
Changing the Counterfactual Changes Which Circuit Heads Are Recovered

Ranveer Gill
Independent researcher

With
Apart Research

Abstract

A causal explanation used in an AI incident investigation should survive reasonable changes to how it is tested. This report audits causal-interp, an open pipeline evaluated on three published circuits in two small transformers. In the docstring task, changing the counterfactual while retaining the model, prompts, metric, and cutoff changes recovery from three of six published heads to five of six. Preserving the answer token exposes two routing heads hidden by the default corruption, although precision falls. Comparing multiple counterfactuals detects disagreement but does not reliably identify which explanation to trust. Ten resamples show that reproducibility alone does not resolve this problem. A separate IOI resample-ablation experiment highlights two heads missed by patching, yet 25 of 200 random head sets pass its recovery threshold. The artifact supports auditing mechanistic evidence; it does not establish incident detection, containment, or reliable causal attribution in frontier agents. All experiments predate this sprint submission.

1. Introduction

The question behind this project is practical: when an intervention produces a plausible circuit, how can an investigator tell whether it captures the mechanism or only what that intervention makes visible? For incident response, an incomplete explanation could direct attention to the wrong component or create unwarranted confidence in a proposed mitigation. This is a methodological risk, not a demonstrated failure in a deployed incident-response system.

The submission fits the **Open track**: its artifact is *causal-interp*, an Apache-2.0 repository of intervention code, experiment plans, measurements, and failure analyses [1]. Its contribution is a checkable sequence of case studies separating three questions: can we measure an effect, does it reproduce, and does it support the intended mechanistic claim? The last question remains unresolved.

Scope and provenance. Experiments were committed from August 13 to September 3, 2026. September 12 commits revise documentation. This submission synthesizes and audits that prior work for the sprint; it claims no new sprint-period model experiments. Appendix B records the evidence boundary.

2. Related Work

The evaluation targets published attention-head sets for indirect object identification (IOI) [2], greater-than prediction [3], and Python docstring completion [4]. These provide reference explanations against which recovery can be measured. Matching their head lists is a benchmark, not proof that a discovered mechanism is complete or uniquely correct.

Sensitivity to experimental design is established prior work. Zhang and Nanda [5] examine how activation-patching metrics and corruptions change localization results; Miller, Chughtai, and Saunders [6] show that circuit faithfulness depends on ablation choices. The contribution here is a concrete, inspectable replication and stress test, including failed attempts to choose trustworthy experiments automatically. It is not the first demonstration of methodological sensitivity.

Chan et al. [7] introduce causal scrubbing through correspondences between model and interpretation graphs. Phase 12 implements a simpler head-set resample ablation, described in the repository as a degenerate interpretation. Its results concern that implementation and its sampling assumptions, not the limits of full causal scrubbing.

3. Methods

3.1 Models, tasks, and recovery criterion

IOI and greater-than use GPT-2 small (12 layers, 12 heads per layer). Docstring uses Attn_Only_4L512W_C4_Code, alias attn-only-4l (4 layers, 8 heads per layer). The principal sweeps use 128 synthetic prompts per scheme and seed 0. The reference sets contain 26, 7, and 6 heads respectively [1-4]. For greater-than, seven-head recovery evaluates the attention-head subset; it is not recovery of the entire computation, which also involves MLPs.

For each head and tested position, the pipeline copies a clean activation into the corresponding corrupted run and measures normalized recovery. If D denotes the batch-mean task metric, the score is:

$$E(\text{head}, \text{position}) = (D_{\text{patched}} - D_{\text{corrupt}}) / (D_{\text{clean}} - D_{\text{corrupt}}).$$

A head is discovered when its largest absolute score over the tested positions is at least 0.02. Negative effects count. IOI uses the indirect-object minus subject logit; docstring uses the answer logit minus the largest competing argument logit. Greater-than uses a probability-difference task metric despite shared code naming. Precision is reference matches divided by all discovered heads; recall is reference matches divided by the reference-set size. Size-matched top-k results are reported separately.

