

What Deterministic Verification Actually Fixes in Financial QA: A Diagnostic Analysis of Rule Heterogeneity Across Datasets

Aaditya Thokal

Universal College of Engineering, University of Mumbai

Mumbai, India

aaditya.thokal24@gmail.com

Abstract

Deterministic, parameter-free post-hoc correction rules are an attractive way to address numerical errors in financial question answering: they are simple, cheap, and fully auditable. We ask whether such rules actually behave like general-purpose correctors, or whether their effect is contingent on properties of the dataset and model output distribution they are applied to. We evaluate a two-rule deterministic verification layer (DVL) — `scale_mul100` and `scale_div100` — on FinQA ($n=873$) and TAT-QA ($n=718$) using Mistral-7B-Instruct-v0.3. This same, fixed model-plus-rule configuration produces a striking cross-dataset divergence: harmful on FinQA (37.34%→16.04%, −21.31pp, McNemar $p=1.76 \times 10^{-39}$) and beneficial on TAT-QA (26.32%→30.64%, +4.32pp, $p=5.15 \times 10^{-3}$). A rule-level transition analysis shows `scale_mul100` is highly precise on TAT-QA (87.95%) but overwhelmingly imprecise on FinQA (3.39%), while `scale_div100` produces zero observed corrections and only harm in both settings; removing it raises TAT-QA accuracy further, to 35.10%. We do not find evidence that these rules correct arithmetic, retrieval, or semantic reasoning failures. We conclude that this specific pair of scaling rules is better understood as a targeted, dataset-contingent representation normalizer than as a general reasoning corrector, and we characterize the boundary conditions under which it helps or hurts.

1 Introduction

Financial question answering over earnings reports and tabular filings routinely requires a language model to locate a figure, apply a scale (thousands, millions, percent), and report a number at the correct magnitude and sign. A single scale error renders a financial answer operationally useless, even when the model has otherwise reasoned correctly.

This has motivated a family of lightweight, post-hoc correction strategies: deterministic rules that inspect a raw model prediction together with surface cues (question keywords, document headers) and rescale or resign it without retraining and without access to the gold label at inference time.

Such rules are attractive precisely because they are simple: they are cheap to compute, fully interpretable, and auditable in regulated settings. But their simplicity raises an immediate question that is easy to overlook when a rule is evaluated on a single benchmark: does a surface-level scale correction correspond to a genuine, recoverable representation error, or does it merely exploit incidental regularities of one dataset’s answer distribution? A rule validated only in one setting offers no evidence either way.

We investigate this question directly with a controlled, cross-dataset evaluation. We apply the same two deterministic scaling rules, `scale_mul100` and `scale_div100`, unchanged, to model output on two financial QA benchmarks, FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021), under an identical scoring and statistical-testing protocol. The result is a striking dataset-level reversal: the intervention is strongly harmful on FinQA and mildly but significantly beneficial on TAT-QA. Aggregate accuracy alone does not explain this reversal; we therefore decompose the intervention rule-by-rule and dataset-by-dataset, examining which corrections succeed, which fail, and how their effects differ across datasets.

Research question. To what extent do deterministic post-hoc verification rules correct structured numerical errors in financial language models, and what performance boundaries are imposed by rule heterogeneity and harmful side effects?

Framing. This is an empirical diagnostic study of the boundaries and failure modes of deterministic post-hoc numerical verification, not a proposal

for a new algorithm. The two-rule DVL evaluated here is deliberately simple; we make no claim to algorithmic novelty. The contribution is the empirical finding that identical deterministic rules behave in opposite ways across datasets, and a rule-level analysis of why.

Our contributions are: (1) a cross-dataset evaluation showing that a fixed deterministic scaling intervention can be strongly harmful on one financial QA benchmark and significantly beneficial on another; (2) a rule-level state-transition analysis showing that this aggregate reversal is driven by `scale_mul100` behaving in opposite ways across datasets, while `scale_div100` is uniformly harmful and never observed to produce a correction; and (3) a provenance-controlled evaluation pipeline supporting this analysis with deterministic scoring, paired significance testing, bootstrap confidence intervals, and artifact-level provenance, including a documented case where we identified and corrected a mislabeled evidence field prior to analysis (§4, §6).

