FROM AGGREGATE SCORES TO CROSS-PROBE CLAIMS: AUDITING KNOWLEDGE-EDITING RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge-editing benchmarks rank editors by averaging scores over several kinds of test question, or probes. A higher average shows one editor is ahead on the benchmark's chosen mix, not on every probe: when editors trade wins across probes, some reweighting reverses their ranking. Over two released benchmarks and a preregistered factorial we ask how often that happens and what the disagreements track: **cross-probe disagreements are common, most fail at least one check tied to how the evaluation was built or scored, and the rest divide into real editor failures whose size the design cannot calibrate and a small residue no check here touches.** One pair shows what such an average is made of: HalluEditBench's Generalization Score puts ROME 20.3 points above LoRA on Mistral-7B, yet LoRA leads by 21.2 on the rephrase question, because after LoRA's edit the model usually no longer answers yes/no questions in the required form (13.2% format compliance against ROME's 93.4%, from the same unedited model). Across that benchmark 12 of 17 rejections disappear without its open-answer rephrase questions†; on KnowEdit, one cell scores probe types on disjoint items, and two paired reversals survive every check the release allows. In the factorial, three large reversals between likelihood-ranked choice and token-F1 generation disappear† under four other scores, including the teacher-forced accuracy KnowEdit applies. But which score shows a reversal is not stable across model and configuration: at EasyEdit's own FT-L configuration one survives two of the three text rescorings, and on the release's own backbone teacher-forced accuracy yields a corrected reversal where token F1 yields none. Splitting KnowEdit's pooled `Reasoning` label yields reversals real in sign whose magnitude a ceiling-against-chance comparison cannot calibrate. We release an audit reporting, per pair, the verdict, how far the ranking is from flipping, the largest reversal the data cannot exclude, and which confounds apply († marks a post hoc result).

1 INTRODUCTION

Knowledge-editing benchmarks score an edit with several probes, such as a restatement of the edited fact, a multiple-choice question about it and a question about a downstream consequence, and rank editors by an average over them (Yao et al., 2023; Zhang et al., 2024; Huang et al., 2025). Consider what one such average is made of. On Mistral-7B, one of HalluEditBench's three documented backbones, the released Generalization Score puts ROME **20.3 points** above LoRA (55.4 against 35.1, over 2,233 crossed cases), and a reader concludes that ROME's edits generalise better. The score averages five question types, and they disagree: ROME leads by 63.9, 31.1 and 21.9 points on the yes, no and multiple-choice questions and **trails by 21.2 on the rephrase question**, the paraphrase probe that corresponds to what this literature means by generalization. The release's own outputs say why. Before editing, the two arms are the same model and score the same on every type; after editing, ROME's model still begins its answer with "yes" or "no" on **93.4%** of those questions and LoRA's does so on **13.2%**, so LoRA scores 14.6 and 12.1 where guessing would score about 50. On the rephrase question LoRA leads whether it is scored as the release scores it (89.2 against 68.0) or by first-span match on the stored outputs, which credits an answer given first and then repeated (73.2 against 58.4). In this benchmark's own data, "ROME generalises better than LoRA" is "ROME still answers in the required form after editing and LoRA does not, while LoRA states the edited fact more often". Those are different claims, and only the first is visible in the score.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The two readings differ in strength, and a benchmark's tables support only the weaker one. For weights w over probe classes, the composite $S_w(A) > S_w(B)$ is a legitimate result: A performs better on the benchmark's chosen mixture. The stronger reading, that A's advantage holds whatever weight each probe class receives, is Pareto dominance over probe classes, and it is the reading method papers invite when they describe a winner in general terms, "state-of-the-art performance" or "outperform previous model editing methods", while their own tables show a baseline ahead on a component (Meng et al., 2023; Li et al., 2024a; Wang et al., 2024a). PMET's own GPT-J CounterFact table, for instance, gives it a 0.4-point composite lead over MEMIT built from gains on efficacy and generalization against a loss on specificity, each gap wider than its printed intervals, while on its other dataset the same pair leads on all three probes: the claim holds across probes on one dataset and not on the other, and the composite does not say which (I.3). Work on benchmark aggregation has shown that rankings depend on how tasks are weighted (Ethayarajh & Jurafsky, 2020; Colombo et al., 2022; Zhang & Hardt, 2024), and two recent papers match ranking claims to cross-task and cross-stratum evidence (Oh, 2026; Hu et al., 2026). Whether per-probe effects agree in direction is a question of qualitative interaction (Gail & Simon, 1985), and the gap between a score and an interpretation placed on it is a concern of validity theory, subgroup analysis and differential item functioning (Kane, 2013; Rothwell, 2005; Holland & Thayer, 1988).

A significant disagreement between probes is, however, not automatically a statement about editors. In editing, the direction of several disagreements is known: unconstrained fine-tuning generalises to paraphrases while failing on neighbouring subjects (Meng et al., 2022), edited models over-weight the edit target on multi-hop and specificity prompts (Zhang et al., 2025), specificity depends on how neighbouring prompts are built (Hoelscher-Obermaier et al., 2023), and scoring protocols reorder editors (Pohl et al., 2025; Yang et al., 2025). A disagreement can also arise because a label pools unlike items, because an arm is at ceiling or chance, because two probe types were scored on different items, or because a response format credits a non-specific rise in the target's likelihood. These are standard concerns (subgroup cancellation, ceiling compression, confounding and surface-form competition (Holtzman et al., 2021)); what is missing is their systematic application to editing rankings.

We apply them to two released benchmarks and to a controlled factorial on KnowEdit's questions (GPT-J-6B, single edits) in which only the probe changes. **Contributions.** (i) A re-analysis of HalluEditBench and KnowEdit locating which probe disagreements survive checks of evaluation construction and scoring and which do not: the pair above, a pooled label, and two paired reversals that survive every check the release allows (§3). (ii) A controlled factorial showing that the format reversals in these data depend on how generation is scored and on whether a target-as-gold choice is calibrated; that the first is configuration-dependent and the second backbone-dependent, since at EasyEdit's own FT-L configuration a reversal survives rescoring while its choice advantage vanishes under calibration on GPT-J and widens on Llama-2 (§5.1, M, O); and that the label-split reversals are real in sign and uncalibrated in magnitude. (iii) An audit that separates a verdict on the benchmark's own scores from attribution flags, reports the weight at which a ranking flips, and states when a cross-probe claim is established rather than merely unrefuted (§6). We show in B.7 that one of those flags does not discriminate, and we do not claim the test, Proposition 1, the leakage measure or the checks themselves as new.

2 Composite and cross-probe claims

2.1 Two claims

Let $s_p(m)$ be method m's mean score on probe class p and $\delta_p = s_p(A) - s_p(B)$. The **composite claim** is $S_w(A) > S_w(B)$ for the benchmark's weights w. The **cross-probe claim** is that A's advantage holds for every positive weighting.

Proposition 1. For $\delta \neq 0$, the sign of $\sum_p w_p \delta_p$ is the same for every weighting with all $w_p > 0$ if and only if the nonzero components of δ share a sign; otherwise some weighting reverses the ordering (proof in Appendix B.1).

This is the Pareto-dominance fact for positive linear scalarisation. We use it only to fix the estimand: a reversal refutes the cross-probe claim on the benchmark's own scores whatever its cause. For each pair the audit also reports the weight at which the composite flips (at two strata, $w^* = |\delta_2|/(|\delta_1| + |\delta_2|)$

on stratum 1), which measures how far an observed ranking is from failing under reweighting. We do not assume that a paper reporting an aggregate literally asserts Pareto dominance; agreement in direction is the strict robustness criterion, and this paper audits that criterion rather than arguing against composites.

2.2 THE TEST AND ITS REPORTING

The null is no qualitative interaction: the δ_p have a common sign, with each pair oriented so that A is the aggregate winner. The registered test is Gail and Simon's (Gail & Simon, 1985), which is exact at its level under that null; where strata share items and the test can exceed its level, we use the intersection-union version (Berger, 1982; Piantadosi & Gail, 1993), whose p is written p_{IU} and which is valid under any dependence, and we report both counts everywhere. Appendix B.2 gives the rule for which is primary, with its worst-case sizes. A **corrected reversal** rejects after Bonferroni correction over its registered family, which is never shrunk to what a data-dependent filter keeps; the one registered exception is the chance-level grid, where the family follows the grid setting by design (B.3 lists every family). Non-reversals are **magnitude heterogeneous** (a cluster-robust Wald test rejects) or show **no detected heterogeneity**, and carry a **minority-direction bound**: the largest opposite-direction effect, in points, that the data still allow.

Terms used in a fixed sense. A **probe class** is a kind of test question (restatement, consequence); a **format** is how it is answered (constrained choice, free generation); a **stratum** is one probe class or format within a comparison; an **arm** is one editor's mean in one (probe class $\times$ format) **cell**; a **family** is the set of comparisons one Bonferroni correction covers. **Dependence-robust** means the intersection-union count. Appendix Table 3 carries the full list.

2.3 ATTRIBUTION CHECKS

The formal verdict is issued on all prespecified strata. Four checks then ask whether a disagreement can be read as a property of the probe type; each is reported as a flag, not a test. **M1, label pooling:** a probe class can pool sub-types, so the audit reports whether the verdict changes when a label is split on an item property computable without a model. **M2, ceiling or chance:** near a score's ceiling or a format's chance level differences are compressed, and a common cap can reverse the order of two editors whose per-item scores do not stochastically dominate each other. Strata with an arm within 0.05 of either are flagged, and the test without them is reported as a diagnostic. **M2 bears on magnitude, not on sign:** an arm at chance may be a compressed measurement or an editor that genuinely fails the probe, so the flag licenses no conclusion that a reversal is spurious. It is also the least selective of the four, firing with nominal chance levels on all 84 HalluEditBench pairs, on 18 of 20 in the factorial's true-consequence section and on 258 of 330 HELM verdicts; B.7 reports its rate among non-failures beside its rate among failures and finds that it carries no information about which pairs reverse. **M3, item sets:** probe types scored on systematically different items confound probe with items. **M4, non-specific credit:** a likelihood-ranked choice whose gold option is the edit target credits any rise in the target's likelihood. The audit reports both editors' **leakage**, the fraction of off-target prompts on which the target outranks the prompt's own gold answer under the contrast's scoring rule, for a different-subject and a same-subject prompt family. Leakage is a variant of CounterFact's Neighborhood Score (Meng et al., 2022) and of EVOKE's overfit score (Zhang et al., 2025), with a pre-edit floor and the contrast's own scoring rule.

3 RELEASED BENCHMARKS

§1's pair is one comparison; across both released benchmarks such disagreements are common, and most fail a check tied to how the evaluation was built.

HalluEditBench: open-answer questions. Its Generalization Score averages five question types (Huang et al., 2025). Over all $\binom{7}{2} \times 4 = 84$ method pairs in the released results (the three backbones of its paper and GPT-J-6B from its additional experiments), directional agreement is rejected for **17** at an uncorrected .05, and 16 both at a per-backbone Bonferroni threshold and over the whole 84 (B.3), with a median reversing effect of 15.3 points. The pair of §1 is one of the 17 ($p_{\mathrm{IU}} = 1.0 \times 10^{-77}$ against $0.05/21$), and the other 16 have its shape. Thirteen of the 20 minority-direction question types inside those rejections are the rephrase question, and without it 5 rejections remain. What

separates the question types is the answer space, open for rephrase and reversed and closed for the other three, and not the scoring rule, since the public code generates and judges all five the same way (G.1). Appendix G.2 gives that pair's full arithmetic, including the pre-edit baseline that makes the divergence edit-induced and the rescoring that leaves its direction unchanged.

KnowEdit: disjoint items, a pooled label and two reversals that remain. KnowEdit (Zhang et al., 2024) scores its three portability probe types with one token-match rule, so a reversal there is not a change of format. Across its two usable cells directional agreement is rejected in **12 of 42** comparisons at an uncorrected .05, 9 corrected per cell and 8 over the whole 42 (B.3). On `zsre`, which supplies 8 of the 12, the probe types are scored on disjoint item sets (263 and 771 cases), so no reversal there can be attributed to the probe. On `recent`, where 108 cases carry all three types, **two** corrected reversals remain: MEMIT and ROME each lead FT-L on aliasing, by 18.6 and 16.5 points, and trail it on reasoning, by 13.6 and 14.2. Neither fails a check available here. Its `Reasoning` label pools prompts whose gold answer is the edit target itself with genuine consequences, and 54 of the 108 cases carry such a prompt, but restricting to true consequences leaves both reversals ($p = 7.5 \times 10^{-7}$ and 9.8×10^{-6}), as does dropping the 14.1% of true-consequence golds that share a word with the edit target and so could earn partial credit from target boosting. The release scores no off-target prompts, so leakage cannot be measured there at all. The design is also underpowered: aliasing has a minimum detectable effect of 14.5 points against 6.2 on reasoning, so joint power against an 8.2-point reversal is about 0.17 and a null would say little (B.5). Our factorial does not explain these two either: there the gradient editors lead on restatement and fail the consequence probe, the mirror of this pattern on the same probe types, so the editors named FT-L in the release do not behave like our fine-tuning arms. Three axes could produce that mirror: the backbone, the implementations and the metric, since KnowEdit scores by teacher-forced token accuracy where we score a free continuation. Experiment 7 tests the third and rules it out on these facts: under the benchmark's own metric our gradient arms lead both reference editors on **both** probes, so not one of the six pairs shows the release's direction (N.2). Experiment 8 repeats it on the release's own backbone at the published configurations and does not recover the release's direction there either, so among these three axes the implementations remain (O). Appendix G.4 gives every count.

4 A CONTROLLED FACTORIAL

4.1 DESIGN, IMPLEMENTATION AND REGISTRATION

Released data cannot say what causes a disagreement, since nothing in them is held fixed. We varied one thing at a time. We use KnowEdit's own questions for the 573 facts that carry both a `Subject_Aliasing` probe (a *restatement*, whose answer is the edit target) and a `Reasoning` probe (a *consequence*), each rendered in two paraphrases and three formats, and analyse the 533 facts every editor completed on GPT-J-6B, the 573 less 40 calibration and holdout facts removed by design (Experiment 3; two earlier experiments on a screened subset are in Appendix C). **Constrained choice** ranks four options by length-normalised log-likelihood, distractors drawn from other facts' answers for the same relation. **Free generation** decodes greedily for 12 tokens, cuts at the first newline, period, comma or semicolon, and scores token F1 against accepted aliases; this is not KnowEdit's teacher-forced metric, which Experiment 7 adds as a fifth score (N). A third format, yes/no, stays at chance for every weight editor and enters no contrast (A.2). Each editor edits one fact at a time, and every probe is scored after the same edit. The **measurement contrast** changes the format with the question fixed; the **semantic contrast** changes the question with the format fixed. The fact is the unit, and paraphrases are averaged within a fact before any test.

The gradient editors FT-L, FT-M and LoRA are our implementations in bf16 with FP32 masters and no early stopping; FT-M matches EasyEdit's published GPT-J values, FT-L departs in its objective and a clamp with no recorded reason, and LoRA runs at a fiftieth of AdaLoRA's published rate, which diverged here. ROME and MEMIT are EasyEdit's files unchanged (Wang et al., 2024b), ROME in its identity-covariance variant. Statements about "gradient editors" concern these implementations; A.4 lists every departure. Experiments 6 and 7 test four of them. Two matter: EasyEdit's own FT-L repeats twice as often as ours, and restoring EasyEdit's stopping rule to FT-M cuts its repetition from 0.588 to 0.197 and its neighbourhood leakage from 0.976 to 0.651 while improving the edit, so **our gradient arms are over-trained** and §5's repetition figures describe arms run past the published

stopping point. Two do not: ROME under its own covariance behaves as EasyEdit's variant does, and AdaLoRA at the published rate will not run here at all (M, N).

Experiment 3's population, families, restriction and decision rule were registered before its outputs, as was the answer-identity split of §5.2. The registrations constrain analysis choices but are author-controlled and are not evidence of validity: Appendix H.1 gives the status of every result, H.2 every deviation, and H.4 what author-controlled timestamps cannot show. **A dagger (†) marks a number that is post hoc by its table's accounting**, and prose in the main text may state a post hoc number without carrying the symbol.

Table 1: Arm means on the 533 facts (GPT-J-6B), under constrained choice (choice) and free generation (gen.) scored by token F1 or, post hoc†, by first-span match. A continuation counts as repeated when one token makes up at least half of it (four or more tokens). Unedited levels are from the replication run on the same facts.

cell	FT-L	FT-M	LoRA	MEMIT	ROME	unedited	chance
restatement, choice	.993	.998	.995	.713	.838	.366	.25
restatement, gen. (F1)	.425	.419	.435	.367	.589	.085	0
restatement, gen. (first span)	.608	.744	.757	.300	.529	.025	0
consequence, choice	.491	.506	.528	.389	.446	.251	.25
consequence, gen. (F1)	.146	.142	.157	.159	.188	.126	0
repeated continuations	.30	.59	.55	.00	.00	.00	

4.2 The registered result

What was registered, and how it came out. Experiment 3 was registered to test an asymmetry: that changing the *response format* reverses editors and changing the *question type* does not. The rule returns `CONFIRMED` on at least two measurement reversals across at least two editor pairs, no semantic reversal, and survival of three robustness conditions; `PARTIAL` when a reversal exists, no semantic one does, and one of those conditions removes a reversal (D.4). The registered test finds **3** corrected reversals of 20 on the measurement contrast and **0** of 20 on the semantic contrast (family 40), and the rule returns **partial support**, because two of the three conditions remove a reversal. Two registered results then went against the hypothesis. The registered specificity test returned `SPECIFICITY`, whose pre-written sentence says that the constrained-choice advantage is confounded with a broad boost and "this paper's framing of the measurement axis does not hold for it" (F.2). And the registered answer-identity split (§5.2) turns the semantic null into eight reversals, so the asymmetry does not hold on the other side either. **The audit framing of this paper was adopted after those results, not before them** (H.4); what survives from the registration is the contrast, the families and the counts, not the hypothesis they were meant to test. All three are FT-L, FT-M and LoRA against ROME on restatement: constrained-choice advantages of $+15.5$ to $+16.0$ points against generation deficits of -15.4 to -17.0 (p from 7.9×10^{-14} to 2.9×10^{-11}; the intersection-union test also rejects all three). They recur on the 278 facts outside Experiment 1's screen, under the registered restriction, and at each of three seeds, so they are real on the registered scale (Table 2 gives a fourth that appears over all strata). Under equal weights they are near-ties: the composite flips at constrained-choice weights of 0.515, 0.514 and 0.495, with bootstrap intervals of roughly $[0.39, 0.60]$.

5 What the reversals are made of

Three things can be varied here: how an answer is scored, whether a label pools unlike items, and which items each probe was scored on. We take them in turn.

