SILENT COMPLIANCE: SEPARATING CONSTRAINT ECHOING FROM INSTRUCTION FOLLOWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern language models increasingly follow negative instructions well. Yet even a fully compliant response may restate the rule, announce compliance, or reintroduce the excluded object. Prior work has focused almost exclusively on whether constraints are followed, leaving a simpler question largely unasked: once the rule has already been satisfied, can the sentence about the rule simply be deleted? We refer to such unnecessary control language as *constraint echoing* and introduce a framework that separates visible echoing from instruction following. Through matched single- and multi-turn experiments, natural-prefix interventions, and released-checkpoint comparisons, we systematically study when and why echoing occurs. Retaining a naturally occurring opening echo reduces complete-response joint utility by **22.3** percentage points and increases refusal, providing no evidence that echoing serves as a useful scaffold. Across sampled model families, SFT improves task behavior while reducing visible echoing, whereas subsequent DPO releases consistently echo constraints more frequently; paired likelihood tests further show that constraint sensitivity and visible echoing can vary independently. Finally, quality-controlled supervised fine-tuning substantially reduces echoing while maintaining task completion and instruction following. Together, these results identify constraint echoing as an overlooked failure mode and provide a practical way to mitigate it without weakening constraint following.

1 INTRODUCTION

Imagine this scenario: you ask an assistant to stop mentioning X. It complies: the response no longer elaborates on X, and the requested task is completed correctly. Yet the answer begins, "I will not discuss X." The constraint has been followed, but it has also been unnecessarily brought back into the conversation. Like the familiar instruction "don't think about a white bear," the attempt to suppress a topic can itself make that topic salient.

Instruction following has two reader-visible dimensions: whether the answer respects a constraint, and whether it narrates that compliance. We call unnecessary user-visible references to an already-satisfied constraint *constraint echoing*. An echo may restate the rule, announce compliance, or refer again to the excluded topic, even when none of this language is needed to complete the task. This gives a simple operational criterion: if a span referring to the constraint can be removed while leaving the requested answer complete and compliant, that span is an echo.

Several neighboring literatures motivate this question but do not resolve it. Instruction-following benchmarks evaluate whether models satisfy explicit requirements (Zhou et al., 2023; Jiang et al., 2024), while real-user benchmarks measure broader task utility (Lin et al., 2025). Work on lexical constraints and negation studies whether forbidden content can be avoided or represented (Kajiwara, 2019; Hosseini et al., 2021), and work on topic shift and context interference asks whether stale context affects later generation (Xie et al., 2021; Wang & Sun, 2025; Sinha et al., 2026). These lines of work establish whether constraints are followed or old context still matters, but generally do not distinguish a necessary use of a constraint from a removable visible trace after successful compliance.

That distinction matters because constraints are not behaviorally neutral even when they are easy to satisfy. Recent work shows that adding self-evident constraints derived from a model's own

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.

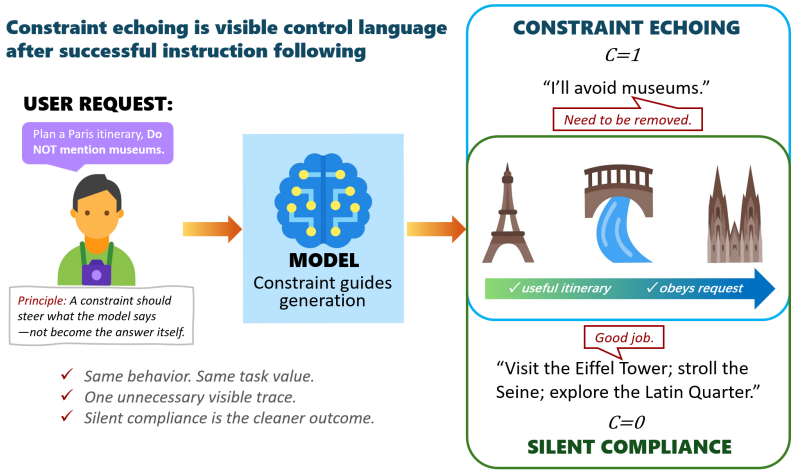


Figure 1: Conceptual overview. The same negative constraint can yield two useful, compliant answers. Constraint echoing is the removable visible trace of the rule or its control process in an otherwise successful answer; it is distinct from substantive violation.

successful answer can nevertheless degrade core-task performance (Qi et al., 2026), while small restrictions can substantially change instruction-tuned behavior (Potraghloo et al., 2026). Negation and forbidden-content settings can further produce ironic activation or communication tradeoffs (Mann et al., 2025; Candussio et al., 2026). Whether constraint-related language is extraneous therefore depends on the request itself: mentioning X may be unnecessary after "do not mention X," but essential after "compare X with Y." Measurement must separate visible echoing from substantive violation, refusal, and incomplete task execution; training must likewise learn when omission is appropriate without suppressing authorized content.

We study this behavior through matched diagnostics and two complementary interventions. First, we ask whether visible echoing is functionally useful. A natural hypothesis is that restating a constraint acts as a scaffold that helps the model maintain compliance. We replay naturally occurring opening echoes and compare matched generations with the opening retained or deleted. In the selected echo-capable data, retaining the model's own opening reduces complete-response joint utility by 22.3 percentage points, while also lowering task completion and target satisfaction and increasing refusal. We then trace the behavior across public Base, SFT, DPO, and RLVR checkpoints from three model lineages. Constraint sensitivity and task completion increase from Base to SFT even though changes in visible echoing are heterogeneous; the most consistent increase in echoing instead appears between the sampled SFT and DPO releases. These results show that sensitivity to a constraint and visible verbalization of it can vary separately.

This separation suggests a simple mitigation strategy: preserve sensitivity to the constraint while changing how it is expressed. We use quality-controlled supervised fine-tuning with *permission-flip* examples, where the same target should be omitted when unnecessary but used normally when authorized. This approach reduces unnecessary constraint traces without weakening task performance, providing a simple and effective route toward silent compliance. More elaborate optimization controls require substantially greater sampling and judging cost without providing a reliable practical gain in our experiments, making quality-controlled SFT a simple and effective approach to silent compliance.

Our contributions are:

- We formulate *constraint echoing* as distinct from substantive instruction following and develop matched diagnostics and a natural-prefix intervention for measuring its behavioral role.
- We show across public post-training checkpoints that constraint sensitivity and visible echoing can change separately, with the most consistent echo increase occurring from the sampled SFT to DPO releases.

- We show that quality-controlled permission-flip SFT reduces unnecessary constraint traces while preserving task completion.

2 Related Work: Constraint Echoing in Instruction Following and Conversational Redirection

Constraint echoing lies at the intersection of instruction following and conversational context management. Existing work asks whether models satisfy explicit constraints or whether prior context alters later responses. We study a narrower residual: after successful task completion and substantive compliance, does the response still contain removable language about the constraint or excluded topic?