2 Related Work

Financial QA and numerical reasoning. FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) both require numerical reasoning over a mixture of prose and tabular financial data, and both report substantial headroom between model and human performance on exactly this kind of scale- and magnitude-sensitive arithmetic. Domain-adapted language models such as BloombergGPT (Wu et al., 2023) and FinBERT (Yang et al., 2020) improve financial text understanding broadly but do not directly target numerical scale errors.

Tool-augmented and program-aided reasoning. Program-aided approaches such as PAL (Gao et al., 2023) route arithmetic through executable code, sidestepping the reliability of free-form numeric generation entirely. This is a complementary regime to the one we study: post-hoc deterministic verification is applicable when tool execution is unavailable, but as our results show, that applicability is itself dataset-dependent.

Retrieval and self-verification. Retrieval-augmented generation (Lewis et al., 2020) and self-consistency or self-refinement methods (Wang et al., 2023; Madaan et al., 2023) improve outputs through additional model computation, at additional inference cost and without a deterministic audit trail. The rules we study take the opposite

approach: zero additional inference cost, full determinism, but — as we show — no guarantee of a positive or even dataset-consistent effect.

Parameter-efficient fine-tuning. QLoRA (Dettmers et al., 2023) makes task-specific adaptation of 7B-scale models practical on consumer hardware; we use the same QLoRA-adapted model (FinVerify-LoRA) for both datasets, but fine-tuning itself is not a variable under study in this paper.

We deliberately do not compare against chain-of-thought prompting in this paper; that comparison, and additional DVL rule types (sign and magnitude correction), were the subject of separate prior work and are out of scope for the two-rule diagnostic evaluated here.

3 Deterministic Verification Layer

The DVL evaluated in this paper consists of exactly two ordered, deterministic scaling rules, applied to ratio-type questions (identified by the presence of keywords such as *ratio*, *margin*, *percent*, *change*, *growth*). Let $\hat{y} \in \mathbb{R}$ denote the raw model prediction and $\hat{y}^* \in \mathbb{R}$ the verified output.

- **scale_mul100.** If $|\hat{y}| < 1$ on a ratio question, set $\hat{y}^* \leftarrow \hat{y} \times 100$. This targets the case where the model outputs a decimal fraction (e.g. 0.28) instead of a percentage-scale figure (e.g. 28).
- **scale_div100.** If $|\hat{y}| > 100$ on a ratio question, set $\hat{y}^* \leftarrow \hat{y}/100$. This targets the inverse case, where the model outputs a percentage-scale figure where a decimal fraction was expected.

Both rules fire only on the raw prediction and the question text; neither has access to the gold label y at intervention time. Every firing is logged with its rule identity, the original value, and the corrected value, so that the intervention is fully auditable after the fact. Gold labels enter the pipeline only in a separate scoring step, after $\hat{y}^*$ has already been produced and logged (Figure 1), using a 5% relative tolerance (and a 1% absolute tolerance when the gold value is zero).

We emphasize what this DVL variant does *not* include: no sign correction, no unit/magnitude-header correction, and no fine-tuning intervention. It is intentionally restricted to the two scaling rules so that any dataset-dependent effect we observe can be attributed to those two rules alone, rather than to interactions among a larger rule set.

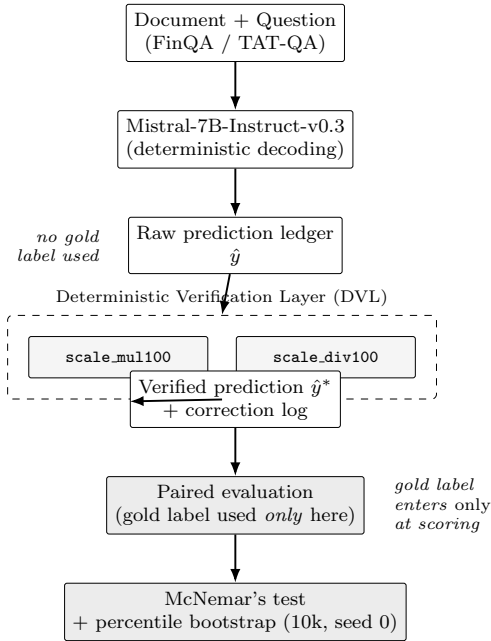


Figure 1: Inference and evaluation pipeline. The DVL intercepts the raw prediction and applies at most one scaling rule using only the prediction and question text. The gold label is used exclusively at the paired-evaluation step, after the verified prediction has already been logged.