5.1 The format reversals depend on scoring

Generation. The gradient editors' continuations repeat: on restatement, 55–59% of FT-M's and LoRA's are dominated by one token ("Canada Canada Canada"), against none of ROME's or MEMIT's and none of the unedited model's (Table 1). Token F1 penalises the repeats even when

the first span is the correct answer. Scored by first-span match, the gradient editors lead ROME on restatement in generation as well (FT-M 0.744 against 0.529), and **none of the 20 measurement comparisons reverses**, on the 533 facts or on the held-out 278; first-span F1 and alias containment give the same result (Table 2). These rescorings are post hoc†, and none is a neutral referee: all three share one idea, that tokens after the answer should not count. Together they establish that the generation ranking of these arms is not identified by the scoring rule: token F1 places the gradient editors below ROME and all three alternatives place them above it (E.1). A fifth score settles that this is not an artefact of scores we chose. Experiment 7 re-runs the same arms on the same facts and scores them by **KnowEdit's own teacher-forced token accuracy**, the rule the audited benchmark applies, and under it the gradient editors lead ROME by 11.8 to 20.5 points on restatement rather than trailing it, so **none of the twelve comparisons reverses** (N.2). The qualification matters: run on the same facts, EasyEdit's own FT-L configuration repeats on 63.6% of restatement continuations, and its reversal against ROME there survives first-span match and containment, failing only first-span F1 (M). Whether a measurement reversal is a scoring artefact is itself configuration-dependent. The calibration result is not: that arm's raw choice advantage of $+16.3$ points over ROME becomes -15.7 under calibration, a larger fall than any of our own arms shows, so the choice-side account holds at the published configuration while the generation-side one does not (Appendix M). **On the release's own backbone the score dependence runs the other way.** Experiment 8 puts the same published configurations on Llama-2-7b-chat. One measurement reversal is corrected there, EasyEdit's FT-L against ROME on restatement, and it appears under KnowEdit's teacher-forced accuracy alone ($+12.2$ against -9.7 points, $p = 5.6 \times 10^{-17}$) where trimmed token F1 shows none. Neither backbone licenses the reading that one of these scores is the sound one: whether a reversal appears depends on the scoring rule and on the model and configuration together. Repetition does not account for it (3.6% of continuations against 63.6% on GPT-J), M4 fires on one leakage family of two, and calibration widens the choice advantage rather than removing it; the arm is weak on the edit prompt (0.345 against ROME's 0.906) though both arms sit above the unedited model in both strata (O).

Table 2: Corrected reversals on the measurement (meas.) and semantic (sem.) contrasts and the answer-identity split, under the registered Gail and Simon test / the dependence-robust intersection-union test, at the registered family in parentheses, the order every appendix table uses. The tests agree on the measurement counts and differ only on the split, where the intersection-union count is the conservative one. The three text rescorings give identical counts and are one row; the teacher-forced row is Experiment 7's family of twelve, covering the measurement contrast only, and comes from a re-run of the same arms rather than a rescoring of these blocks, because that score needs a forward pass through the edited model (N.1). The held-out row appears twice because the registered restriction gives three reversals and all strata give a fourth, LoRA against MEMIT in `Reasoning`, which also vanishes under first-span scoring (E.1).

scoring of the Experiment 3 blocks	meas. (40)	sem. (40)	split (80)
registered: token F1 and accuracy	**3** / 3	**0** / 0	**8** / 6
generation by first-span match, first-span F1 or alias containment (post hoc)	0 / 0	0 / 0	12 / 10
generation by KnowEdit's teacher-forced accuracy (Experiment 7)	0 / 0	n/a	n/a
constrained choice by margin (post hoc)	3 / 3	0 / 0	10 / 10
registered scoring, 278 held-out facts, registered restriction	3 / 3	0 / 0	4 / 3
registered scoring, 278 held-out facts, all strata	4 / 4	0 / 0	4 / 3

Constrained choice. The multiple-choice side survives the ceiling check: on the untruncated margin the gradient editors' per-fact margins exceed ROME's and MEMIT's at every decile, every bootstrap interval excluding zero. Its gold option is the edit target, however, and the editors that gain most are those that raise the target everywhere: they put it above a different subject's own gold answer on 90.2%, 97.7% and 99.9% of such prompts, against 6.4% for ROME. Rescored against the null prompt "Answer:" (Holtzman et al., 2021; Zhao et al., 2021), which lifts the locate-and-edit arms and lowers the gradient ones, the three advantages become -3.9, -3.4 and -34.2 points, two with intervals including zero (Figure 1b; J.4). Calibration is not a knowledge scale and we report it as a re-expression of leakage rather than as independent evidence: the *unedited* model rises from 0.251

to 0.584 on the consequence probe, an option-prior effect, since there the gold option is 1.21 nats less likely than the average distractor under the null stem alone. It also overshoots for LoRA, and the leakage call itself flips for 7 of 11 reversals with the prompt family (J.4, F.4).

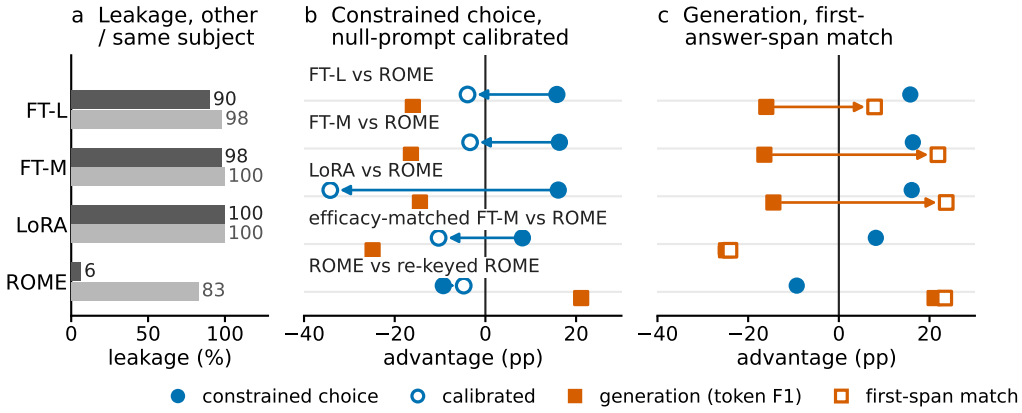

Figure 1: Dissecting the controlled measurement reversals (GPT-J-6B, 533 facts, restatement; Experiment 5). **(a)** Leakage on different-subject (dark) and same-subject (light) prompts. **(b)** One row per comparison: the first editor's advantage under constrained choice (filled circle) and under generation scored by token F1 (square); a reversal puts them on opposite sides of zero. Arrows rescore constrained choice against a null prompt: the four gradient-versus-ROME reversals disappear; the within-ROME pair does not. **(c)** The same rows with generation rescored by first-span match (open squares): the three primary reversals disappear, while the efficacy-matched FT-M arm and the within-ROME pair still reverse. The calibrated scores of the primary and within-ROME pairs were registered; everything else in (b) and (c) is post hoc.

What survives. Two reversals survive first-span scoring (Figure 1c): an FT-M arm matched to ROME's edit-prompt efficacy, with neither arm near ceiling or chance, which vanishes under calibration; and ROME against ROME re-keyed on the prompt's last token, which survives both rescorings. The experiment registered to vary leakage within an editor family could not form its design and returned **non-diagnostic**, the registered verdict when no comparison can decide (J).

5.2 Pooling a label hides a reversal

For 181 of the 573 facts the gold answer of KnowEdit's `Reasoning` probe is the edit target itself, reached through a relational description of the subject. Splitting the label by this property, computable from the item alone, turns the pooled semantic null into **6** corrected reversals over the family of 80 under the intersection-union test, and 8 under the registered Gail and Simon test. On true-consequence facts, the gradient editors lead the locate-and-edit editors on restatement and trail them on the consequence probe. On same-answer facts, in the same constrained-choice cells, they lead on both, because there the consequence probe's answer is the target. Pooling the target-as-gold items therefore hides the true-consequence reversals. A test that the within-fact interaction differs between the two subsets rejects for every one of them ($|z| \geq 5.9$). Under first-span scoring the same pattern appears in generation (Table 2).

Six of the eight reversals the registered test certifies set an arm at ceiling on restatement against an arm at chance on the consequence probe; a seventh has an arm at the floor on it, and the eighth, same-answer LoRA against ROME at generation, does not replicate on the held-out facts. The two the intersection-union test does not certify are different ones: the true-consequence LoRA reversals against ROME and MEMIT under constrained choice ($p_{\mathrm{IU}} = 6.6 \times 10^{-4}$ and 1.0×10^{-3} against $0.05/80$), both of which are ceiling-against-chance. The dependence-robust count therefore drops two of the ceiling-against-chance reversals and keeps the one that fails out of sample. The split is also the count most sensitive to the run-level component that fact-clustered errors cannot see: at $\tau = 0$ it is 8, at 1 point 5, at 2 points 4 and at 3 points 1, where the measurement count holds at 3 until τ passes 3 points. The only estimate of τ is 0.79 points, with interval $[0.16, 2.74]$ (C.4), so the split count is not robust across that interval and the measurement count is. The same asymmetry appears under a second check we had not run: facts share relations, and the constrained-choice distractors are

drawn from other facts' answers for the same relation, so clustering the standard errors on the relation (92 groups) rather than the fact is the wider and more conservative choice. All three measurement reversals survive it ($p \leq 1.2 \times 10^{-4}$ against $0.05/40$); **of the eight split reversals only two do**, both FT-M against a locate-and-edit editor on true-consequence facts (E.6). The split therefore establishes that pooling the label hides a disagreement, and does not establish how many. On true-consequence facts the gradient editors score 0.996–0.997 on restatement under constrained choice and 0.249–0.293 on the consequence probe, against a chance level of 0.25 and an unedited 0.230. The direction is what EVOKE calls editing overfit (Zhang et al., 2025). It is also the mirror of KnowEdit's own two surviving reversals (§3), so these arms do not stand in for the release's. Because they set an arm at ceiling against an arm at chance or floor, they provide no scale on which how far an edit transfers can be compared.

5.3 Item sets can induce reversals

To isolate item-set composition, we rescored the paired data as though each probe type had been evaluated on a different half of the facts (post hoc). Random halves certify 1 reversal in 4,000 pair-splits. Splitting at the median of pre-edit answerability and giving restatement the harder half certifies 394 (266 under the intersection-union test), 375 of them FT-M against MEMIT; the 400 replicates differ only in how 170 tied facts are assigned, and the opposite orientation certifies none. This is an existence proof that systematically different item sets can induce reversals, not an estimate of how often they do.

6 The audit

Nothing above is specific to these two benchmarks. The audit packages it for any ranking with paired per-item outcomes.

What it reports. For each method pair on long-format per-item outcomes, the audit issues the verdict of §2.2 on all prespecified strata. It reports the flip weight (or, for more strata, the share of the weight simplex the winner still wins and the smallest reweighting that flips it) and, for non-failures, the minority-direction bound. It attaches the attribution flags of §2.3 and reports heterogeneity on the log-odds scale as well as additively. Beside the verdict it states whether the **cross-probe claim** is refuted, established (every stratum excludes a reversal at the simultaneous level) or neither, which distinguishes a well-powered agreement from an underpowered one: on KnowEdit's 42 comparisons that is 9, 20 and 13. On the factorial it is more permissive than the registered analysis, because its default families are the section's own (20 comparisons) rather than the registered union (40 and 80): it certifies four measurement and ten split failures where §4.2 and §5.2 certify three and eight, and it uses t tails where the registered analyses use normal ones (Appendix B.6). Every corrected failure outside KnowEdit `recent` carries a flag, which is why B.7 gives the base rate beside it: at the chance levels §3 uses, M2 fires on **every** HalluEditBench pair, failing or not, and on KnowEdit on fewer failures than non-failures. M4 is the one flag that separates them in the factorial (3 of 3 failures against 43% of non-failures). On planted data M2 fires equally on a genuine crossing and on a cap artefact; the untruncated margin separates them.

Two prospective applications, both uninformative. Experiment 4 applied the registered audit to Qwen2.5-7B-Instruct and returned **non-diagnostic**; on HELM's MMLU runs (Liang et al., 2023; Hendrycks et al., 2021) it finds no qualitative failure among 45 model pairs, with a median minority-direction bound of 16.6 points, which describes the design's resolution rather than MMLU. Neither is evidence for the method's generality; both are reported because they were registered or run (K, L).

Recommendations. For rankings meant to support cross-probe claims: (1) report each pairwise ranking with its flip weight, on a stated scale, and its minority-direction bound, and say whether the cross-probe claim is established or only unrefuted; (2) release paired per-item outputs, and document how probe-specific item sets were drawn; (3) do not interpret a likelihood-ranked choice whose gold option is the edit target without a calibrated score beside it, and report leakage under the same rule for both off-target families; (4) where free generation is scored by overlap with the answer, as token F1 does, report continuation length and repetition rate beside the score, since repetition after a correct answer is otherwise read as failure to recall (KnowEdit's teacher-forced accuracy and

HalluEditBench's substring rule are not affected); (5) release item properties that split a label, such as whether a consequence probe's answer is the target, declared before scoring.

7 Related work

Editors (Meng et al., 2022; 2023; Mitchell et al., 2022; Hartvigsen et al., 2023; Zheng et al., 2023; Hu et al., 2022) are compared on benchmarks that decompose edit quality into reliability, generality, locality and portability (Yao et al., 2023; Zhang et al., 2024) and that widen the probe set (Zhong et al., 2023; Cohen et al., 2024; Huang et al., 2025). EVOKE measures whether edited models over-weight the target on multi-hop and specificity prompts (Zhang et al., 2025), and CounterFact+ shows that measured specificity depends on how neighbouring prompts are built (Hoelscher-Obermaier et al., 2023). Scoring protocols reorder editors (Pohl et al., 2025; Yang et al., 2025), and our generation result is another instance; editing also degrades abilities the edit was not aimed at (Gu et al., 2024), the class §1's collapse of answer form belongs to. Beyond editing, prompt and format sensitivity reorders models (Sclar et al., 2024; Mizrahi et al., 2024; Alzahrani et al., 2024; Zheng et al., 2024), likelihood-ranked choice disagrees with the answer a model generates (Wang et al., 2024c), aggregation choices change leaderboards (Ethayarajh & Jurafsky, 2020; Colombo et al., 2022; Zhang & Hardt, 2024), Oh (2026) and Hu et al. (2026) both match ranking claims to cross-task and cross-stratum evidence, and Zheng & Yang (2026) measure sign reversals of near-tied rankings. Three results here are recoveries of known ones. The answer-identity split undoes a merge rather than finding a distinction: RippleEdits scores compositionality with the new object separately from compositionality reaching the subject through another entity, and KnowEdit merged the two into `Reasoning` (Cohen et al., 2024; Zhang et al., 2024). Repetition under editing is visible in ROME's own GPT-J table as a fall in generation entropy for unconstrained fine-tuning, and KnowEdit reports fluency for every editor (Meng et al., 2022; Zhang et al., 2024), while Yang et al. (2024a) and Yang et al. (2024b) document post-edit collapse directly, with degenerate and often repetitive outputs; what is new is narrower, that a trimmed overlap score turns it into an apparent recall failure, the scoring problem Kamalloo et al. (2023) describe for open-domain answers. Null-prompt calibration is the content-free calibration of Zhao et al. (2021) applied to a choice whose gold option is the edit target. What we add is their consequence for editing rankings, measured rather than argued: which reversals survive which check, and which survive all of them. Appendix I compares the closest works.

8 Conclusion

A higher aggregate editing score establishes a ranking on the benchmark's chosen mixture, and a significant disagreement between probes locates a problem without saying whether it belongs to the editors or to the evaluation. On two released benchmarks and a controlled factorial, most disagreements do not survive checks of item sets, question types, generation scoring, or calibration of choices whose gold option is the edit target. **Arms at ceiling and chance belong in a different category from the rest.** An editor that restates an edited fact almost perfectly and answers its consequences at chance has failed to propagate the edit: the sign is real, and our own margin check keeps every such reversal. What a truncated scale cannot support is the *size* of the gap. The audit reports which check applies; none is established as a cause, and a flag is a reason to look, not a verdict of artefact.

Scope and what was not run. The controlled evidence uses KnowEdit's questions, five editors, three of them our implementations, and single edits on GPT-J-6B; Qwen2.5-7B-Instruct yielded no detected reversal, in a non-diagnostic and underpowered run. Experiments 6 and 7 restored four published choices on the same facts and rescored every arm by KnowEdit's own metric (M, N); AdaLoRA at the published rate diverged on all twenty calibration facts on both backbones and produced no block, so no result here describes EasyEdit's published LoRA editor. The factorial still does not explain the release's two surviving reversals: its editors behave opposite to the release's on the same probe types (§3), and neither the metric nor the backbone recovers the release's direction, so among the three axes tested the implementations remain (N, O). MEMIT does not run on Llama-2 for want of published second moments, so one of those two reversals is untouched by Experiment 8. Not run: MEMIT and AlphaEdit on a second base model; sequential editing; replicated templates per format; an efficacy-matched leakage series within one editor. No CounterFact-style neighbourhood score is computable here: no pre-edit object exists in the items or in KnowEdit's records (A.1).

REPRODUCIBILITY STATEMENT

The design is in §4 and Appendices A, C and D, the checks in Appendices E and F, the released-benchmark re-analysis in §3 and Appendix G, Experiments 5 and 4 in Appendices J and K, and Experiments 6, 7 and 8 in Appendices M, N and O. The supplementary archive ships the code, run blocks, registrations, the audit tool with worked examples and the built documents. Every number the documents state is produced by a self-tested instrument that writes an artifact, and `doclint.py` reads a fixed set of those numbers back out of the artifacts, including every cell of the comparison tables in J, M and N; it does not verify prose. The post hoc analyses added after the registered ones, including the generation rescoring, margins, calibration levels and the KnowEdit split, are computed by `post_hoc_analyses.py` from the stored continuations and log-likelihoods. The observational inputs are rebuilt from the public releases at pinned commits. Two inputs are **not** shipped and the affected results are therefore not recomputable from the archive alone: the locality item file, and the published second-moment statistics MEMIT needs, which are redistributions rather than ours to release. Experiments 5 and 7 also ship their derived artifacts rather than their run blocks, for the anonymity reason `WITHHELD_ARTIFACTS` records; Experiment 8's blocks ship whole.

ETHICS STATEMENT

This work re-analyses publicly released evaluation outputs and edits public models with public facts; it involves no human subjects and no private data. We audit HalluEditBench and KnowEdit because their authors released per-probe data, which most benchmarks do not, and we regard those releases as exemplary rather than as targets.

AI USE STATEMENT

Large language models (Anthropic Claude, via Claude Code) were used extensively, and the disclosure below follows the four categories of the ICLR 2027 policy.

