SOAR: Recovering Operational Overlays from Shared Semantics in Frozen Language Models

Ali Uyar
Independent Researcher

Abstract

Post-training adapters often mix task semantics with operational output behavior, which makes behavior reuse hard to study cleanly. We consider a deliberately controlled version of that problem. Starting from filtered SQuAD 1.1 question answering, we retain only examples with exactly one gold support sentence, compile targets deterministically, and evaluate with parsers rather than an LLM judge. On this fixed QA substrate, we train a semantic scaffold adapter for plain answers and direct single-overlay adapters for two deployment-style output contracts: strict JSON and exact support quotation. We then construct *Semantic-Operational Adapter Residuals* (SOAR) by subtracting the scaffold in effective dense-delta space and recompressing the result.

On the corrected full-scale study, both overlays recover cleanly. For JSON, $S + R_J$ matches the direct adapter on strict schema validity (0.9986 vs. 0.9986) with essentially unchanged answer F1 (0.8344 vs. 0.8346). For exact quotation, $S + R_Q$ matches or slightly exceeds the direct adapter on quote exactness (0.9414 vs. 0.9413) and answer F1 (0.8122 vs. 0.8121). Under the same output contracts, prompt-only control on the scaffold is near zero on the active overlay metric in both cases, so the recovered behavior is not explained by prompting alone.

We present these results as evidence for *partial operational recoverability*, not universal modularity. The validated claim surface is intentionally narrow: the main contribution of this paper is a semantics-fixed protocol, a dense-delta residualization method, and full-scale single-overlay recovery for two operational behaviors with a deterministic, auditable evaluation pipeline.

1 Introduction

Modern language-model deployments often care less about open-ended chat behavior than about reliable *output contracts*: emit valid JSON, provide the exact supporting sentence, or otherwise satisfy a downstream interface without post-hoc repair. Those contracts are operational requirements, not new semantic tasks. Yet standard post-training usually learns task semantics and output behavior together, making it difficult to ask a clean scientific question: *how much of a narrow output behavior can be isolated from the shared task semantics that support it?*

This paper studies a deliberately controlled version of that question. We hold the underlying task fixed to evidence-grounded question answering and vary only the required output contract. The substrate is a filtered derivative of SQuAD 1.1 (Rajpurkar et al., 2016) in which every retained example has exactly one gold support sentence. Contexts are sentence-labeled, targets are compiled deterministically, and evaluation is parser-based. The resulting setup is narrow by design: it sacrifices breadth in exchange for a clean object of study.

Our method, SOAR (*Semantic-Operational Adapter Residuals*), separates a shared semantic scaffold adapter from overlay-specific behavior. We first train a scaffold adapter S for plain QA and direct single-overlay adapters F_J and F_Q for two deployment-style contracts: strict JSON output and exact support quotation. We then subtract the scaffold in *effective dense-delta space* and recompress the remainder into deployable residual

adapters R_J and R_Q. The central empirical question is simple: can $S + R_k$ recover the target behavior of F_k without materially degrading answer quality?

The answer, on the corrected full-scale study, is yes for both validated overlays. For JSON, $S + R_J$ matches the direct adapter exactly on strict schema validity while remaining effectively identical on answer F1. For exact quotation, $S + R_Q$ matches or slightly exceeds the direct adapter on both quote exactness and answer F1. In both overlays, prompt-only control on the scaffold collapses under the same output contract, which is the key empirical signature that the residual is carrying genuine operational behavior rather than merely restating the prompt.

The scope of the paper is intentionally narrow, and we preserve that narrowness throughout. We do *not* claim universal behavior atoms, pairwise or triple composition success, or full equivalence to direct multi-behavior training. The broader project also included a citation overlay; direct full-scale citation training exists, but the clean downstream full-scale evidence chain does not, so citation appears only as scope context in the appendix rather than as a main result. Within that claim-safe boundary, the paper makes three contributions:

- A **semantics-fixed protocol** for studying operational overlays on top of a shared QA substrate, with deterministic compilation and parser-based evaluation.
- A **dense-delta residualization method** that extracts deployable single-overlay residuals from direct adapters while avoiding the representation ambiguity of raw LoRA factor arithmetic.
- **Corrected full-scale evidence** that JSON and exact quote behaviors recover cleanly at full scale, together with an auditable artifact trail that makes the results easy to inspect and hard to overstate.