4 Experimental Protocol

Data. We evaluate on the full FinQA development set restricted to numeric answers ($n=873$) and on TAT-QA arithmetic questions ($n=718$). Both are evaluated under the same DVL implementation and scoring harness.

Models. Both FinQA and TAT-QA predictions were generated from the identical model configuration: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) with the same QLoRA-adapted checkpoint (FinVerify-LoRA) (Detmers et al., 2023), the same base-model and tokenizer snapshot, and the same deterministic decoding settings (greedy decoding, no sampling). We state this explicitly because the evidence record’s summary field names the adapter only under its FinQA entry; we verified from the underlying execution provenance (the ledger-generation commands that produced the frozen prediction files for both datasets) that the same locally-resolved adapter directory was passed to both runs, so the two datasets are not confounded by different fine-tuning states. This matters for interpretation: because the model and adapter are held fixed, the cross-dataset reversal we report in §5–§6 reflects an interaction between this fixed model-output regime and each dataset’s answer distribution, not a differ-

ence in which model produced the predictions.

Scoring. A prediction is scored correct if it falls within 5% relative tolerance of the gold value (1% absolute tolerance when gold is exactly zero). Significance between the baseline and post-DVL paired outcomes is assessed with McNemar’s test: continuity-corrected chi-square when the discordant count is ≥ 25 , and an exact binomial test otherwise. Confidence intervals use percentile bootstrap resampling (10,000 resamples, seed 0, 95% confidence level).

Provenance. The evaluation pipeline is provenance-locked: dataset revision, base model revision, adapter revision, and implementation commit are recorded for each run. The TAT-QA loading pipeline underwent a documented serialization fix mid-project after an earlier run produced an implausibly low baseline consistent with a table/context serialization defect rather than genuine model behavior; that earlier run is not used anywhere in this paper and is mentioned here only for provenance completeness.

CI provenance asymmetry. We disclose an asymmetry in how confidence intervals were obtained for the two datasets. TAT-QA’s bootstrap confidence intervals were saved directly by the original experimental run. FinQA’s bootstrap confidence intervals were derived after the fact from recovered original prediction ledgers, using the same frozen bootstrap implementation, because the live bootstrap output from the original FinQA run was not preserved. Both procedures use identical resampling parameters; we report this only in the interest of full evidentiary transparency, not because we have reason to think it affects validity.

A note on evidence correction. During manuscript preparation we cross-checked every quantity in our evidence record against the underlying experiment artifacts. One quantity — the accuracy pair reported for the TAT-QA `scale_div100` ablation (§6) — was found to have been mislabeled: the recorded “post-ablation” accuracy and its confidence interval had been copied from the unrelated full-DVL result rather than computed from the ablated condition. We traced this to the original scoring output, corrected the two mislabeled fields, and removed a bootstrap interval that could not be verified for the ablated condition rather than approximate one. No experiment was rerun or reconstructed; this was a correction to how an already-completed result had been recorded. We disclose it here because it is directly

relevant to the diagnostic conclusion in §6, and because we would want any paper we reviewed to disclose the same; full correction detail is in Appendix A, §A.1.

5 Results

The intervention has opposite aggregate effects on the two datasets (Table 1). On FinQA, DVL reduces accuracy from 37.34% to 16.04%, a drop of 21.31 percentage points (McNemar $p = 1.76 \times 10^{-39}$; 192 examples move from correct to incorrect against only 6 in the opposite direction). On TAT-QA, DVL increases accuracy from 26.32% to 30.64%, a gain of 4.32 percentage points (McNemar $p = 5.15 \times 10^{-3}$; 73 examples move from incorrect to correct against 42 in the opposite direction). Both effects are statistically significant; they point in opposite directions.

