark evolving framework. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9574–9590, 2026. doi: 10.18653/v1/2026.acl-long.433. URL https://aclanthology.org/2026.acl-long.433/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In International Conference on Learning Representations (ICLR), 2026. URL https://arxiv.org/abs/2505.06120.

Ming Li, Pei Chen, Zhenhao Zhang, Tao Yang, Xinyang Zhang, Han Li, Tianyu Cao, Ming Zeng, Zhuofeng Wu, Meng Jiang, Huasheng Li, Lihong Li, and Bing Yin. Mitigating lost in multi-turn conversation via curriculum RL with verifiable accuracy and abstention rewards. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 33364–33381, 2026. doi: 10.18653/v1/2026.acl-long.1540. URL https://aclanthology.org/2026.acl-long.1540/.

Shu Tong Luo, Wenqin Liu, Rui Liu, Mingming Gong, and Jiaxian Guo. Multi-turn reasoning when context arrives in pieces: Scalable sharding and memory-augmented RL. arXiv preprint arXiv:2606.12941, 2026. URL https://arxiv.org/abs/2606.12941.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction

following. In Advances in Neural Information Processing Systems, volume 38, 2025. doi: 10.52202/085713-1645. URL https://arxiv.org/abs/2507.02833.

Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. SysBench: Can large language models follow system messages? In International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=KZWaxtzIRx.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems, volume 36, 2023. doi: 10.52202/075280-2338. URL https://arxiv.org/abs/2305.18290.

Qingyu Ren, Qianyu He, Powei Chang, Jie Zeng, Zeye Sun, Fei Yu, Jiaqing Liang, and Yanghua Xiao. Instructions are all you need: Self-supervised reinforcement learning for instruction following. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4755–4776, 2026. doi: 10.18653/v1/2026.acl-long.217. URL https://aclanthology.org/2026.acl-long.217/.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13484–13508, 2023. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang, and Qi He. SimRAG: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 11534–11550, 2025. doi: 10.18653/v1/2025.naacl-long.575. URL https://aclanthology.org/2025.naacl-long.575/.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 57905–57923, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In Advances in Neural Information Processing Systems, volume 35, 2022. doi: 10.52202/068431-1126. URL https://arxiv.org/abs/2203.14465.

Yangyang Zhao, Linfan Dai, Li Cai, Bowen Xing, and Libo Qin. Bridging reasoning and action: Hybrid LLM–RL framework for efficient cross-domain task-oriented dialogue. In Findings of the Association for Computational Linguistics: ACL 2026, pp. 12987–13003, 2026. doi: 10.18653/v1/2026.findings-acl.633. URL https://aclanthology.org/2026.findings-acl.633/.

Haoyu Zheng, Yun Zhu, Shu Yuan, Shangming Chen, Qing Wang, Wenqiao Zhang, Jun Xiao, and Yueting Zhuang. MAIGO: Mitigating lost-in-conversation with history-cleaned on-policy self-distillation. arXiv preprint arXiv:2605.27186, 2026. URL https://arxiv.org/abs/2605.27186.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911, 2023. URL https://arxiv.org/abs/2311.07911.

Haoyuan Zhu. Dead text or binding clause? Measuring and restoring constraint influence in black-box LLM dialogues. arXiv preprint arXiv:2608.12599, 2026. URL https://arxiv.org/abs/2608.12599.

A Constraint Families and Trajectories

Training families (12): all_caps, all_lower, no_commas, end_phrase, start_phrase, keyword_count, json_only, bullets_exact, max_words, forbidden_word, paragraphs_exact, wrap_quotes. Held-out families (6, test only): no_digits, sentence_count, title_brackets, postscript, two_responses, numbered_sections. Each family has an executable checker and one to three natural-language phrasings per operation (add, amend, revoke); the paraphrase set uses three further author-written phrasings per family and operation and new task prompts. Measured chance-satisfaction rates p_0 on the backbone's unconstrained responses are below 0.05 for every family except forbidden_word (0.94–0.95), which is therefore excluded from relapse, and paragraphs_exact (0.04) and bullets_exact (0.02), which are retained. Trajectories are generated from a seed with 4–8 turns, at most three active constraints, incompatible pairs never jointly active, a distract/return pair in half of the trajectories, and at least one revoke or amend per trajectory. Test sets: held-out seed 777 (300 conversations), in-distribution seed 20260916 (240), paraphrase seed 999 (300); training trajectories use disjoint seeds.

B Training Details and Filter Statistics

SFT stages. LoRA rank 64, $\alpha = 128$, dropout 0.05 on all attention and MLP projections; learning rate 10^{-4}, 3% warm-up, one epoch, effective batch 16 sequences, maximum length 4,096; identical for every SFT arm, only the data differs. Stage I (Llama): 9,238 examples, 578 steps, 55–61 minutes on two A100s (loss 0.29 $\rightarrow$ 0.07 for the state arm). Stage II (state arm): 9,920 kept turns, 620 steps, 68 minutes. Single-turn arm A1: 2,237 examples, 140 steps.