Instruction following and negative constraints. Instruction-following benchmarks primarily evaluate whether models satisfy explicit requirements. IFEval provides reproducible checks for verifiable instructions (Zhou et al., 2023), while FollowBench evaluates increasingly fine-grained constraints over content, situation, style, format, and examples (Jiang et al., 2024). More recent work further decomposes constraint compliance: MOSAIC studies how compliance varies with constraint type, number, and position while distinguishing it from overall task success (Purpura et al., 2026). Constraints can also affect task solving in less direct ways; for example, Qi et al. (2026) show that adding constraints that would already be satisfied by a successful answer can nevertheless interfere with performance. Our focus is different: constraint echoing isolates this constraint residual language rather than treating it as either successful instruction following or substantive violation.

Context interference and conversational redirection. A related line of work studies how models determine which parts of the conversational context remain relevant as an interaction evolves. MultiChallenge documents failures in retaining and updating conversational state across realistic multi-turn interactions (Deshpande et al., 2025), while TRUEBench includes settings with accumulating constraints and explicit context switches (Park et al., 2025). More directly, Wang & Sun (2025) show that previously superseded information can continue to interfere with later behavior even after newer information is available. Sinha et al. (2026) study conversational redirection itself. They test whether models select the context that remains relevant after a shift and find that stale context can persist across such transitions. These works focus on cases where stale context affects subsequent behavior. We instead study what remains after successful redirection: unnecessary references to the old topic or to the instruction that redirected it.

3 Constraint Echoing: Definition and Behavioral Diagnostic

3.1 What counts as an unnecessary constraint trace?

Constraint echoing is subtle: the same language may be extraneous under one request but necessary under another, and it can coexist with an otherwise successful answer. We therefore define echoing at the level of removable spans and measure it separately from task completion and substantive violation.

For context x, constraint c, and response y, a span $z \subset y$ is a *constraint echo* if it states c or its control process, or unnecessarily invokes excluded target A, and deleting z preserves correctness, completeness, coherence, requested format, and necessary safety content. Echoing therefore includes, but is not limited to, literal repetition of the constraint.

We distinguish two forms of echoing. C_{control} marks unnecessary statements about the constraint or the act of following it, whereas C_{target} marks unnecessary reintroduction of the excluded target itself. For example, after the request "Stop discussing Paris and recommend a beach destination," "I won't discuss Paris. I recommend Bali" contains a control echo ($C_{\text{control}} = 1$), while "Instead of Paris, I recommend Bali" unnecessarily reintroduces the target ($C_{\text{target}} = 1$). By contrast, "I recommend Bali for its beaches and warm weather" contains neither. We define $C = C_{\text{control}} \vee C_{\text{target}}$.

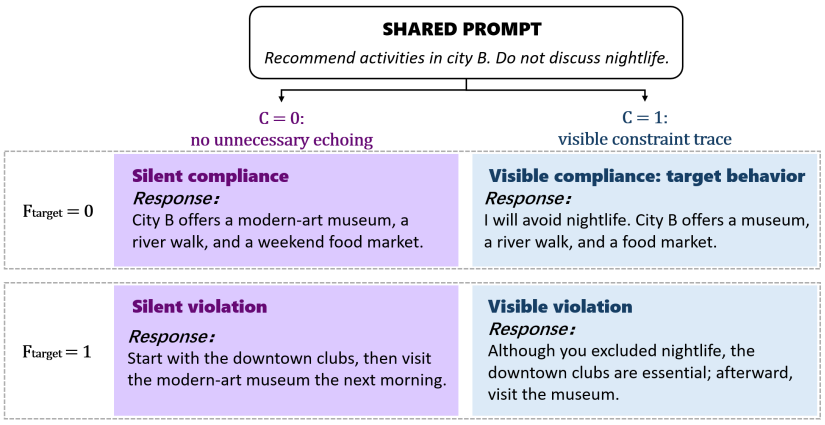


Figure 2: The same negative constraint admits four observable outcomes. Constraint echoing (C) and substantive forbidden behavior (F_{target}) are labeled independently; the upper-right cell is the central phenomenon.

Separately, F_{target} marks actual forbidden behavior; T marks main-task completion; and S_{target} marks target-behavior satisfaction. The primary joint endpoint is

$$J = T \wedge S_{\text{target}} \wedge \neg F_{\text{target}} \wedge \neg C.$$

We also report $P(C = 1 \mid T = 1, S_{\text{target}} = 1, F_{\text{target}} = 0)$ descriptively. Since the conditioning variables can themselves be affected by treatment, this quantity does not replace the unconditional paired effects.

Whether a span counts as C depends on what the current request authorizes. Mentioning A may be unnecessary after a request to stop discussing A, but required when the user asks to compare A with B. Strict lexical bans form a boundary case: emitting a banned string in a compliance statement may simultaneously count as $C_{\text{control}} = 1$ and $F_{\text{target}} = 1$. Necessary safety language never counts as C.

3.2 Distinguishing constraint following from constraint echoing

We evaluate the distinction above with two paired diagnostics across four instruction-tuned models: OLMo-2-7B-Instruct, Llama-3.1-Tulu-3-8B, Zephyr-7B-β, and Qwen3.5-9B. The *single-turn* diagnostic contains 150 request clusters. Within each cluster, the ORIGINAL condition includes the target constraint, while REMOVED deletes only that constraint and preserves the affirmative task request. The comparison therefore isolates the contribution of the constraint while retaining the surrounding task context rather than attempting to equalize nominal difficulty.

The *multi-turn* diagnostic contains 120 families. H_{clean} presents the new task B without prior discussion of target A; H_{focus} makes the earlier discussion of A accessible but omits the stopping instruction; H_A retains the same history and explicitly instructs the model to stop discussing A before carrying out B; and H_{swap} replaces the earlier history with an unrelated, length-matched context. Together, these conditions separate history-induced carryover from the effect of explicit conversational redirection.

Across the two diagnostics, we collect 3,120 answers. Qwen3.5-122B-A10B and Kimi-K3 independently score each response without condition labels, and the main analysis uses strict consensus. Full execution details, judge-specific results, agreement analyses, and robustness checks appear in Appendix B; bootstrap intervals are reported in Appendix F.

Single-turn: successful constraint following can coexist with visible echoing. Figure 3 shows that adding the negative constraint sharply improves substantive task behavior. Relative to REMOVED, the ORIGINAL condition reduces forbidden behavior by 14.63 pp and improves task completion, target satisfaction, and joint utility by 15.26, 20.86, and 15.53 pp, respectively. The constraint therefore provides useful boundary information in these requests rather than merely making generation more restrictive.

