One Frozen Edit, Four Ways to Ask

November 18, 2026 — Research update v1.2

Prompt-conditioned functional effects in a transformer

Harmain Dastgir • Independent Researcher

What the new experiment adds

The original archived analysis evaluated 32 questions in a 1,024-cell factorial panel. A subsequent, fully completed GPU study replays the decisive historical contrast and evaluates 64 additional GSM8K train question identities on the same pinned Qwen2.5-1.5B-Instruct checkpoint and the same frozen rank-four model edit.

Under the new input-token-count matching condition, the signed-edit contrast averaged **+0.193491** (95% question-bootstrap interval **[+0.166896, +0.219002]**) and was positive for **62/64** questions. The matched orthogonal edit control averaged **+0.004060**. The 128-cell original replay matched exactly.

What these results do — and do not — say

They strengthen a **within-model functional response-geometry observation**. They do **not** measure mathematical answer accuracy, demonstrate a general-purpose model repair method, or independently replicate the result on a different model or hardware stack. Prompt-length matching uses explicit newline-token insertion and cannot remove all formatting confounds.

How to read this version

Part I: the complete earlier editorial manuscript (v1.1, November 16), preserved as published. **Part II:** the dated independent-question GPU confirmation supplement and two figures. This is an additive research update; it does not retroactively insert the new data into the original publication.

Original archival record: https://zenodo.org/records/24284827

Public repository: https://github.com/Harmain11/fluid-tensors

Working preprint / supplementary study • Not peer reviewed • Version-specific Zenodo DOI pending deposit

One Frozen Edit, Four Ways to Ask: Prompt-Conditioned Functional Effects in a Transformer

A large model-response effect under one wording changed sign after a rewrite—without changing the weight edit.

Harmain Dastgir | Independent Researcher | harmainmughal215@gmail.com

Research study • 16 November 2026 • Editorial revision v1.1 (author review; not peer reviewed)

EVIDENCE LEVEL | 1,024 archived factorial cells; 32 question-level units; one historical 1.5B-parameter model/edit setting; measured final-token logit response geometry, not answer accuracy.

Abstract

Model-editing effects can appear stable when measured with one prompt, while changing substantially under alternative formulations. We audit the preserved response-geometry outputs of one frozen, historically selected rank-4 edit to Qwen2.5-1.5B-Instruct. The complete CI V2.1 panel contains 1,024 records: 32 evaluation questions × two edit directions × four wordings × four instruction templates. All 32 questions were drawn from GSM8K's training split, but held out from the experiment's 12 calibration questions and earlier excluded IDs. For the historical signed edit, the mean math-plus-format versus generic attenuation contrast was +0.205542 in original wording and −0.020673 in reworded wording; their paired 32-question difference was +0.226216 (original 10,000-resample question-bootstrap 95% interval [+0.183538, +0.275323]). Attenuation measures the norm of a last-token logit displacement relative to its finite-difference local tangent; it is not answer accuracy. The 1,024 values match archived original-run records, all six plots regenerate exactly, and the 90-degree rotated control is norm matched and orthogonal in recorded effective-weight geometry. Because the original edit was historically selected, prompt lengths differ and no new model inference was performed for this audit, the finding is a bounded source-verified case study—not evidence of mathematical reasoning repair or a unique causal pathway.

1. A deceptively stable object: the frozen weight edit

An edited model may appear to have acquired a reliable behavior because a fixed prompt elicits a striking numerical response. But the weights are only one half of a transformer computation; the prompt controls the sequence of token embeddings and positions entering the same fixed network. A difference observed under one instruction template may be an interaction between the edit and the prompt, not an intrinsic property that transfers to every formulation of the task.

We ask a deliberately constrained question: when a weight edit is frozen, how much does its functional response depend on instruction wording and task framing? This inquiry grew from the broader Fluid Tensors program, in which activation pathways, localized projection edits and runtime interventions were studied across many versions. Most later general-purpose-repair and learned-editor attempts did not validate broad external task benefit. We isolate the small, better-supported phenomenon here rather than retrospectively presenting the whole program as a successful autonomous model repair system.

Editing robustness under paraphrase, contextual changes and task-format changes has already been studied, as has language-model sensitivity to prompt formatting (Ma et al., 2024; Sclar et al., 2024; Yan et al., 2025; Yang et al., 2025; Kang et al., 2026). We do not claim to discover prompt sensitivity or propose a new editing method. Our narrower contribution is an archived, source-traceable measurement of one selected low-rank edit's last-token logit-response geometry across a complete paired wording-by-instruction panel and a matched orthogonal direction control.

2. Research question, estimand and scope

For each question, we compare the archived attenuation score under the combined math-plus-format instruction with the generic instruction, holding direction and wording fixed. Define M(v,d) = mean_q [attenuation(q,d,v,math_format) − attenuation(q,d,v,generic)]. Positive contrasts mean that this measured attenuation index is higher under math-plus-format; they do not indicate more correct answers or stronger edit utility. The index is the relative difference between an observed final-token logit displacement norm and the norm predicted by its local linear tangent.

The primary wording contrast is C = mean_q { [attenuation(q,signed,original,math_format) − attenuation(q,signed,original,generic)] − [attenuation(q,signed,reworded,math_format) − attenuation(q,signed,reworded,generic)] }. Each question remains the unit of pairing; 1,024 factorial cells are not 1,024 independent questions. We also describe formal and terse wording as secondary contrasts and rotated direction as a direction control.

The inference scope is a single historically chosen Qwen2.5-1.5B edit, a fixed GSM8K-train-split evaluation cohort and one last-token logit-response statistic. The 32 evaluation question IDs were held out from this experiment's calibration/earlier-exclusion process, not from the benchmark's training split or necessarily from the base model's pretraining. The result does not generalize automatically to independently learned directions, different models, answer quality, reasoning improvement or neural mediation.

3. Design and audit methods

3.1. Complete factorial panel

The frozen CI V2.1 experiment used 12 calibration and 32 evaluation questions, all selected from GSM8K train under its historical exclusion rules. The evaluation panel crossed each of 32 question IDs with two directions (historical signed; norm-matched rotated_90), four wordings (original, reworded, formal, terse), and four instruction templates (generic, math_only, format_only, math_format). Every 32 × 2 × 4 × 4 = 1,024 stored cell is present and uniquely identified. The response table is repeated-measures evidence across 32 questions, not 1,024 independent trials. The protocol's 'held-out' label refers to study-level question separation rather than the public GSM8K test split; pretraining exposure cannot be ruled out.