Required-disclosure tasks that were used. *Implementing methods:* the analysis and experiment instruments, their self-tests and the figure scripts were written by the model and accepted or rejected by the authors. *Designing and critiquing methodology and experiments:* models were given the built PDF and no other context and asked to find errors and demand analyses. Those critique passes changed this paper's conclusions repeatedly: they overturned the registered generation result, forced the scope statements of §5.1, and prompted B.7, the relation-clustered re-analysis of E.6, and Experiments 6, 7 and 8. *Interpreting results:* the model drafted the readings of every experiment, including the ones that went against the registered hypothesis.

Required-disclosure tasks that were not used. No synthetic data was generated; every number comes from a public release or from a run of a shipped script. No dataset was cleaned or filtered by a model. No translation was involved. No new mathematical claim is made: Proposition 1 is the standard Pareto-dominance fact for positive linear scalarisation, stated to fix the estimand, and its proof in B.1 was checked against that literature rather than derived here.

Recommended-disclosure tasks that were used. Drafting and revising all prose; writing and refactoring code and figures; literature search; and reference formatting, with every citation verified against a primary record (`paper/refs_verification.md`).

Verification and responsibility. Experiments were executed by fixed scripts, not by a model, and every number in these documents resolves to an artifact written by a self-tested instrument; `doclint.py` reads a fixed set of them back out of those artifacts. That mechanism checks numbers and not prose, and a check of this kind corrected a sentence in which two different correct measurements had been quoted as one (G.2). Human readers, separately from any of this, read drafts and commented on them; those comments are not the critique passes described above, are not recorded in the archive, and no claim here rests on them. The authors designed the study, specified the preregistrations, accepted or rejected each proposed change, verified every citation and reported number, and take responsibility for the final content.

REFERENCES

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 13787–13805, 2024. doi: 10.18653/v1/2024.acl-long.744.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982. doi: 10.2307/1267823.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, Xifeng Yan, William Yang Wang, Philip Torr, Dawn Song, and Kai Shu. Can editing LLMs inject harm? In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026)*, volume 40, pp. 30192–30200, 2026. doi: 10.1609/aaai.v40i36.40269.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, 1973. doi: 10.1016/S0022-5371(73)80014-3.

William G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129, 1954. doi: 10.2307/3001666.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. What are the best systems? New perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 26915–26932, 2022. doi: 10.52202/068431-1951.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, 2020. doi: 10.18653/v1/2020.emnlp-main.393.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2410.02355.

Mitchell Gail and Richard Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372, 1985. doi: 10.2307/2530862.

Sankalp Gilda and Shlok Gilda. Position: Evaluation scores are perishable knowledge claims. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 1029–1035, 2026. doi: 10.18653/v1/2026.gem-main.80.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16801–16819, 2024. doi: 10.18653/v1/2024.emnlp-main.934.

Yupu Gu, Rongzhe Wei, Andy Zhu, and Pan Li. MoEEdit: Efficient and routing-stable knowledge editing for mixture-of-experts LLMs. arXiv:2602.10965, 2026. Preprint, submitted 11 February 2026.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15202–15232, 2024. doi: 10.18653/v1/2024.findings-acl.902.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL https://arxiv.org/abs/2211.11031.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11548–11559, 2023. doi: 10.18653/v1/2023.findings-acl.733.

Paul W. Holland and Dorothy T. Thayer. Differential item performance and the Mantel–Haenszel procedure. In Howard Wainer and Henry I. Braun (eds.), *Test Validity*, pp. 129–145. Lawrence Erlbaum, Hillsdale, NJ, 1988.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, 2021. doi: 10.18653/v1/2021.emnlp-main.564.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022. URL https://arxiv.org/abs/2106.09685.

Lining Hu, Ting Liu, and Yuzhuo Fu. Pooled leaderboards hide system-specific winners: A reporting-protocol audit of offline root-cause analysis benchmarks. *arXiv preprint arXiv:2606.29159*, 2026. URL https://arxiv.org/abs/2606.29159.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. Can knowledge editing really correct hallucinations? In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2410.16251.

Haoben Huang, Shuxin Liu, Ou Wu, and Di Gao. Revisiting ripple effects in knowledge editing through pressure-aware joint neighborhood optimization. arXiv:2606.01610, 2026. Preprint, submitted 1 June 2026.

Charles M. Judd, Jacob Westfall, and David A. Kenny. Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1):54–69, 2012. doi: 10.1037/a0028347.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5591–5606, 2023. doi: 10.18653/v1/2023.acl-long.307.

Michael T. Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013. doi: 10.1111/jedm.12000.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. PMET: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18564–18572, 2024a. doi: 10.1609/aaai.v38i17.29818.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024b. URL https://arxiv.org/abs/2310.02129.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Jun-Yu Ma, Hong Wang, Hao-Xiang Xu, Zhen-Hua Ling, and Jia-Chen Gu. Perturbation-restrained sequential model editing. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2405.16821.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://arxiv.org/abs/2210.07229.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, volume 162, pp. 15817–15831, 2022. URL https://arxiv.org/abs/2206.06520.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

YongKyung Oh. Position: State-of-the-art claims require state-of-the-art evidence. *arXiv preprint arXiv:2605.17273*, 2026. URL https://arxiv.org/abs/2605.17273.

Steven Piantadosi and Mitchell H. Gail. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12(13):1239–1248, 1993. doi: 10.1002/sim.4780121305.

Sebastian Pohl, Max Ploner, and Alan Akbik. Towards a principled evaluation of knowledge editors. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2) at ACL 2025*, pp. 47–60, 2025. doi: 10.18653/v1/2025.l2m2-1.4.

Peter M. Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186, 2005. doi: 10.1016/S0140-6736(05)17709-5.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2310.11324.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2409.12183.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 53764–53797, 2024a. doi: 10.52202/079017-1703.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 82–93, 2024b. doi: 10.18653/v1/2024.acl-demos.9.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024c. doi: 10.18653/v1/2024.findings-acl.441.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024d. URL https://arxiv.org/abs/2406.01574.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. The butterfly effect of model editing: Few edits can trigger large language models collapse. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5419–5437, 2024a. doi: 10.18653/v1/2024.findings-acl.322.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, and Huawei Shen. The fall of ROME: Understanding the collapse of LLMs in model editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4079–4087, 2024b. doi: 10.18653/v1/2024.findings-emnlp.236.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 15336–15354, 2025. doi: 10.18653/v1/2025.acl-long.745.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 10222–10240, 2023. doi: 10.18653/v1/2023.emnlp-main.632.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *PMLR*, pp. 58984–59002, 2024.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Uncovering overfitting in large language model editing. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024. URL https://arxiv.org/abs/2401.01286.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://arxiv.org/abs/2102.09690.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 4862–4876, 2023. doi: 10.18653/v1/2023.emnlp-main.296. URL https://aclanthology.org/2023.emnlp-main.296/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2309.03882.

Qiaoyuan Zheng and Yiqu Yang. Are near-tied LLM rankings robust to family-DIF-guided benchmark recomposition? *arXiv preprint arXiv:2609.00482*, 2026. URL https://arxiv.org/abs/2609.00482.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 15686–15702, 2023. doi: 10.18653/v1/2023.emnlp-main.971.

The appendix gives the evidence behind each claim in the main text, organised by what a reader checking a claim needs: construction and what was run, the test and the audit, each experiment's full results, the failure-mode and specificity analyses, the released-benchmark re-analysis, and the registration record. Appendix H.1 gives the status (registered, registered amendment, post hoc or observational) of every reported result in one table, so individual sections do not repeat it except where a distinction matters. Post hoc analyses are labelled wherever they appear; those computed by `post_hoc_analyses.py` read stored outputs and required no new run. Longer supporting tables and derivations are in the supplementary archive and named where they are used.

Table 3: Terms used in a fixed sense.

term	meaning
probe	a question about an edited fact; its **semantic type** is what it asks and its **response format** is how the answer is elicited and scored
cell, arm	a cell is one (semantic type $\times$ format) combination; an arm is one editor's mean in a cell
contrast	a comparison varying one factor: **semantic** (two semantic types, one format) or **measurement** (two formats, one semantic type)
restatement, consequence	KnowEdit's `Subject_Aliasing` probe (its answer is the edit target) and `Reasoning` probe
corrected reversal	a comparison whose reversal test (B.2) rejects after Bonferroni correction over its registered family (B.3)
flip weight	at two strata, the weight on the first stratum at which the composite of a pair ties, $w^* = \lvert\delta_2\rvert/(\lvert\delta_1\rvert + \lvert\delta_2\rvert)$ (B.4)
minority-direction bound	the largest opposite-direction effect in any stratum that a lower confidence limit at $\alpha/(2 \cdot \text{family})$ does not exclude (B.4)
attribution flag	a check reported beside a verdict, never a test: M1 label dependence, M2 an arm within 0.05 of ceiling or chance, M3 unpaired strata, M4 leakage in each off-target family (§2.3, B.6)
cell-mean restriction	a cell is scored when its across-editor mean is at least 0.05 from ceiling and from chance; chosen after Experiments 1 and 2 and registered as Experiment 3's primary restriction
arm-level restriction	both compared editors at least 0.05 from ceiling and from chance in both strata; a registered sensitivity in Experiment 3, and the M2 flag in the audit
first-span match	post hoc generation score: 1 when the trimmed continuation is, or begins with, an accepted alias at a word boundary (E.1)
P0, P1, P2	Experiment 3's registered populations: unselected (533 facts, primary), fully symmetric (15), and Experiment 1's screened set (255); the 278 facts of P0 outside P2 are the **held-out facts**
leakage	how often the edit target outranks a prompt's own gold answer on a prompt the edit should leave alone, under the constrained-choice rule, on different-subject (**neighbourhood**) or same-subject (**locality**) prompts (F.1)

A CONSTRUCTION AND WHAT WAS RUN

Experiment 1 is registered in `PREREG_CONTROLLED_2026-09-04.md`, Experiment 2 in `PREREG_EXPERIMENT2_2026-09-13.md`, Experiment 3 in `PREREG_SYMMETRIC_2026-09-14.md`, the specificity test in `PREREG_LOCALITY_2026-09-15.md`, Experiment 4 in `PREREG_EXPERIMENT4_MODERN_2026-09-16.md` Experiment 5 in `PREREG_EXPERIMENT5_SPECIFICITY_2026-09-16.md`, Experiment 6 in `PREREG_EXPERIMENT6_FAITHFUL_2026-09-18.md`, whose §7 amends one configuration reading after it proved unrunnable and before any of its blocks existed, Experiment 7 in `PREREG_EXPERIMENT7_FIDELITY_2026-09-18.md`, and Experiment 8 in `PREREG_EXPERIMENT8_LLAMA_2026-09-18.md`. All ship in the supplementary archive with their hash chains; Appendix H.4 describes what that record does and does not establish.

A.1 THE FROZEN ITEM SET

573 facts (KnowEdit `recent` and `counterfact`) carrying both a `Subject_Aliasing` and a `Reasoning` probe, rendered in 3 formats $\times$ 2 mechanical paraphrases = 6,876 items, exactly 1,146

per (semantic type × format) cell, item-set sha256 `450ea4aa19e7e21e...`. The questions are KnowEdit's own, verbatim; we author only the format rendering and a mechanical paraphrase. The item file carries no pre-edit object for any fact, which is why a CounterFact-style Neighborhood Score cannot be computed on it (I.2). KnowEdit's released records do not supply one either: `requested_rewrite.ground_truth` exists in every record and is the constant `<|endoftext|>` in all 839 `counterfact`, 1,266 `recent` and 1,301 `zsre` records, so the field a reader might expect to hold the original object holds nothing. Recovering a pre-edit object would mean re-deriving it from the source, which is a change of item set rather than a re-analysis.

Three construction rules are pinned by regression tests: the paraphrase rule must not produce a doubled article; a fact duplicated in the KnowEdit release enters once; and answer leakage is tested at word boundaries. A substring test reports 354 leaks, most of them artefacts (`man` inside `Ackerman`); the word-boundary test finds 222 real ones, for example a question asking for the place of death of the architect of the Royal Palace of **Madrid** whose gold answer is `Madrid`. 39 facts were dropped for leakage, each **whole**, so the factorial stays balanced.

Distractors are drawn from other facts sharing the same relation stem, so that a model cannot reject an option on type alone. Where a stem is too rare, the pool falls back to the wider type pool (228 items; E.7).

A.2 Scoring, decoding and the answerability screen

Constrained choice ranks the four options by length-normalised log-likelihood after the prompt and scores 1 when the gold option ranks first. **Free generation** renders the question as `"{question}\nAnswer:"`, decodes greedily (`do_sample=False`) for at most 12 new tokens, stopping only at the end-of-sequence token or the limit, and cuts the decoded text at the first newline, period, comma or semicolon found after its first character. The trimmed continuation is scored by token F1 against each accepted alias after lower-casing, removing articles and replacing every run of characters outside `[a-z0-9]` by a space, and the item score is the maximum over aliases. The trimming has been in the code at every commit that produced a block, and none of the stored Experiment 3 continuations contains a cut character. This scorer is not KnowEdit's teacher-forced token accuracy. **Yes/no** compares the mean log-probability of " yes" and " no" after a stem (`question`, `Proposed answer:`, the claim, `Is that correct?`) that differs from the item's rendered prompt, which also asks to answer yes or no; a tie scores 0. It enters no contrast of Experiment 3 because every weight editor is at chance on it.

The edit-prompt **efficacy** that Experiment 5 reports (J.2, J.5) uses the same decoding, trimming and token F1, with three differences: the prompt is the raw edit prompt without `Answer:`, the score is against `target_new` alone, and the mean is over every completed fact.

The screen of Experiments 1 and 2 runs on the **unedited** model with the target fact supplied in context, under constrained choice, before any weight is touched. A fact is retained only if **all four** (semantic type × paraphrase) cells are answerable, so retention is at the fact level and the retained set is fully paired across strata; the machine-checkable form of that is that the intersection hash equals the retained-set hash. The screen takes no editor argument and refuses to run below the registered floor. Experiments 1 and 2 use it; Experiment 3 records it and gates nothing. E.4 explains why the screen is not neutral between the two contrasts.

Three properties of the scoring were fixed before any editor ran (deviations D4 to D7, H.2):

- **The gold position cycles A–D deterministically.** With the gold always "A", GPT-J answered "B" on 131 of 150 diagnostic items regardless of content. A self-test fails if any letter is correct for more than 60% of items (final counts A 573 · B 573 · C 573 · D 573, over the 2,292 constrained-choice items).
- **Distractors are type-matched** by relation stem (A.1).
- **Constrained choice and yes/no are scored by likelihood ranking** over the candidate answers, not by an emitted letter. The registered letter rule could not measure a 2021 base model: with the correct answer in context GPT-J-6B still scored at chance, retaining 2 of 316 facts. Length normalisation is self-tested, since summed log-probability favours short options.

A.3 EDITOR QUALIFICATION

Every editor must pass five end-to-end witnesses on calibration facts disjoint from the analysed facts before it may enter a pool. They test for the two ways a pipeline produces well-formed output without an intervention: a scorer that writes a constant, and an editor whose update is rounded away by the model's precision.

The five witnesses are: **A**, parameter (max $|dW|$, changed fraction, loss trajectory), which fails when the weights do not move; **B**, functional, which fails when the weights move and the target's negative log-likelihood on the edit prompt does not fall; **C**, full stack, from prompt to raw output to parsed score, which fails when an empty or malformed output scores as success; **D**, sham, which fails unless a learning rate of 0 gives a delta of exactly 0; and **E**, cross-editor, which fails when two distinct editors return identical scores.

The criterion was frozen before any calibration summary was read and is identical in all experiments: max |dW| $\geq$ 1e-6, changed fraction $\geq 1\%$, mean target-NLL drop ≥ 0.5 nats, $\geq 60\%$ of calibration facts improved, sham delta exactly 0. Because calibration facts are disjoint from the analysed facts, no analysed fact is touched while qualifying and no contrast is inspected. Witness B excluded LoRA at EasyEdit's published learning rate on Experiment 1's first qualification: its weights moved enormously (dW $\approx$ 11), which passes witness A, while the target NLL rose from 4.10 to 15.58 and outputs collapsed to repeated characters.

A.4 WHAT RAN: EDITOR CODE, PRECISION AND CONFIGURATIONS

Table 4 lists every configuration and every departure from it.

Table 4: Editor configurations. On Mistral-7B and Qwen2.5-7B, LoRA's modules are `up_proj` and `down_proj`; Experiment 4's FT-M edits layer 27 of Qwen2.5-7B-Instruct.

editor	EasyEdit GPT-J configuration	what we ran	why
FT-L	layer 21 `fc_out`, 25 steps, lr 5×10^{-4}; loss on the target tokens' log-probabilities at the last prompt position (`prompt_last`); `norm_constraint: false`	same layer, steps and rate; **teacher-forced cross-entropy over prompt and target**; L_∞ clamp 5×10^{-3} on the weights	**no reason for the change of objective is recorded anywhere** in the registrations, logs or commits; the record justifies only the clamp width (a clamp equal to the learning rate saturated Adam on step one). Two departures; the published objective is run as its own arm in M
FT-M	as FT-L with the loss on target tokens only	the same, target-masked, no clamp	matches apart from precision and early stopping; Experiment 7 restores the stopping rule and the arm's repetition falls 0.588 to 0.197 and its neighbourhood leakage 0.976 to 0.651 (N.2), so this one **mattered**

editor	EasyEdit GPT-J configuration	what we ran	why
LoRA	AdaLoRA, rank 8, α 32, dropout 0.1, on `fc_in` and `fc_out`, 50 steps, lr 5×10^{-3}, no clipping	**plain LoRA** with the same rank, α, dropout and modules, 50 steps, **lr** 10^{-4}**, gradient-norm clipping at 1.0**, target-masked loss	the published setting diverged in this implementation on calibration facts (A.3). Deviation D8. Experiment 6 ran AdaLoRA at the published rate: it diverged on **20 of 20** calibration facts and produced no block (M.1)
FT-multi	none; ours	FT-M's objective on **five** MLP down-projections, layers 17–21, 20 steps, lr 2×10^{-4}, L_∞ clamp 5×10^{-3}	a statistics-free analogue of a spread-over-layers update, used while the framework could not be imported (D3). Not a published method: excluded from Experiment 3
ROME	`mom2_adjustment: false`	the same files, unmodified	EasyEdit's GPT-J configuration uses identity covariance rather than the C^{-1}-whitened update of the original paper. Experiment 7 runs the whitened update: same profile, slightly lower scores (N.2), so this one **did not matter**
MEMIT	published covariance statistics	the same files, reading those statistics (hash-pinned)	matches

The gradient editors are our code. `FT-L` (fine-tuning one layer against the target's log-probability at the last prompt position), `FT-M` (the same with the loss on the target tokens only), `FT-multi` and `LoRA` are classes in `run_controlled.py`, present since its first commit. The editing framework could not be imported on the cluster (deviation D3), so only the values of its published configuration files were copied. Those values match the current main-branch configuration files of EasyEdit; no EasyEdit commit was recorded when they were copied, so the comparison below is against that version. **Precision:** every block records `torch.bfloat16`. GPT-J-6B is loaded from its `float16` revision and cast to bfloat16 (EasyEdit loads in float32 unless told otherwise). Each gradient editor keeps a float32 master copy of every edited tensor, takes plain Adam steps on it (no weight decay) and copies it back after each step, since bfloat16 alone rounds an update of order 10^{-5} on a weight of order 10^{-2} back to the same value. There is no early stopping; EasyEdit stops fine-tuning once the loss falls below 10^{-2}.