Stage III (DPO). Base and frozen reference: the Stage-II state model; LoRA as above; learning rate 2×10^{-5}, 5% warm-up, one epoch, effective batch 16 pairs, $\beta = 0.1$, NLL weight 0.5 on the chosen response, at most 30% active-violation pairs. 931 pairs encoded (663 relapse, 284 active-violation; 16 dropped for length), 59 steps, 17 minutes; training preference accuracy 0.50 $\rightarrow$ 1.00 and loss 0.77 $\rightarrow$ 0.31. Ablations: relapse-only 649 pairs / 41 steps; active-only 922 pairs / 58 steps; label-shuffled 931 pairs / 59 steps. All values were fixed in the protocol amendment of 2026-09-16 16:50 and never tuned.

Filter yields. A turn is rejected when none of its $K = 4$ sampled candidates passes, which also stops the trajectory; a rejected turn is counted under every distinct reason among its candidates, so reason counts slightly exceed rejected-turn counts. Stage-I data generation (backbone as teacher, 4 shards): 10,723 turns attempted, 9,238 kept (yield 0.86); stopping reasons 485 active-constraint violations vs. 1,027 relapses. Stage-II generation with the state model: 11,243 attempted, 9,920 kept (0.88); reasons 411 active violations, 887 relapses, 158 state mismatches. Relapse is the dominant rejection reason at every stage and on both backbones (Qwen Stage I: 1,253 relapse vs. 446 active-violation stops), which is the data-side signature of the result in Section 6.2.

C Preregistration Summary

The protocol fixed, before any training run, the arms of Table 2, the metrics, the test seeds, and the gates: G0.1 (backbone final-turn persistence ≤ 0.80; pass, 0.653), G0.2 (prompted state exact-match ≥ 0.5; fail, 0.003), G0.3 (state prompt does not hurt; fail), G1 (A4 above A0 on success and below on relapse, intervals excluding zero; fail), G2 (A4 below A3 on relapse; narrow pass), H2 (state-free A5 retains half of the A4 gain; fail), H3 (A4 within the interval of A6b; fail). Amendment of 2026-09-16 16:50 (before the H5 runs): A9 and A7, gate H5 (pass for A7, fail for A9). Amendment of 2026-09-17 22:50: seed 2 and Qwen replication (direction replicates; strict two-part gate passes on Qwen, fails narrowly on Llama seed 2), Multi-IF direction (reversed), IFEval no-regression within 2 points (fail, -6.5), MT-Bench loss-minus-win within 10 points (boundary; treated as fail), paraphrase relapse below backbone with interval excluding zero (upper bound -0.000; marginal miss).

Amendment of 2026-09-18 22:50: Stage-III ablations (both gates pass) and interpolation (gate fails; both α reported). No gate or hyper-parameter was changed after results were seen.

D Additional Results

In-distribution replication. Llama seed 2 (240 conversations): A0 success 0.321 / relapse 0.092; A6b 0.450 / 0.101; A4 0.379 / 0.264; A7 0.475 / 0.074. Qwen2.5-7B: A0 0.133 / 0.254; A6b 0.254 / 0.275; A4 0.279 / 0.284; A7 0.358 / 0.179. Stage-III ablations in distribution (Table 4): A7 vs. A7-act relapse -0.266 [-0.309, -0.224]; vs. A7-shuf -0.199 [-0.240, -0.161]; vs. A7-rel -0.061 [-0.090, -0.033] with success $+0.092$ [0.042, 0.146].

Per-turn behavior. Backbone persistence by turn index on the diagnostic set: 0.887, 0.842, 0.803, 0.771, 0.741, 0.712, 0.681, 0.653; relapse on post-edit turns 0.09–0.13 with no trend. Trained arms shift the whole persistence curve upward by 4–9 points; A7 lowers relapse below the backbone at every depth, A4 raises it at every depth. Per-family relapse counts for A4 concentrate on no_digits, sentence_count, no_commas and paragraphs_exact; Stage III reduces each by more than half.

Interpolation vs. oracle ledger. $\alpha = 0.5$ vs. A6b (held-out, paired): success $+0.023$ [-0.037, 0.083], relapse $+0.004$ [-0.023, 0.031], persistence $+0.008$ [-0.015, 0.034]. $\alpha = 0.75$ vs. A0: success $+0.107$ [0.047, 0.170], relapse -0.005 [-0.038, 0.027], persistence $+0.058$ [0.026, 0.089]; IFEval prompt-strict 0.686, Multi-IF turn 3 0.374.

Prompts. The state-mode system prompt asks the model to begin its reply with a fenced JSON block containing active (list of {family, params}) and revoked (same schema) and then to answer; the block is stripped before checking. State-free mode uses the backbone's default system prompt. The ledger baselines prepend the compiled specification (A6b: rules in force; A6: rules in force plus a "do not follow" list of withdrawn rules) to every user turn, matched in token count to the state block.