The result is a controlled, high-integrity paper about single-overlay recovery. Its contribution is not breadth. Its contribution is that the object under study is unusually well specified, the evaluation path is auditable, and the claims stop exactly where the evidence stops.

2 Related Work

Parameter-efficient adaptation and weight-space composition. SOAR lives in the broader ecosystem of parameter-efficient adaptation methods such as LoRA (Hu et al., 2022). It is also adjacent to work on combining task-specific updates or adapters, including AdapterFusion (Pfeiffer et al., 2021) and task arithmetic (Ilharco et al., 2023). Those lines of work establish that narrow updates can sometimes be useful building blocks. Our contribution is narrower and more controlled: we study whether *operational output contracts* can be separated from shared semantics on a fixed QA substrate, rather than proposing a general-purpose composition system across many tasks.

Structured and grounded output control. A second neighboring area studies structured generation and grounded outputs directly. JSONSchemaBench, for example, highlights how operationally important and empirically nontrivial structured generation remains for modern language models (Geng et al., 2025). Our paper is not a structured-generation benchmark and does not attempt to outperform constrained decoding systems. Instead, strict JSON and exact quotation serve as two concrete contracts through which to study operational recoverability.

Reading or decomposing behavior from updates. Recent work has begun to read behavior from trained low-rank updates or to decompose broader behavioral structure from training signals. W2T shows that LoRA checkpoints already encode rich behavioral information once factorization ambiguity is removed (Han et al., 2026). Gradient Atoms discovers interpretable behavior-like components through sparse decomposition of training gradients (Rosser, 2026). Those papers are philosophically close to ours, but the scientific object is different. We do not try to infer behavior from weights alone or discover behaviors unsupervised. We ask a more operational question: after training a direct overlay adapter, can we subtract a shared semantic scaffold and still recover a deployable behavior residual at full scale?

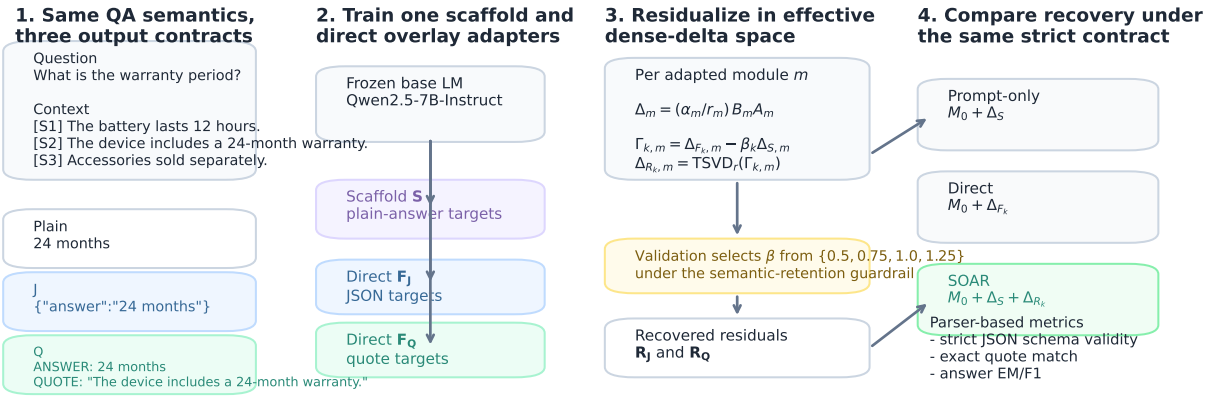

Figure 1: **Semantics-fixed single-overlay study.** The question, labeled context, and support sentence stay fixed; only the output contract changes. We train one scaffold adapter for plain answers and one direct adapter per overlay, convert both to effective dense deltas, subtract the scaffold in dense-delta space, and recompress the residual with truncated SVD. The paper validates the resulting single-overlay systems $S + R_J$ and $S + R_Q$, not broader pairwise or triple compositions.

Table 1: **Operational contracts in the validated paper path.** The same labeled context and answer semantics underlie all conditions; only the required output format changes.

Cond.	Target form	Primary metric
Plain	answer text only	answer F1
`J`	canonical JSON object with exactly one key, `"answer"`	strict JSON schema validity
`Q`	two lines: `ANSWER: <text>` and `QUOTE: <JSON string literal of exact support sentence>`	exact support quote match

Novelty boundary. The novelty claim we make is deliberately narrow. We do *not* claim the first decomposition of behavior from adapters, the first adapter composition method, or universal modularity. The paper contributes a semantics-fixed protocol, a dense-delta residualization method for single-overlay recovery, and evidence that two deployment-style behaviors recover cleanly under that protocol.