3.2 Controlled comparison and later checks

The docstring generator draws the same clean examples at a fixed seed across schemes. *random_random* changes definition and docstring arguments; *random_answer* changes the answer; *random_def* preserves the answer while changing non-answer definition arguments. The model, clean prompts, metric, tested positions, and cutoff remain fixed. Normalization spans differ, so scores are conditional on the corruption, not directly comparable raw causal effects.

Phase 8 compares discovered sets across registered schemes. Phase 9 replaces the shared cutoff with per-scheme shuffled-source null floors. Phase 11 redraws prompts and corruption instances jointly over seeds 0-9, with 128 prompts per run. Phase 12 tests IOI head sets on clean prompts using resampled activations outside each set. Details, deviations, and file-level evidence appear in Appendices A-C.

4. Results

4.1 Counterfactual choice changes which heads are recovered

The strongest controlled result is the docstring contrast in Table 1. Default corruption recovers 3/6 reference heads; preserving the answer recovers 5/6. The extra heads are 1.4 and 2.0. This is a gain of two reference heads, not an across-the-board improvement: the discovered set grows from 9 to 17 heads and precision decreases from 0.333 to 0.294 [1, phases 7-8].

Counterfactual	Answer retained?	Found / reference	All found	Precision
random_random	No	3 / 6	9	0.333
random_def	Yes	5 / 6	17	0.294
random_answer	No	3 / 6	11	0.273

Table 1. Docstring activation-patching recovery at $|E| \geq 0.02$, 128 prompts, seed 0. Counts recalculated from phase7_sweep.json and checked against phase8_docstring.json. Reference heads are benchmark labels, so non-matches are not proven irrelevant.

Published docstring heads recovered at the same cutoff

	0.5	1.2	1.4	2.0	3.0	3.6
random_random						
random_def						
random_answer						

Figure 1. Filled cells indicate a reference head above cutoff; empty cells indicate a measured head below cutoff. The answer-preserving scheme exposes routing heads 1.4 and 2.0; head 1.2 remains missed. Data: phase7_sweep.json.

Observation. Head 1.4 changes from $E = -0.0156$ to $+0.6023$, and head 2.0 from $+0.0057$ to $+1.0906$. The clean metric is 0.6332 in both cases; corrupted values are -5.8142 and -1.9466. Recovery can exceed one and is not an accuracy percentage.

Interpretation. The published circuit routes attention to an argument and moves its content [4]. Restoring routing alone need not recover the answer when the answer token itself has been replaced. Keeping that token available makes the routing contribution measurable. The code and results support this explanation, but do not isolate every intermediate causal edge.

4.2 Disagreement detects sensitivity, not correctness

Across five docstring schemes, the union finds all six reference heads but includes 26 heads in total (precision 0.231). The primary-scheme warning flags 17 additional heads, only three of which match the reference. On greater-than, the primary already finds 7/7; four-scheme union adds no reference heads and lowers precision from 0.778 to 0.280. The same warning can signal a useful omission or an unhelpful expansion [1, phase 8].

4.3 Calibration and replication do not settle which result to trust

Per-scheme null calibration improves the docstring union precision to 0.625, with recall 5/6, but still leaves a warning on greater-than without an additional reference head. It controls one source of inflated scores without validating the counterfactual itself [1, phase 9]. Task induction also falls short: the registered greater-than run recovers 3/7 at a size-matched cutoff, versus 6/7 for the hand-built comparator. A disclosed repair reaches 5/7 and remains a separate post-hoc result [1, phase 10].

Across nine circuit-scheme rows in phase 11, dividing the mean effect magnitude by its replication standard deviation reduces reference-head AUC by a median 0.045 relative to mean magnitude. The repository reports exact Wilcoxon $p = 0.0078$. Seven of eight nonempty scheme rows have median rediscovery frequency 10/10 under the phase-9 thresholds; only four rediscover every seed-0 selected head in all runs. This tests reproducibility under the specified resampling axis, not correctness or robustness to arbitrary distribution shift.