This reversal cannot be explained by coverage alone: DVL fires on a similar share of examples in both datasets (57.50% of FinQA, 48.61% of TAT-QA). The difference is in what happens when it fires.

6 Diagnostic Analysis

Transition Precision, formally defined. The "Precision" column in Table 2 is not standard classification precision; we define it explicitly to prevent misreading. For a given rule and dataset, let $|I \rightarrow C|$ and $|C \rightarrow I|$ be the counts of fired examples that moved from incorrect to correct and from correct to incorrect, respectively. We define **Transition Precision** as

$$\text{Prec}_{\text{trans}} = \frac{|I \rightarrow C|}{|I \rightarrow C| + |C \rightarrow I|}. \quad (1)$$

Firings that leave the correctness state unchanged ($I \rightarrow I$) are excluded from the denominator: they neither help nor harm the reported accuracy, so a metric intended to capture "when this rule changes the answer's correctness, how often is the change beneficial?" should not be diluted by firings that did not change correctness at all. Transition Precision is therefore a directional, state-change-conditional metric, not an estimate of how often the rule's output is correct among all examples it touches; we use this specific definition throughout the paper.

Rule-level precision is dataset-dependent. Table 2 (top) breaks the aggregate result down by rule and dataset. `scale_mul100` fires 443 times on FinQA, converting only 6 incorrect predictions to

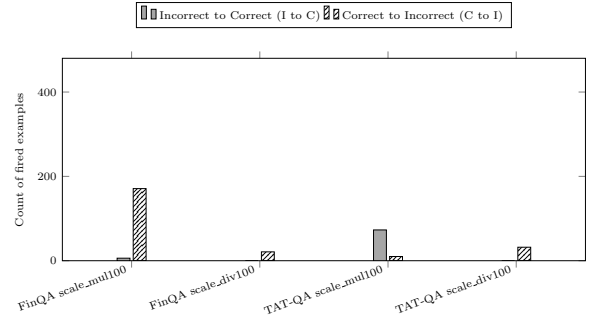


Figure 2: Fired-example state transitions by rule and dataset. `scale_mul100` is dominated by beneficial corrections ($I \rightarrow C$) on TAT-QA and by harmful regressions ($C \rightarrow I$) on FinQA. `scale_div100` produces zero $I \rightarrow C$ corrections in either dataset.

correct while converting 171 correct predictions to incorrect — a Transition Precision of 3.39% (95% CI [1.09, 6.32]). The same rule fires 283 times on TAT-QA, converting 73 incorrect predictions to correct against only 10 regressions — a Transition Precision of 87.95% (95% CI [80.72, 93.98]). This is not a small effect-size difference; it is a rule behaving as a near-reliable corrector on one dataset and as a near-reliable *corruptor* on the other.

`scale_div100` never corrects, only harms.

Across both datasets combined, `scale_div100` fires 125 times and is never observed to convert an incorrect prediction to correct ($I \rightarrow C = 0$ in both FinQA and TAT-QA). It does, however, convert 21 correct FinQA predictions and 32 correct TAT-QA predictions to incorrect. Its precision is 0.00% in both settings (Table 2, top; Figure 2).

Isolating `scale_mul100`'s ceiling on TAT-QA. Because `scale_div100` is uniformly harmful, we ask what TAT-QA accuracy would be if it were removed and only `scale_mul100` were applied. Table 2 (bottom) reports this ablation. Full DVL reaches 30.64% accuracy on TAT-QA; `scale_mul100` alone reaches 35.10% (McNemar $p = 1.01 \times 10^{-11}$ for the mul100-only intervention against the no-DVL baseline). That is, `scale_div100` is suppressing approximately 4.46 percentage points of accuracy that `scale_mul100` would otherwise deliver on TAT-QA. This is, in our view, the strongest single diagnostic result in the paper: it isolates a specific, quantifiable cost of including an indiscriminately harmful rule alongside a genuinely useful one, within the same rule family and the same evaluation.