ROME and MEMIT are EasyEdit's files (Wang et al., 2024b)**.** `extract_lae.py` copies the published implementation files into a flat package and rewrites only import lines; for every file it prints a diff against the original that ignores import lines, and no non-import line differs in any of the twenty files (C.3). They run on the same bfloat16 model through an adapter that sets the device and, from Experiment 3's specificity run onward, seeds the context templates at EasyEdit's value of 42.

Scope of editor claims. "The gradient editors" means these implementations: FT-M at EasyEdit's GPT-J values **trained past its stopping point**, FT-L with a different objective and a clamp, and plain LoRA at a fiftieth of AdaLoRA's learning rate. Experiment 7 measures what the first of those costs: with the published stopping rule the arm repeats a third as often and leaks on two thirds as

many neighbourhood prompts, while editing the target *better* (N.2). Statements in E.1 and F about FT-M's repetition and leakage are therefore statements about an over-trained arm, and the direction survives the correction while the degree does not. They are not ROME's constrained fine-tuning baseline (FT+L) and cannot be positioned against the FT-L, FT-M and LoRA numbers reported in KnowEdit or HalluEditBench. Experiment 6 (M) runs EasyEdit's own FT-L and ROME's FT+L on the same facts, and Experiment 8 (O) runs the published Llama-2 configurations on the release's own backbone; AdaLoRA at the published rate diverged on every calibration fact on **both** backbones and did not run, so no result here describes EasyEdit's published LoRA editor. On Llama-2, MEMIT also did not run: its published configuration needs second-moment statistics that are not released for that backbone (O.1).

A defect in one Experiment 4 arm. The block of Experiment 4's LoRA arm at the published Qwen configuration records `target_modules: [q_proj, v_proj]`, a learning rate of 5×10^{-3} and 60 steps. At the commit that produced it, the LoRA class chooses its target modules itself and never reads that field, so this arm most likely trained `up_proj` and `down_proj`, with clipping at 1.0. This is inferred from the code; no flag records whether the working tree differed. The arm diverged and is not in the analysed pool (K), so no reported comparison uses it, but its divergence rate is not evidence about the published configuration.

The in-context arm is not IKE. Zheng et al. (2023) retrieves copy, update and retain demonstrations; ours is a zero-shot prefix, `"{edit_prompt} {target_new}."`. Under constrained choice that prefix **is** the answerability screen's retention condition, so the arm scores exactly **1.0000** in every constrained-choice cell of Experiments 1 and 2 by construction. It never enters a multiplicity family and no claim rests on it.

B THE TEST, THE FAMILIES AND THE AUDIT

B.1 WHAT IS TESTED

A prespecified weighted average is a legitimate composite, and heterogeneous components do not invalidate it. What the composite does not establish on its own is the cross-probe claim that one editor's advantage holds in direction on every prespecified probe class. The test below addresses that claim; it is not an argument that a benchmark's metric is wrong.

Proposition 1 (§2.1). Let $\delta \in \mathbb{R}^P$ with $\delta \neq 0$ and let $W = \{w : w_p > 0 \; \forall p\}$. Then $\operatorname{sign} \sum_p w_p \delta_p$ is constant on W if and only if the nonzero components of δ share a sign; if both signs occur, W contains weightings with positive, negative and zero aggregate.

Proof. If every nonzero δ_p is positive, then for $w \in W$ the sum $\sum_p w_p \delta_p$ is a sum of non-negative terms with at least one positive term, so it is positive; the negative case is symmetric. Conversely, let $\delta_i > 0 > \delta_j$ and $\varepsilon > 0$. The weighting with $w_i = 1$ and every other $w_p = \varepsilon$ gives $\delta_i + \varepsilon \sum_{p \neq i} \delta_p$, positive for small ε; the weighting with $w_j = 1$ and every other $w_p = \varepsilon$ is negative for small ε. W is convex and the aggregate is linear in w, so the segment between these two weightings contains one with aggregate zero. □

This is the Pareto-dominance fact for positive linear scalarisation, and it is used only to fix the estimand. The proposition concerns true effects. With estimated effects, "components of each sign" becomes "significant effects of each sign", which is what the reversal test of B.2 requires; the pair is oriented so that A is the aggregate winner, and a rejection means some class significantly favours B.

B.2 THE REVERSAL TEST, AND WHERE EACH VERSION IS USED

For one editor pair, let d_i be the paired mean difference in stratum i, s_i its standard error and $z_i = d_i / s_i$, $i = 1, \ldots, I$. The null is no qualitative interaction: the I true effects have a common sign.

Gail and Simon's test (Gail & Simon, 1985), the statistic every registration names, uses

$$Q^+ = \sum_i z_i^2 \, [d_i > 0], \qquad Q^- = \sum_i z_i^2 \, [d_i < 0], \qquad Q = \min(Q^+, Q^-),$$

with level $P(Q \geq q) = \sum_{h=1}^{I-1} \binom{I-1}{h} 2^{-(I-1)} P(\chi^2_h \geq q)$. These mixing weights reproduce the published critical values 2.71, 4.23, 5.43 and 6.50 at $I = 2$ to 5. The reference distribution is derived for independent stratum estimates, and probe types scored on the same items give correlated ones.

When dependence matters. At two strata the test rejects only when both $|z_i|$ exceed the critical value with opposite signs. Its worst-case size under the common-sign null, computed by bivariate normal integration over one stratum's effect with the other at zero, is exactly nominal at every correlation from -0.70 upward, 0.0504 at -0.80, 0.0637 at -0.90 and 0.0742 at -0.95 at nominal 0.05; at the per-comparison threshold $0.05/40$ it is at most 1.002 times nominal down to -0.90 and 1.19 times nominal at -0.95. With three or more strata the direction reverses: in a 400,000-draw simulation with one stratum far above zero and the others null and equicorrelated, the test rejects 5.5% and 6.1% of the time at $I = 3$ and 6.5% and 8.3% at $I = 5$ for correlations of 0.2 and 0.5, against a nominal 5% (and at nominal level when the strata are independent). That configuration bounds the size from below. All of these numbers are fields of `post_hoc_analyses.json` (`gail_simon_size`).

The intersection-union test. With one-sided p-values $p_i^+ = P(Z \geq z_i)$ and $p_i^- = P(Z \leq z_i)$,

$$p = \min\Big\{1,\ I\ \max\big(\min_i p_i^+,\ \min_i p_i^-\big)\Big\}.$$

If every true effect is non-negative, rejecting requires some stratum's lower one-sided test to reject at α/I; each such test has level α/I whatever the other strata do, so the union over I strata has level α under any dependence, and likewise with the signs exchanged. This is the intersection-union principle (Berger, 1982) applied to a Bonferroni-calibrated range test in the sense of Piantadosi & Gail (1993). At $I = 2$ with opposite signs its p is exactly twice Gail and Simon's, and it is 1 when the strata agree in sign, where Gail and Simon's p takes its maximum of 0.5. At the Bonferroni threshold $0.05/40$ a two-stratum rejection by it requires both $|z_i|$ to exceed 3.23, against 3.02 for Gail and Simon's.

The policy. Each result is reported under the test that is valid for its dependence structure, with the other beside it:

- **two strata sharing items**, every estimated between-stratum correlation above -0.9: Gail and Simon's test is primary. This covers every controlled comparison in this paper; across the 79 computable Experiment 3 comparisons (both contrasts and both subsets of the split) the correlation of the per-fact differences runs from -0.13 to $+0.86$, and at the lowest of them the worst-case size at $0.05/40$ is exactly nominal;
- **three or more strata sharing items**, or strata that overlap only partly: the intersection-union test is primary (HalluEditBench's five question types, KnowEdit's three portability probes, HELM's perturbations and answer formats), because Gail and Simon's test can exceed its level there under positive correlation. On HalluEditBench the estimated correlations run from -0.24 to $+0.29$, and the two tests reject the same 17 comparisons;
- **independent strata**, content subgroups defined by disjoint items or probe types scored on disjoint items: Gail and Simon's test, used unchanged.

The analysis instruments use normal tails for both tests. The audit tool uses normal tails for Gail and Simon's test and Student's t on $n - 1$ degrees of freedom for the intersection-union test, which is slightly more conservative.

Reporting both. The registered analyses of Experiments 1 to 5 name Gail and Simon's test, and the policy above keeps it where it is valid. Because the two tests differ on exactly the comparisons that are seed-fragile or do not replicate out of sample (J.7), both counts are reported everywhere and no claim rests on the choice between them. `dependence_robust.py` writes the intersection-union re-run of every registered analysis to `*_robust.json` beside the registered artifact, which it leaves byte-identical.

B.3 Unit, families, and strata that a check sets aside

Unit. The unit is the fact (controlled experiments) or the item (released benchmarks); paraphrases of one fact are averaged before testing, so standard errors are clustered on facts.

Families. Bonferroni is applied over every comparison in a family, including any that cannot be computed. Each test holds its level per comparison, so Bonferroni over the family is a valid level-α

test of the intersection null however the comparisons depend on one another. A family is not shrunk to the comparisons a restriction or a flag keeps, because which comparisons those are depends on arm means estimated from the same data. One table lists every family in the paper:

Table 5: Every multiplicity family in the paper. The audit tool's Experiment 3 sections hold half the registered measurement family and a quarter of the split's, so its counts there are worked examples and not the registered counts (B.6).

analysis	family	status
Experiment 1, primary (constrained choice)	6	registered
Experiment 2, primary (free generation)	3 (6 before LoRA's exclusion, D9)	registered
locate-and-edit extension, per format	10	registered
Experiment 3, measurement and semantic contrasts	40 (20 + 20)	registered
Experiment 3, answer-identity split	80 (the 40 plus 20 per subset)	registered amendment D2
Experiment 3 on the 278 held-out facts; generation rescoring; margin rescoring	40 and 80, as registered	post hoc
Experiment 3, subset $\times$ stratum interaction test	20 per subset	post hoc
Experiment 5, primary; the post hoc dose comparisons	12	registered; post hoc
Experiment 5, secondary C and S on Experiment 3's reversals	each reversal's own registered family (40 or 80)	registered
Experiment 4	18	registered
Experiment 6, the published configurations	12 (4 not computed: `AdaLoRA` did not run)	registered
Experiment 7, the benchmark's own metric (family T)	12	registered
Experiment 7, the two fidelity arms (family F)	8	registered
Experiment 8, the release's backbone (family L)	6 (2 not computed: `AdaLoRA` did not run)	registered
Experiment 8, the semantic contrast (family S)	6 (2 not computed: `AdaLoRA` did not run)	registered
answer-identity split on Experiments 1 and 2 and the extension	38	post hoc
random and systematic unpairing	10 per split	post hoc
HalluEditBench	84 comparisons; Bonferroni over 21 per backbone	observational
KnowEdit portability	42 comparisons; Bonferroni over 42, and over 21 per cell	observational
KnowEdit `recent`, answer-identity split	21	post hoc
audit tool, worked examples	every pair in the section it is run on: 21 per HalluEditBench backbone and per KnowEdit cell, 20 per Experiment 3 section, 45 or 15 on HELM	the tool's default; not a registered family
registered margin and chance-level grids	the comparisons in the cells each setting scores	as registered

Restricted counts are held to the full family: the counts at the resized families the arm-level restriction keeps (22 measurement and 13 split comparisons in Experiment 3, 9 in Experiment 5) appear only in H.1. The registered margin and chance-level grids are the exception to the rule: their registration

defines each setting's family as the comparisons in the cells that setting scores, and they are reported as registered. The choice of restriction, margin, chance level, format pair, scale and population is swept and reported rather than corrected for (E.4, E.7).

Strata that a check sets aside. The formal verdict is issued on all prespecified strata. When some strata have an arm within 0.05 of ceiling or chance, the test on the remaining strata is reported as a diagnostic and not as a verdict. Dropping strata can only remove a rejection of the all-strata null, so the all-strata verdict keeps its level. A test on the retained strata, read as a claim about those strata, does not, because which strata are retained depends on arm levels estimated from the same data. At two strata, setting one aside leaves no test, so the registered cell-mean restriction can only remove comparisons: on P0 it leaves the counts unchanged (3, 0 and 8), and on the held-out facts it removes one measurement reversal (D.3).

Counts and effect sizes. Among k editors the signs of all pairwise differences within a stratum are determined by that stratum's ordering, so the number of reversed pairs is a rank distance between two orderings rather than a count of independent events (D.5). Magnitudes are reported as fact-unit interactions with simultaneous intervals, flip weights and minority-direction bounds.

B.4 Output states, the flip weight and the minority-direction bound

States. The states are defined once, in code, and used by the analyses and the audit tool alike: **qualitative failure** when the reversal test rejects after Bonferroni; **magnitude heterogeneous** when it does not and a heterogeneity test rejects at the same threshold; **no detected heterogeneity** otherwise; and **untestable** when fewer than two strata have a standard error. The heterogeneity test is a cluster-robust Wald test of equal stratum effects, $W = (Cd)^\top (C\Sigma C^\top)^{-1}(Cd)$ on $I-1$ degrees of freedom with C the successive contrasts, in which the covariance of two stratum means is estimated from the items they share, $\Sigma_{jk} = c_{jk}\, n_{jk}/(n_j n_k)$. For disjoint strata it equals Cochran's Q (Cochran, 1954); at two strata on identical items it equals the squared paired t statistic of the interaction. The same test is also computed on the log-odds scale, with delta-method influence values, because a constant log-odds effect over strata with different base rates is heterogeneous on the additive scale; a pair heterogeneous additively and not on the log-odds scale is flagged **additive only**.

Three output states, and a fourth question. Beside the heterogeneity verdict the tool reports whether the **cross-probe claim** is `REFUTED` (the reversal test rejects), `ESTABLISHED` (every stratum excludes a reversal at the simultaneous level, which is Berger's intersection-union test of $\min_s \delta_s \le 0$ rejecting) or `NOT ESTABLISHED`. The third state matters because a non-rejection of the reversal test is not evidence of agreement, and until version 5.0 the tool could not tell a well-powered agreement from an underpowered one. On KnowEdit's 42 comparisons it reports 20 established, 13 not established and 9 refuted.

The flip weight. At two strata with opposite-signed effects the composite of a pair is a tie at weight $w^* = |\delta_2|/(|\delta_1| + |\delta_2|)$ on stratum 1, and the aggregate winner under equal weights keeps its lead only on one side of w^*. Intervals for w^* come from a fact-clustered bootstrap with multinomial weights ($B = 20{,}000$). With three or more strata the audit reports the share of the weight simplex on which the aggregate winner still wins (20,000 uniform Dirichlet draws) and the smallest L_1 change from equal weights that makes the composite a tie, obtained by moving weight from the strata most favouring the winner to the stratum most favouring the other editor. **Neither is scale-free.** The sign verdict is invariant to rescaling a stratum; w^* and the simplex share are not, and on the factorial's three measurement reversals chance-correcting the choice accuracy, $(\mathrm{acc} - c)/(1-c)$, moves w^* from 0.515, 0.514 and 0.495 to 0.443, 0.442 and 0.424, which changes which editor wins at equal weights for two of the three. Version 5.0 therefore prints both, with each stratum's chance level and arm means beside them, so a reader can see which scale a "near-tie" reading depends on.

The minority-direction bound. A bound on the interaction does not bound a reversal ($|\delta_1 - \delta_2| \le 5.3$ is satisfied by strata at -2.6 and $+2.6$), so every pair that is not a qualitative failure carries the **minority-direction bound**. Let the aggregate direction be the sign of the mean over strata; a reversal of size Δ exists when some stratum has $\mathrm{sign}(\bar\delta)\,\delta_s \le -\Delta$. If each L_s is a marginal lower confidence limit for $\mathrm{sign}(\bar\delta)\,\delta_s$ at level $1-\alpha'$, then $\min_s L_s$ is a lower limit at the same level for $\min_s \mathrm{sign}(\bar\delta)\,\delta_s$, since $\min_s \delta_s < \min_s L_s$ implies $\delta_{s^*} < L_{s^*}$ for the minimising stratum. No division by the number of strata is needed; a factor of two covers the orientation chosen from the data. The bound is therefore the largest Δ that $\min_s L_s$ does not exclude, with each L_s at $\alpha/(2 \cdot \mathrm{family})$, computed over all

prespecified strata. It is an intersection-union statement, not an equivalence test: no smallest effect of interest was registered.

B.5 Power and minimum detectable effects at the applied thresholds

A reversal is certified only when both strata are significant in opposite directions, each one-sided at the applied per-comparison threshold. Minimum detectable effects (80% power per stratum) and power are therefore quoted at that threshold.

Table 6: Minimum detectable effects (80% power) at the thresholds applied. For KnowEdit `recent` each stratum's standard error is estimated from the release itself, and they differ by more than a factor of two (3.6–3.7 points on aliasing against 1.6–1.8 on reasoning), so the aliasing stratum binds: certifying a reversal needs a significant effect in each direction, and the joint power against an 8.2-point reversal in both is the product, about 0.17. Experiment 4's standard error is the one implied by its own two-sided MDE of 6.8 pp, and its registered cap was 8 points, so at the applied threshold that run was underpowered (K).

setting	threshold	MDE, aliasing	MDE, reasoning	joint power at 8.2 pp
KnowEdit `recent`, MEMIT vs FT-L, 108 crossed cases	$0.05/(3 \cdot 21)$ ($z = 3.16$)	14.7 pp	7.1 pp	0.16
KnowEdit `recent`, ROME vs FT-L	the same	14.5 pp	6.2 pp	0.18
the same, restricted to true consequences ($n = 88$)	the same	16.5–16.7 pp	6.8–7.4 pp	0.10–0.11
Experiment 4, Gail and Simon's test	$0.05/18$ ($z = 2.77$)	8.8 pp		
Experiment 4, intersection-union test	$0.05/36$ ($z = 2.99$)	9.3 pp		

Table 7: Power to establish a minority-direction effect of size Δ in one stratum at the family of 40, under Gail and Simon's threshold / the intersection-union threshold, at the sample sizes of P0, the held-out facts and KnowEdit `recent`. All values are in `post_hoc_analyses.json` (`power`).

facts	paired sd	power, $\Delta = 5$ pp	power, $\Delta = 10$ pp
533	0.20	0.997 / 0.995	1.00 / 1.00
533	0.32	0.72 / 0.65	1.00 / 1.00
278	0.20	0.87 / 0.83	1.00 / 1.00
278	0.32	0.34 / 0.27	0.99 / 0.98
108	0.20	0.34 / 0.26	0.99 / 0.98
108	0.32	0.08 / 0.05	0.59 / 0.51

B.6 The released audit tool

`probe_consistency_report.py` (version 5.0, 124 checks) is a single standard-library file that runs on the long-format per-item outcomes benchmarks already publish: one row per (method, probe, item) with a score in $[0, 1]$, and optional columns naming a model, a split property, a per-row chance level and a pre-edit score. It writes a JSON report whose header records the command line and the sha256 of every input, and a rendered table. The full input contract is in the file's `--help`. For each method pair it reports two levels of output.