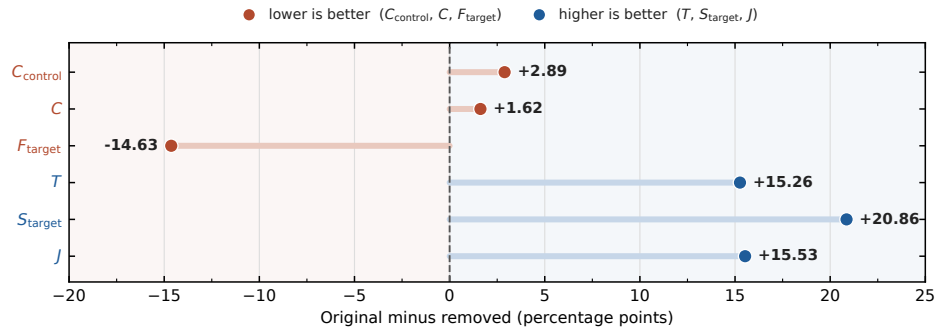


Figure 3: Single-turn paired diagnostic under strict consensus. Adding the constraint sharply improves task and target behavior while reducing forbidden behavior, but also introduces a smaller increase in visible rule/control echoing. Judge-specific results and bootstrap uncertainty analysis appear in Appendix B and Appendix F.

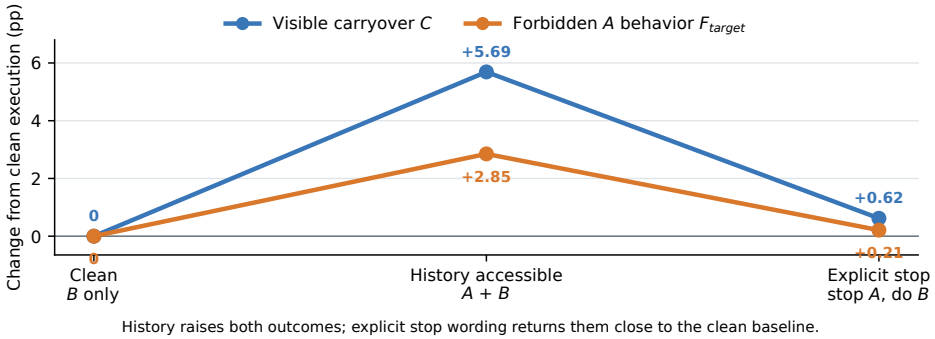


Figure 4: Multi-turn redirection trajectory under strict consensus. Making old-topic history accessible increases visible carryover and forbidden behavior. Explicit stop wording returns substantive behavior close to the clean-context baseline, while a small visible trace remains after redirection.

Yet successful constraint following does not eliminate visible traces of the constraint itself. Strict consensus estimates a $+2.89$ pp increase in rule/control echoing when the constraint is present. The broader C effect is smaller and varies more across judges (Appendix B). The two effects are therefore behaviorally distinct: the constraint produces large changes in substantive task behavior while also introducing a smaller, separately measurable tendency to surface constraint-related language in the response.

Multi-turn: successful redirection can leave a residual visible trace. The multi-turn diagnostic shows the same separation in a conversational setting. As shown in Figure 4, making the old-topic history accessible increases both visible carryover and substantive forbidden behavior: relative to H_{clean}, H_{focus} raises C by 5.69 pp and F_{target} by 2.85 pp. Adding explicit stop wording largely reverses both effects. Relative to H_{focus}, H_A reduces C by 5.90 pp and F_{target} by 3.12 pp.

After redirection, substantive behavior is nearly restored to the clean-context baseline: $H_A - H_{\text{clean}}$ is only $+0.21$ pp in F_{target}. Visible language, however, retains a small residual difference of $+0.62$ pp in C. Thus, even when the stop instruction successfully redirects substantive behavior, a detectable trace of that redirection can remain in the user-visible response.

4 Do Natural Opening Echoes Help the Answer?

4.1 Experimental design: replaying naturally occurring opening echoes

Explicit constraints can interfere with task solving even when they encode properties already satisfied by a model's successful answer (Qi et al., 2026). This raises a related but distinct question on the output side: once the model itself verbalizes a constraint as an opening echo, does that echo scaffold

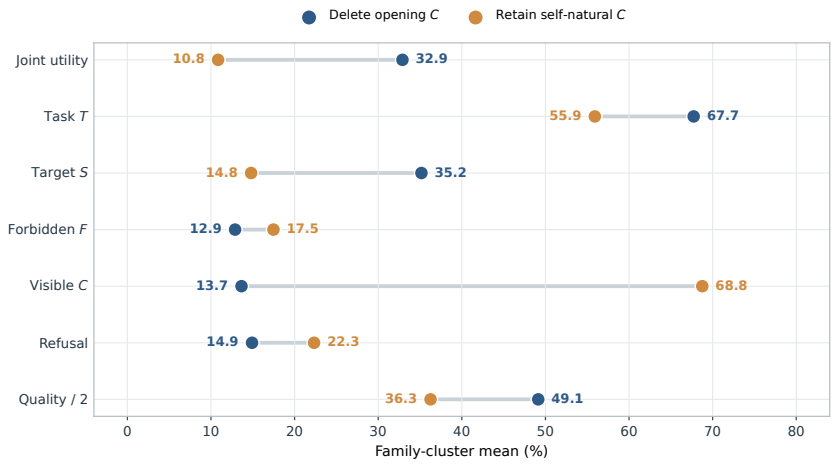


Figure 5: Absolute outcome rates after deleting or retaining a naturally occurring opening echo C. Retaining the echo sharply increases visible C while lowering joint utility, task completion, target behavior, and quality.

the continuation, or does it burden the answer? We test this with a replay intervention that deletes or retains naturally occurring model-generated echoes while holding the remainder of the observed response fixed.

We recruit prompt–model strata in which OLMo-2-7B-SFT or Tulu-3-8B-SFT naturally produces a complete, removable opening echo C. For each selected response, we split the observed output into the opening echo C and the remainder of the response, which we call the original suffix. The echo-retained condition restores C before this suffix, whereas the echo-deleted condition begins directly from the same suffix. The suffix is held byte-for-byte fixed, so the two conditions differ only in whether the natural echo is present before the model continues generation.

The frozen cohort contains 285 usable strata from 272 families. We generate the continuation under four matched seeds (temperature 0.7, top-$p = 0.9$), yielding 2,280 responses. We evaluate the intervention in two ways: the complete-response analysis measures the answer the user ultimately receives, while a continuation-only analysis excludes the replayed echo itself from evaluation and asks whether retaining it changes what the model generates afterward. Cohort composition, judging, and resampling details are reported in Appendix C.

4.2 Retaining the echo worsens the answer and alters the continuation

Figure 5 reports the complete-response outcomes. Retaining the natural echo reduces pooled joint utility by 22.34 pp, with the same direction for OLMo (-20.59 pp) and Tulu (-23.72 pp). At the pooled level, task completion declines by 12.20 pp, target satisfaction by 20.49 pp, and quality-at-maximum by 13.26 pp, while forbidden behavior rises by 4.59 pp. Retaining the model's own echo therefore substantially worsens the answer the user receives.

Part of the complete-response effect on C is mechanical: the echo-retained condition explicitly restores a span already labeled as C. The continuation-only analysis separates this direct contribution from the echo's effect on subsequent generation by excluding the replayed echo itself from evaluation. If the echo acts as a useful scaffold, retaining it should improve the behavior of the subsequent continuation.