Interpretation. We separate what the experiments establish from what remains speculative,

Dataset	N	Baseline	Post-DVL	Δ (pp)	McNemar p	Bootstrap Δ 95% CI
FinQA	873	37.34%	16.04%	-21.31	1.76×10^{-39}	$[-24.17, -18.44]$
TAT-QA	718	26.32%	30.64%	+4.32	5.15×10^{-3}	$[+1.39, +7.24]$

Table 1: Cross-dataset evaluation of the two-rule DVL (`scale_mul100`, `scale_div100`). FinQA bootstrap CIs are derived post-hoc from recovered ledgers using the frozen bootstrap implementation; TAT-QA CIs were saved directly by the original experimental run (see §4).

Rule	Dataset	Firings	$I \rightarrow C$	$C \rightarrow I$	Trans. Prec. [†]	95% CI
<code>scale_mul100</code>	FinQA	443	6	171	3.39%	[1.09, 6.32]
<code>scale_div100</code>	FinQA	59	0	21	0.00%	[0.00, 0.00]
<code>scale_mul100</code>	TAT-QA	283	73	10	87.95%	[80.72, 93.98]
<code>scale_div100</code>	TAT-QA	66	0	32	0.00%	[0.00, 0.00]

<i>TAT-QA ablation: removing <code>scale_div100</code> (<code>scale_mul100</code> only)</i>						
Condition	Accuracy	Δ vs. no-DVL	$I \rightarrow C$	$C \rightarrow I$	McNemar p	
No DVL	26.32%	—	—	—	—	—
Full DVL	30.64%	+4.32pp	73	42	5.15×10^{-3}	
<code>scale_mul100</code> only	35.10%	+8.77pp	73	10	1.01×10^{-11}	

Table 2: Top: per-rule Transition Precision on each dataset ([†]Transition Precision = $|I \rightarrow C| / (|I \rightarrow C| + |C \rightarrow I|)$; see Eq. 1, §6 — $I \rightarrow I$ firings are excluded since they do not change correctness). Bottom: TAT-QA ablation isolating `scale_mul100`. Removing `scale_div100` raises TAT-QA accuracy from 30.64% (full DVL) to 35.10%, indicating `scale_div100` suppresses approximately 4.46pp that `scale_mul100` alone would otherwise deliver.

since the most interesting part of this result — *why* the reversal occurs — is not something we designed an experiment to test.

Established. The reversal is real, large, and statistically robust (Table 1). It is not attributable to a difference in model or adapter, since both datasets were scored from the identical Mistral-7B-Instruct-v0.3 + FinVerify-LoRA configuration (§4). `scale_mul100` is a near-reliable corrector on TAT-QA and a near-reliable corruptor on FinQA (Table 2); `scale_div100` is uniformly harmful and, across 125 combined firings on both datasets, is never observed to produce a single beneficial correction. Because the model is held fixed, the divergence is an interaction between this fixed model-output regime and each dataset’s answer distribution, not a between-model artifact.

Plausible hypothesis, not tested here. One account consistent with, but not established by, our results is that `scale_mul100`’s trigger condition ($|\hat{y}| < 1$ on a ratio question) coincides with a genuine representation-level tendency in this model’s output on TAT-QA-style questions, while the same surface pattern is largely coincidental on FinQA’s answer distribution — so the rule is “correcting” noise rather than signal in the FinQA setting. We did not test this directly (e.g. by characterizing the two datasets’ gold-answer scale distributions or the

model’s pre-DVL error structure), so we present it as a hypothesis for future work, not as an explanation supported by the current evidence.

Explicitly untested. We do not know, and do not claim to know, which specific property of the two datasets or their answer distributions drives the reversal; whether it would persist under a different model, adapter, or prompt format; or whether a context-aware version of these rules (rather than the fixed scalar thresholds used here) would resolve the FinQA harm without losing the TAT-QA benefit. We also do not find evidence in this data that either rule corrects arithmetic reasoning, retrieval failures, or semantic misunderstanding of the question; our claims are restricted to representation-level scale correction.