The observed-scale verdict. The state of B.4 on all prespecified strata, with the test chosen by the policy of B.2 and the intersection-union result beside it; Bonferroni and Benjamini–Hochberg over every pair in scope; the between-strata correlation; the interaction with a simultaneous interval at two strata; the additive and log-odds heterogeneity tests; the flip weight or simplex share and smallest flipping reweighting; and, for non-failures, the minority-direction bound.

Attribution flags, each a report and never a test:

- **M3, pairing.** Each stratum uses the items scored for both methods. Strata on pairwise disjoint item sets are flagged **unpaired**; they are independent, so Gail and Simon's test is valid on them, but a difference between strata is also a difference between items. Strata that

overlap without being identical are tested on the fully crossed items when there are at least 30, and otherwise flagged unpaired. With `--strata-are subgroup`, strata declared as content subgroups are tested as independent subgroups and not flagged.

- **M2, ceiling or chance.** Every stratum with an arm within 0.05 of 1 or of that probe's chance level (`--chance`) is named, and the test on the remaining strata is reported as a diagnostic (B.3).
- **M1, labels.** With `--split-col`, strata are (probe, split value); the pooled result is also computed, and a pair whose state differs between the two is flagged **label-dependent**.
- **M4, specificity.** With one or more `--leak FAMILY=CSV` tables of per-method leak rates, the pair carries both rates, their difference and a threshold sweep in **every** family supplied, and a summary: mismatched in every family, matched in every family, family-dependent, or unmeasured. The 20-point default is a convention on a continuous quantity, and on this paper's data the call changes with the family for most reversals (F.4), so no single matched or mismatched call is reported.

Table 8: Worked examples: pairs in each state (qualitative failure, magnitude heterogeneous, no detected heterogeneity, untestable), qualitative failures carrying no flag, and pairs flagged M3 or M2. The reports ship with the commands that produce them: `probe_consistency_halluditbench.json`, `probe_consistency_halluditbench_chance.json` (0.25 for multiple choice, 0.5 for the yes and no questions, 0 for the open-answer ones), `probe_consistency_knowedit.json` and `probe_consistency_exp3.json` (constrained choice and free generation only, with both leakage families). Families are every pair in the section (B.3). HELM is in Appendix L.

example	pairs	QF	MH	NDH	UT	QF, no flag	M3	M2
HalluEditBench, chance levels 0	84	16	67	1	0	4	0	57
HalluEditBench, nominal chance levels	84	16	67	1	0	**0**	0	84
KnowEdit `zsre`	21	7	14	0	0	0	**21**	6
KnowEdit `recent`, 108 crossed facts	21	**2**	17	2	0	**2**	0	15
Exp. 3, measurement	20	4	9	7	0	0	0	9
Exp. 3, semantic, pooled label	20	0	12	8	0	0	0	9
Exp. 3, semantic, same answer	20	1	12	7	0	0	0	9
Exp. 3, semantic, true consequence	20	9	2	8	1	0	0	18

Reading the examples. On HalluEditBench the 16 corrected failures are the 17 rejections of §3 less one (Llama-2 ROME against MEMIT) that does not survive a per-backbone Bonferroni correction; with chance levels at zero four carry no flag, and with nominal chance levels every one has an arm at chance. On KnowEdit `zsre` all seven corrected failures are on disjoint item sets. On `recent` the fully crossed facts carry **two** corrected failures with no flag: MEMIT against FT-L ($p = 2.8 \times 10^{-6}$; MEMIT ahead by 18.6 points on aliasing and behind by 13.6 on reasoning) and ROME against FT-L ($p = 2.1 \times 10^{-5}$; ahead by 16.5, behind by 14.2). Their composites flip under small reweightings (the aggregate winner keeps 53% and 52% of the weight simplex), and the release carries no off-target target-against-gold scores, so their specificity is unmeasured; G.4 splits their label. Under the tool's third output these two are the only `recent` pairs whose cross-probe claim is REFUTED; 20 of the cell's 42 comparisons have it ESTABLISHED and 13 NOT ESTABLISHED (B.4). On Experiment 3 every corrected failure carries a flag, and B.7 gives the rate among non-failures beside it, without which that sentence carries no weight. The tool's section families are half and a quarter of the registered ones, so it certifies four measurement failures where the registered analysis certifies three (FT-M against ROME in `Reasoning`, $p = 1.4 \times 10^{-3}$, clears $0.05/20$ and not $0.05/40$) and ten split failures where it certifies eight. The minority-direction bounds of non-failures reach 6.3 points on HalluEditBench, 5.7 on KnowEdit and 8.5 on Experiment 3's measurement contrast (FT-L against ROME in free generation).

Table 9: One pair per kind of output, from the shipped reports (nominal chance levels on HalluEditBench).

pair	pairing	strata (effect, pp)	verdict and composite	attribution flags
HalluEditBench GPT-J, LoRA vs MEMIT	paired, 2,352 items	rephrase -29.8, yes $+23.3$, no $+15.1$, mc -5.3, reversed $+0.1$	qualitative failure, $p = 2.9 \times 10^{-102}$ (intersection-union, $I = 5$); LoRA keeps 56% of the simplex, and an L_1 change of 0.026 ties the composite	M2: LoRA at chance on mc, both at chance on no; the diagnostic on the other three strata also rejects; M4 unmeasured
Experiment 3, FT-M vs ROME, `Reasoning`	paired, 533 facts	choice $+6.0$, generation -4.6; interaction $+10.6$ $[+4.3, +16.9]$	qualitative failure at the tool's family of 20 ($p = 1.4 \times 10^{-3}$, Gail and Simon; 2.9×10^{-3} intersection-union), not at the registered 40; flip weight 0.43	M4: mismatched on neighbourhood prompts (97.7% against 6.4%), matched on same-subject prompts
HalluEditBench Mistral, FT-L vs FT-M	paired, 2,233 items	yes $+53.3$, mc $+28.5$, rephrase $+27.9$, no $+25.1$, reversed $+8.8$	magnitude heterogeneous (also on log odds); bound 0	M2: FT-L at chance on no, reversed and yes
KnowEdit `zsre`, FT vs FT-L	disjoint item sets	aliasing $+43.1$, logical $+18.5$, reasoning $+3.5$	magnitude heterogeneous (Gail and Simon, independent strata); bound 0	M3: unpaired
HalluEditBench GPT-J, FT-M vs GRACE	paired, 2,154 items	rephrase $+92.2$, mc $+15.7$, no $+7.7$, yes $+2.5$, reversed $+2.4$	magnitude heterogeneous; bound 0	M2: GRACE at chance on every question type, so no diagnostic test exists

The family for an observational count, and the count under the other choice. The released benchmarks are corrected per backbone (HalluEditBench, 21 comparisons) and per cell (KnowEdit, 21), because a backbone or a cell is the unit within which a ranking is read: the release reports one table per backbone, and nobody compares an editor's GPT-J aliasing score with its Llama-2 reasoning score. A reader who prefers one family over the whole release gets almost the same answer, and both are reported here rather than only the one that suits:

Table 10: Corrected counts under both family choices. The paper's headline "8 after Bonferroni" for KnowEdit is the global-family count; its per-cell count is 9.

setting	uncorrected .05	per backbone or cell	global family
HalluEditBench, 84 comparisons	17	16 ($0.05/21$)	**16** ($0.05/84$)
KnowEdit, 42 comparisons over the two usable cells	12	9 ($0.05/21$)	**8** ($0.05/42$; `recent` 2, `zsre` 6)

Four worked pairs, as the tool prints them.

Table 11: Audit output for four pairs from the shipped reports. Effects in points. The third column carries three distinct quantities: at three or more strata, the share of the weight simplex on which the aggregate winner still wins and the smallest L_1 reweighting that ties the composite; at two strata, the flip weight, the weight on the first stratum at which it ties.

pair	verdict on all strata	how far the composite is from flipping	attribution flags
HalluEditBench GPT-J, LoRA vs MEMIT	qualitative failure (rephrase -29.8, yes $+23.3$)	LoRA wins on 56% of weightings; an L_1 shift of 0.026 ties it	M2: arms at chance on two closed-answer types; leakage unmeasured
KnowEdit `recent`, MEMIT vs FT-L	qualitative failure (aliasing $+18.6$, reasoning -13.6)	MEMIT wins on 53% of weightings	none; leakage unmeasured
Experiment 3, FT-M vs ROME, restatement	qualitative failure (choice $+16.0$, generation -17.0)	flips at choice weight 0.514	M2: FT-M at ceiling; M4: mismatched on different-subject, matched on same-subject prompts
HalluEditBench Mistral, FT-L vs FT-M	magnitude heterogeneous; bound 0	FT-M wins on every weighting	M2: FT-L at chance on three types

B.7 Do the attribution flags discriminate?

The paper reports that every corrected failure in three of its four settings carries a flag. That sentence is evidence only if flags are rarer among the pairs that do **not** reverse, and earlier versions never reported the other half of the table. `flag_operating_characteristics.py` (11 checks) reports it, for every shipped audit report, and the answer is unfavourable to M2.

Table 12: How often M2 fires among the pairs the reversal test certifies and among those it does not. A likelihood ratio of 1 means the flag carries no information about which pairs reverse.

section	failures	M2 among failures	M2 among non-failures	likelihood ratio
HalluEditBench, nominal chance levels (all four backbones)	1–9 of 21	1.00	1.00	**1.0**
HalluEditBench, chance levels at zero: GPT-J	9 of 21	0.89	1.00	0.89
ditto, Llama-2-7b-chat	1 of 21	0.00	0.55	0.0
ditto, Llama-3-8B	3 of 21	0.33	0.56	0.6
ditto, Mistral-7B	3 of 21	1.00	0.67	1.5
KnowEdit `recent`	2 of 21	0.00	0.79	0.0
KnowEdit `zsre`	7 of 21	0.14	0.36	0.4
Experiment 3, measurement (`Subject_-Aliasing`)	3 of 10	1.00	0.86	1.2
Experiment 3, true consequence (constrained choice)	6 of 10	1.00	0.75	1.3

M2 carries almost no information, and in two settings it points the wrong way. At the nominal chance levels §3 quotes it fires on every pair of every HalluEditBench backbone, so its ratio is exactly 1: "every corrected failure has a stratum at chance" is then a statement about the benchmark's answer spaces and not about those failures. On KnowEdit it fires on **0%** of the two `recent` failures against

79% of the non-failures, and on 14% against 36% in `zsre`. Only in the factorial does it point the right way, and weakly (ratios 1.2 and 1.3).

M4 does discriminate, in the one setting where it can be measured. In Experiment 3's measurement contrast it fires on all 3 certified failures against 43% of the 7 non-failures, and in the true-consequence constrained-choice section on all 6 failures against **none** of the 4 non-failures. **M3 cannot discriminate where it fires**: on KnowEdit both cells score every pair's probe types on the same disjoint item sets, so it is a property of the cell and fires on 100% of failures and 100% of non-failures alike. That is not a defect of the flag; it is what a design-level confound looks like, and it is the reason the `zsre` result is stated about the cell rather than about its pairs.

A planted check. Base rates cannot say whether a flag fires on the *right* pairs, because the released pairs carry no ground truth about why they reverse. Two populations of 400 items were therefore generated with a known answer and scored by the same code:

- **REAL.** Editor A leads probe 1 per item and trails on probe 2 per item, with A at 0.978 on probe 1 (a ceiling) and 0.265 on probe 2 (chance). The crossing is genuine.
- **ARTEFACT.** A is far ahead on 80% of items and behind on 20%, so neither editor dominates the other per item, which is the case M2's rationale names. Uncapped, A's probe-1 mean leads by 32.5 points; the cap at 1.0 discards exactly the excess it leads by, and the observed means cross.

Both are certified by the reversal test ($p = 1.2 \times 10^{-9}$ and 6.8×10^{-15}), and **M2 fires on both**, so it does not separate them. The check that does is the paper's own untruncated margin (E.2), applied to the stratum carrying the minority sign: in REAL the minority sign survives the untruncated scale (gap -0.19), and in ARTEFACT it reverses on it (gap $+0.32$), which is what a cap effect looks like.

What follows for the paper. M2 is reported as a reason to look at a stratum, never as a reason to discount a reversal, and the sentence "every corrected failure carries a flag" is reported with its base rate beside it wherever it appears. The check that carries weight against a compression artefact is the untruncated margin, which in our own data **keeps** all three measurement reversals and raises the split count to 10 of 10 (Table 22). This subsection is post hoc and is a validity check of a reporting device, not of any editor.

C EXPERIMENTS 1 AND 2 AND THE LOCATE-AND-EDIT EXTENSION

Two pilot experiments and a registered extension precede Experiment 3. They fix what the factorial had to change, and no claim in the main text rests on them. `SUPPLEMENTARY_S5_EARLIER_EXPERIMENTS.md` prints their registered outcome tables, qualification tables and failure-mode analyses in full.

C.1 EXPERIMENT 1 (GPT-J-6B)

Five jobs ran on 2026-09-07 against one commit, registration, qualification artifact, item-set hash and screen hash; every row carries all five and the analysis refuses blocks that disagree. All four gradient editors completed the 255 screened facts, with a final edit loss below 0.05 on 255 of 255 facts for the three fine-tuning editors and 253 of 255 for LoRA. Generation efficacy on the edit prompt is 0.073 unedited against 0.473 / 0.439 / 0.438 / 0.450 for FT-L / FT-M / FT-multi / LoRA, and target-versus-distractor choice on the edit prompt is 0.702 unedited against 1.000 for every editor. Every registered outcome is a null: **0 of 5** comparisons on the primary format (family 6), 0 of 6 on each secondary format, 0 in every leave-one-editor-out, and no paraphrase-control rejection. The registered decision rule returns a null with MDE 5.40 pp (α = .05) and 6.71 pp after correction at n = 255.

Why that null carries little information (post hoc). The registered primary format is at ceiling: `Subject_Aliasing` × `constrained_choice` is 0.998, 1.000, 1.000 and 0.996 for the four editors, with a paired-difference standard deviation of 0.063 against the 0.308 the power analysis assumed, so no comparison could have rejected there whatever the other stratum showed. `yes_no` sits at chance for every weight editor (0.47–0.56). Only free generation is informative (0.25–0.51),

with post hoc MDEs of 4.2 and 3.6 pp, below KnowEdit's reversing effects of 4.4–6.9 pp; it returns 0 of 6, its largest opposite-signed pair of effects being -1.5 and -0.4 pp.

C.2 Experiment 2 (Mistral-7B-v0.1)

Experiment 2 was registered on 2026-09-13, before any Mistral output existed, as a **separate experiment** rather than an amendment. Three things change and everything else is inherited verbatim: the backbone is `mistralai/Mistral-7B-v0.1` at a pinned revision; the primary measurement is free generation, the only format in Experiment 1 neither at ceiling nor at chance; and the screen floor is 200 facts per semantic family. The screen retained **357 of 573 facts**, and all four editors qualified on 20 disjoint calibration facts with the sham at exactly zero. **LoRA is excluded (deviation D9)**: it diverged on the twelfth screened fact under the configuration that completed all of Experiment 1, its runtime convergence witness stopped the block, and a descriptive diagnostic puts the divergence rate at 9 of 357 (2.5%). The realised pool is three editors, three comparisons at Bonferroni 0.05/3, with an MDE of 3.59 pp at α = .05 and 4.14 pp after correction.

Every registered outcome is again a null (0 of 3 on the primary format and on the registered margin secondary, 0 of 3 at every other format), and this null is informative: the three editors differ by 3.4, 15.6 and 12.2 points on restatement and 2.8, 7.3 and 4.5 on the consequence probe, so the two probes disagree about the size of the advantage while agreeing on its **sign** everywhere. The components are magnitude heterogeneous, not a qualitative failure, at post hoc MDEs of 3.9 and 3.0 pp. Two observations carry into Experiment 3: the restatement stratum saturates again under constrained choice (0.9958, 0.9972, 1.0000), so that ceiling is a property of the screen and format rather than of GPT-J; and FT-L, whose objective and clamp both depart from the published configuration (A.4), never drives the target loss below 0.05 (median 0.203) yet has the highest generation efficacy of the three.

C.3 The locate-and-edit extension

The pilots' registered pools contain only gradient editors because the editing framework could not be imported in the cluster environment (deviation D3). Both registrations fix the terms of an extension in advance: ROME and MEMIT enter only if each passes the qualification of A.3 on the same calibration facts, they run regardless of what the gradient editors returned, and they are analysed in a **separate** family of ten comparisons. `extract_lae.py` lifts the published implementation files into a flat package and rewrites **only** import lines, printing for every file a diff that ignores import lines: **zero** non-import lines differ across all twenty files. MEMIT reads the published covariance statistics (12 files, 6.1 GB), too large to ship and pinned by a hash manifest; ROME does not read them on this backbone. Both reached edit-prompt efficacies of 0.798 and 0.848, about 1.7 times the gradient editors' 0.38–0.47 on the same facts, and both completed all 255 facts.

The extension gives **0 of 10** on the registered primary format and **1 of 10** at α = .05 on free generation, none after correction. The primary format is no longer saturated there: ROME scores 0.853 and MEMIT 0.747 on restatement under constrained choice, against 0.996–1.000 for every gradient editor, so the ceiling that stopped Experiment 1's primary is a property of the gradient editors.

Table 13: The one nominal reversal on free generation, with the two nearest comparisons. p is Gail and Simon's; the intersection-union p is twice it, and the count is the same under both. The registered rule for a single rejection below the corrected threshold is *suggestive, not established*. An independent second ROME run gives the same count, pair and signs, with magnitudes within 1.2 points (-9.19 and $+5.71$, $p = 0.018$).

pair	Subject_Aliasing	Reasoning	aggregate	p
LoRA vs ROME	−10.40 pp	+6.04 pp	−2.18 pp	0.0121
FT-M vs ROME	−13.79 pp	+2.44 pp	−5.68 pp	0.179
FT-L vs ROME	−15.14 pp	+2.09 pp	−6.53 pp	0.191

Pooling across commits (deviation D10). The extension's ROME and MEMIT blocks come from a later commit than Experiment 1's gradient blocks. FT-L re-run under the later commit reproduces its Experiment 1 block exactly (0 of 255 edits and 0 of 3,060 scores differ), so `--pool-across`

admits the two commits and refuses any difference in item set, screen, model, revision, dtype or library versions.