Instead, retaining the echo reduces later C by 7.96 pp but increases refusal by 20.00 pp; task completion falls by 16.67 pp, target satisfaction by 17.47 pp, and joint utility by 15.24 pp. The echo therefore does affect subsequent generation, but not by providing a useful scaffold: less later repetition comes together with substantially worse task behavior. The reduction in later echoing should consequently not be interpreted as an improvement in the continuation.

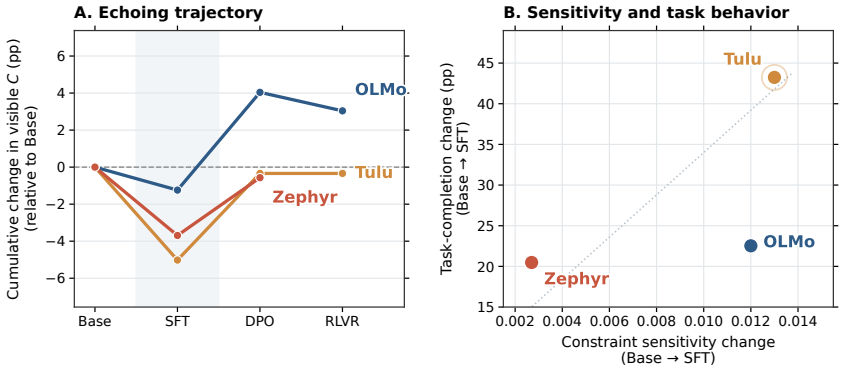


Figure 6: Post-training trajectories across public model lineages. From Base to SFT, visible echoing decreases while constraint sensitivity and task completion increase; visible echoing then rises consistently at the SFT-to-DPO transition. Panel A accumulates strict-consensus changes relative to each Base checkpoint, while Panel B shows the corresponding Base-to-SFT changes in teacher-forced sensitivity and task completion. Bootstrap uncertainty analysis appears in Appendix F.

5 Post-Training Separates Constraint Sensitivity from Visible Echoing

5.1 Visible echoing falls from Base to SFT and rises after DPO

We compare 11 public checkpoints across three lineages: OLMo-2-7B (Base, SFT, DPO, RLVR), Tulu-3-8B (Base, SFT, DPO, RLVR), and Mistral/Zephyr-7B (Base, SFT, DPO). The analysis contains 906 family-target records and 904 independent normalized prompt clusters. As in Section 3, Qwen and Kimi independently judge checkpoint outputs and the primary analysis uses strict consensus.

Figure 6 shows a consistent two-stage pattern across the three lineages. From Base to SFT, visible echoing decreases, while task completion improves substantially. The direction then reverses at the SFT-to-DPO transition: strict-consensus C rises by 5.28 pp, 4.68 pp, and 3.12 pp for OLMo, Tulu, and Zephyr, respectively. Both judges show the same SFT-to-DPO direction. DPO-to-RLVR, available for OLMo and Tulu, does not show a stable continuation of this pattern.

A narrower endpoint for active constraint-focused discussion near the beginning of the response shows the same SFT-to-DPO direction. Meanwhile, Base-to-SFT substantially improves T and S_{target} in all three lineages. The reduction in visible echoing therefore occurs alongside stronger task and target behavior rather than a general loss of capability.

Because these are bundled public-release comparisons, training data, templates, duration, formatting, objectives, and other recipe details co-vary. We therefore localize the most consistent increase in both aggregate C and active constraint discussion to the sampled SFT-to-DPO releases, without attributing the increase causally to DPO itself.

5.2 Constraint sensitivity rises as visible echoing falls

The Base-to-SFT reduction in visible echoing leaves an important question: does SFT make the model less sensitive to the constraint itself, or does it make the constraint less visible while preserving its influence on generation? We distinguish these possibilities with a teacher-forced sensitivity analysis.

The analysis contains 91 paired examples. Within each pair, we make a single local edit to the input that removes the target control cue while leaving the paired answer fixed. For each checkpoint, we then compare the token-mean log probability assigned to that same answer with versus without the cue. This contrast measures how strongly the local constraint changes the likelihood of a fixed continuation, separately from whether constraint-related language appears in free generation.

Teacher-forced sensitivity increases from Base to SFT in all three lineages, both in the original independent cohort and in a separately labeled extension. Visible echoing therefore moves in the

opposite direction from constraint sensitivity over this transition: the models become more sensitive to the constraint while making it less visible in their responses. Stronger conditioning on a constraint need not make that constraint more visible in the answer. Full cohort construction, coverage, judge-specific estimates, and the narrower constraint-discussion endpoint are reported in Appendix D.

6 Training for Silent Compliance: SFT and Optimization Controls

6.1 Permission-flip supervision targets selective omission

The preceding experiments suggest that the desired objective is selective omission: the model should omit target A when it is unnecessary while continuing to use it when authorized. Penalizing constraint-related language or response length directly would not enforce this distinction: low echoing alone is insufficient, since it can also be achieved by refusing or incompletely answering the request. Section 4 further shows that visible echoing need not benefit task performance, while Section 5 shows that constraint sensitivity and visible echoing can move separately. Together, these observations motivate training that reduces unnecessary echoing without weakening task completion or authorized target use.

We implement this objective with quality-controlled SFT and permission-flip coverage. Each training family contains a natural negative condition, where mentioning or using target A is unnecessary, and a minimally changed authorized condition, where A is required. For every prompt, we retain answers that satisfy safety, T, S_{target}, $F_{\text{target}} = 0$, and a minimum quality threshold. A deterministic hierarchy first matches the best primary-outcome level and quality, and only then chooses a concise target; brevity is not rewarded by itself. The resulting demonstrations therefore teach selective omission in the negative condition while preserving appropriate target use in the authorized partner.

We use two SFT variants. Ordinary SFT trains on quality-controlled targets, whereas Matched SFT additionally selects paired silent and authorized targets from the permission-flip families. The latter makes the omission/use boundary explicit in the supervision itself.

6.2 Frozen evaluation of selective omission

We evaluate the two SFT designs on a frozen 800-family permission-flip cohort disjoint from the final SFT training and validation bundle. As shown in Table 1, the Base checkpoint has $C = 4.77\%$ and $T = 90.60\%$. Across three training seeds, Ordinary SFT reduces the mean negative-condition echo rate to 3.77% but also slightly reduces task completion to 89.80%. Matched SFT performs best on both primary metrics, reducing echoing to 3.04% while increasing task completion to 92.55%. Thus, explicitly matched permission-flip supervision reduces unnecessary echoing without sacrificing task completion on the frozen evaluation cohort.

6.3 Do additional optimization controls improve on Matched SFT?

Given the stronger frozen-evaluation results of Matched SFT, we use it as the starting point for additional optimization controls. Table 1 reports offline preference learning and single-token and trajectory-wide credit assignment on Matched SFT checkpoints. The separate PCSE parameter-editing control and broader recipe sweep use distinct cohorts or estimands and are reported in Appendix E.