A caveat on the TAT-QA result. The TAT-QA no-DVL baseline is itself low (26.32%), reflecting a relatively weak operating point for this model/pipeline on TAT-QA-style questions. The statistically significant improvement we report is real at this operating point, but it should not be read as evidence that this intervention would benefit a stronger or better-calibrated model on similar data; a model that already outputs correctly-scaled answers more often would give `scale_mul100` fewer, and possibly different, opportunities to fire.

The evidence therefore supports treating this spe-

cific pair of scaling rules, on this model configuration, as targeted, dataset-contingent representation normalizers rather than general-purpose numerical reasoning correctors — a claim about these two rules as evaluated here, not about deterministic scaling correction as a general class of technique.

7 Limitations

Our evaluation is intentionally narrow in scope, and the following limitations bound what can be concluded from it.

Two datasets. We evaluate on FinQA and TAT-QA only. The reversal we document establishes that dataset-dependence exists and can be large; it does not establish how common such reversals are across the broader space of financial QA benchmarks, nor does it identify, beyond the qualitative pattern above, a general predictive rule for when a scaling correction will help versus hurt on an unseen dataset.

Narrow intervention family. We study exactly two deterministic scaling rules. We make no claim about sign correction, unit/magnitude-header correction, or other post-hoc intervention types; those are separate rule families with separate failure modes, evaluated elsewhere and out of scope here.

Model/configuration specificity. All results are tied to Mistral-7B-Instruct-v0.3 and the FinVerify-LoRA adapter under the reported decoding and prompting configuration. We do not evaluate cross-architecture or cross-configuration generalization of the specific findings in this paper.

CI provenance asymmetry. As disclosed in §4, FinQA’s bootstrap confidence intervals are derived post-hoc from recovered prediction ledgers rather than saved directly from the original run, unlike TAT-QA’s. We used an identical bootstrap implementation and parameters in both cases, but readers should weigh this provenance difference when comparing interval widths across datasets.

No evidence of reasoning correction. Nothing in our evidence supports a claim that these rules repair arithmetic reasoning, retrieval errors, or semantic misinterpretation of a question. Our claims are restricted to representation-level scale correction, as described in §6.

Excluded reconstruction attempt. A later attempt to reconstruct the original TAT-QA execution environment produced results inconsistent with the frozen experimental evidence and was excluded

from this evidence package because it did not reproduce the original environment; we do not report its specific figures since we cannot independently verify them against the original notebook artifacts available to us.

8 Conclusion

We evaluated a deliberately simple, two-rule deterministic verification layer across two financial QA datasets under an identical, provenance-controlled protocol, and found that it produces opposite aggregate effects: a 21.31-point degradation on FinQA and a 4.32-point improvement on TAT-QA. A rule-level transition analysis shows this reversal is concentrated in `scale_mul100`, which is highly precise on TAT-QA (87.95%) and highly imprecise on FinQA (3.39%), while `scale_div100` is uniformly harmful and, when removed, raises TAT-QA accuracy by a further 4.46 points. We find no evidence that either rule corrects reasoning, retrieval, or semantic errors. We conclude that this specific pair of scaling rules, on this model configuration, is best understood as a targeted, dataset-contingent representation normalizer rather than a general reasoning corrector, and that evaluating such rules on a single dataset risks materially overstating — or understating — their reliability.

Ethics Statement

This work uses only publicly available research datasets (FinQA, TAT-QA) under their respective research licenses and a publicly released base language model. No human subjects were involved, no personally identifiable financial data was used, and no new data was collected. The deterministic verification rules studied here are diagnostic research artifacts; we explicitly do not claim they are suitable, in their current two-rule form, for unsupervised deployment in real financial decision-making, given the harmful behavior documented on FinQA in this paper.

References

Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10764–10799. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock  tschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Bodhisattwa Prasad Majumder, Ximing Lu, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prafulla Kam-badur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. FinBERT: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Provenance and Evidence Correction Detail

Table 3 summarizes the provenance tier of each headline quantity reported in this paper, following the classification used during evidence audit.

Tier 1 — directly saved experimental output: FinQA and TAT-QA baseline/post-DVL accuracy and transition counts; both datasets’ McNemar statistics; TAT-QA per-rule precision for `scale_mul100`; the corrected TAT-QA ablation accuracy pair and its McNemar statistic.