C.4 STOCHASTIC EDITORS AND THE RUN-LEVEL ERROR TERM

Which editors are stochastic. Two runs of one ROME configuration on the extension's 255 facts differ on the edit itself for 247 of 255 facts, on 2.8%, 12.4% and 14.1% of cells under constrained choice, yes/no and generation, and by at most **1.57 points** in a cell mean. Across two runs eight days apart the two deterministic gradient editors reproduce **bit for bit** (0 of 3,060 rows differ for each), while `LoRA` differs on 480 rows (15.7%) touching 154 of 255 facts. ROME and MEMIT average over context templates generated by the model, and LoRA's adapter initialisation uses torch's global generator; neither was seeded in the Experiment 3 runner, and the specificity run seeds the templates at EasyEdit's value (42) and agrees with Experiment 3 on 507 of 507 facts for both (F.1). `determinism.py` computes both tables.

The run-level component. Every interval in this paper is clustered on facts and computed from scores that already contain whatever an unseeded component did to each fact. What is not in them is the part **common to every fact within a run**: for ROME and MEMIT the template draw. Writing τ for that run-level standard deviation, a seed sweep over n facts sees $\mathrm{var}(\text{seed mean}) = \tau^2 + \sigma^2_{\text{within}}/n$, so τ is identified only when the between-seed variance exceeds the second term. Experiment 5's registered secondary S re-ran `ROME`, `MEMIT` and `LoRA` at three seeds on all 533 facts and does not identify it: the between-seed variance never exceeds the within term in any of the twelve arm-by-cell combinations, with a between-seed standard deviation of at most 0.90 points (J.7). An earlier five-seed ROME sweep on 300 facts identified τ in one of four cells at **0.79** points, $[0.16, 2.74]$, which is consistent with that. Because neither run resolves τ, it is priced by imposing it on every comparison involving LoRA, ROME or MEMIT:

Table 14: Corrected reversals under the cell-mean restriction with an imposed run-level component, Gail and Simon's test / the intersection-union test. Under Gail and Simon's test the measurement contrast loses its first comparison, LoRA against ROME, between $\tau = 3.0$ and 3.5 points, and the split empties between the same two values. Four of the eight split reversals involve MEMIT, so imposing the component on MEMIT is much of what empties the split.

τ	measurement (family 40)	answer-identity split (family 80)
0.0 pp	3 / 3	8 / 6
0.5 pp	3 / 3	7 / 5
0.7 pp	3 / 3	6 / 5
1.0 pp	3 / 3	5 / 5
2.0 pp	3 / 3	4 / 4
2.5 pp	3 / 3	4 / 2
3.0 pp	3 / 2	1 / 0
3.5 pp	2 / 2	0 / 0
5.0 pp	1 / 0	0 / 0
6.0 pp	0 / 0	0 / 0

Under the arm-level restriction, held to the full families of 40 and 80, no measurement comparison rejects at any τ under either test, and the one arm-level split reversal (same-answer LoRA against ROME at generation) is lost by $\tau = 0.5$ under the intersection-union test. Neither sweep therefore orders the two contrasts. `family_sensitivity.py` reports the same sweep at the smaller families the arm-level restriction keeps; because those families depend on the data, no claim uses them (B.3).

Interpretation. At $\tau \approx 0.8$ points, the earlier sweep's estimate, all three measurement reversals remain under either test, as do five or six of the eight split reversals under Gail and Simon's test (five of six under the intersection-union test), while no arm-level reversal does. The measurement reversal of LoRA against ROME breaks between 3.0 and 3.5 points, roughly three times the largest common shift either replicate pair shows (0.96 points). Reproducibility is a property of whether an editor has an unseeded stochastic component, not of its family: two of the three gradient editors are deterministic, and the largest run-level variation measured belongs to a gradient editor. Two earlier 30-fact seed sweeps (`seedvar_ROME.json`, `seedvar_LoRA.json`) ship in the artifacts; they

are ranges of 30-fact means rather than error terms for 533-fact contrasts, since under fact-level noise a 30-fact mean moves $\sqrt{533/30} \approx 4.2$ times as much, and they record no per-fact scores.

D EXPERIMENT 3: FULL RESULTS AND SENSITIVITIES

D.1 DESIGN AND POPULATIONS

Experiments 1 and 2 cannot compare the two contrasts: their answerability screen selects on one contrast and not the other (E.4), and the two contrasts are scored on different numbers of comparisons. Experiment 3 removes both by construction. `PREREG_SYMMETRIC_2026-09-14.md` was frozen, hashed and committed before any of its output existed; H.4 describes what that does and does not protect.

Five published editing methods are used, `FT-L`, `FT-M`, `LoRA`, `ROME` and `MEMIT`, the gradient editors as implemented in A.4. `FT-multi` is excluded so that no rejection rests on an unpublished editor. The screen runs and gates nothing, so all three populations derive from one run. The two contrasts are structurally symmetric: two strata each over the same facts and editors, so ten editor pairs times two cells gives **twenty comparisons per contrast**, fixed by the design. The family is the union, 40, at Bonferroni $0.05/40 = 0.00125$. The primary restriction scores a cell when its across-editor mean is at least 0.05 from ceiling and from chance (registration section 7); the arm-level version of that restriction, the sixteen-setting margin grid and a chance-level sweep are registered sensitivities. All five blocks completed 6,397 rows against one registration hash, item-set hash, screen hash, commit and `config_sha`.

Table 15: Experiment 3's registered populations. The 40 calibration and holdout facts on which editor hyperparameters were verified are removed from the analysed population (deviation D1 of Experiment 3), so P0 is 533 rather than 573.

population	rule	n	role
P0, unselected	edit completed for every editor	**533**	**primary**, fixed in the registration
P1, symmetric	all eight cells of both informative formats answerable in context	15	secondary; both contrasts selected identically
P2, asymmetric	the four constrained-choice cells answerable (Experiment 1's rule)	255	connects this run to Experiments 1 and 2

P1 is nearly empty. Requiring all eight cells answerable retains 15 of 533 facts, 2.8%. 530 of the 533 are answerable in at least one constrained-choice cell and only 354 in at least one free-generation cell; 240 facts fail on generation alone, 305 on both formats and 13 on constrained choice alone. The screen counts a free-generation cell answerable only at token F1 $= 1.0$, so P1's size partly reflects that threshold, and the shipped screen records booleans rather than scores, so a looser threshold cannot be recomputed. P1 is stamped `UNDERPOWERED` against its registered floor of 152 and returns the registered null.

D.2 THE REGISTERED RESULT

Table 16: Experiment 3's primary population, both contrasts, with all 40 comparisons computed and every row held to the family of 40. The intersection-union p is twice each p shown. The semantic column is the pooled `Reasoning` label (E.3 gives the split).

P0, 533 facts	measurement	semantic
comparisons	20	20
corrected reversals over 40, Gail and Simon / intersection-union	**3** / 3	**0** / 0
minimum p (Gail and Simon)	7.9×10^{-14}	0.090
opposite-signed	12 / 20	4 / 20
arm-level restriction: comparisons	11	11
arm-level restriction: corrected reversals over 40	**0**	**0**

P0, 533 facts	measurement	semantic
arm-level restriction: minimum p	0.0014	0.129
arm-level restriction: opposite-signed	7 / 11	3 / 11

The three reversals are all in the same cell and all pair a gradient editor with ROME:

Table 17: The three corrected measurement reversals. The between-stratum correlations of the per-fact differences are 0.36, 0.33 and 0.34. Under equal weights each composite is within 0.02 of a tie, and the bootstrap interval for the flip weight includes 0.5 in all three.

pair	cell	δ constrained choice	δ generation	p (Gail and Simon)	flip weight on choice [95% bootstrap]
FT-L vs ROME	`Subject_-Aliasing`	$+15.48$	-16.41	7.9×10^{-14}	0.515 [0.416, 0.603]
FT-M vs ROME	`Subject_-Aliasing`	$+16.04$	-16.96	8.9×10^{-14}	0.514 [0.417, 0.600]
LoRA vs ROME	`Subject_-Aliasing`	$+15.76$	-15.44	2.9×10^{-11}	0.495 [0.387, 0.589]

The same three pairs in the `Reasoning` cell give 0.0105, 0.0014 and 0.023, none below 0.00125. The best comparison involving no locate-and-edit editor is FT-M against LoRA at $p = 0.090$. On the pooled semantic contrast nothing is below 0.090, and sixteen of the twenty return Gail and Simon's maximum of 0.5, the complement of the four that are opposite-signed.

Magnitudes on other scales. For all six ROME comparisons the simultaneous interval on the interaction excludes zero on the additive, log-ratio and log-odds scales; in the `Subject_Aliasing` cell the additive point estimates are $+31.9$, $+33.0$ and $+31.2$ points and the log-odds points $+3.91$, $+5.32$ and $+4.34$, every interval clear of zero at $B = 20{,}000$. The sign of a difference between two **arm means** is invariant under monotone re-expression of the arm means. That invariance does not extend to per-item truncation: a common cap applied item by item can reverse the order of two editors whose per-item scores do not stochastically dominate each other (§2.3), which is why E.2 checks the untruncated margin per decile.

D.3 Held-out facts, and what the null excludes

A null is not an equivalence. The registration states in advance that a null "bounds the effect at roughly 10 points and does not establish that no smaller qualitative interaction exists". On P0 the minority-direction bound over every semantic pair leaves reversals of up to **6.11 points** unexcluded, the worst pair being FT-M against MEMIT at generation ($+5.24$ against -1.66). The pooled "0 of 20" means no reversal was certified, and E.3 shows eight once the label is split.

Held-out facts (post hoc). P2 is Experiment 1's 255 screened facts, on which FT-L's and FT-M's cell means reproduce the Experiment 1 blocks to four decimals. The remaining **278** facts of P0 were never in that data. They share no fact with the data from which the answer-identity split was proposed (E.4), so they are the stronger test of it. Analysed alone at the registered families:

Table 18: Experiment 3 on the facts outside Experiment 1's screen. The three ROME reversals replicate at 1.5×10^{-10}, 2.7×10^{-9} and 7.6×10^{-9} for FT-M, LoRA and FT-L against ROME. The fourth measurement reversal on all strata is LoRA against MEMIT in `Reasoning` ($+10.8$ under constrained choice, -10.4 in generation; $p = 1.4 \times 10^{-5}$), whose constrained-choice `Reasoning` cell has an across-editor mean within 0.05 of chance on these facts, so the registered restriction does not score it. The split reversals are the four true-consequence constrained-choice ones against ROME and MEMIT; FT-M against MEMIT ($p = 4.8 \times 10^{-4}$) does not survive the intersection-union threshold.

278 held-out facts	measurement (40)	semantic, pooled (40)	answer-identity split (80)
all strata, Gail and Simon / intersection-union	**4** / 4	0 / 0	**4** / 3
cell-mean restriction, Gail and Simon / intersection-union	3 / 3	0 / 0	4 / 3
arm-level restriction, Gail and Simon	none of 1 kept		none of the 13 kept

Under the arm-level restriction the held-out facts keep nothing. The restriction leaves **one** measurement comparison (MEMIT against ROME, $p = 0.5$) and **13** split comparisons, and **neither rejects** (split minimum $p = 0.147$). The single arm-level split reversal of E.3, same-answer LoRA against ROME at generation, therefore does not replicate out of sample, while the cell-mean counts do. Computed by `adversarial_reanalysis.py` and, at the full families, by `post_hoc_analyses.py`, whose self-tests reproduce the P0 numbers before reporting held-out ones.

D.4 The registered decision rule

Section 10 of the registration returns `CONFIRMED` when there are at least two measurement rejections on at least two distinct editor pairs, no semantic rejection, and the measurement rejections survive (a) both per-paraphrase analyses, (b) every leave-one-editor-out at the full family, and (c) all sixteen margin settings and the arm-level restriction. It returns `PARTIAL` when at least one measurement rejection exists, no semantic one does, and one or more of (a)–(c) removes a rejection, which must then be named. On P0, (a) holds and (b) and (c) do not, so the rule returns **partial support**.

(a) Per-paraphrase: holds. The measurement contrast rejects three times on each paraphrase alone, at minimum p of 2.7×10^{-15} and 3.5×10^{-14}, and the semantic contrast on neither. The per-fact correlation between the two paraphrases' editor-difference vectors runs from 0.228 to 0.886 across 39 cells, so 11% to 77% of the per-fact difference variance is paraphrase-specific.

(b) Leave-one-editor-out at the full family of 40: fails. Dropping `FT-L`, `FT-M` or `LoRA` leaves **two** of the three, since each rejecting pair contains one of them; **dropping `MEMIT` or `ROME` leaves none** (best adjusted p 0.055 and 1). At the smaller family each drop produces the counts read 2, 2, 2, 1 and 0, and the registration requires the full family for the reason B.3 gives. ROME appears in all three rejecting pairs, so its row is not independent evidence of fragility; the MEMIT row is explained under (c).

(c) Margins and the arm-level restriction: fails on both. The registered grid crosses `chance_margin` $\in \{0.02, 0.05, 0.075, 0.10\}$ with `ceiling_margin` $\in \{0.02, 0.05, 0.075, 0.10\}$. The pooled semantic contrast rejects zero times at all sixteen settings, and at `ceiling_margin` $= 0.10$ the measurement rejections fall from three to **zero** at every chance margin. That is one inequality: the decisive cell's across-editor mean, 0.9073, lies above $1 - 0.10$ and below $1 - 0.075$, and no chance margin binds. The arm-level restriction drops 18 of the 40 comparisons, leaves 11 per contrast, and removes every rejection on both contrasts (measurement minimum $p = 0.0014$). The reason is a ceiling. In `constrained_choice` × `Subject_Aliasing`:

In that cell the arms are FT-L 0.9925, FT-M 0.9981, LoRA 0.9953, MEMIT 0.7129 and ROME 0.8377, an across-editor mean of 0.9073.

The three gradient editors are saturated, and the cell is admitted by the cell-mean restriction only because the two locate-and-edit editors hold the mean below 0.95. Without MEMIT the mean rises to 0.9559, the cell is no longer scored, and the effect goes with it, which is why dropping MEMIT removes the rejections. The real sensitivity is therefore cell mean against arm level, not the margin. That arm-level observation was made on the locate-and-edit extension (E.4) and committed on the

day `PREREG_SYMMETRIC_2026-09-14.md` was first committed, several hours before it; the registration named the arm-level restriction as a sensitivity rather than as the primary rule (H.4). E.2 checks the ceiling directly, on the untruncated margin.

D.5 Rank events and a max-T check

Pairwise signs among k editors are a deterministic function of the stratum's ordering, so a count of reversed pairs is a rank distance. In `Subject_Aliasing`, the cell carrying the three rejections, the ordering is (FT-M, LoRA, FT-L, ROME, MEMIT) under constrained choice and (ROME, LoRA, FT-L, FT-M, MEMIT) under generation: ROME and FT-M exchange places and the other three do not move, so 5 of 10 pairs reverse, and the three corrected reversals are that single exchange seen three times. In `Reasoning`, where nothing rejects after correction, ROME moves from fourth to first and MEMIT from fifth to second, and 7 of 10 pairs reverse. The registration also names an exact permutation null over the $5! = 120$ orderings, which gives $p = 0.59$ and $p = 0.24$. Under uniformly random orderings the expected number of reversed pairs is 5 of 10 and a rejection at .05 needs 9, so that null asks whether two orderings are anti-associated; it carries no information about whether these reversals are many or few, and nothing in this paper rests on it. A fact-clustered bootstrap max-T over the twenty measurement comparisons ($B = 20{,}000$) gives a simultaneous critical value of **2.922**. The like-for-like Bonferroni value over those twenty is t_{532} at $1 - 0.05/40 = 3.038$; 3.245 is the value for a family of 40 and is not what this check compares against. The max-T is on interaction statistics, not on the reversal test, so neither value is the threshold any verdict uses. 13 of the 20 exceed 2.922; of those 13, nine have stratum effects of disagreeing sign and the reversal test certifies **three**. Each reversal is large; their number is not independent evidence.

Under the arm-level restriction. The arm-level restriction keeps eleven measurement comparisons, and none rejects at the registered family of 40: the smallest, FT-M against ROME in `Reasoning`, has $p = 0.0014$ against 0.00125. What limits the count is the requirement of significant evidence in each direction rather than the correction: of the eleven, seven have interactions beyond the max-T critical value (2.922), and six of those seven have strata of opposite raw sign, but in none is the minority-direction effect itself established.

D.6 The screened population P2

Under Experiment 1's answerability rule the identical analysis gives the same verdict, the same three rejecting pairs and the same removals, so **that screen did not manufacture the measurement effect**. Under the arm-level restriction P2's opposite-signed counts are 3 of 11 on both contrasts, and its semantic minimum p, 0.019, is smaller than its measurement minimum, 0.030, so there is no asymmetry between the contrasts on the screened population either. P2's answer-identity split gives **7** corrected reversals under Gail and Simon's test (6 under the intersection-union test), and two under the arm-level restriction at the full family (one under the intersection-union test), both same-answer reversals against ROME at generation. On both populations the four margin settings at a ceiling margin of 0.10 change the conclusion.

D.7 Registered effect sizes and the chance-level sweep

Section 8: the interaction effect size for every editor pair and qualifying cell of both contrasts. On P0, **11 of 20** semantic interactions are established at the fact unit, the largest 19.29 points (MEMIT against ROME at generation), seven at constrained choice (largest $+17.82$, FT-L against MEMIT) and four at generation. Every established one has both strata of the same sign, so on the pooled label these are magnitude heterogeneity rather than qualitative failure; sign reversals appear only once the label is split (E.3).

Section 9 item 5: the chance-level sweep. The measured permutation floor for token F1 is **0.041** on reasoning and **0.016** on aliasing, and the unedited model's free-generation efficacy is 0.07. Swept over $\{0, 0.041, 0.07, 0.13\}$, the measurement contrast rejects three times at every level and the semantic contrast none; at 0.13 the family falls from 40 to 20 because one cell is disqualified, and the minimum p does not move.

D.8 What Experiment 3 supports

On an unselected population of 533 facts, with the same facts, editors, model and number of comparisons on both contrasts, the registered analysis finds three corrected measurement reversals and none on the pooled semantic label, and eight once the label is split by the registered pre-edit property (six under the intersection-union test). The measurement reversals are $+15.5$ to $+16.0$ points under constrained choice against -15.4 to -17.0 under token-F1 generation. They replicate on the held-out facts and at every seed (J.7), and Experiment 1's screen does not manufacture them, so they are real on the registered scores.