None of the three optimization controls shown here improves on Matched SFT across both primary metrics. Offline preference learning increases C by 3.50 pp and decreases T by 6.94 pp. Single-token credit assignment reduces C by 0.74 pp but also reduces T by 0.75 pp, while trajectory-wide credit assignment increases T by 1.34 pp but also increases C by 0.67 pp. The two credit-assignment variants therefore move the two objectives in opposite directions rather than improving both. Given their additional signal-acquisition and rollout costs, these optimization controls offer no clear advantage over quality-controlled Matched SFT in our setting.

7 Conclusion

Constraint use and constraint echoing are separable behaviors. Matched prompts show that models can successfully use a constraint while still leaving residual visible echoes. Our natural-prefix

Table 1: Supervised configurations and optimization alternatives. Offline preference learning uses pre-collected scored contrasts; single-token and trajectory-wide credit assign the optimization signal to one selected token or across the response, respectively. The panels are not a cross-method ranking.

Approach	**Echo result** ($C \downarrow$)	**Task outcome** ($T \uparrow$)
Frozen 800-family evaluation: absolute negative-condition rates		
Base	4.77%	90.60%
Ordinary SFT	3.77%	89.80%
Matched SFT	$\mathbf{3.04}\%$	$\mathbf{92.55}\%$
Study-specific optimization effects: paired ΔC and ΔT (pp) based on Matched SFT checkpoint		
Offline preference learning	$+3.50$ pp	-6.94 pp
Single-token credit assignment	-0.74 pp	-0.75 pp
Trajectory-wide credit assignment	$+0.67$ pp	$+1.34$ pp

intervention further shows that retaining an observed opening echo does not improve joint utility and instead degrades the measured answer. Across public post-training checkpoints, constraint sensitivity and task completion increase from Base to SFT, while the most consistent increase in visible echoing appears at the subsequent SFT-to-DPO transition, showing that stronger constraint sensitivity need not make the constraint more visible in the response.

Finally, in an independent 800-family evaluation across three training seeds, both ORDINARY SFT and MATCHED SFT achieve low echoing while maintaining high task utility. Together, these results show that models can learn to use constraints without unnecessarily verbalizing them, and that quality-controlled SFT provides a simple and practical route toward this form of silent compliance.

8 LIMITATIONS

Our results rely primarily on automated evaluators. Before experimental use, we calibrated the evaluation protocol on a small human-labeled development set of approximately 50 examples and froze the evaluators only after their decisions aligned with the human annotations. As an additional check on the resulting evaluation pipeline, we manually reviewed 100 randomly sampled strict-consensus decisions per experimental stage and found complete agreement with human labels. These checks support the reliability of the automated evaluation, but do not replace exhaustive human evaluation. In particular, C is a low-prevalence and semantically demanding outcome, so the reported absolute rates are best interpreted as measurements under our experimental conditions rather than estimates of deployment prevalence.

The scope of our data and interventions also limits generalization. The diagnostic data emphasize actionable omission constraints, the natural-prefix intervention primarily studies removable opening C_{control}, and the corpus is predominantly English with coverage limited to a set of open model families. The checkpoint analysis compares bundled public releases, for which training data, objectives, formatting, and other recipe details may co-vary; it therefore localizes where echoing changes across released checkpoints rather than attributing those changes causally to a particular post-training algorithm. Likewise, our optimization experiments establish quality-controlled SFT as an effective practical solution in the evaluated setting, but are not designed as a controlled general comparison between SFT and RL. These limitations bound the generality, rather than the central empirical conclusions, of the study.

ETHICS STATEMENT

Reducing visible control language is not always desirable. Safety refusals, uncertainty statements, legal disclosures, and explanations requested by users may be necessary. Our definition exempts necessary language, and MATCHED SFT selects only primary-feasible targets with authorized-use counterparts. Silent compliance may also reduce output-level auditability; the method should not be used to hide policy-relevant reasoning or transparency obligations.

Reproducibility Statement

Appendix F identifies data isolation, generation, evaluation, missingness, resampling, and replacement requirements. Frozen manifests, model revisions, score files, and artifact hashes are retained in the accompanying repository.

References

Sara Candussio, Francesca Padovani, Daniel Scalena, and Malvina Nissim. "Don't Say It!": Constraints, compliance, and communication when language models play taboo, 2026. URL https://arxiv.org/abs/2607.00601.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1301–1312, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.102. URL https://aclanthology.org/2021.naacl-main.102/.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257/.

Tomoyuki Kajiwara. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6047–6052, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1607. URL https://aclanthology.org/P19-1607/.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, pp. 47852–47870, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/771155abaae744e08576f1f3b4b7ac0d-Abstract-Conference.html.

Logan Mann, Nayan Saxena, Sarah Tandon, Chenhao Sun, Savar Toteja, and Kevin Zhu. Don't think of the white bear: Ironic negation in transformer models under cognitive load, 2025. URL https://arxiv.org/abs/2511.12381.

Jiho Park, Jongyoon Song, Minjin Choi, Kyuho Heo, Taehun Huh, and Ji Won Kim. TRUEBench: Can LLM response meet real-world constraints as productivity assistant? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14383–14409, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.776. URL https://aclanthology.org/2025.findings-emnlp.776/.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. One token away from collapse: The fragility of instruction-tuned helpfulness, 2026. URL https://arxiv.org/abs/2604.13006.

Alberto Purpura, Li Wang, Sahil Badyal, Gene Beaufrand, and Adam Faulkner. Deconstructing instruction-following: A new benchmark for granular evaluation of large language model instruction compliance abilities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1334–1349, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.62. URL https://aclanthology.org/2026.eacl-long.62/.

Yunjia Qi, Hao Peng, Xintong Shi, Amy Xin, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. On the paradoxical interference between instruction-following and task solving, 2026. URL https://arxiv.org/abs/2601.22047.

Aditya Sinha, Harald Steck, Vito Ostuni, and Matteo Rinaldi. Beyond continuity: Challenges of context switching in multi-turn dialogue with LLMs, 2026. URL https://arxiv.org/abs/2605.09268. Accepted to the ICBINB Workshop at ICLR 2026.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive interference reveals working memory limits in LLMs beyond context length, 2025. URL https://arxiv.org/abs/2506.08184. Accepted at the ICML 2025 Workshop on Long Context Foundation Models (ICFM).

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1684–1690, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.145. URL https://aclanthology.org/2021.findings-emnlp.145/.

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

A Annotation Specification

We annotate task behavior and visible constraint traces separately. Annotators first score main-task completion T, target-behavior satisfaction S_{target}, and substantive forbidden behavior F_{target}. They then identify candidate spans that may constitute constraint echoing. A positive C_{control}, C_{target}, F_{target}, or refusal decision requires a short verbatim quote that can be relocated in the visible response; otherwise the relevant label is abstain.

