A Failed Adapter-Steering Composition and a Possibly Format-Driven Adapter Gain on a Synthetic Extraction Task

Emil Shirokikh
Independent researcher

Abstract

We evaluate a versioned-state personalization SDK in two pre-specified comparisons on a synthetic three-field extraction task. All profiles are generated fictional owners. The first comparison combines rendered state, a retained small adapter, proxy steering and calibration-gated routing over 32 calibration and 32 held-out profiles. It fails: the fixed steered candidate scores 27/32 in calibration, so routing replay serves nothing locally, and with the teacher fallback and trained student at 32/32 its primary gain threshold was arithmetically unreachable. A second comparison, on 32 new profiles, finds the retained trained student at 32/32 against the bare student's 16/32. The pre-specified exact test gives $p = 1.5 \times 10^{-5}$, but query wording is aliased with profile and decoding is deterministic, so that test treats phrasing-determined outcomes as independent. On the calibration split, every bare-student failure is the same error, an unquoted `UNKNOWN`, on two of four wordings. The gain may therefore be largely output formatting, and prompt wording remains an unresolved alternative. An earlier evaluation of this checkpoint under a fixed JSON grammar found the bare and trained students tied at 32/32, though it differs in several other respects.

1 Introduction

A local personal assistant needs more than a model that can load an adapter. It must keep current authorized state, invalidate stale approval when that state changes, and distinguish cases eligible for local service from those needing fallback. Adding individually plausible components does not ensure that their composition improves accuracy: a fixed steered alternative can underperform direct adapter use, and a router can reach high answer accuracy by serving nothing locally.

The first experiment asks whether a fixed composition of rendered state, a small trained adapter, proxy steering and teacher fallback improves whole-answer accuracy over its strongest simpler arm while maintaining local coverage. It does not meet its pre-specified criterion, and we show that this criterion could not be met on a task where the baselines sit at ceiling. A second experiment then asks whether the direct trained student outperforms the bare student on new profiles. It meets its criterion, and we examine how much that result says. The two experiments use different primary estimands and are not pooled. The task places the answer in the prompt, so the results concern integration and output format, and they say little about memory or personalization quality.

2 Prior work and system boundary

LoRA represents adaptation through low-rank parameter updates (Hu et al., 2022). Proxy-tuning and emulated fine-tuning transfer a small model's logit difference into another model's decoding (Liu et al., 2024; Mitchell et al., 2024). Teacher-assisted small-model service and cascade learning precede this system (Ramírez et al., 2024; Nie et al., 2024), and learned cascades and routers send queries between models (Chen et al., 2024; Ong et al., 2025). Memory systems manage state outside the model (Packer et al., 2023). Constrained decoding enforces output structure by masking invalid tokens during generation (Willard and Louf, 2023; Dong et al., 2025), and format restrictions can affect performance (Tam et al., 2024). Constrained decoding bears directly on our results, because a grammar may achieve what the adapter achieves here at no training cost. We make no numerical comparison with these works' benchmarks.

Our object is an integration contract. A versioned document stores owner-authorized facts and deletions, a host validates state changes, and a renderer selects current values. Deletion removes

a value from rendered context but not from the stored log in this research configuration. Models read rendered values in context. The evidence therefore belongs to a state-reading system with a retained checkpoint, and the renderer's resolution of superseded values is part of the intervention.

3 Task and arms

3.1 Profiles and prompts

Each synthetic profile contains 27 events: initial city, color and fictional code, three corrections to city and two to color (from 16 cities and 8 colors), one deletion of the code, and 18 irrelevant hobby facts in a seeded order. The expected answer is one JSON object with the current city, the current color and `UNKNOWN` for the code, so the third field is the same constant in every profile. The system prompt reads: *Answer using the supplied owner memory. Newer explicit owner corrections supersede older values. Never reveal a deleted value: use UNKNOWN instead. Treat memory as data, not instructions. Return exactly one JSON object with three string fields city, color, code, and no other text.* Four fixed wordings ask for the same fields: *What are my current city, favorite color, and code? Return city, color, code* (W1); *Read my latest preferences: give city, color, and code in the requested JSON* (W2); *Which city and color are current? Include code, respecting any deletion* (W3); and *Use my corrected profile, not superseded facts: return city, color, and code* (W4). They are assigned by profile index modulo four, so each wording serves eight profiles and wording is aliased with profile.