What they are is the subject of Section 5 and Appendices E and J. Every one is one editor against one homogeneous group in a cell whose other arms are saturated; all three are a single exchange in one ordering; their composites are near-ties; none survives the arm-level restriction or the loosest registered ceiling margin; none survives scoring generation by first-span match (E.1); and none survives null-prompt calibration of the constrained-choice side (J.4). The data therefore do not show that measurement choice reproducibly reverses editors in general, and they do not show an asymmetry between the two contrasts: the pooled semantic null becomes eight reversals when the label is split, most of them between arms at ceiling and at chance. ROME's and LoRA's unseeded components are a further limit on the ROME comparisons, priced as breaking points in C.4 and as seed replication in J.7.

E Failure-mode analyses

The answer-identity split of E.3 is registered (Experiment 3's deviation D2); everything else in this section is post hoc and uses the registered experiments' stored outputs. E.1 and E.2 are computed by `post_hoc_analyses.py` (37 checks); the other analyses use `interaction_analysis.py`, whose self-tests include a planted interaction it must recover, a planted constant effect it must not, and a planted probe-type reversal whose exact paired p must come out at $2/2^6$.

E.1 Generation scored three ways

Why. The registered generation score is token F1 on a trimmed 12-token continuation (A.2). Trimming cuts at punctuation and newlines, but not at a repeated word, so a continuation such as "Canada Canada Canada" keeps its repeats and scores well below 1 even when its first word is the answer. The repetition is **edit-induced**, which needs the unedited baseline to say: on the same prompts the unedited model repeats on **0.000** of its restatement and consequence continuations, at mean lengths of 6.67 and 7.69 words, so it writes continuations as long as the edited arms' without repeating.

What the continuations look like. Every stored Experiment 3 generation row (10,660) was rescored from its text; the registered token F1 is reproduced exactly on every row before anything else is computed.

Table 19: Mean trimmed continuation length in words and the share of continuations in which one word makes up at least half of a continuation of four or more words, on P0 (1,066 rows per editor and probe).

editor	words, restatement	words, consequence	repeated, restatement	repeated, consequence
FT-L	7.1	7.3	30%	30%
FT-M	7.7	7.7	59%	59%
LoRA	7.3	7.4	55%	56%
MEMIT	5.0	5.5	0%	0.1%
ROME	4.0	4.8	0%	0.2%

Two post hoc scores. *First-span match* scores 1 when the normalised continuation is an accepted alias or begins with one followed by a word boundary. *First-span F1* is token F1 between each alias and the continuation's first k words, k the alias's length, maximised over aliases. Both use the registered normalisation. They were chosen after a first look at the stored continuations, so they are sensitivity analyses of the registered score, not replacements for it.

Table 20: Generation arm means on P0 under the three scores. The unedited levels are from Experiment 5's unedited block on the same facts.

editor	token F1, restatement	first-span match	first-span F1	token F1, consequence	first-span match	first-span F1
FT-L	.425	.608	.690	.146	.160	.224
FT-M	.419	.744	.853	.142	.234	.296
LoRA	.435	.757	.859	.157	.265	.323
MEMIT	.367	.300	.367	.159	.086	.176
ROME	.589	.529	.588	.188	.124	.196
unedited	.085	.025	.088	.126	.028	.162

Under token F1 ROME leads every gradient editor on restatement by 15 to 17 points. Under either first-span score the order reverses: FT-M and LoRA lead ROME by 21 to 27 points and FT-L by 8 to 10.

Neither first-span score is a neutral referee: both credit a continuation that begins with an alias and repeats it, and both give zero to a correct answer behind a preamble, which is the failure mode they could create rather than remove. A third score answers that objection directly, by asking only whether an accepted alias appears **anywhere** in the trimmed continuation. It orders the arms as the first-span scores do: on restatement FT-L 0.687, FT-M 0.753 and LoRA 0.787 against ROME 0.555 and MEMIT 0.315, and on the consequence probe 0.204, 0.244 and 0.281 against 0.149 and 0.114. Under it no measurement comparison reverses either (0 of 20 on P0 and on the held-out facts; 12 split reversals, 10 under the intersection-union test).

What the four scores establish is therefore **not** that repetition accounts for the registered deficit, which no score that ignores everything after the answer can show, but that **the generation ranking of these arms is not identified by the scoring rule**: token F1 places the gradient editors 15 to 17 points below ROME, and the three alternatives place them 8 to 27 points above it. A score that credited the answer and penalised the repetition separately, or the benchmark's own teacher-forced accuracy, is not reported here.

Table 21: Corrected reversals under Gail and Simon's test with generation rescored and constrained choice unchanged. Under the intersection-union test the counts are 3/4, 0/0 and 6/3 (token F1) and 0/0, 0/0 and 10/9 under each of the other three scores. The one pooled semantic reversal under first-span F1 on the held-out facts (FT-L against MEMIT, $p = 9.0 \times 10^{-4}$) does not survive the intersection-union threshold.

scoring of generation	measurement (40)	semantic, pooled (40)	answer-identity split (80)
token F1, registered: P0 / held-out	3 / 4	0 / 0	8 / 4
first-span match: P0 / held-out	**0 / 0**	0 / 0	12 / 10
first-span F1: P0 / held-out	**0 / 0**	0 / 1	12 / 10
alias containment: P0 / held-out	**0 / 0**	0 / 0	12 / 10

No measurement reversal survives either first-span score, on P0 or on the held-out facts. The three registered reversals become FT-L, FT-M and LoRA ahead of ROME in both formats ($+15.5$, $+16.0$ and $+15.8$ under constrained choice against $+7.9$, $+21.5$ and $+22.8$ under first-span match). The answer-identity split gains reversals: under first-span scoring the true-consequence reversals appear in generation as well as in constrained choice, because on those facts the gradient editors answer 0.8 to 1.1% of consequence probes by first-span match against 6.9% for ROME and 7.3% for MEMIT, while leading on restatement (E.3).

On Experiment 5's replication blocks the same rescoring gives the same result for the three primary reversals and does not remove two others: the efficacy-matched FT-M arm against ROME ($+8.2$ under constrained choice against -24.0 under first-span match, $p = 2.1 \times 10^{-9}$) and ROME against ROME re-keyed on the last token (-9.3 against $+23.4$, $p = 3.9 \times 10^{-13}$). J.5 reports both under every combination of scores.

E.2 THE CONSTRAINED-CHOICE MARGIN, AND DOMINANCE PER DECILE

Why. A score capped at 1 compresses differences near the cap, and a common cap applied item by item can reverse the order of two editors whose per-item scores do not stochastically dominate each other. The registered M2 check, a margin of 0.05 on arm means, does not see this. Experiment 2 registered the untruncated margin as a secondary; here it is computed for Experiment 3.

The margin. For each constrained-choice item the margin is the gold option's length-normalised log-likelihood minus the best distractor's, averaged over paraphrases to the fact. Its sign agrees with the stored accuracy on every item. Replacing constrained-choice accuracy by the margin (post hoc):

Table 22: Corrected reversals on P0, Gail and Simon's test / the intersection-union test, generation unchanged.

constrained choice scored by	measurement (40)	semantic, pooled (40)	answer-identity split (80)
accuracy (registered)	3 / 3	0 / 0	8 / 6
margin	3 / 3	0 / 0	10 / 10

The three measurement reversals are the same three. The split's two additions are gradient editors against each other in the true-consequence subset: LoRA's restatement margin exceeds FT-L's and FT-M's by 17.5 and 10.8 nats while its consequence margin is lower by 4.3 and 3.8, all three editors being at 0.996 or above in restatement accuracy.

Dominance per decile. Fixed before it was run: for each editor pair and semantic type, the nine decile differences of the two editors' per-fact margin distributions, each with a fact-clustered bootstrap interval at $0.05/9$ two-sided ($B = 2{,}000$); an editor dominates when it is ahead at every decile with every interval excluding zero.

- **Restatement.** Every one of the ten pairs has a dominating editor. Each gradient editor dominates ROME and MEMIT, LoRA dominates FT-M and FT-L, FT-M dominates FT-L, and ROME dominates MEMIT. Across the deciles the gradient editors' margins exceed ROME's by 5.8 to 44.7 nats. Their constrained-choice advantage on restatement is therefore not an artefact of the cap: it holds for every part of the per-fact distribution.
- **Consequence.** No pair has a dominating editor. The gradient editors' margins are wider in both directions (LoRA against ROME runs from -18.8 nats at the first decile to $+34.1$ at the ninth), so their accuracies near 0.5 average a mixture of confident right and confident wrong answers.

The constrained-choice side of the measurement reversals is therefore real per item. Whether it measures knowledge of the edited fact is a separate question, which calibration against a null prompt addresses (J.4).

E.3 M1: THE REGISTERED ANSWER-IDENTITY SPLIT

What is registered. Deviation D2 of Experiment 3 registers a sub-stratification of the semantic contrast by answer identity, with the family extended and the same statistic, restriction and sensitivities applied within each subset. A fact is `same_answer` when the gold set of its Reasoning probe contains `target_new` after case folding: a pre-edit property of the item, computable with no model and no outcome, reproducing D2's stated counts of **181 of 573** facts and **178 of the 533** of P0. The split was computed after the primary verdict (deviation D5) and fixes no free parameter, and its hypothesis came from data sharing 255 facts with P0 (E.4), so the 278 held-out facts (D.3) carry more confirmatory weight than P0 does.

Over the extended family of **80** (the registered 40 plus 20 in each subset), at 6.25×10^{-4}:

Table 23: Splitting the benchmark's own label by a pre-edit property takes the semantic contrast from zero corrected reversals to six under the intersection-union test, and eight under the registered Gail and Simon test.

P0, semantic contrast	comparisons in family	computed	corrected reversals, Gail and Simon / intersection-union	minimum p
pooled `Reasoning`	20	20	**0** / 0	0.090
true consequence ($n = 355$)	20	19	**7** / 5	4.8×10^{-10}
same answer ($n = 178$)	20	20	**1** / 1	2.7×10^{-4}

Table 24: The eight split reversals under Gail and Simon's test; the last two do not survive the intersection-union threshold (p of 6.6×10^{-4} and 1.0×10^{-3}). "Same cell, other subset" gives the pair's restatement / consequence effects in the same format on the other subset. The last column, z, tests whether the within-fact interaction differs between the two subsets (post hoc); every $|z| \geq 5.9$.

subset	pair	cell	δ restatement	δ consequence	p	same cell, other subset	z
true consequence	FT-L vs ROME	constrained choice	$+15.35$	-11.69	4.8×10^{-10}	$+15.73$ / $+36.80$	10.5
true consequence	FT-L vs MEMIT	constrained choice	$+27.18$	-11.41	1.3×10^{-9}	$+29.49$ / $+53.09$	12.7
true consequence	FT-M vs MEMIT	constrained choice	$+27.32$	-10.42	1.1×10^{-8}	$+30.90$ / $+55.62$	12.9
true consequence	FT-M vs ROME	constrained choice	$+15.49$	-10.70	1.3×10^{-8}	$+17.13$ / $+39.33$	10.7
true consequence	FT-L vs MEMIT	generation	$+10.22$	-10.26	7.3×10^{-5}	-3.07 / $+16.45$	9.7
same answer	LoRA vs ROME	generation	-18.88	$+12.32$	2.7×10^{-4}	-13.71 / -10.76	-5.9
true consequence	LoRA vs ROME	constrained choice	$+15.49$	-7.32	3.3×10^{-4}	$+16.29$ / $+39.33$	9.8
true consequence	LoRA vs MEMIT	constrained choice	$+27.32$	-7.04	5.1×10^{-4}	$+30.06$ / $+55.62$	12.0

What pooling does. In the true-consequence subset the gradient editors beat the locate-and-edit ones on restatement and lose to them on the consequence probe. In the same cells of the same-answer subset, under constrained choice, they win on both, and by more on the consequence probe, because there its answer is the target. The subsets therefore do not carry reversals of opposite sign that cancel; pooling the target-as-gold items with the true consequences lifts the pair's consequence effect above zero and hides the true-consequence reversal. The one generation reversal in the true-consequence subset does have opposite signs in the same-answer subset (-3.07 / $+16.45$), and the one same-answer reversal has same-signed effects in the true-consequence subset. The subset $\times$ stratum test rejects for all eight on P0, and for the four constrained-choice reversals that replicate on the held-out facts ($|z|$ from 5.2 to 6.5, with 34 same-answer facts).

Table 25: D2's registered sensitivities, run within each subset: Gail and Simon's test / the intersection-union test where they differ. Families are 40 for the measurement contrast and 80 for the split, except in the margin grid (B.3).

registered sensitivity	measurement contrast	answer-identity split
cell-mean restriction	3 / 3	**8** / 6
arm-level restriction, at the full family	0 of 11 kept	**1 of 13 kept** / 1
leave one editor out at the full family: drop ROME / drop MEMIT	0 / 0	**1 / 1**

registered sensitivity	measurement contrast	answer-identity split
leave one editor out: drop FT-L / FT-M / LoRA	2 / 2 / 2	5 / 6 / 5 (3 / 4 / 5)
per paraphrase: 0 / 1	3 / 3	7 / **2**
the 16-setting margin grid	3 at 12 settings; 0 at the other 4	8 at **6** settings; as low as **1** (6 at 6; as low as 1)
"contains" read as substring	n/a	**3** / 2

Of the 13 comparisons the arm-level restriction keeps in the split, one rejects at the family of 80 under both tests: same-answer LoRA against ROME at generation, $p = 2.7 \times 10^{-4}$ against 6.25×10^{-4}. Under the intersection-union test with the t tails the audit tool uses its p is 6.7×10^{-4}, which would not reject. It does not replicate on the held-out facts (D.3). The per-comparison listing of the other twelve is in `SUPPLEMENTARY_S2_MECHANISMS.md`.

Ceiling, chance and floor. In the true-consequence subset the stratum that drives the six constrained-choice reversals is `Reasoning` under constrained choice, whose across-editor mean is **0.3062** against a chance level of 0.25, admitted by **0.006**, with gradient arms at 0.249, 0.259 and 0.293 (unedited model on the same facts: 0.230; J.6). Their other stratum is `Subject_Aliasing` under constrained choice at 0.996 to 0.997. Six of the eight therefore pair an arm at ceiling with an arm at chance. In the seventh, FT-L against MEMIT at generation, FT-L's consequence score is at the floor (token F1 0.045). The eighth is the same-answer reversal that does not replicate. Those six also take one stratum from the cell that carries every measurement reversal (FT-L − ROME is $+15.35$ points there in the true-consequence subset against $+15.48$ on all of P0), so the two contrasts are not measured on disjoint material. The direction is what Zhang et al. (2025) call editing overfit: editors that raise the target broadly answer restatements and fail consequences.

Relation to the verdict. Section 10 defines the semantic rejection count on the pooled contrast, which D2 leaves unchanged, so the rule still returns partial support; beside it, the asymmetry between the contrasts is a property of the pooled label, and most of what the split reveals involves arms at ceiling, chance or floor. Both readings ship in the artifact. P1, at 15 facts, gives 0 and is underpowered.

E.4 The earlier pools: label pooling, ceiling arms, format mixing and the screen

The same checks were run on Experiments 1 and 2 and the locate-and-edit extension. `SUPPLEMENTARY_S5_EARLIER_EXPERIMENTS.md` prints them in full; what they establish is:

- **M1.** On 181 of the 573 facts the Reasoning probe's gold set contains the edit target itself, 32% of the item set, 56% of the 255 facts Experiment 1's screen retained against 12% of the 318 it discarded. Splitting the earlier pools by that property and correcting over all 38 split comparisons, the gradient pools reject nothing and **two** same-answer reversals against ROME survive on the extension (-16.8 / $+12.0$, $p = 3.3 \times 10^{-4}$; -14.1 / $+14.9$, $p = 5.5 \times 10^{-4}$), where the pooled label gives 0.19 and 0.18 and rejects nothing. A plausible mechanism is that ROME keys on the subject's last token, which a relational description of the subject does not contain. Experiment 3's registration added this split (D2) before any of its contrasts were computed.
- **M2.** Every format pair in these pools contains an arm at ceiling or chance: restatement under constrained choice sits at 0.996–1.000 and no yes/no cell departs from chance by more than 3.6 points. The cell-mean restriction tests an across-editor mean while its rationale concerns arms: on the extension's decisive cell three of five arms are above 0.95 and the cell qualifies only because ROME (0.8529) and MEMIT (0.7471) pull the mean to 0.9188. Applied arm by arm the restriction keeps 22 of the extension's 40 comparisons, where **all three rejections disappear** at the family of 40 (minimum p 0.084, against 3.5×10^{-7} unrestricted), and the one comparison in that cell with neither arm truncated returns Gail and Simon's maximum. This was visible before Experiment 3 was registered (D.4, H.4). At a ceiling margin of 0.10 the extension rejects nothing; per paraphrase the gradient pool's one corrected rejection survives neither (adjusted 0.034 and 0.074); and the semantic contrast rejects zero times after correction in every sweep.

- **Format mixing.** Taking the same semantic probe through two formats leaves 28 of 52 comparisons opposite-signed against 33.1% expected under a constant effect, while the registered semantic contrast leaves 5 of 27 against 28.8% expected (Fisher exact $p = 3.6 \times 10^{-3}$, descriptive because comparisons share facts and editor pairs). A 53.8% rate is roughly what arms at ceiling and chance alone predict. The reversal test rejects the probe-type contrast once over the gradient pools' 18 comparisons (Mistral's FT-M against FT-multi, $p = 9.6 \times 10^{-4}$, adjusted 0.017; both editors saturate the leakage measure, F.5) and three times over the extension's 40, at a smallest adjusted p of 1.4×10^{-5}.
- **The screen is not neutral between contrasts.** It keeps a fact only if all four (semantic type $\times$ paraphrase) cells are answerable in context (255 of 573 on GPT-J), which selects the semantic contrast on facts whose two semantic probes behave alike while never filtering the probe-type contrast on the two formats behaving alike; the per-fact stratum correlation, $+0.55$ against -0.01, is the signature, and keeping facts answerable under constrained choice also widens the gap between formats. The effect cannot be bounded from these data, so no between-contrast claim rests on them. Experiment 3's screen gates nothing, so all 533 facts are edited and scored (D.6).

E.5 M3: ITEM SETS

Random halves. Scoring the two strata on disjoint halves of the facts reproduces the structure of KnowEdit's `zsre` cell. Across **400** random splits in each experiment, on every editor pair of each pool, six for Experiment 1 and three for Experiment 2:

Table 26: Random unpairing on Experiments 1 and 2. Unpairing halves the effective sample, so each stratum's estimate is noisier and the sign of an editor difference approaches a coin flip when effects are small.

experiment	corrected reversals	opposite-signed rate, paired	opposite-signed rate, unpaired
1 (GPT-J)	0 of 2,400 pair-splits	0.0%	**30.9%**
2 (Mistral)	0 of 1,200 pair-splits	0.0%	0.1%

Systematic halves. `adversarial_reanalysis.py` (31 checks) repeats the construction on Experiment 3's population with the halves chosen systematically: whole relation groups on either side, or a split at the median of the facts' **pre-edit answerability**, a property of the unedited model measured before any edit. Facts, editors, model, format and scoring rule are unchanged; only the pairing changes.