Tier 2 — derived from saved artifacts: FinQA bootstrap confidence intervals (recomputed from recovered ledgers with the frozen bootstrap implementation); the full-*N* FinQA transition matrix (fired-subset counts combined with pass-through non-fired examples); the FinQA per-rule firing breakdown (aggregated from per-example correction logs; internally consistent with, but not printed as, a standalone table in the source notebook); TAT-QA `scale_div100` counts (obtained by subtraction from dataset-level and `scale_mul100` totals).

Explicitly excluded: a later, non-matching environment reconstruction of the TAT-QA pipeline; an early, pre-serialization-fix TAT-QA run retained in the source notebook only as engineering background. Neither is used as evidence in this paper.

Table 3: Provenance classification of reported quantities.

A.1 Ablation Correction

The TAT-QA `scale_div100` ablation reported in §6 was corrected prior to drafting. The originally recorded accuracy pair (21.87%→30.64%) had been constructed by copying the full-DVL post-intervention accuracy into the ablation’s “post-ablation” field and back-solving a baseline from the (correct) delta, rather than reflecting the literal ablated-condition scoring output. The corrected pair (26.32%→35.10%) is the literal output of scoring the intervention log with all `scale_div100` corrections reverted to the raw prediction. The associated bootstrap confidence interval could not be verified for the ablated condition and has been omitted rather than estimated.

Response to Reviewer Comments

We thank the reviewers for their careful evaluation and constructive comments. The following clarifications address the main concerns while preserving the scope of the accepted diagnostic study. The accepted manuscript is intentionally a diagnostic study of a fixed two-rule DVL rather than a proposal for a new correction algorithm; the clarifications below rely exclusively on information already present in that manuscript.

1. Scope and contribution. The study intentionally evaluates exactly two deterministic scaling rules (`scale_mull100`, `scale_div100`) under a fixed model configuration. Its contribution is the cross-dataset reversal, the rule-level Transition Precision analysis, the `scale_div100` ablation, and the provenance-controlled evaluation supporting them (§1, §6). The paper does not claim to introduce a new correction algorithm or to characterize deterministic verification methods generally.

2. Interpretation and causal mechanism. §6 deliberately separates established findings from plausible hypotheses and explicitly untested explanations. The scale-distribution account is presented throughout as a hypothesis, not an established causal explanation. Because model and adapter are held fixed across both datasets (§4), the current evidence supports a dataset-contingent interaction under the evaluated model/output regime, rather than a universal dataset-only explanation (§4, §6, §7).

3. Model scope and role of the LLM. Both datasets use the identical Mistral-7B-Instruct-v0.3 + FinVerify-LoRA configuration. The DVL itself is deterministic post-hoc logic operating on the model’s raw prediction and question text; the LLM’s role is upstream, generating the candidate answer that the DVL then verifies. The paper evaluates whether this fixed rule pair is safe to apply to those outputs, not whether an LLM is required for the correction step. Cross-model generalization remains outside the scope of this study (§3, §4, §6, §7).

4. Provenance and statistical considerations. The manuscript intentionally discloses several provenance details — the FinQA/TAT-QA confidence-interval derivation asymmetry, the corrected TAT-QA ablation labeling, the excluded pre-fix run, and the excluded reconstruction attempt — so that readers can distinguish directly saved outputs from derived or corrected quantities (Appendix A). The zero-correction `scale_div100` result is based on the logged state transitions already reported in Table 2 and Figure 2. We note that the degenerate $[0.00, 0.00]$ percentile-bootstrap intervals for these zero-success cases are a limitation of the reported interval presentation; an exact-interval method would be more informative in a future methodological revision. Table 2 and all reported intervals are unchanged.

5. Context-aware alternatives and future work. Context-aware or scale-aware rule variants are a natural extension we did not evaluate here. A gold-scale oracle on TAT-QA would be an offline diagnostic, since gold scale information is unavailable to the DVL at inference time on either dataset. Cross-model replication and an explicit comparison of gold-answer scale distributions across datasets are natural next steps for determining how broadly the observed interaction generalizes; none of these were performed for the accepted study.