A candidate span is not labeled as echoing if it is requested by the user, necessary for safety, required to disambiguate the answer, or needed to preserve correctness, completeness, coherence, or requested format. This is a deletion test: only a span whose deletion preserves all of those properties is eligible for a C label. Otherwise, an unnecessary statement of the rule, compliance process, response focus, or transition away from excluded content is labeled C_{control}. Examples include "I will not discuss A," "As requested, I will avoid that topic," and "I will now focus only on B." An unnecessary comparison, reference, or organizational use of the excluded target is labeled C_{target}, e.g., "B is unlike the earlier A" when no comparison was requested. We define

$$C = C_{\text{control}} \vee C_{\text{target}}.$$

A target name occurring only inside a pure control statement is not double-counted as C_{target}, although distinct spans in the same response may receive distinct labels.

The criterion is functional rather than lexical. A comparison with A is not C when the user has requested an A/B comparison; similarly, a requested explanation of why the assistant will not discuss A is not an unnecessary trace merely because it contains the target. Necessary safety language is never counted as C. Strict lexical bans form a boundary case: if the user explicitly forbids the literal string A, emitting A in a compliance statement both verbalizes and violates the constraint and may therefore satisfy C_{control} and F_{target}.

Table 2: Judge-specific and strict-consensus diagnostic effects (percentage points). Each row reports the first condition minus the second. Bootstrap uncertainty for the primary contrasts is reported in Appendix F.

Setting	Outcome or contrast	Qwen	Kimi	Consensus
Single turn: Original − Removed	C	+1.94	+5.22	+1.62
	F_{target}	-12.50	-17.56	-14.63
	T	+16.22	+17.17	+15.26
	S_{target}	+18.44	+20.53	+20.86
	J	+15.72	+15.66	+15.53
Multi-turn: $H_{\text{focus}} - H_{\text{clean}}$	C/F_{target}	+6.94/+3.75	+6.32/+4.86	+5.69/+2.85
Multi-turn: $H_A - H_{\text{focus}}$	C/F_{target}	-4.79/-3.75	-3.33/-4.31	-5.90/-3.12
Multi-turn: $H_A - H_{\text{clean}}$	C/F_{target}	+2.08/0.00	+2.99/+0.62	+0.62/+0.21

B Diagnostic Design and Robustness

B.1 Construction and execution

The single-turn diagnostic contains 150 independent real-request clusters. Each has an ORIGINAL prompt containing the target negative constraint and a REMOVED prompt in which that constraint span is deleted while the affirmative task is retained. The multi-turn diagnostic contains 120 accepted real-source families. For each family, A is the old topic in the source history and B is the curated final request. Let H be the compacted real history ending immediately before B. The four frozen conditions are

$$H_A = H + \text{stop}(A) + B, \quad H_{\text{focus}} = H + B,$$
$$H_{\text{clean}} = B, \quad H_{\text{swap}} = H' + \text{stop}(A') + B,$$

where H' and A' come from a different, adjacent length-matched donor family. Donors are formed by sorting histories by length and reciprocally pairing adjacent families, preventing self-donation. The fixed stop prefix instructs the model to stop discussing A, not mention, compare with, allude to, or continue it, and then to complete B directly.

We execute every condition with OLMo-2-7B-Instruct, Llama-3.1-Tulu-3-8B, Zephyr-7B-β, and Qwen3.5-9B under native chat templates, greedy decoding, and explicit thinking disabled, producing 3,120 answers. The formal cap is 1,500 tokens. A condition-blind completeness rule reruns cap hits at 3,000 tokens. The final set contains no empty answers and 44 unresolved truncations; these remain unresolved rather than being mapped to negative labels.

B.2 Evaluation, consensus, and robustness

Before experimental use, we calibrate the evaluation protocol on a small human-labeled development set of approximately 50 examples. During judge development, the evaluation criteria and judge prompts are refined against these human annotations and frozen before experimental evaluation. This development set is used for evaluator calibration rather than as an independent estimate of evaluation accuracy.

Qwen3.5-122B-A10B and Kimi-K3 then evaluate every experimental response independently without condition labels. The judge receives the complete prompt, family-specific rubric, visible answer, and finish reason. Positive C, F_{target}, and refusal decisions must be supported by a relocated verbatim span; malformed, uncertain, or unlocatable cases remain abstentions. Main-text diagnostic results use strict consensus: a yes label is retained only if both judges independently return yes, while disagreement and any abstention remain unresolved. A length-truncated response cannot retain a positive T label, because completion is unknown.

Table 2 reports the judge-specific estimates underlying the strict-consensus effects in the main text. The two judges agree on the direction of the principal single-turn and multi-turn contrasts, although the magnitude of visible echoing varies across evaluators.

Table 3: Agreement between the two automatic judges over all 3,120 diagnostic responses.

Outcome	Agreement (%)	Cohen's κ	Qwen yes	Kimi yes
C_{control}	95.6	.379	122	58
C_{target}	96.7	.369	57	59
C	94.4	.445	168	109
F_{target}	92.3	.681	352	424
T	82.1	.616	1,869	2,202
S_{target}	88.2	.644	2,413	2,551

Nineteen multi-turn families exceed the prespecified lexical-relatedness threshold between A and B. Excluding these families preserves both the positive history effect and the negative stop-wording effect on C, indicating that the redirection pattern is not driven by highly related old and new topics.

As a secondary robustness check, the unrelated, length-matched swap control remained close to clean execution: the strict-consensus macro effect for $H_{\text{swap}} - H_{\text{clean}}$ was $+0.63$ pp for C and -0.56 pp for F_{target} (Qwen: $+1.25/-0.63$ pp; Kimi: $+0.83/-0.63$ pp). Thus, within this diagnostic, adding an unrelated stopped history does not reproduce the carryover observed when the accessible history concerns the target topic.

Each individual judge identifies approximately 2% union-C in H_A, yet no response is positive for union-C under strict consensus. This illustrates the conservativeness of the consensus endpoint for low-prevalence visible echoes. Paired effects are computed only on units resolved under both conditions and therefore need not equal differences between marginal condition rates.

Table 3 reports agreement between the two automatic judges over all 3,120 diagnostic responses. As a post-development check separate from the calibration set, for each experimental stage we randomly sample 100 strict-consensus decisions for human review. Human labels agree with every sampled strict-consensus decision (100% agreement). All abstentions remain missing. The diagnostic contrasts and endpoint set were specified jointly; model, condition, and checkpoint decompositions are descriptive.

C Natural-Opening Replay Cohort and Secondary Analysis

The effective cohort contains 285 prompt–model strata from 272 families (90 OLMo and 195 Tulu). It includes 184 nonempty original suffixes and 101 cases ending at the opening boundary. Four paired replay seeds yield 2,280 full responses; 25 generations are length-ended, and 26 full-response rows remain unresolved. The continuation-only view uses the prespecified seed 17 and contains 570 responses, seven of which remain unresolved.

Table 4 reports the secondary decomposition used to interpret the main result. The full-response analysis measures the answer the user ultimately receives; the continuation-only view removes the mechanically restored opening and asks whether retaining the echo changes the subsequent generation. Absolute rates, source breakdowns, and token-level audit details are retained in the project artifact rather than repeated here.