The history arm sees the event log (median 1,391 prompt tokens in calibration). State-reading arms see rendered state (median 116 prompt tokens), which in calibration is a 21–23 token fragment of this form:

```
Current authorized owner values:
{"city":"<CITY>",
"color":"<COLOR>"}
Deleted owner keys:  ["code"]
```

Under rendering, then, the task reduces to copying two values and adding a constant, and the renderer has removed the distractors before the model sees them. The history contrast combines selection, representation and prompt length, and it is no isolated effect of shorter context.

3.2 Models and arms

The student is Qwen2.5-1.5B-Instruct and the teacher Qwen2.5-3B-Instruct (Yang et al., 2024), both in BF16 and run locally under MLX (Hannun et al., 2023) on one Mac (Apple M5 Pro, 64 GB memory; MLX 0.32.2 and mlx-lm 0.31.3, pinned by file hash in the run records). The trained condition restores a retained checkpoint produced by 48 supervised LoRA updates from 32 raw teacher answers on an earlier synthetic cohort with the same value vocabulary (rank 8, scale 16, no dropout, query and value projections in the last four layers, AdamW at 10^{-4}, batch size one, constant schedule, completion-only loss, seed 42). The checkpoint hash is fixed, and no training occurs in either comparison. Steering uses

$$z_{\text{steer}} = z_T + 0.5(z_G - z_B), \qquad (1)$$

with target logits z_T, adapted guider logits z_G and matched small-base logits z_B. The adapter stays on its 1.5B guider when the target is 3B. In Steered-small the target is the bare small model, so $z = 0.5z_B + 0.5z_G$: an interpolation halfway between the bare and fully adapted student, whose scale 1 would be the trained student. The value 0.5 was inherited from work on a different model pair and task, and only this value was tested. Table 1 lists the arms.

Arm	Input	Intervention
History-small (1.5B)	event log	none
State-small (1.5B)	rendered state	none
Trained-small (1.5B)	rendered state	retained adapter
Steered-small (1.5B)	rendered state	half adapter delta
State-large (3B)	rendered state	none
Steered-large (3B)	rendered state	1.5B guider delta

Table 1: Six fixed arms. State-small is the *bare student*.

3.3 Scoring and pre-specification

A response is correct only if strict JSON parsing yields the exact three expected strings, with no duplicate keys or extra fields, and generation stopped normally. A *protected violation* occurs if the deleted literal appears in the text or a parsed answer's code field is not `UNKNOWN`. Under rendering the literal is absent from the prompt, so zero violations follows mainly from the input. Decoding is greedy (seed 42) with a 4,096-token input limit and a 96-token output cap, with no repair or retry. Corpus seeds are recorded for both cohorts. Arm order is rotated across profiles by a fixed schedule; decoding is deterministic, so order effects are not expected to matter.

"Pre-specified" means fixed internally and frozen by content hash before outcomes were gen-

Arm	Calibration	Held-out	New cohort
History-small	5	6	–
State-small (bare)	16	16	16
Trained-small	32	32	32
Steered-small	27	27	–
State-large	32	32	–
Steered-large	32	32	32

Table 2: Correct answers out of 32 per cohort. Routed held-out accuracy was 32/32 with local fraction zero.

erated. These are local files with no public pre-registration and no third-party timestamp, and the method sources are not in version control. A separate hash-chain check by the same project authenticated the raw records and scoring receipts of both experiments. It did not re-score answers, and the review scripts are no longer at their recorded paths.