A post-hoc diagnosis finds positive correlations between replication spread and effect magnitude in all nine rows (Spearman 0.577-0.863). Dividing by this spread can suppress informative ranking differences. The rows share circuits and models; their reported p-value should not be read as nine independent demonstrations of generality [1, phase 11].

4.4 Resample ablation adds evidence, but a pass is permissive

Phase 12 retains a proposed IOI head set and resamples other heads from eight aligned clean-prompt batches. MLPs remain active in the primary condition. The reference set recovers 1.022 ± 0.025 of the normalized logit difference (mean $\pm$ SD across 20 draws). Yet 25/200 uniformly random 26-head sets exceed the registered 0.70 logit-recovery bar, and eight meet or exceed the reference score. Those counts are post-hoc diagnostics of that bar, not passes on every registered criterion.

IOI head-set condition	Logit recovery	What the comparison shows
Published 26 heads	1.022	High score under primary ablation
Published minus 12 heads	0.911	Mean of 20 random deletions
Greedy 3-head set	0.968	High recovery need not imply completeness
Published 26; MLPs also resampled	0.050	Result depends strongly on scaffolding

Table 2. Phase-12 normalized logit recovery, recalculated from phase12_scrub.json. Deletion mean aggregates 20 sets; named sets use 20 draws each. The MLP condition is secondary and changes what is preserved. The score is not raw task accuracy.

Leave-one-out ablation from the reference set ranks previous-token heads 4.11 and 2.2 first and third, with drops 0.316 and 0.177. Their maximum phase-1 patching magnitudes are only 0.0028 and 0.0023, below cutoff. This uses the reference circuit as context. Separately, adding each head to an empty set yields a reference-label-free ranking with AUC 0.799 and Spearman 0.240 against patching magnitude; patching AUC is higher, at 0.936. Low correlation indicates a different ranking, not statistical independence [1, phase 12].

5. Discussion and Limitations

These experiments support a narrower and more useful claim than autonomous circuit validation: **an intervention pipeline can produce stable, incomplete mechanistic evidence**. In docstring, a targeted counterfactual reveals two reference heads that the default misses. Agreement checks make that dependence visible, but neither unioning findings nor ranking their stability supplies a general rule for choosing the right explanation.

For an investigator with authorized access to model internals, the repository offers an experimental audit instrument. It can record which components respond to which interventions and where conclusions change. It cannot infer an agent's intent, establish why a real intrusion occurred, or certify that a mitigation prevents recurrence. No incident logs, frontier-agent trajectories, or containment tests were evaluated in these experiments.

An evidence standard suggested by the results

A mechanistic incident claim should retain its experimental context: the checkpoint and prompt set, the behavior metric, what the counterfactual changes and preserves, the tested positions, raw baselines, normalization, and selection rule. Reports should distinguish a zero effect from an intervention that cannot test the proposed role. In IOI, phase 1 records 576/576 exact-zero cells at four positions before the altered S2 token; identical causal prefixes explain this specific blind spot [1].

A second, differently designed test can expose missing evidence, but should be reported alongside the first. The phase-12 leave-one-out result depends on a known candidate set, whereas its all-head single-addition ranking does not. Combining those as if they were one discovery result would erase an important condition on the claim. Similarly, a high ablation recovery score should be compared with matched random sets and more than one behavior metric.

Limits and a concrete next month of work

The evidence covers three benchmark circuits and two small models. Reference lists are incomplete proxies for causal truth; prompt distributions are narrow; some analyses are post-hoc. The resample-ablation results also depend on retaining MLP computation. Appendix A specifies these limits and their consequences.

The next experiment should encode an explicit IOI interpretation graph, including candidate information flow into name-mover queries and S-inhibition values, and test edge-conditioned resampling. Precommit the candidate graph, metric, reference-free acceptance rule, and matched null before evaluating a held-out prompt family. Then repeat on docstring and greater-than. A useful outcome would be rejecting incomplete explanations without merely penalizing all small circuits; a failure would be continued acceptance of matched null explanations. This is proposed work, not a result reported here.