Table 4: Pooled effects of retaining the natural opening (percentage points). Full-response results average four paired seeds; continuation-only results use prespecified seed 17. The full-response C effect includes the mechanically restored opening.

Outcome	Full response	Continuation only
J	-22.34	-15.24
T	-12.20	-16.67
S_{target}	-20.49	-17.47
F_{target}	+4.59	-0.93
C	+55.36	-7.96
Refusal	+6.81	+20.00
Quality at maximum	-13.26	–

D Post-Training Checkpoint and Sensitivity Analysis

D.1 Checkpoint evaluation

The checkpoint comparison covers 11 public releases from three lineages: OLMo-2-7B (Base, SFT, DPO, RLVR), Tulu-3-8B (Base, SFT, DPO, RLVR), and Mistral/Zephyr-7B (Base, SFT, DPO). The cohort contains 906 family-target records and 904 independent normalized prompt clusters. Within a lineage, the same frozen lineage chat formatting is used across stages, including Base checkpoints, so that changes are not attributed to raw-completion versus chat serialization.

Qwen and Kimi evaluate checkpoint outputs independently, and the primary stage-localization analysis uses strict consensus. Strict-consensus usable coverage is 85.28% for C, 82.20% for F_{target}, 74.82% for T, and 69.92% for S_{target}. Because consensus therefore does not form a complete-case sample, we retain both individual-judge analyses alongside consensus in the analysis artifact. Both individual judges show an increase in C from SFT to DPO in all three lineages, matching the direction of the strict-consensus result reported in the main text.

The narrower endpoint for active constraint-focused discussion in the first assistant segment provides a complementary localization check. Its strict-consensus SFT-to-DPO changes are +6.07 pp for OLMo, +3.68 pp for Tulu, and +3.41 pp for Zephyr; unresolved-case worst-case bounds weaken the evidence for the Zephyr increase.

D.2 Teacher-forced sensitivity cohort

The teacher-forced analysis provides a complementary measurement of whether the target control cue continues to affect relative likelihood when visible echoing changes. It contains 91 frozen answer pairs: an independent 44-pair cohort frozen before checkpoint likelihood scoring, and a separately labeled 47-pair post-hoc extension. The extension passed an independent automated eligibility audit, but did not meet its separate 180-pair frozen-cohort gate; the pooled analysis is therefore retrospective and automatic-only, not confirmatory or human-validated. The 44-pair cohort contains single-turn prompts, whereas the 47-pair extension contains multi-turn prompts in which only the final user turn is edited.

In every pair, x_{removed} is obtained by deleting one contiguous target cue from the final user message of x_{original} while holding the preceding message history fixed. Each pair contains two constructed, frozen continuations. y_+ follows the target cue and y_- violates it while answering the same main task; the two continuations are identical outside one contiguous answer-side replacement. Let s_+ and s_- denote the respective replacement spans in y_+ and y_-.

For each prompt variant $x \in \{x_{\text{original}}, x_{\text{removed}}\}$, we teacher-force both continuations and exclude prompt tokens. If $\ell_x(y)$ denotes the mean answer-token log probability, let $\ell_x^{\text{local}}(y_+)$ and $\ell_x^{\text{local}}(y_-)$ denote the corresponding means over answer tokens whose character offsets overlap s_+ and s_-, respectively. The primary score is

$$A_{\text{local}} = [\ell^{\text{local}}_{x_{\text{original}}}(y_+) - \ell^{\text{local}}_{x_{\text{original}}}(y_-)] - [\ell^{\text{local}}_{x_{\text{removed}}}(y_+) - \ell^{\text{local}}_{x_{\text{removed}}}(y_-)].$$

Thus, the estimand measures the cue-induced change in the relative likelihood of the cue-following versus cue-violating continuation. We also retain full-answer mean and summed-log-probability variants. Character-offset overlap is applied separately within each frozen answer to identify tokens intersecting its answer-side replacement span, avoiding tokenizer boundary merges or splits that can arise from prefix-token indexing.

Base-to-SFT sensitivity change is computed pairwise as the SFT score minus the Base score for the same answer pair. The original 44-pair cohort and the 47-pair automatic-only extension show the same positive direction for local sensitivity, bringing the post-hoc pooled analysis to $n = 91$. This analysis is complementary to, rather than a replacement for, free-generation echoing.

E METHOD CONFIRMATION AND RESOURCE COST

E.1 TRAINING PROTOCOL

Each permission-flip family contains a negative condition, in which target A is unnecessary, and a minimally changed authorized condition, in which A is required. Training targets pass safety, task completion, target satisfaction, zero forbidden behavior, and a quality threshold. ORDINARY SFT uses quality-controlled targets selected without conditioning target selection on echoing, whereas MATCHED SFT additionally requires a silent primary-feasible target in the negative condition and a primary-feasible authorized partner from the same family.

The actual three-arm SFT bundles are constructed from 820 primary-feasible families selected from a scored 1,000-family pilot. Of these, 700 families are assigned to training and 120 to validation. Each family contributes a negative and an authorized condition, yielding 1,400 training rows and 240 validation rows per arm. The Silent-unmatched control uses the same silent negative targets as Matched SFT but cyclically permutes authorized partners within each split, removing the within-family permission match. All three trained configurations use seeds 0–2.

The frozen 800-family evaluation reported in the main text is disjoint from the final 820-family SFT bundle. Full source-pool provenance, exclusion rules, and data-partition details are reported in Appendix F.

E.2 ADDITIONAL OPTIMIZATION CONTROLS AND RESOURCE COST

Additional optimization methods require an informative contrast between feasible echoed and silent trajectories. The frozen signal gate scored 16,000 trajectories but found only 123 informative families (12.3%), below its prespecified threshold. We nevertheless evaluated additional optimization controls from Matched SFT checkpoints. As reported in Table 1, offline preference learning increases C by 3.50 pp and decreases T by 6.94 pp; single-token credit assignment reduces C by 0.74 pp but also reduces T by 0.75 pp; and trajectory-wide credit assignment increases T by 1.34 pp but also increases C by 0.67 pp. None therefore improves on Matched SFT across both primary metrics. Together with the signal-acquisition and rollout requirements reported below, these results provide no clear practical advantage over quality-controlled Matched SFT in our setting.

Method identities. Permission-contrastive stage editing (PCSE) is a parameter-efficient LoRA edit trained on permission-flip pairs and compared with an ordinary LoRA arm from the same initialization and budget. It is a parameter-editing control, not a replacement for the separately evaluated full-SFT configurations. Offline preference learning optimizes pre-collected scored preferred/dispreferred contrasts. Single-token credit assignment attaches the optimization signal to one selected output token, whereas trajectory-wide credit assignment applies it across eligible response tokens. These explorations use their own matched controls and cohorts; their effect sizes are consequently not comparable to the common frozen SFT rates.