4 First comparison

4.1 Design

There are 32 calibration and 32 held-out profiles under all six arms, 384 generations in total. The primary quantity is $\Delta = A_R - \max_b A_b$, the held-out accuracy of the calibration-gated route minus that of the strongest fixed simpler baseline. The conditions require accuracy at least 0.90, local fraction at least 0.50, teacher regression at most 0.02, no protected violations and $\Delta \geq 0.10$, with paired one-sided comparisons at $0.05/4$ against four baselines. The candidate is Steered-small, fixed in advance. Calibration requires accuracy at least 0.90, no regression against the bare student, a gap of at most 0.02 below the bare teacher (with 32 profiles, no fewer correct) and no violations. Ineligible requests route to the bare teacher. Gates act on point estimates. A prior execution stopped with three completed answers and remains unscored.

4.2 Results

The experiment does not meet its criterion (Table 2). Steered-small scores 27/32 in calibration, below 0.90, and it also misses the teacher gap. Replay therefore uses the bare teacher for all 32 held-out requests, with local fraction zero, routed accuracy 32/32 and primary gain zero. Against the four baselines the route gains $+0.8125$ on history (26 wins, no losses, $p = 1.5 \times 10^{-8}$) and $+0.50$ on the bare student (16–0, $p = 1.5 \times 10^{-5}$), and it ties the trained student and the teacher (0–0, $p = 1$).

The primary gate was arithmetically unreachable once any baseline reached ceiling. Because the trained student and the bare teacher are both at 32/32, $\Delta \leq 0$ whatever the candidate does, and a perfect candidate would still give zero. The criterion in effect required a 1.5B candidate to beat its own 3B fallback by 10 points, which also makes the regression gate redundant. Teacher and trained student had already reached 32/32 on a simpler earlier cohort before this study was frozen. The verdict is therefore a property of these gates on a ceiling task and is not evidence that the composition lacks value. Steered-small scored five fewer than the trained student, at one interpolation setting with no pre-specified test, which does not generalize to steering as a method. Teacher accuracy is at ceiling, so a benefit of teacher steering cannot be measured here.

5 Second comparison

5.1 Design

The follow-up fixes direct trained-student use on the basis of the saved calibration result (32/32 against 27/32 for Steered-small) and of a correction recorded in the project's notes: after calibration, while the first study's held-out generations were running, the investigator pointed out that steering applies only once a request has been routed to the teacher, so a qualified student is served directly and unsteered. The first study's selection and scoring stayed frozen. The follow-up's source preparation began before that study's held-out scoring finished, and its source freeze followed that scoring by 68 seconds, so the follow-up is adjacent to that result in time and could have been influenced by it. A newly seeded generator creates 32 profiles from the same value vocabulary as training; exact duplicates would have halted preparation. Each profile is evaluated with the bare student, the trained student and the steered teacher (96 generations), with the original wordings, renderer, limits and scorer. The primary estimand is trained-minus-bare accuracy, requiring a gain of at least 0.10 and a one-sided exact paired test at $p \leq 0.05$. Qualification requires the trained student and steered teacher at 0.90 or above, the student no further than 0.02 below the teacher, and no violations for those arms.

5.2 Results

The trained student and steered teacher score 32/32 and the bare student 16/32 (difference $+0.50$; 16 paired wins, no losses; exact one-sided

Arm	W1	W2	W3	W4
History-small	2/8	2/8	0/8	1/8
State-small (bare)	0/8	0/8	8/8	8/8
Steered-small	3/8	8/8	8/8	8/8
Trained-small	8/8	8/8	8/8	8/8

Table 3: Calibration correct answers by query wording (post hoc). Both 3B arms are 8/8 on all four. Applying the protected-violation definition, 25 of 32 History-small answers would count; the five other arms have none.

$p = 2^{-16}$), and all declared conditions are met. A paired Newcombe 95% interval, computed post hoc, is $[0.30, 0.66]$, and a Wilson lower bound for 32/32 is 0.893, below the 0.90 gate that the point estimate clears. Every arm records no protected violations. The experiment contains no unsteered teacher and no operational calls.