Axis	Levels	Role
Held-out question	32 distinct questions	Paired uncertainty unit
Edit direction	Signed, rotated 90°	Specificity control
Wording	Original, reworded, formal, terse	Prompt sensitivity
Instruction	Generic, math, format, math+format	Functional factorial contrast
Total observations	1,024 archived cells	Repeated cells; not n=1,024 independent

3.2. Original edit and scoring discipline

The recovered run identifies Qwen/Qwen2.5-1.5B-Instruct (revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306), evaluated in FP32 on an NVIDIA L4 with TF32 disabled. A historical rank-4, alpha-8 low-rank additive edit was applied to model.model.layers[19].mlp.down_proj, with an effective delta norm cap of 0.008 × ||W||. The signed edit's selection originated in earlier studies and was not prospectively repeated for this audit. The recovered frozen 90° control uses a matched effective Frobenius norm and is nearly orthogonal (absolute cosine around 3.8×10^{-17}) to the signed direction; this validates a geometric negative control but does not identify a mechanism.

For edit strength s > 0, let r_s be the edited-versus-unedited final-token output-logit displacement vector, and t the local logit-response tangent estimated by symmetric finite differences (smallest stored step 0.005). The original code defines attenuation as follows. The reported full-strength rows have s = 1 and equal 1 − full_displacement_l2 / logit_tangent_l2. A positive value means less normed displacement than the local linear approximation; a negative value means more. This is neither probability of a correct answer nor a calibrated measure of reasoning performance.

$$attenuation(s) = 1 - \frac{\|r_s\|_2}{s\ \|t\|_2}$$

3.3. Bootstrap and source integrity

The original locked study's question-level bootstrap used 10,000 draws (seed 20261130), yielding the 95% percentile interval [+0.1835380530, +0.2753226289] for the 32-question paired original-minus-reworded contrast. The public reanalysis independently recalculates the same archived contrasts with 20,000 draws (seed 20261116), yielding [+0.1842555448, +0.2741488261]. These are different Monte Carlo summaries of the same fixed data, not replications with new independent model runs. A SHA-256 record identifies the public table; question IDs, primary pairing and resampling unit remain frozen.

3.4. Reanalysis pipeline and inferential unit

Reanalysis starts with the archive's 1,024-row CI_V2_RAW_RESPONSE_GEOMETRY.csv. Rows are keyed by question ID, edit direction, wording and instruction template; a deterministic loader checks complete factorial coverage, unique keys and finite values. Separately, the recovered original-run ci_v2_heldout_full.jsonl holds all 1,024 complete source records. Exact key, score, displacement norm, tangent norm, cosine and token-count comparisons agree with every released CSV row. All source-record attenuation calculations match the recovered code equation at stored precision. This verifies the saved measurement chain, not a new inference pass from the underlying 1.5B model.

For each question q, direction d and wording v, the functional contrast M(q,d,v) is the archived attenuation for math_format minus attenuation for generic. The primary per-question interaction is D(q) = M(q,signed,original) minus M(q,signed,reworded). The reported mean averages D over the 32 questions. A nonparametric percentile interval resamples entire question-level D values with replacement, using 20,000 draws and fixed seed 20261116. It is conditional on this observed question panel and historical edit selection.

3.5. Descriptive visualizations and control boundaries

Four principal figures show signed-edit wording contrasts, the rotated direction control, individual paired question differences, and the complete 4 × 4 wording-by-template mean response matrix for each direction. Supplementary figures show the resampling distribution and recorded prompt lengths. The matrix heatmaps use mean raw attenuation, whereas the first three figures use differences between template conditions; the two measurement scales should not be conflated. The plots use no new model inference calls.

The intervals and panels are descriptive. We do not use the 32 related question values as evidence that the historically chosen edit generalizes across architectures, task families or independently learned edit directions. Secondary contrasts were not prospective familywise-error-controlled tests. The plotted prompt-token counts identify imbalances but cannot identify which lexical, semantic, position or attention-context change produces the measured response.

3.6. Source-record reconciliation and archived control geometry

The recovered original-run archive verifies 22 file hashes and three checkpoint hashes. Its 1,024 held-out JSONL records reproduce all published CSV fields exactly, and the original fixed-seed 10,000-draw bootstrap reproduces the archived primary decision. Independent decoding of saved float32 adapter tensor storages confirms that the rotated_90 control has the same effective Frobenius norm as the historical edit (both 5.667346 at recorded numerical precision) and negligible directional cosine. These independent checks are at the stored-data and saved-tensor levels; they do not establish a new run of transformer forward inference.

4. Results

4.1. Wording changes the signed-edit response

Wording variant	Signed direction M	Interpretation within metric
Original	+0.205542	Positive functional contrast
Reworded	−0.020673	Near zero / weak negative
Formal	−0.045819	Negative contrast
Terse	−0.002337	Approximately zero
Rotated, original	−0.023924	Control direction negative

Within this stored panel, the historical signed edit's math-plus-format minus generic attenuation score averaged +0.205542 for original wording, but −0.020673, −0.045819 and −0.002337 for reworded, formal and terse wording. These are changes in a final-token response-norm *index*, not changes in answer correctness or an intervention-success rate. The independently constructed single 90° rotated direction did not reproduce the same original-wording contrast (rotated original mean −0.023924).

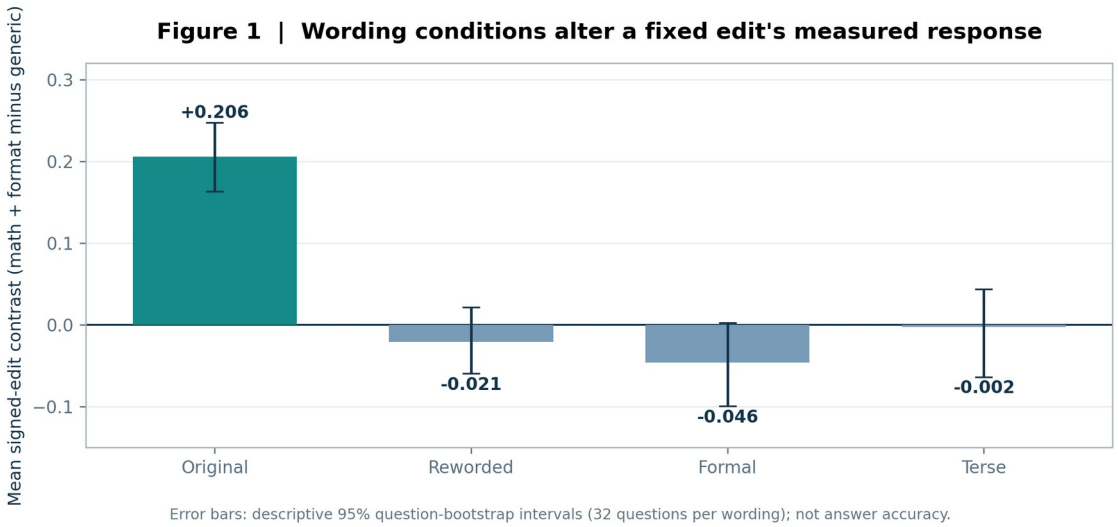


Figure 1. Historical signed-edit math-plus-format minus generic attenuation response under four wordings. Error bars are descriptive 95% question-bootstrap intervals (32 questions per wording), not accuracy intervals.