3 Semantics-Fixed Protocol and SOAR

3.1 Semantics-fixed QA substrate

Each processed example contains a question, a sentence-labeled context, one gold answer span, and exactly one gold support sentence. The operational contract is the only thing that changes across conditions. Table 1 summarizes the contracts used in the validated paper path.

This controlled compilation matters. By keeping the question, evidence, and gold answer fixed, the study turns "JSON" and "quote" into operational overlays rather than new semantic tasks.

3.2 Dense deltas as the method object

Let M_0 denote the frozen base model. We train a scaffold adapter S on plain-answer targets and direct single-overlay adapters F_J and F_Q on the same examples under their respective contracts. For each adapted

module m with LoRA factors $(A_m, B_m, \alpha_m, r_m)$, we work with the effective dense update

$$\Delta_m = \left(\frac{\alpha_m}{r_m}\right) B_m A_m. \tag{1}$$

This choice is methodological rather than cosmetic. Raw LoRA factor pairs are not unique: different factor pairs can represent the same update. The effective dense delta is the stable object that actually perturbs the frozen backbone. In this paper we therefore treat dense-delta residualization as the core method object, while explicitly stopping short of a stronger empirical superiority claim over factor-space subtraction.

3.3 Single-overlay residualization

For overlay $k \in \{J, Q\}$ and adapted module m, SOAR computes the raw residual target

$$\Gamma_{k,m} = \Delta_{F_k,m} - \beta_k \Delta_{S,m}, \tag{2}$$

then recompresses $\Gamma_{k,m}$ with truncated SVD to obtain a deployable residual update

$$\Delta_{R_k,m} = \mathrm{TSVD}_r(\Gamma_{k,m})\,. \tag{3}$$

In practice, β_k is selected on validation from the locked grid $\{0.5, 0.75, 1.0, 1.25\}$ under the standard semantic-retention guardrail. On the full-scale validated path, both overlays selected $\beta = 1.0$.

The resulting single-overlay system is

$$M_k^{\mathrm{SOAR}} = M_0 + \Delta_S + \Delta_{R_k}. \tag{4}$$

For each overlay we compare three systems under the same required contract:

1. **Prompt-only**: $M_0 + \Delta_S$ with the overlay contract specified in the prompt.
2. **Direct**: $M_0 + \Delta_{F_k}$.
3. **SOAR**: $M_0 + \Delta_S + \Delta_{R_k}$.

The broader SOAR program also defines composition by adding multiple residuals on top of the scaffold. That idea motivated the project, but it is *not* the validated contribution of this submission. The paper claims only single-overlay recovery.

4 Experimental Setup

4.1 Data

The core dataset is a filtered derivative of SQuAD 1.1 (Rajpurkar et al., 2016). We retain only answerable examples whose gold answer span maps fully inside exactly one sentence under a pinned sentencizer. The resulting processed dataset, `squad_jcq_v1`, contains 78 098.0000 training examples, 9353.0000 validation examples, and 10 554.0000 test examples. A total of 164.0000 raw examples were dropped because the answer span did not map to exactly one sentence. The saved split audit reports no title overlap across train, validation, and test, and the dataset manifest reports zero round-trip compilation failures over 98 005.0000 checked processed examples.

4.2 Model and training

All main-path experiments use the frozen backbone `Qwen/Qwen2.5-7B-Instruct` (Yang et al., 2024). Adapters target `o_proj` and `down_proj`, with LoRA rank 16.0000, LoRA alpha 16.0000, dropout 0.0500, and bf16 training. We report the corrected full-scale artifact bundle for one seed (#17). This is therefore a single-seed study, and we treat it that way in the claims and discussion.

Table 2: **Corrected full-scale single-overlay recovery.** In both validated overlays, prompt-only under the same output contract collapses on the active operational metric, while SOAR matches or slightly exceeds the direct single-overlay adapter and remains effectively identical on answer F1.

Overlay	Primary metric	Prompt	Direct	SOAR	Direct F1	SOAR F1	Recovery ratio	β
`J`	JSON strict valid	0.0000	0.9986	0.9986	0.8346	0.8344	1.0000	1.0000
`Q`	Quote exact	0.0000	0.9413	0.9414	0.8121	0.8122	1.0001	1.0000

Prompt-only answer F1 is also 0.0000 in both overlays because malformed outputs count as failures under the paper's strict parser-based evaluation.