6. Conclusion

Changing a counterfactual changes which published docstring heads are recovered, from 3/6 to 5/6 at a fixed cutoff, while precision falls. Resampling shows that stable measurements do not resolve the interpretation problem. Head-set ablation supplies a complementary view but a permissive circuit-level score. The contribution is an auditable tool and a concrete warning about the evidence it produces, not a solved method for causal incident attribution.

Code and Data

github.com/ranveerlabs/causal-interp [1]. Snapshot: **c36289a**. Code, generated-task fixtures, measurements, plans, and reports are public under Apache-2.0. Appendix C maps claims to files.

References

- [1] Ranveer Gill. 2026. *causal-interp*. GitHub repository, snapshot c36289a (September 12, 2026). [Repository and results](#).
- [2] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small*. arXiv:2211.00593. [arXiv:2211.00593](#).
- [3] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. *How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model*. NeurIPS 2023. [arXiv:2305.00586](#).
- [4] Stefan Heimersheim and Jett Janiak. 2023. *A circuit for Python docstrings in a 4-layer attention-only transformer*. LessWrong. [Original circuit write-up](#).
- [5] Fred Zhang and Neel Nanda. 2024. *Towards Best Practices of Activation Patching in Language Models: Metrics and Methods*. ICLR 2024. [arXiv:2309.16042](#).
- [6] Joseph Miller, Bilal Chughtai, and William Saunders. 2024. *Transformer Circuit Faithfulness Metrics are not Robust*. CoLM 2024. [arXiv:2407.08734](#).
- [7] Lawrence Chan et al. 2022. *Causal Scrubbing: a method for rigorously testing interpretability hypotheses*. Redwood Research, AI Alignment Forum. [Method and interpretation-graph formalism](#).
- [8] Apart Research and CeSIA. 2026. *AI Incident Response Sprint*. Event guidelines and official submission template, checked September 13, 2026. [Sprint guidelines](#).

Official template: [Apart Research hackathon submission template](#). The report preserves the template's US Letter format, one-inch margins, Old Standard TT typography, title treatment, and section sequence. The event-specific abstract limit and mandatory appendix take precedence over generic template guidance.

Appendix A. Limitations and Dual-Use Considerations

A.1 Scope and validity

Small, selected benchmarks. The three tasks are established interpretability benchmarks with available head lists, not a representative sample of models or incidents. IOI and greater-than share GPT-2 small. The study cannot estimate how often counterfactual sensitivity occurs in frontier models. Seven-head greater-than recovery omits its MLP computation.

Reference labels. Precision, recall, and AUC measure agreement with published explanations. A head outside a list may still contribute, and an alternative valid circuit may differ from that list. The results therefore establish reference recovery and experimental sensitivity, rather than a unique ground-truth mechanism.

Selection and normalization. The headline cutoff is 0.02 after maximizing over tested positions. Counterfactuals change both the intervened computation and the clean-corrupt span. The 3/6 versus 5/6 contrast is not a formal statistical test of general superiority, and precision decreases. Per-scheme calibration is a later experiment, not a retroactive replacement of the original result.

Replication. Phase 11 varies dataset seed and redraws both prompts and corruption instances. It does not test new model checkpoints, adversarial prompt distributions, or independent laboratories. Its nine scheme rows are clustered within two circuits. The reported Wilcoxon result is descriptive evidence within this design; inference to a population of tasks is limited.

Resample-ablation assumptions. Phase 12 uses 128 IOI prompts, seed 0, with eight replacement batches at seeds 1000-1007. Replacement rows preserve template and token-position alignment. Each ablated head gets an independently selected source per row in the primary condition. Kept heads continue computing with potentially altered upstream inputs; MLPs are retained. These choices may preserve useful scaffolding and permit incomplete head sets to pass. This is not a full interpretation-graph test.