4.2. A paired contrast, not 1,024 independent trials

The 32-question mean paired original-minus-reworded signed contrast is +0.2262157365. The public 20,000-draw question-bootstrap 95% interval is [+0.1842555, +0.2741488], and all 32 paired differences are positive. This documents consistency within the observed fixed question set. It is not a prospective hypothesis test accounting for historical edit selection, benchmark dependence or correlated prompt families.

The original 10,000-draw locked-run decision reported [+0.183538, +0.275323]. Recomputing its frozen resampling scheme and original seed from the recovered detailed run records yields the same endpoints. The small difference from the later 20,000-draw interval reflects the changed seed and Monte Carlo stream, not a second model execution.

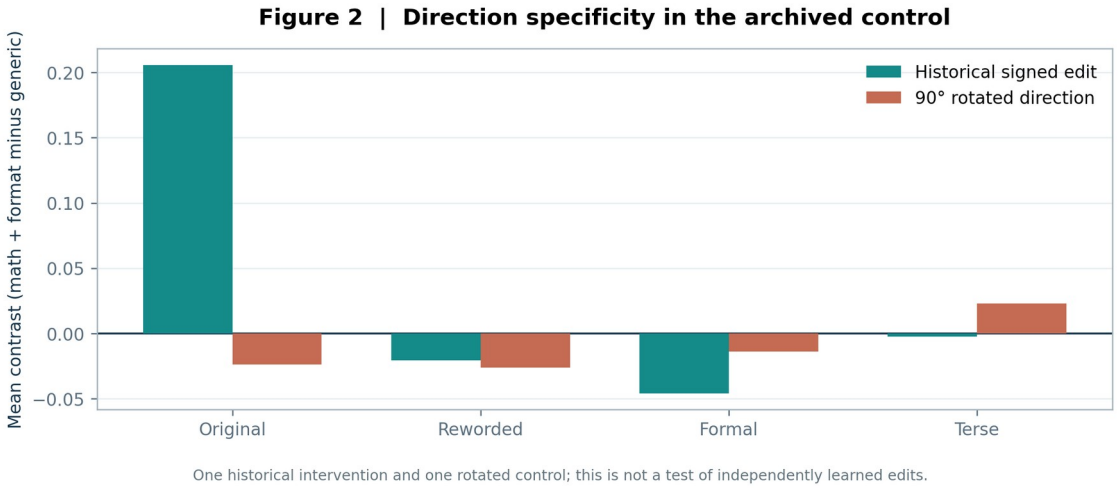


Figure 2. Signed historical edit versus a 90-degree rotated control on the same question and wording families. The comparison does not identify a causal circuit or a learned-editor replication.

4.3. Prompt-length imbalance is a concrete confound

In original wording, the mean stored prompt lengths were 100.25 tokens for generic, 102.25 for math_only, 112.25 for format_only and 114.25 for math_format. The math_format minus generic comparison adds 14 tokens for original and terse wording and 15 tokens for reworded and formal wording. Changes in instruction semantics, length and token positions therefore cannot be causally separated by this panel.

Original wording template	Mean recorded prompt tokens
Generic	100.25
Math only	102.25
Format only	112.25
Math + format	114.25

4.4. Question-wise effect heterogeneity

All 32 archived questions have a positive signed original-minus-reworded paired difference, but the magnitudes are heterogeneous. The mean difference is +0.226216 and the median is +0.241814. The observed minimum is +0.008832 and maximum is +0.712422; the interquartile range is [+0.155603, +0.289917]. Figure 3 displays each question as a single point, ordered by its observed paired effect. This is stronger evidence for within-panel consistency than the means alone, but the points share a benchmark source and edit-selection history.

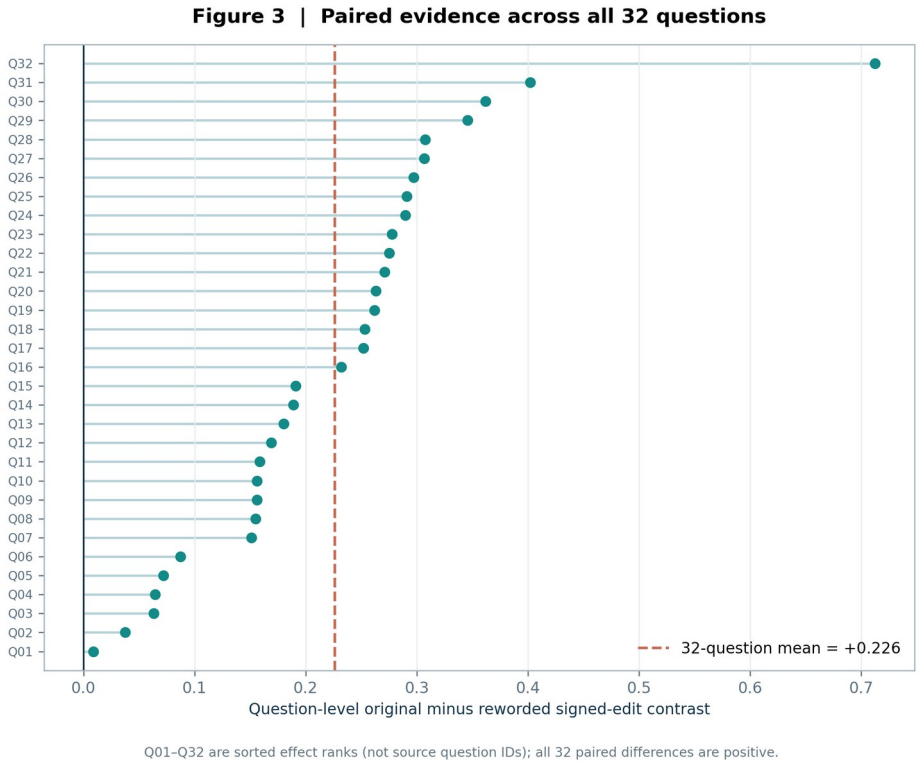


Figure 3. Thirty-two question-level signed-edit paired contrasts, ordered by effect rank (Q01–Q32 are ranks, not the original identifiers). Every value is positive; the dashed line is their mean.