The scaffold adapter S is trained on plain-answer targets. Direct overlay adapters F_J and F_Q are trained on the same processed examples with only the output contract changed. Residualization selects β on the validation set using the locked grid above and the project's standard semantic-retention guardrail; both validated overlays select $\beta = 1.0$.

4.3 Deterministic evaluation

Evaluation is parser-based and fully deterministic. JSON outputs are parsed and scored for validity and strict schema compliance. Quote outputs are parsed into an answer field and an exact support sentence string; the active quote metrics are exact sentence match and quote token F1. Semantic retention is measured by answer EM/F1 after parser-based answer extraction. Greedy decoding is used throughout, and no LLM judge appears in the critical claim path.

For the main text we focus on two outcomes:

1. the **active overlay metric** (`json_strict_schema_valid` for `J`, `quote_exact` for `Q`), and
2. **answer F1** as the primary semantic-retention metric.

Prompt-only control is deliberately strong in one sense and deliberately weak in another: it shares the same scaffold semantics as SOAR, but it must satisfy the overlay contract through prompting alone. If prompt-only remains competitive, the paper's recovery interpretation weakens. If prompt-only collapses while SOAR remains near the direct adapter, the residual is doing meaningful work.

4.4 Scope discipline

The broader project defined three overlays (`J`, `C`, `Q`). The current paper path validates only `J` and `Q` at full scale. Direct full-scale citation training exists, but the downstream full-scale citation residualization/evaluation chain was not completed cleanly enough for a claim-safe main-text row. We therefore keep citation strictly as appendix-style scope context rather than as main empirical evidence.

5 Results

5.1 Two overlays recover cleanly at full scale

Table 2 and Figure 2 show the main empirical pattern. In both validated overlays, the prompt-only scaffold is near zero on the active operational metric, while the residualized system lands directly on top of the direct adapter. That is the pattern the paper needs. It means the recovered behavior is not being explained away by prompt conditioning alone.

JSON (`J`). The JSON result is exceptionally clean. SOAR matches the direct adapter exactly on strict schema validity (0.9986 vs. 0.9986) and remains effectively identical on answer F1 (0.8344 vs. 0.8346). Under

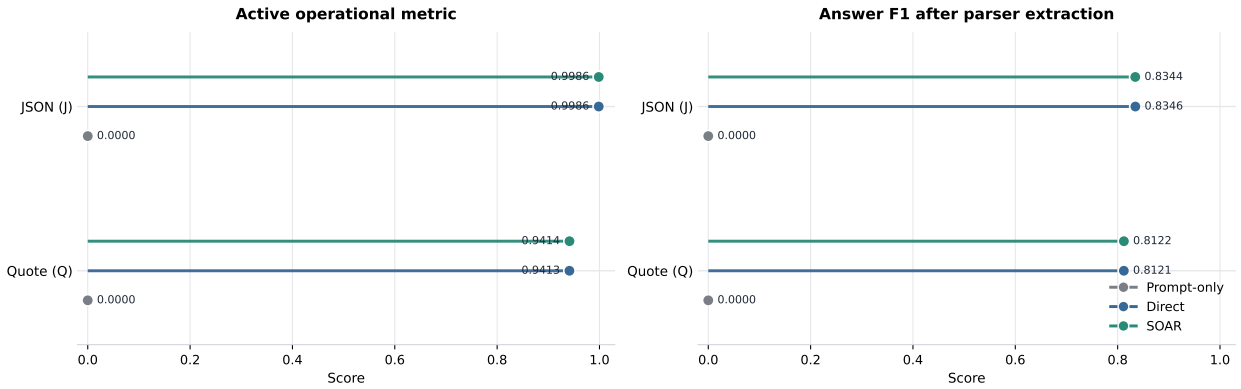

Figure 2: **Prompting alone is not enough; residualization is.** Each point is a full-scale test result. Left: the active operational metric for each overlay. Right: answer F1 after parser-based extraction. For both overlays, prompt-only collapses under the required contract, while SOAR overlaps with the direct single-overlay adapter.

the same contract, prompt-only remains noncompetitive. This is the strongest case in the paper that a formatting-oriented operational requirement can be separated from shared QA semantics without paying a meaningful semantic penalty.