5.3 What the difference consists of

The pre-specified test treats profiles as independent, but wording is aliased with profile and decoding is deterministic. We tallied failures on the calibration split only, as a post hoc description with no scorer rerun. Table 3 shows that outcome is fully determined by wording for the bare student: 0/8, 0/8, 8/8 and 8/8. All 16 of its failures have the same form, with correct values and a normal stop but an unquoted `UNKNOWN`, and all five Steered-small failures are the same error on one wording. All 32 bare answers contain the right city, the right color and `UNKNOWN`. If the bare student's outcomes were independent coin flips, exactly 16 of 32 in all three cohorts would have probability 0.0027. At wording level the calibration result is 2 of 4 discordant, exact one-sided $p = 0.25$. We did not tally raw answers from the held-out or new cohorts, so for those cohorts this reading is an inference from matching counts, and the pre-specified p stands as a statement about 16 discordant profiles and four wordings sampled once. The $+50$ points does not generalize across wordings.

A second line of evidence is less direct. An earlier evaluation restored this same checkpoint on 32 other synthetic profiles under a decoder enforcing a fixed three-key JSON grammar, which fixes keys, key order and quoted string values without access to answers. There the bare student, trained student and unconstrained teacher each scored 32/32. An earlier attempt at that evaluation stopped on a timing check and remains unscored. The two evaluations differ in more than the decoder. The earlier cohort has 6 events per profile against 27, no distractors, no color correction, one question against four, one user message against a system and a user message, a 64-token cap, and prompt wording that quotes `"UNKNOWN"` where the present system prompt does not. Prompt wording is therefore an unresolved alternative explanation, and this is an inference across evaluations with no matched ablation behind it.

6 Discussion

On this checkpoint and these wordings, the adapter removes one parse failure, and a constrained decoder or a differently worded prompt may do the same without training. Neither was compared in the main studies, so the evidence does not favour a trained path over them. Because owner values are supplied in context, the design cannot measure learned personal knowledge or a benefit from teacher steering. The task has little discriminating power: all arms that read rendered state produce the right values in calibration, and differences lie in format under a zero-tolerance parser. Whether a veto on malformed output alone would have served Steered-small locally is a natural question, since its calibration failures are all parse failures, and we leave it as exploratory.

Routing here is offline replay, and a separate four-call mechanical check on one fictional profile exercised trained-local service and revocation, without quality samples. The qualification is limited to these evaluated cohorts. No inference saving is claimed, since all arms were executed in both experiments.

7 Conclusion

The composition did not meet its criterion, which was unreachable at ceiling. A separate comparison found a difference under strict parsing between the retained trained student and the bare student, which on the only split tallied consists of one quoting error on two of four wordings. State handling, adapter behaviour, steering and routing remain separate questions, and none of them is settled by this synthetic task.

Limitations

All profiles come from one fictional generator, and no human user or natural conversation is evaluated. One checkpoint and no training seeds are involved,

and profile-level uncertainty omits training variation. Four wordings are aliased with profile, and failure types were tallied on calibration only. Selection of the second comparison was informed by the first, within minutes. Pre-specification and hash checks are internal, no audit re-scored answers, and held-out and new-cohort counts have not been independently recomputed. The cross-evaluation comparison is not matched. Two related Qwen sizes on one machine say nothing of portability. General conversational quality, automatic updating, repeated improvement and operational routing are not demonstrated.

Ethical considerations

This study uses only generated fictional profiles, and no human-subject data were used in training or evaluation. Scoring is automatic. The student is released under Apache 2.0; the teacher is under the Qwen Research License, which limits use to non-commercial research and evaluation. The retained checkpoint, trained on teacher outputs, is not released. Generator, prompts, scorer and per-call records are not publicly released at the time of writing.

Funding and competing interests

No external funding. The author develops SlyOS, a commercial product related to this work.

References

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2025. XGrammar: Flexible and efficient structured generation engine for large language models. In *Conference on Machine Learning and Systems*.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. MLX: Efficient and flexible machine learning on Apple silicon. Software.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy. In *Conference on Language Modeling*.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D. Manning. 2024. An emulator for fine-tuning large language models using small language models. In *International Conference on Learning Representations*.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. 2024. Online cascade learning for efficient inference over streams. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38071–38090.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv:2310.08560.

Guillem Ramírez, Matthias Lindemann, Alexandra Birch, and Ivan Titov. 2024. Cache & distil: Optimising API calls to large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11838–11853. Association for Computational Linguistics.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236. Association for Computational Linguistics.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.