4.5. Full factorial response patterns

Figure 4 expands beyond the single contrast to all four instruction templates and four wordings under both tested directions. The signed edit's original wording and combined math-plus-format template produced mean raw attenuation +0.205168, versus approximately zero for its original generic template (−0.000375). Under reworded questions, the signed format-only template had mean attenuation −0.121681, whereas the signed combined template was −0.007398. Those are raw archived responses, not independently calibrated reasoning scores. The rotated control occupies a different response region across these cells, demonstrating that direction and prompt context are entangled in the observed factorial data.

Figure 4 | The full wording-by-template response map

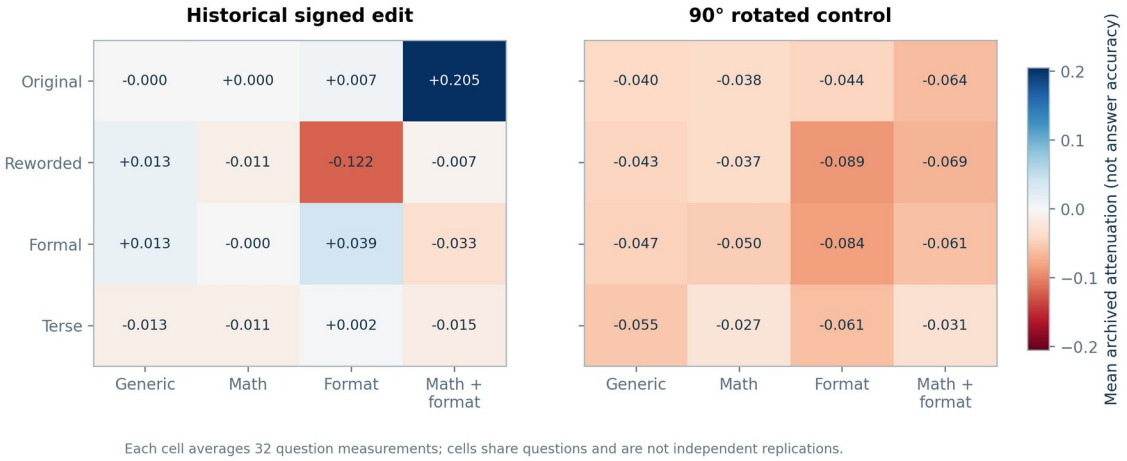


Figure 4. Mean raw attenuation across the complete 4 × 4 wording-by-instruction matrix for each direction. All cells average 32 question values and use the same color scale; this is not a map of answer correctness.

4.6. What the expanded plots cannot identify

The factorial map makes the historical interaction visually clear, but factorial variation is not by itself a mediation experiment. A changed instruction alters content, token count and token positions together. The original wording's average math-plus-format prompt length is 114.25 tokens versus 100.25 for generic, a difference of 14 tokens. Reworded and formal conditions differ by 15 tokens under the same comparison, while terse differs by 14. Supplementary Figure S2 exposes all four by four length means. No length-matched lexical placebo or causal-token-position control is available from this deposit, so the word 'prompt-conditioned' describes an observed interaction rather than an isolated semantic mechanism.

5. Relation to earlier Fluid Tensors stages

The CI V2.1 panel belongs to a longer sequence of Fluid Tensors experiments involving different response metrics, edit directions and exploratory controller ideas. The present audit includes complete source-level checks for the frozen CI V2.1 archive, but it does not verify every numerical claim from the preceding Grand V4/V5 trials. Those earlier studies should be evaluated separately before their detailed numbers are used as independent evidence for this manuscript.

The CI V2.1 logit-response effect should not be read as mathematical-answer improvement. Earlier task-level checks in the broader program were mixed or unfavorable, and no new blinded answer-correctness comparison was performed as part of this archival reanalysis. Keeping the response-norm measurement distinct from task utility is a central interpretive boundary of this study.

Later Fluid Tensors studies attempted to diagnose failure mechanisms and adapt interventions over generation time. Those questions are not settled by the frozen CI V2.1 panel. Its defensible contribution is a specific, source-audited response interaction, rather than a newly validated general-purpose repair method.

6. Competing explanations and mechanistic restraint

At least four explanations are consistent with the panel. First, the historical edit may align with token features peculiar to the original prompt. Second, additional or removed instruction tokens may shift the location where its effect is sampled. Third, different instruction templates may alter model confidence or response dynamics without changing correctness. Fourth, the selected historical edit could have been optimized, implicitly or explicitly, for a subset of prompt contexts, making its apparent sensitivity partly a selection artifact.

The rotated control makes an unconstrained ‘all directions behave the same’ explanation less plausible in the measured setting, but it does not identify a specific computational circuit. Establishing mediation would require controlled activation measurements, matched-length lexical placebos, systematic finite-strength sweeps, perturbations at rival layers, and cross-model prospective replication. No such complete mechanistic chain is claimed as a new observation here.

We also avoid using ‘phase transition’ for the response sign flip: four discrete wordings with different token lengths provide no evidence for a discontinuity as a smooth control parameter is varied. The observed phenomenon is a conditional functional interaction with a large contrast under one prompt family.

The recovered rotated-90 control is norm matched in effective-weight geometry: a direct tensor-storage recomputation yields a norm ratio of 1.0 and an absolute cosine near 3.8×10^{-17}. Thus a simple difference in effective edit norm does not explain the signed-versus-rotated comparison. Nevertheless, one nearly orthogonal control does not exclude other vector directions, layer interactions or prompt-position effects, nor does it identify a privileged causal circuit.

7. Threats to validity and negative results

The chief limitation is one historically selected edit in one small open-weight model, with one train-split-based question cohort. The archived evaluation questions were disjoint from experimental calibration and earlier excluded IDs but not from the GSM8K training split, and possible pretraining exposure cannot be assessed here. The unknown broader edit-selection search makes standard inferential intervals conditional rather than prospective discovery intervals.

Second, attenuation concerns a last-token logit-response norm rather than graded solutions or free-running generation. Larger scores need not imply better predictions. Third, wordings and templates change semantics, token counts and positions together. Fourth, a single norm-matched 90° rotated vector is insufficient to identify a unique mechanism or an edit effect general to independent directions.