Quote (`Q`). The quote result is also clearly positive. On the corrected canonical path,[1] SOAR reaches 0.9414 quote exactness versus 0.9413 for the direct quote adapter, while slightly exceeding it on answer F1 (0.8122 vs. 0.8121). This matters because quotation is a stricter and more semantically coupled contract than JSON formatting alone. The positive quote result therefore makes the single-overlay story materially stronger.

5.2 What the current evidence supports

The corrected full-scale bundle supports a strong version of the paper's single-overlay claim: in this semantics-fixed QA setting, at least two operational behaviors can be recovered from a shared scaffold plus an operational residual with negligible semantic degradation. It also supports the systems claim that the evaluation path is deterministic, parser-based, and auditable.

What it *does not* support is just as important. The paper does not claim pairwise or triple composition success, an ordering over which overlays are most composable, transfer beyond the SQuAD-derived core dataset, or universal superiority of dense deltas over factor-space subtraction in empirical terms. Dense deltas remain the principled method object, but the factor-subtraction ablation is absent, so the stronger version of that methods claim remains out of scope.

5.3 Why the result matters despite the narrow scope

The paper's value does not come from breadth. It comes from how tightly the scientific object is controlled. The question, support sentence, and answer semantics are fixed; only the output contract changes. The parser-based evaluation prevents the paper from leaning on a soft judge. And the prompt-only control makes the residual interpretation falsifiable. Within that controlled setting, recovering both JSON and exact quotation at full scale is a meaningful result.

[1]The initial full-scale quote path exposed a metric aggregation bug. The tainted artifacts were archived, the metric path was repaired, and the canonical reports in this bundle were regenerated from corrected metrics. Appendix B summarizes that audit trail.

6 Discussion and Limitations

The central lesson of the paper is not that operational behavior is universally modular. It is that *some* operational behavior can be isolated cleanly enough to recover at full scale on top of shared semantics. JSON and exact quotation are already informative contrasts. JSON is the more formatting-local contract, and unsurprisingly it yields an almost exact recovery signature. Exact quotation is stricter and more semantically coupled, yet it still recovers cleanly in the corrected bundle. Together they provide strong evidence for partial recoverability.

That said, the paper is deliberately narrow, and the narrowness matters for interpretation.

Single-overlay only. Although SOAR was designed with composition in mind, composition is not part of the validated claim path for this submission. The paper therefore supports single-overlay recovery, not pairwise or triple overlay algebra.

One substrate, one backbone, one seed. All reported experiments use one filtered SQuAD-derived dataset, one frozen backbone, one module subset, and one seed. This is enough for a controlled methods paper, but not for broad generalization claims. In particular, we do not know from this paper whether the same recovery pattern holds on other tasks, with other overlay grammars, or with other backbones.

Dense deltas as a principled choice, not a settled empirical theorem. The method works in effective dense-delta space because raw LoRA factor pairs are representation-dependent. That is a strong design rationale, and recent work on weight-space representations reinforces the importance of removing factorization ambiguity (Han et al., 2026). But this paper does not include the missing factor-subtraction ablation, so it should not be read as an empirical proof that dense-delta residualization always dominates every factor-space alternative.

Citation remains scope context. The broader project included citation behavior. Direct full-scale citation training exists, but the clean downstream full-scale evidence chain does not. We treat that honestly: citation appears only in the appendix as a scope note, not as a positive or negative main-text result.

These limitations do not undercut the main contribution. They clarify its intended use. The right reading of the paper is that a carefully controlled protocol can reveal clean recovery signatures for at least two operational overlays, and that such signatures are worth studying before claims of broader compositionality are made.

7 Conclusion

SOAR asks a narrow question and gives a clear answer. On a semantics-fixed QA substrate with deterministic compilation and parser-based evaluation, two deployment-style output requirements—strict JSON and exact support quotation—can be recovered from a shared semantic scaffold plus a residual overlay at full scale. Prompting alone does not explain the result, and the recovered systems remain effectively identical to direct single-overlay adapters on answer quality.

That is enough to justify a strong single-overlay paper. It is not enough to justify universal modularity, and we do not claim it. The contribution is instead a clean protocol, a dense-delta residualization method, and a corrected full-scale evidence bundle showing that operational recoverability is real for at least two behaviors in a frozen language model. For a problem area that is usually studied through looser prompts, softer judges, or broader but blurrier composition stories, that narrower result is precisely the point.

References

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*.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models. *arXiv preprint arXiv:2501.10868*.