Table 5: Parameter-editing control on a separate 600-family OLMo-DPO/LoRA cohort: absolute negative-condition rates. Ordinary LoRA and PCSE LoRA share the same OLMo-2-7B-DPO initialization and LoRA budget.

Method	$C \downarrow$	$T \uparrow$
Ordinary LoRA	4.51%	32.55%
PCSE LoRA	3.52%	36.79%

Recipe sweep. A separate development-only sweep evaluates ten trajectory-wide training recipes at the same 333-update, two-pass exposure budget on 149 families. It changes the data mixture, contrast exposure, unlikelihood weight, retention weight, learning rate, or selected span. It has no single pooled paired effect and is therefore not condensed into one main-table row. The observed trade-off is representative: real-SFT has $C = 4.70\%$, $T = 67.79\%$; a weaker-unlikelihood recipe has lower $C = 2.68\%$ but lower $T = 64.43\%$ and higher forbidden behavior; an eight-contrast recipe has $C = 6.04\%$ and $T = 69.13\%$. Thus the sweep did not identify a stable recipe improving echoing without an offsetting outcome cost. Arm-level records are retained in the artifact.

Table 6: Signal acquisition and rollout cost for additional optimization. Trajectories within a family are not independent statistical units.

Stage	Candidates or families	Trajectories	Outcome
Natural pool proposal	12,000	—	11,341 eligible
Independent review	11,341	—	5,680 accepted; 5,000 frozen
Frozen signal-gate pilot	1,000	16,000	123 informative (12.3%); failed gate
Contrast recruitment	1,600	19,200	203 informative families
Full-trajectory exploration	1,104	52,992 per arm	18 fresh-rollout updates per arm

F Reproducibility Details

Family or normalized prompt cluster is the resampling unit throughout. Diagnostic and training analyses use 10,000 cluster-bootstrap draws; scaffold analyses use 20,000. Abstentions, judge failures, invalid evidence spans, and unresolved truncations are never converted to negative labels. The diagnostic, scaffold, and method families follow their frozen execution and selection manifests; held-out evaluation data are not used for model or checkpoint selection.

Models, inference, and judging. The diagnostic uses OLMo-2-7B-Instruct, Llama-3.1-Tulu-3-8B, Zephyr-7B-β, and Qwen3.5-9B under each model's native chat template, greedy decoding, and disabled explicit thinking. The released-checkpoint and natural-prefix studies use frozen public model revisions and execution manifests. Exact revision identifiers, templates, decoding settings, canonical generations, and SHA256 checksums are included in the artifact.

Qwen3.5-122B-A10B-FP8 is the local primary judge; Qwen3.5-122B-A10B and Kimi-K3 independently judge the diagnostic and checkpoint studies. Judge calls use temperature 0 and disabled thinking. Positive C/F labels require a re-locatable verbatim span; malformed output, low confidence, invalid evidence, truncation that prevents assessment, and service failure yield abstention. The scoring prompts, raw judge responses, retry logs, and the exact model snapshot manifests are included in the artifact.

Data and training. The diagnostic and permission-flip families are controlled transformations of `allenai/WildChat` conversations (ODC-BY); the source snapshot and all derived files are hash-recorded.

The natural permission-flip pipeline uses a frozen 5,000-family source pool. A 1,000-family pilot from this pool is used for trajectory generation and feasibility screening. Of these pilot families, 820 satisfy the paired primary-feasibility requirements used to construct the three SFT arms. A deterministic split assigns 700 families to training and 120 to validation. Each family contributes a negative and an authorized condition, yielding 1,400 training rows and 240 validation rows per arm.

Ordinary targets use primary-outcome and quality criteria without conditioning target selection on C. Matched targets require a primary-feasible silent negative target and a primary-feasible authorized partner from the same family. Silent-unmatched uses the same silent negative targets as Matched SFT but a fixed cyclic permutation of authorized partners within each split. All targets pass safety, T, S, $F = 0$, and the prespecified quality threshold; the selection manifests and source SHA256 values are included in the artifact.

An earlier frozen v1 partition bundle contains 1,149 families from the same source pool, assigned by SHA256 of `family_id` to 800 target-generation, 149 development, and 200 nominal held-out families. This partition bundle is not the final SFT training split. The frozen 800-family evaluation reported in the main text is produced by a later evaluation freeze from the same 5,000-family source pool after excluding the 1,000-family pilot and all families in the v1 partition bundle. The resulting evaluation cohort is disjoint from the final 820-family SFT bundle.

All three SFT arms start from Qwen3.5-9B and use seeds 0/1/2, full-parameter BF16 training, four CUDA processes, DeepSpeed ZeRO-2 with CPU optimizer-state offload, microbatch 1 per GPU, no gradient accumulation, gradient clipping 1.0, AdamW/DeepSpeedCPUAdam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01), cosine scheduling, learning rate 10^{-6}, 5% warmup, maximum length 32,768, one epoch, and 175 optimizer updates. Greedy evaluation uses temperature 0, top-$p = 1$, top-$k = -1$,

maximum model length 32,768, and at most 16,384 new tokens. The training command, DeepSpeed configuration, environment versions, execution logs, and per-run wall-clock records are included in the artifact; no LoRA adapters are used for these full-SFT arms.

Table 7: Experiment and evaluation summary for the retained results.

Experiment	Independent unit	Evaluation design	Missing-data rule
Single-turn diagnostic	150 families	Independent Qwen and Kimi judges	Abstain retained; consensus reported
Multi-turn diagnostic	120 families	Independent Qwen and Kimi judges	Abstain retained; consensus reported
Natural-prefix scaffold	272 families / 285 strata	Qwen; Kimi for unresolved cells	Abstain retained; family-equal pooling
Frozen SFT evaluation	800 families	Automatic local 122B judge	Resolved family-paired labels
Additional optimization	Study-specific training families	Automatic checkpoint evaluation	Metric-specific resolved denominators

All paired effects are computed only on units resolved under both conditions. Marginal condition rates use all resolved labels for that row and therefore need not subtract to the reported paired effect.

Table 8: Bootstrap uncertainty summary for the primary empirical claims. Intervals use family-cluster resampling; fuller output is retained in the project artifact.

Claim	Point estimate	95% bootstrap interval
Single-turn constraint reduces forbidden behavior	$-14.63\,\mathrm{pp}$	$[-20.28, -9.27]$
Single-turn constraint improves joint utility	$+15.53\,\mathrm{pp}$	$[9.40, 21.87]$
Natural-opening retention reduces joint utility	$-22.34\,\mathrm{pp}$	$[-27.66, -17.12]$
History raises visible carryover	$+5.69\,\mathrm{pp}$	$[3.06, 8.75]$

G LLM USAGE

In accordance with the ICLR author guidelines, a language model assisted with organizing existing experiment notes and drafting LaTeX prose. The authors are responsible for verifying every claim, citation, numerical result, and interpretation. Experimental design, data, executions, and final scientific decisions remain the authors' responsibility. Training comparisons identified as automatic-only evidence are not represented as human-gold observations.