Finally, there are only 32 question-level resampling units from a single benchmark family, and exploratory secondary contrasts were not corrected for historical selection or multiple tests. This audit independently reproduced archived metrics, table joins, statistics, six figures and saved-tensor control geometry; it did not rerun the 1.5B-parameter Qwen model or recover full-vocabulary baseline logits from an independently executed forward pass.

8. What a stronger test would need

To distinguish semantic prompt conditioning from length or token-position artifacts, a future preregistered study should match sequence length and positions while changing just the target semantic clause; then reverse the control by holding semantics fixed and varying length. It should freeze edit selection before held-out wording construction, evaluate several independently trained edit vectors and model families, and report both functional response and blinded final-answer correctness.

To support a causal pathway rather than a statistical interaction, it would also need internal activation mediation and carefully matched rotated, random and norm-equivalent controls. None of these future experiments is represented as complete here. The results support publishing a conditional functional effect with a candid limits section, not a new general-purpose model repair mechanism.

9. Conclusion

Under one frozen historical parameter-edit direction, the observed *last-token logit attenuation* depended strongly on the combined question wording and instruction template. The paired original-minus-reworded contrast was positive for all 32 evaluation questions, and a single norm-matched orthogonal direction did not reproduce the original-wording pattern. This is a source-supported within-panel measurement statement, not a prompt-invariant property of edits in general.

Before describing a localized edit as repairing reasoning, it is necessary to test matched-length prompts, alternate edit vectors and independently evaluated answer correctness. This paper establishes none of those broader outcomes. Its value is the transparent documentation of a bounded, conditional response interaction, and the demonstration of why apparently large edited-model response metrics require careful interpretation.

Data and reproducibility

The current public v1.0.0 repository contains the complete 1,024-row CI_V2_RAW_RESPONSE_GEOMETRY.csv, deterministic reanalysis, six figures and figure generation scripts. The exact table SHA-256 is 303c7a42509d2480107591bfd3c5c3467f657233aa272fba546fa73a2d374f77. The released source reproduces the reported primary table summaries and six figures; all six regenerated PNGs match the release byte for byte. Repository: https://github.com/Harmain11/fluid-tensors; preprint: https://doi.org/10.5281/zenodo.24280041.

Separately from the public v1.0.0 release, the author supplied a recovered original-run archive containing frozen protocol, exact source cells, runtime manifest, 44 question identities/texts, three hashed adapter checkpoints, four JSONL logs and the original decision report, as well as the source Colab notebook. All 22 manifest digests, 1,024 full held-out rows and their stored attenuation equations were verified. This material is not yet comprehensively deposited in the public release and must undergo

third-party redistribution-rights checks before publication. Neither public reanalysis nor the present archive audit independently repeated full GPU inference.

Declarations and evidence status

Author and correspondence: Harmain Dastgir, Independent Researcher; harmainmughal215@gmail.com. Funding: self-funded; no external research grant. Competing interests: the author reports none. Publication status: working preprint and editorial v1.1 draft, not peer reviewed. Original paper text, figures and numeric tables are offered under CC BY 4.0 and original repository code under its stated license; third-party benchmarks and pretrained model weights retain their own terms. AI tools assisted writing, statistical checking, figure scripting and document production; the author remains responsible for methodological decisions and claims.

References

[1] Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. NeurIPS. https://rome.baulab.info/

[2] Mitchell, E. et al. (2022). Fast Model Editing at Scale. ICLR. https://openreview.net/pdf?id=0DcZxeWfOPt

[3] Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. ICLR. https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html

[4] Ma, X., Ju, T., Qiu, J., Zhang, Z., Zhao, H., Liu, L., and Wang, Y. (2024). On the Robustness of Editing Large Language Models. EMNLP, 16197–16216. https://doi.org/10.18653/v1/2024.emnlp-main.906

[5] Yang, W. et al. (2025). The Mirage of Model Editing: Revisiting Evaluation in the Wild. ACL. https://aclanthology.org/2025.acl-long.745/

[6] Yan, J., Wang, F., Luo, Y., Li, Y., and Zhang, Y. (2025). Keys to Robust Edits: From Theoretical Insights to Practical Advances. ACL. https://aclanthology.org/2025.acl-long.1099/

[7] Kang, C., Baek, J., Kim, Y., and Choi, J. (2026). Beyond Paraphrases: Generalization Across Task Formats in Parametric Knowledge Editing. Findings of EMNLP. https://aclanthology.org/2026.findings-emnlp.1188/

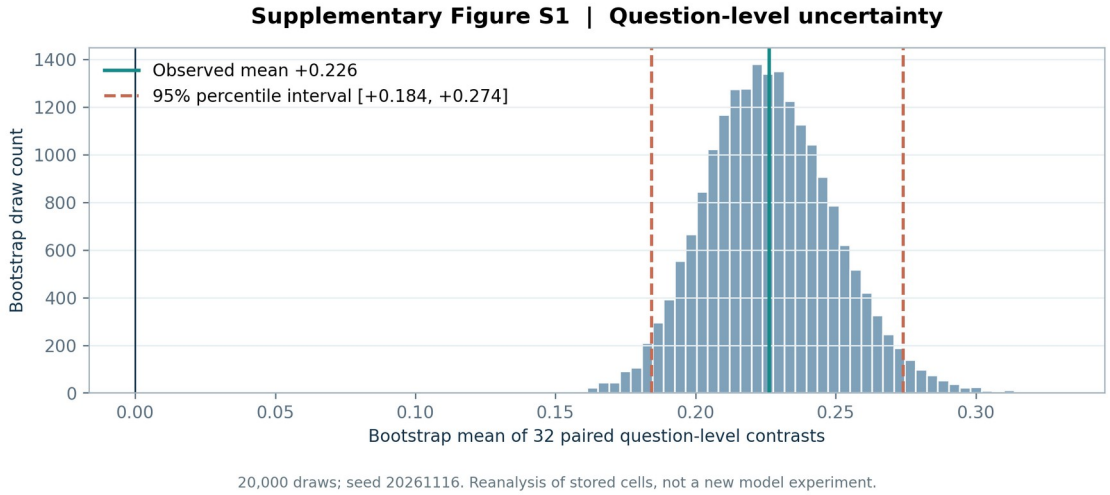
Appendix A. Scope ledger

Claim	Available evidence	Status
Frozen edit + wording dependence	1024-cell CI V2.1 matrix	Supported within setup
Direction specificity	One 90-degree rotated control	Supported in tested context
Prompts token/position matched	Mean lengths differ	Not established
Reasoning accuracy gain	Response attenuation, not correctness	Not established
Independently learned edit generality	Historic edit and failed new-arm gates	Not established
Identified causal circuit	No complete mediation chain	Not established