Metrics and deviations. Published-set normalized logit recovery is 1.022, KL recovery 0.870, and normalized top-1 recovery 0.604. These are distinct behavior summaries, not interchangeable accuracy rates. Named candidates and deletion/addition experiments use 20 draws; the 4,200 random and 200 layer-matched sets use five, deviating from the planned 20. Random-set threshold counts and combined-ranking diagnostics are post-hoc. The layer-matched null yields $p = 0.124$ for exceeding the reference score, versus $p = 0.045$ for the uniform null. A uniform-null advantage does not establish mechanistic completeness.

A.2 Dual-use implications and release scope

The immediate intended use is defensive: auditing causal explanations before relying on them in a safety investigation. The repository contains small-model experiments and public benchmark tasks. This report adds no intrusion technique, exploit, persistence mechanism, or novel installation recipe.

General intervention tools can also help identify or alter model components, and knowledge of evaluation blind spots could be used to optimize against an oversight test. This work does not demonstrate such evasion. A further practical risk is false assurance: treating a recovered circuit as proof of intent or a scrub score as a safety certificate. Applications to security-sensitive models should retain the qualifications above and evaluate behavior directly. Any future vulnerability or installation result would require a separate disclosure assessment before release.

Appendix B. Provenance and Reproducibility

Prior work. The inspected repository ends at commit c36289a, September 12, 2026. The phase-1 result commit is 88c2c75 (August 13); phase 7 is 2cf9b8a (August 17); phase 11 is 6a9667c (August 24); phase 12 is c396d3a (September 3). September 12 commits 1d4ced and c36289a update notes and presentation. No new model experiment during September 11-13 is evidenced by this snapshot. The submission's new work is synthesis, verification of recorded summaries, and discussion of incident-response relevance.

Verification boundary. Preparation included inventorying the full repository, checking code and experiment plans, parsing its stored JSON/CSV artifacts, and recalculating the central recovery counts, random-set pass counts, leave-one-out drops, AUCs, and rank correlation from committed measurements. These checks establish internal consistency. Model inference was not rerun, so they are not an independent replication of the original GPU experiments.

Recorded environment. The main runs report Python 3.12.10, PyTorch 2.11.0+cu128, TransformerLens 3.7.1, and an NVIDIA GeForce RTX 5060 Laptop GPU. requirements.txt provides the declared dependencies. Phase 12 discloses an uncommitted three-function datasets stub used to avoid a blocked pyarrow dependency. That workaround is absent from the repository and limits exact environment reproduction. The results are retained as recorded, without claiming a clean-checkout end-to-end rerun.

Plans and later analyses. Plans precede major phases, but use knowledge from earlier results; the whole sequence is not prospectively fixed. Post-hoc analyses retain that label. Two narrative overstatements are corrected: 25/200 random sets passing is 12.5%, not a majority; seven scheme rows have median rediscovery 10/10, but only four rediscover every initial head in all runs.

LLM Usage Statement

Codex was used to inspect the repository and official template, recalculate selected summary statistics from stored outputs, draft and edit this report, and generate and visually check the PDF. No new model-inference results were generated in preparing the report. This statement describes report preparation; it does not establish the extent of LLM use in the original project development.

Appendix C. Evidence Map

Table 3. Links below point to the inspected snapshot. The repository README lists commands for rerunning experiments in sequence. The stored files allow summary verification without rerunning model inference.

Claim / experiment	Measurement and implementation
IOI 20/26 across two schemes	Measurements ; Patching code
Docstring 3/6 to 5/6; Figure 1	Measurements ; Task generator
Scheme disagreement and calibration	Docstring ; Greater-than ; Calibration
Task induction and disclosed repair	Results ; Amendment
Ten-seed stability comparison	Stability ; Tests ; Post-hoc checks
IOI ablation, nulls, and Table 2	Measurements ; Tests ; Post-hoc checks ; Ablation code