Xiaolong Han, Ferrante Neri, Zijian Jiang, Fang Wu, Yanfang Ye, Lu Yin, and Zehong Wang. 2026. W2T: LoRA Weights Already Know What They Can Do. *arXiv preprint arXiv:2603.15990*.

Gabriel Ilharco, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Yejin Choi, and Ali Farhadi. 2023. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

J. Rosser. 2026. Gradient Atoms: Unsupervised Discovery, Attribution and Steering of Model Behaviors via Sparse Decomposition of Training Gradients. *arXiv preprint arXiv:2603.14665*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

A Citation Overlay Scope Note

The broader project scope included a citation overlay (`C`) in addition to JSON (`J`) and exact quotation (`Q`). The current submission does not report citation as a main empirical result because the full downstream full-scale evidence chain was not completed cleanly enough for a claim-safe recovery row.

Table 3: **Artifact status for the citation overlay.** Direct full-scale citation training exists, but the clean downstream full-scale recovery chain does not. The correct manuscript treatment is therefore appendix-style scope context, not a main-text positive or negative claim.

Stage	Status	Safe interpretation
Direct full-scale `C` training	Completed	The citation overlay was run at the direct-training stage.
Full-scale `C` residualization	Off-path archived attempt	No clean full-scale citation residual result exists.
Full-scale citation eval bundle	Absent	No full-scale citation test metrics should be reported.
Full-scale citation recovery report	Absent	Citation does not belong in the main recovery table.

This appendix note matters for honesty. It prevents two opposite misreadings at once: that citation was never run at full scale, and that citation recovery either clearly succeeded or clearly failed. The saved artifacts support neither extreme reading. They support the narrower truth that the reduced-scope one-GPU paper path prioritized the cleaner JSON/quote evidence and stopped citation short of a complete claim-safe downstream chain.

B Audit Trail and Reproducibility Notes

The paper is built from a reduced-scope artifact bundle rather than from informal notebook outputs. Three details are especially important for reviewers.

Deterministic data and evaluation. The processed dataset is versioned, sentence splitting is pinned, and the split audit reports no title overlap across train, validation, and test. The bundle's dataset manifest also reports zero round-trip compilation failures over all non-empty processed splits. Evaluation is parser-based throughout; malformed outputs score as failures under the strict metrics rather than being rescued by ad hoc manual inspection.

Corrected quote path. The canonical quote artifacts in this paper are *corrected* artifacts. An earlier full-scale quote path surfaced a condition-metric aggregation bug that could silently undercount parse failures. That bug was repaired, the tainted canonical quote artifacts were archived rather than overwritten, and the canonical quote reports in the current bundle were regenerated from corrected metrics. The corrected full-scale quote path selects $\beta = 1.0$ and yields the positive recovery result reported in the main text.

Single-seed evidence. All reported full-scale numbers come from seed 17. We therefore present the paper as a controlled, artifact-backed single-seed study. The contribution lies in the protocol, the auditable result chain, and the clarity of the recovery signature, not in broad variance estimates.

Table 4 provides the secondary metrics saved in the artifact bundle.

Table 4: **Secondary full-scale metrics from the canonical bundle.**

Overlay	System	Ans. EM	Ans. F1	Aux. 1	Aux. 2
J	Direct	0.6803	0.8346	0.9986[a]	0.9986[b]
	SOAR	0.6810	0.8344	0.9986[a]	0.9986[b]
Q	Direct	0.6507	0.8121	0.9413[c]	0.9497[d]
	SOAR	0.6504	0.8122	0.9414[c]	0.9499[d]

[a] `json_strict_schema_valid`
[b] `json_valid`
[c] `quote_exact`
[d] `quote_f1`

C Prompt Contract and Target Examples

The logical prompt content is fixed across conditions except for the format specification.

Prompt template

```
Use only the provided context.  Context sentences are labeled [S1], [S2], ...
Do not use outside knowledge.
Required output format:  FORMAT_SPEC

Question:
QUESTION

Context:
LABELED_CONTEXT

Response:
```

For a toy example with answer $a = 24$ months, support sentence ID $s = 2$, and support sentence `The device includes a 24-month warranty.`, the compiled targets are:

Cond.	Compiled target
Plain	`24 months`
`J`	`{"answer":"24 months"}`
`Q`	`ANSWER: 24 months` `QUOTE: "The device includes a 24-month warranty."`

For strict metrics, malformed outputs are scored as failures rather than repaired by auxiliary heuristics. This strictness is intentional: it makes the paper's recovery claim more falsifiable and the evaluation trail easier to audit.