Appendix B. Supplementary uncertainty and length diagnostics

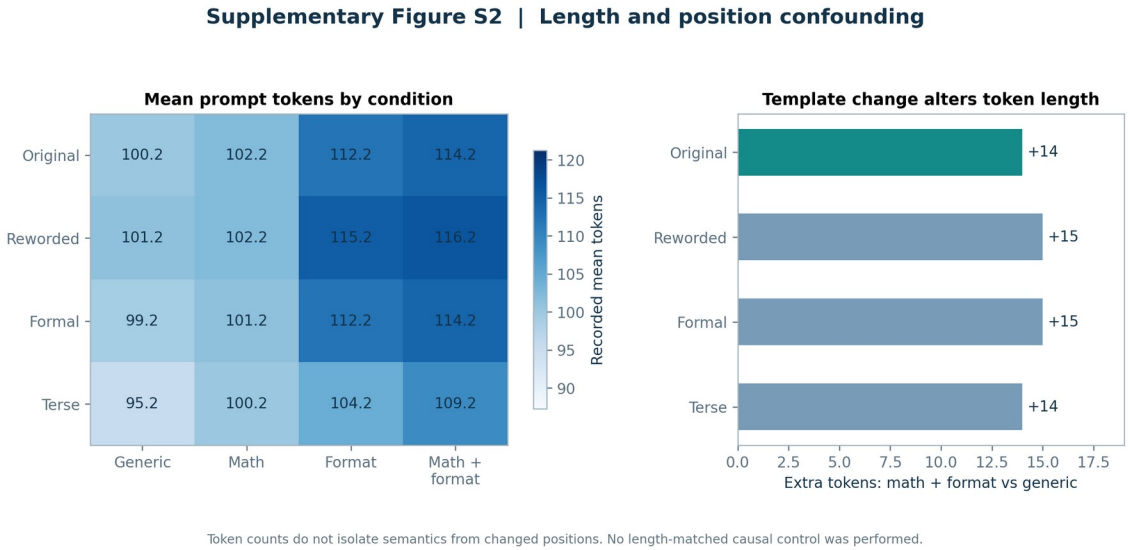
Supplementary Figure S1 shows 20,000 question-level bootstrap means for the original-minus-reworded signed-edit contrast. Its percentile 95% interval is [+0.184256, +0.274149], and its sampling unit is the

question, not the 1,024 stored factorial cells. It measures internal sampling variability conditional on this dataset and cannot adjust for prior edit selection.



Supplementary Figure S1. Resampled question-level paired contrast means (20,000 draws, fixed seed 20261116); vertical dashed lines mark the percentile interval.

Supplementary Figure S2 reports the recorded mean number of prompt tokens for each wording and instruction template. The math-plus-format versus generic contrast adds 14 or 15 tokens per question depending on wording. Neither an adjustment for this small difference nor the visual display demonstrates that the effect is due to tokens alone; meaning and position change together, and prospectively matched control inputs are needed.



Supplementary Figure S2. Average prompt tokens for 4 × 4 conditions and the measured increase from generic to combined math-plus-format templates. This documents length confounding, not a causal length model.

Paper 1 — Independent-Question GPU Confirmation Addendum

Associated publication: One Frozen Edit, Four Ways to Ask: Prompt-Conditioned Functional Effects in a Transformer

Author: Harmain Dastgir, Independent Researcher **Associated archival record:** https://zenodo.org/records/24284827 **Confirmation analysis date:** November 18, 2026 **Status:** Complete within-model confirmation experiment; this document is a supplementary manuscript draft, **not** a separately peer-reviewed publication.

Abstract

We conducted an additional within-model confirmation of a previously identified prompt-conditioned response-geometry effect under a frozen low-rank modification. A pinned Qwen2.5-1.5B-Instruct checkpoint and the original rank-four, layer-19 modification were preserved. We independently replayed the preexisting primary contrast on the archived 32-question cohort, then evaluated the same contrast on 64 further GSM8K questions selected deterministically after exclusion of the frozen historical question registry. The new cohort was tested under both ordinary prompts and explicit token-count matching, alongside a norm-matched orthogonal modification. Across 1,152 completed measurement cells, all prespecified numerical reliability checks passed. On 64 token-count-matched questions, the primary signed-edit contrast averaged **0.19349**, with a question-cluster bootstrap 95% interval of **[0.16690, 0.21900]** and positive paired contrasts on **62/64 questions**. The norm-matched orthogonal control averaged **0.00406**, with interval **[−0.00461, 0.01263]**. The paired difference between signed and orthogonal contrasts averaged **0.18943**, interval **[0.16204, 0.21651]**. The result confirms the narrowly defined prompt-conditioned last-token functional response pattern within this frozen model/edit configuration. It does **not** establish improved task accuracy, an isolated causal circuit, unseen-model generalization, or a general-purpose repair mechanism.

1. Research question and scope

Does the previously observed difference in the full-strength response to a fixed model edit across mathematical prompt formats persist on additional questions when the final input token counts are matched? Is the same difference comparatively small for an equally sized edit pointing in an orthogonal parameter direction?

Scope statement: This is a **new-question GPU confirmation within one model**, not a fully independent cross-model or independent-laboratory replication. The same frozen edit, model revision, and software/GPU configuration were used. Claims in the associated paper must retain this distinction.

2. Frozen model and intervention

- Base model: Qwen/Qwen2.5-1.5B-Instruct.
- Model revision: 989aa7980e4cf806f80c7fef2b1adb7bc71aa306.
- Device and runtime: NVIDIA L4 GPU; Python 3.13.16; PyTorch 2.14.1+cu130; Transformers 4.57.3; float32 model; TF32 disabled.
- Target: model.model.layers[19].mlp.down_proj.
- Editing mechanism: original frozen rank-four low-rank factor intervention using the archive's checkpoint and runtime hook implementation; no retraining in this study.
- Direction control: an output-factor rotation of 90 degrees with equal Frobenius norm. The recorded relative norm was **1.000000** and the inter-direction cosine **3.63×10^{-17}**.
- Frozen base-weight digest: 56c3c552a24bfae10426abad4e538cf81497b59ea97d277e70c404d0801441e5 (SHA-256).

3. Question cohorts and experimental design

Two distinct sets of question identities were evaluated:

1. **Original replay cohort:** the 32 archived held-out GSM8K questions. Only the previously decisive signed-edit contrast was replayed: 2 wording variants × 2 prompt templates = **128 cells**. These were genuinely recomputed by the archived model pipeline, not used in place of new-question measurements.
2. **New cohort:** 64 deterministically selected GSM8K train questions whose SHA-256 identities did not overlap the 2,975 identities in the study's frozen historical exclusion list. The notebook recorded zero intersections with those excluded identities. The new cohort was tested on signed and orthogonal edit directions, original and reworded wording variants, and generic and math-format prompts: **512 raw-prompt cells** plus **512 token-count-matched cells**.

Total: **1,152 distinct recorded measurement cells**.

Data-exposure qualification: The new set is independent of the archived historical study identities, but it comes from the GSM8K **training** split, which may have been included in model pretraining. In addition, overlap with any later unpublished question sets not present in that archival exclusion registry cannot be ruled out without their hashes.

Prompt-length control

For each new question, all four wording/template conditions were adjusted to have the same number of input tokens by inserting an exact number of explicit **newline token IDs immediately after the chat role header**. The original tokenizer output for each question/prompt was preserved after this padding. All 64 questions passed the token-count equality checks; 256 prompt-condition inputs were matched, requiring between 0 and 16 additional token IDs per condition.

Interpretive qualification: This input-level operation may produce a noncanonical byte-pair-encoding sequence. It changes formatting and model context. It equalizes the final input length; it does not guarantee semantic neutrality or eliminate every prompt-position/content/attention confound. Results should be described as persistence under this specific token-count-matching intervention.

4. Outcome and statistical analysis

The recorded measurement was **last-token logit response attenuation** relative to a centered finite-difference tangent at zero intervention. In shorthand,

$$A = 1 - ||z(1) - z(0)||_2 \,/\, ||z'(0)||_2$$

where z(s) is the final-token logit vector under edit strength s, and the local derivative was estimated with centered steps 0.005 and 0.01. This is a response-geometry statistic, **not the correctness rate on a math task**.

For every question, the primary paired contrast was

$$\Delta(q) = [A(\text{original, math}) - A(\text{original, generic})] - [A(\text{reworded, math}) - A(\text{reworded, generic})]$$

Inference used **one question as the resampling unit** with 10,000 deterministic bootstrap draws (seed 20261118). The frozen decision rule required a matched signed-edit sample mean of at least 0.10, positive lower 95% bound, complete cells, acceptable numerical diagnostics, and a matching archive replay. The matched result passes those gates.

5. Results

Condition	Questions	Mean question-level contrast	95% question-bootstrap interval	Positive question pairs
Original 32-question replay, signed edit	32	**0.226216**	[0.183103, 0.274155]	32/32
New raw prompts, signed edit	64	**0.229246**	[0.197980, 0.259130]	61/64
New matched prompts, signed edit **(primary)**	64	**0.193491**	[0.166896, 0.219002]	62/64
New matched prompts, orthogonal control	64	**0.004060**	[−0.004614, 0.012629]	35/64
New raw prompts, orthogonal control	64	**0.008959**	[−0.001434, 0.019347]	36/64

Two paired secondary contrasts:

- Signed minus orthogonal direction under matched prompts: **0.189431**; 95% interval **[0.162041, 0.216506]**.
- Matched minus raw signed-edit question contrast: **−0.035755**; 95% interval **[−0.048599, −0.023462]**.

The raw-to-matched reduction is approximately **15.6%** of the raw-prompt mean. It is evidence that prompt length/formatting choices affect the measured magnitude, **not** that they fully explain the effect. The primary contrast remains positive under the particular length-matching operation evaluated.

Numerical and provenance checks

- 1,152/1,152 expected cells present; no duplicate or nonfinite grid entries in the final audit.
- The original 128 replayed attenuation values matched their archive comparators exactly in the recorded GPU environment (maximum absolute deviation 0). This is **same-runtime reproducibility**, not external hardware/laboratory independence.
- 0/1,152 numeric-quality violations; maximum centered-tangent convergence ratio approximately **0.00334** (threshold **0.03**).
- The baseline repeat displacement was 0 in the recorded run, and layer-19 linearity diagnostics passed.
- All 26 archived file checksums in the results ZIP matched the included SHA-256 manifest.

6. Additional robustness analyses (post hoc; descriptive only)

The following analyses were performed **after inspecting the finished experimental results**. They are diagnostic checks, not additional preregistered hypothesis tests.

- Matched signed-edit contrast median: **0.198922**.
- Trimmed mean after dropping the six smallest and six largest question contrasts: **0.193911**.
- Leave-one-question-out mean range: **0.189599–0.198096**.
- By question text length (shortest 22 / middle 21 / longest 21 questions), the matched signed-edit means were **0.187585 / 0.212618 / 0.180552**, respectively. Each group was positive in the sample; this is not evidence of invariant effects in every text-length population.
- The matched and raw signed-edit question contrasts had Pearson correlation **0.9129**. This shows within-sample rank-order stability, not causal invariance under arbitrary formatting changes.

These checks reduce concern that the average is the result of one or two influential questions. They **cannot** resolve independent-dataset contamination, cross-model transportability, or the effects of a noncanonical padding procedure.

7. Interpretation, novelty, and appropriate claims

The new results strengthen the original within-model observation by adding a fixed fresh-question cohort, a measured length-matching condition, a norm-matched direction control, complete numerical checks, and a preserved audit trail. The most specific defensible conclusion is:

> A frozen rank-four intervention in Qwen2.5-1.5B-Instruct shows a reproducible prompt-conditioned difference in last-token response geometry across additional GSM8K question identities. The prespecified contrast persists under explicit input-token-count matching, while an equally sized orthogonal edit shows a much smaller contrast.

This is a **functional sensitivity finding**. It does not establish that the intervention improves model accuracy, identifies a single causal computation pathway, predicts results on other model families, or provides an operational general-purpose failure-repair system.

8. Study limitations and what is, and is not, resolved

Concern	What this confirmation changes	Residual limitation
Limited questions in original result	Adds 64 disjoint archived-history identities	New cohort is GSM8K train; not independent of all model pretraining
Original result might be numerically unstable	Archive replay matched; all numerical gates passed	Same L4/software environment, not external laboratory replication
Finding might reflect any equal-sized weight edit	Orthogonal control is near zero	Only one fixed edit and one constructed orthogonal comparator
Prompt length may affect the result	Effect persists with equal input-token counts	Padding itself changes formatting/tokenization; not full isolation
Outlier-driven average	62/64 matched contrasts positive; trimmed and leave-one-out means robust	Sample-level diagnostics do not prove population universality
Cross-model transfer	Not addressed	Requires new models and likely newly characterized interventions
Math reasoning improvement	Not measured	Requires answer generation and separate correctness evaluation

9. Reproducibility and ethics statement

The complete evidence archive supplied with this addendum contains raw JSONL response measurements, frozen question identities, archived exact original source, the replacement confirmation source, model and checkpoint hashes, environment information, paired question-level data, plots, numeric gate results, a verifier script, and an SHA-256 manifest. The analysis did not use GSM8K gold answer fields or retrain the model in this confirmation run. No participant data or human-subject claims are involved. All model and dataset provenance and licenses should be checked against the associated original publication before public redistribution of checkpoints.

10. Final study decision

Planned within-model independent-question GPU confirmation: COMPLETE.

The notebook's prespecified primary criterion passed. For the narrowly stated Paper 1 claim, **no additional GPU run is presently required to report this study**. A cross-model, external-benchmark, or independent-laboratory replication would be a distinct expansion, not an unperformed step hidden in this protocol.

Publication handling: Retain the already published GitHub and Zenodo original. Add this clearly dated confirmation as a versioned supplement or related follow-up artifact after checking the existing record/repository linkage and

journal-review status. Do not silently overwrite a fixed archival version or imply that these November 18 results were available at the time of the original publication.

Figures from the archived confirmation run

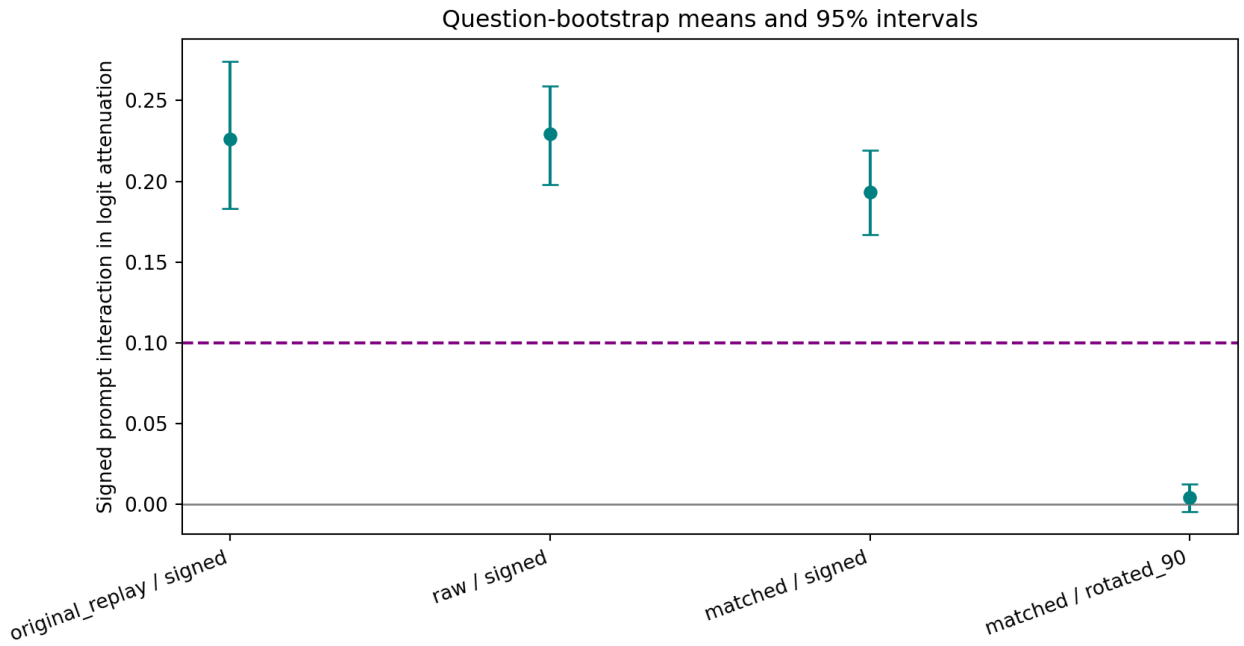


Figure A1. Comparisons in the new cohort. The primary signed intervention remains positive under the recorded length matching, whereas the orthogonal direction produces a substantially smaller contrast. Bootstrap intervals resample questions, not individual factorial cells.

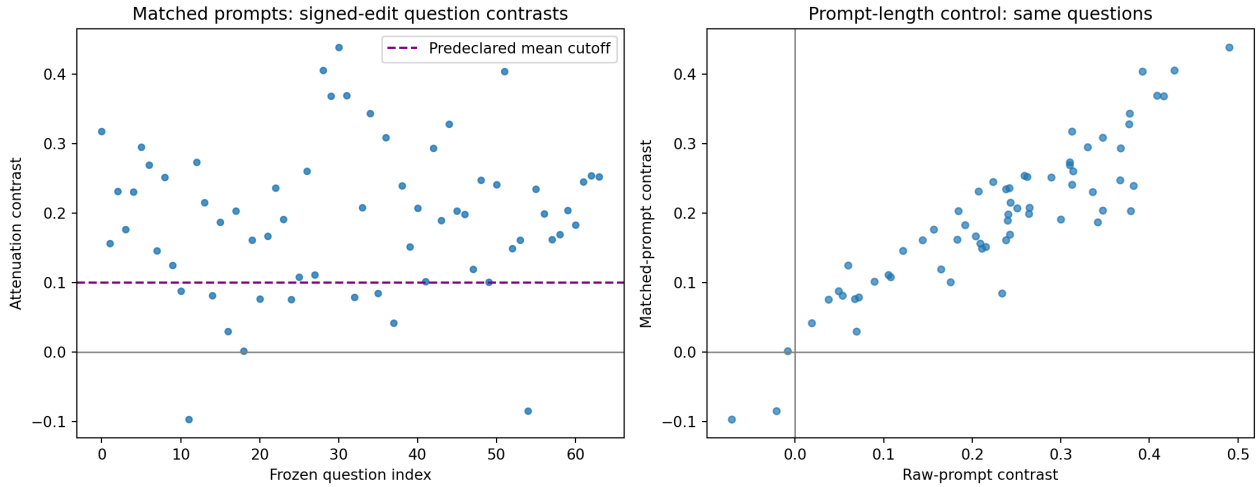


Figure A2. Raw and token-count-matched inputs. The difference in magnitude is itself part of the scientific finding: the prompt matching manipulation did not leave the effect entirely unchanged.