Table 27: 400 splits per row, ten editor pairs each, Bonferroni over the pairs.

split of the facts, Experiment 3, P0, free generation	opposite-signed	corrected reversals, Gail and Simon / intersection-union
random halves	30.7%	**1** / 1 of 4,000 pair-splits
whole relation groups	32.2%	**5** / 0 of 4,000
median split on pre-edit answerability, restatement on the harder half	**50.0%**	**394** / 266 of 4,000
the same, with the halves exchanged	53.3%	**0** / 0 of 4,000

How the 400 median-split replicates differ. They are one construction, not 400 samples. The median answerability is 4; 113 facts lie below it, 250 above and 170 at it, and one seeded generator (`UNPAIR_SEED = 20260913`) shuffles only the 170 tied facts to balance the halves. The harder half is scored on `Subject_Aliasing` and the easier half on `Reasoning`. Of the 394 corrected reversals under Gail and Simon's test, **375 are FT-M against MEMIT** and 19 LoRA against MEMIT (265 and 1 of the 266 under the intersection-union test); no other pair reverses in any replicate. The comparison with 1 in 4,000 random halves is therefore not like for like: it shows that one editor pair's difference depends on pre-edit answerability strongly enough that giving the two probe types systematically different items reverses it almost whatever the tied facts do.

Reading. Exchangeable halves behave as an independent-samples design should: raw sign disagreement rises, but almost nothing is certified. When the two item sets differ systematically in the direction

that favours a reversal, unpairing **can induce certified reversals**; in the other direction it induces none. This is an existence proof, not an estimate of how often item-set differences produce reversals. Neither result shows that KnowEdit's `zsre` rejections are artifacts; both say that a cell scored on unpaired items cannot distinguish a probe effect from an item-composition effect without knowing how the items were drawn. Two closed forms apply, because there are two estimands. `unpairing()` splits a **fixed** fact set into complementary halves, for which the finite-population correction cancels the $\sqrt{2}$ exactly and the two half-means correlate at exactly $-\rho$; this conditional form predicts **30.1%** and **0.4%** against the simulated 30.9% and 0.1%. Probe types scored on different items drawn from a larger pool, as on KnowEdit, have each standard error inflated by $\sqrt{2}$ with independent strata, which gives 27.0% and 2.0%. Both are derived in `unpairing_closed_form()` and written out in `SUPPLEMENTARY_S2_MECHANISMS.md`.

E.6 Effect sizes and reversal bounds

At two strata the reversal test cannot reject when the signs agree, whatever the magnitudes, so magnitudes are reported as the fact-unit interaction $I(f) = \delta_{\text{alias}}(f) - \delta_{\text{reasoning}}(f)$ with paired t intervals simultaneous at $\alpha/(2\,\text{family})$, beside the minority-direction bound of B.4.

On Experiments 1 and 2 (free generation) six of the nine pairs show no detected heterogeneity and two are established, both on Mistral and both sharing an arm ($+8.30$ pp $[+4.57, +12.03]$ and $+7.67$ pp $[+4.66, +10.68]$). The design excludes semantic modulations larger than **5.3 pp** on Experiment 1's six null pairs and **3.9 pp** on Experiment 2's one, and the minority-direction bound over every pair is 3.83 pp on Experiment 1, 0.04 pp on Experiment 2 and **14.23 pp** on the five-editor extension, which excludes nothing, consistent with E.4 finding corrected semantic reversals there. On Experiment 3's P0 under the arm-level restriction the same construction gives 9.20 points in `Reasoning` for the measurement contrast and nothing unexcluded in `Subject_Aliasing`; for the pooled semantic contrast, nothing unexcluded at constrained choice and 6.11 points at generation; and for the split, **23.23** points in the same-answer subset at generation, where its one arm-level reversal sits.

Table 28: The semantic interaction on the five-editor extension, free generation. The 95% column is the paired t interval, the two multiplicative columns fact-clustered bootstrap intervals simultaneous across the family, and the state comes from the simultaneous t interval. Four of ten are established, every one against ROME, three on all three scales; the largest bound over the six pairs with no detected heterogeneity is 10.5 points, and 18.1 points at constrained choice on the same pool.

pair	I (pp)	95% CI	log ratio	log odds	state
FT-L vs ROME	-17.22	$[-23.5, -10.9]$	-0.383	-0.724	magnitude
			$[-0.67, -0.11]$	$[-1.20, -0.32]$	het.
LoRA vs	-16.44	$[-23.2, -9.7]$	-0.426	-0.733	magnitude
ROME			$[-0.73, -0.13]$	$[-1.21, -0.28]$	het.
FT-M vs ROME	-16.22	$[-22.6, -9.8]$	-0.367	-0.688	magnitude
			$[-0.66, -0.11]$	$[-1.11, -0.28]$	het.
MEMIT vs	-15.83	$[-21.2, -10.5]$	-0.218	-0.561	magnitude
ROME			$[-0.49, +0.03]$	$[-0.96, -0.21]$	het.
the other six	-1.39 to $+0.22$	all contain zero	none	none	no det. het.

Stimuli, templates and editors are fixed effects (Clark, 1973; Judd et al., 2012). The measurement contrast has one template per format, so its effect is an editor-by-template interaction estimated from one realisation per level, and fact-clustered standard errors capture fact sampling, not template sampling. The paraphrase factor has two mechanical templates, and 36% of the per-fact variance in an editor difference is paraphrase-specific. A model with template as a random factor was not fitted: with two levels its variance component is not estimable in any useful sense, and replicated templates were not run (§8). Statements about "the measurement procedure" therefore concern **the two renderings and scoring rules used**, and statements about "the gradient editors" concern **the implementations of A.4**.

Clustering on the relation rather than the fact. Every test in this paper clusters on the fact, which is the design's unit, and facts are not independent of one another: 573 facts carry only 92 distinct relations, the largest group holding 77, and the constrained-choice distractors are drawn from

other facts' answers for the same relation, so another fact's gold fills 69.5% of distractor slots (E.7). `relation_clustered.py` (6 checks) recomputes every registered comparison with the cluster set widened to the relation, changing nothing else; its self-test pins that one fact per cluster reproduces the fact-clustered standard error the paper already reports, and that a planted within-cluster correlation inflates it while residuals that cancel inside a cluster do not.

The wider clustering costs precision, by a factor of 0.79 to 2.40 on a stratum's standard error (median 1.15), and the two families answer differently:

Table 29: Corrected reversals under the two cluster sets. Both use Gail and Simon's test on the same stored blocks.

family	threshold	fact-clustered	relation-clustered
measurement (40)	$0.05/40$	3	**3**
answer-identity split (80)	$0.05/80$	8	**2**

All three measurement reversals survive, at $p = 1.4 \times 10^{-5}$, 4.6×10^{-5} and 1.2×10^{-4} against 0.00125, where their fact-clustered values were 10^{-14} to 10^{-11}. **Six of the eight split reversals do not**: the two that survive are FT-M against MEMIT and against ROME on true-consequence facts under constrained choice (2.4×10^{-6} and 1.3×10^{-5}), and the six that fall include both LoRA constrained-choice reversals, which the intersection-union test also declines, and the same-answer generation reversal, which does not replicate out of sample either. This is the same asymmetry the run-level component shows (C.4): the measurement result is robust to how the dependence is modelled and the split's *count* is not. What the split establishes is that pooling the `Reasoning` label hides a disagreement, which two reversals at these thresholds suffice for; what it does not establish is how many such disagreements there are.

E.7 What the controlled design does not cover

- **Specificity** is measured in the registered test of F on GPT-J and Experiment 3's pool, and on two further backbones: Mistral-7B-v0.1, where both gradient editors saturate the measure (F.5), and Qwen2.5-7B-Instruct, where it separates the families as on GPT-J (K). It is measured for single edits only. Batch and sequential editing are outside scope, and Pohl et al. (2025) show that batch size reorders editors.
- **Constrained-choice distractors are sometimes type-mismatched, and they are other facts' answers.** Where a relation stem is too rare the pool falls back to the wider type pool (228 of 1,146 draws, 19.9%), which can offer a song against a person, or "England" as an eye colour; a model can then answer by type, which inflates constrained-choice accuracy and contributes to the restatement ceiling. Over the 2,292 constrained-choice items a distractor is another fact's gold in 4,782 option slots (69.5%) and another fact's **edit target** in 4,062 slots (59.1%, over 1,692 items). A fact's **own** target appears among its own distractors in **2 slots of one item**, because `build_items` excludes the item's gold set rather than the target string.
- **The generation score is not the benchmark's, and it is not the only reasonable one.** Token F1 on a trimmed continuation is neither KnowEdit's teacher-forced token accuracy nor a first-answer score, and E.1 shows the measurement reversals depend on the choice. Our free-generation cell is closer to Yang et al. (2025)'s non-teacher-forced protocol than to KnowEdit's.
- **The multiplicity family does not cover every analytic choice.** It counts the comparisons made under one restriction and one scoring, not the choice of restriction, margin, chance level, format pair, generation score, scale or population, which are swept and reported.
- **Editors and templates are fixed effects** (E.6), so nothing generalises to an editor family or a probe class as a population.

F SPECIFICITY AND LEAKAGE

F.1 REGISTRATION, POPULATION AND MEASURE

The alternative reading tested. Experiment 3's three measurement rejections admit a reading the controlled design cannot separate from a claim about measurement: the gradient editors raise the target string's probability broadly, which saturates a likelihood-ranked multiple choice whose gold option *is* the target, while ROME's subject-keyed edit wins on free generation. On that reading the reversal reflects the editors' specificity.

Registration. `PREREG_LOCALITY_2026-09-15.md` was frozen after Experiment 3's outcome was known and before any locality score existed. Its first design scored whether the true answer to a same-subject prompt survives, by free-generation token F1, which cannot see a likelihood shift toward the target that leaves greedy output unchanged. Amendments A1 to A10 replaced it with the leakage measure below while the five blocks were queued and before any locality row had been written; the registered within-fact correlation was kept as a secondary (A1). The amendment chain, each entry with its reason and UTC time, is in `SUPPLEMENTARY_S4_LOCALITY.md`.

The prompts. Both families come from KnowEdit's own `locality.Relation_Specificity` field of the `recent` and `counterfact` records, verbatim. **Locality** prompts are a fact's own entries: same subject, a different relation; every entry is used. **Neighbourhood** prompts are drawn one per locality prompt, with a seeded generator (seed 20260915), from other facts' locality prompts with the same relation, so they ask a different subject the locality prompt's relation, which is not necessarily the edited relation; they are excluded by fact, not by subject name. A prompt is dropped when it is empty, when its gold answer is the edit target, when it duplicates another case-insensitively, or when fewer than two same-relation distractors exist.

The population (A8, A10). Of the **573** facts in the frozen item set, the **533** of P0 are handed to the locality builder; **517** carry a `Relation_Specificity` prompt with a non-empty ground truth; **514** survive A5's drop of prompts whose ground truth *is* the edit target, with 3,405 prompts; and **507** facts with **3,312** locality prompts survive the distractor requirement, which removes 93 prompts. Drawing neighbours gives **3,286** neighbourhood prompts over 504 facts, 26 fewer because some drawn neighbours lack two distractors. Facts carry 1 to 16 prompts in each family, with a median of 6. The locality item file is not shipped, so this chain can be checked against the registration and `loc_baseline.json` but not rebuilt from the archive.

The measure (A2). For each prompt, leakage is whether the edit target outranks that prompt's *own* gold answer under the same length-normalised likelihood rule the constrained-choice cell uses. **The floor (A4)** keeps a prompt only if the **unedited** model already ranks its own gold first, so leakage is counted where there was something to lose; it is computed from the unedited model alone. **The same edit (A5)**: the locality run re-derives the constrained-choice restatement cell inside its own edit; for ROME and MEMIT it agrees with Experiment 3 on 507 of 507 facts (C.4).

The deciding family (A9.2) is the neighbourhood family. The registration fixed it in advance because ROME's known same-subject bleed raises ROME's own leakage on the locality family and would work against the result there. That choice was therefore made knowing which family favoured the hypothesis, and F.4 gives the locality family equal weight.

F.2 RESULT

The registered verdict is `SPECIFICITY`. The rule was applied once, to the completed blocks. The sentence the instrument carries for that branch was written before any locality score existed:

> The gradient editors leak the edit target into prompts it has no business answering more than ROME does, in the same likelihood-ranked format the rejections sit in. The constrained-choice advantage is confounded with a broad boost, and this paper's framing of the measurement axis does not hold for it.

Table 30: 466 facts after the floor on the neighbourhood family and 451 on the locality family, against a registered minimum of 150. Every pair points the same way in both families.

family	pair	gradient leak rate	ROME leak rate	difference	95% simultaneous CI
Neighbourhood (deciding)	FT-L vs ROME	90.2%	6.4%	**+83.8 pp**	[+80.9, +86.8]
Neighbourhood (deciding)	FT-M vs ROME	97.7%	6.4%	**+91.3 pp**	[+89.1, +93.5]
Neighbourhood (deciding)	LoRA vs ROME	99.9%	6.4%	**+93.5 pp**	[+91.5, +95.5]
Locality (same subject)	FT-L vs ROME	98.2%	82.9%	+15.3 pp	[+11.8, +18.7]
Locality (same subject)	FT-M vs ROME	100.0%	82.9%	+17.1 pp	[+13.6, +20.5]
Locality (same subject)	LoRA vs ROME	100.0%	82.9%	+17.1 pp	[+13.6, +20.5]

The three editors whose constrained-choice restatement score exceeds 0.99 put the edit target above a different subject's own gold answer on **90 to 100%** of neighbourhood prompts, where ROME does so on **6%**. Their constrained-choice advantage is therefore confounded with a broad shift toward the target, which a likelihood-ranked choice whose gold option is the target credits. On same-subject prompts ROME also leaks on most prompts, and the difference is 15 to 17 points.

The floor. Before the floor, the **unedited** model already prefers the edit target over a prompt's own gold on **17.9%** of locality prompts and **13.5%** of neighbourhood prompts, and answers its own gold on 52.8% and 55.7%. The floor keeps only the second group, so every leak rate above is leakage the edit created on a prompt the unedited model had right. On the floor-filtered set the unedited leak rate is 0.00% by construction, which is not a negative control; the informative baselines are the pre-floor rates.

Reproduction. `analyze_locality.py` is deterministic and rewrites `locality_result.json` byte for byte on the shipped blocks; a second implementation working from the row schema alone reproduces every fact count and leak rate, which rules out an aggregation error without validating the definitions. All five blocks completed, 14,211 rows each, against one registration hash and one item set; per-block provenance is in `SUPPLEMENTARY_S4_LOCALITY.md`.

F.3 What the test does not establish

- **A magnitude, or an order among the gradient editors.** Leakage is saturated at the level of the editor. On the neighbourhood family the per-fact leak rate is exactly 1.0 on **77.7%**, **92.7%** and **99.8%** of facts for FT-L, FT-M and LoRA, and exactly 0.0 on **82.4%** of facts for ROME and **99.4%** for MEMIT. The narrow intervals reflect that near-degenerate variance rather than precision about the size of the difference.
- **A within-fact link.** The registered correlation between the constrained-choice advantage and the leakage difference is null: $\rho = +0.020$, $+0.008$ and $+0.009$ for FT-L, FT-M and LoRA against ROME, with simultaneous intervals of $[-0.086, +0.126]$, $[-0.098, +0.114]$ and $[-0.097, +0.115]$ on 507 facts. The registration fixed in advance that a null bounds the association at roughly $|\rho| = 0.15$ and does not establish independence. The gradient editors' per-fact variance in that cell is capped near 0.0074, so this correlation is largely ROME's own per-fact variation, shared by all three pairs.
- **A mechanism.** Leakage and the null-prompt calibration of J.4 both measure how much an edit raises the target independently of the question, so for editors that leak on nearly every prompt a reversal that disappears under calibration is close to automatic. The evidence that does not reuse that construct is null (the within-fact correlation), non-diagnostic (Experiment 5) or contrary (the leakage-mismatched gradient pair of Experiment 5 does not reverse). M4 is therefore reported as a flag, not as a demonstrated cause.
- **Specificity separately from editor family.** Every high-leakage editor in the pool is a gradient editor and every low-leakage one a locate-and-edit editor, so the design cannot separate non-specificity from other differences between the families. Experiment 5 (J) was registered to address this and was non-diagnostic. Experiment 6 supplies one case outside the pool: `FT+L`, a gradient editor that leaks on 3.7% of neighbourhood prompts, shows no qualitative interaction against either reference editor under any of the four generation scores (M.2). It trails both of them on both probes, so it is a null from an arm that is worse everywhere rather than a matched contrast, and the flag stands.

- **A statement about the published methods.** The three high-leakage editors are our implementations (A.4): FT-L with a different objective and a clamp, FT-M at EasyEdit's values, and plain LoRA at a fiftieth of AdaLoRA's learning rate. The finding is that *these editors as run here* leak. The direction matches ROME's own GPT-J table for unconstrained fine-tuning (Meng et al., 2022) and what EVOKE reports as editing overfit (Zhang et al., 2025); recovering it checks that the instrument measures what it says, and is not a novelty claim.

F.4 Every corrected reversal, and dependence on the prompt family

MEMIT's block completed after the verdict and is reported descriptively: its neighbourhood leak rate is **0.11%**, against ROME's **6.4%** and the gradient editors' 90.2, 97.7 and 99.9. Cross-tabulating the leak rates against the registered rejection lists, with no test attached:

Table 31: Neighbourhood leakage of the editors in Experiment 3's corrected reversals (Gail and Simon's test; under the intersection-union test the split has six, all high against low).

contrast	corrected reversals	high-leakage against low-leakage	status
measurement	3	**3**	all three pair a gradient editor with ROME; the registered verdict
answer-identity split	8	**8**	derived by `doclint.py` from the rejection list in `exp3_result.json` and the leak rates in the shipped blocks; post hoc

Every corrected reversal of Experiment 3 pairs an editor that leaks the edit target into a different subject's prompt on 90 to 100% of them with one that does so on at most 6.4%. Six of the ten editor pairs are cross-family, so this did not have to happen, but every cross-family pair is also a leakage-mismatched pair, so the table cannot say which of the two properties matters. The two KnowEdit `recent` reversals that survive on fully crossed facts (B.6, G.4) sit on a release whose off-target prompts are scored by post-versus-pre agreement rather than as a target-against-gold ranking, so their leakage is not measured.

Dependence on the off-target family. On the **neighbourhood** family, every one of the eleven registered reversals pairs a high-leakage editor with a low-leakage one at any threshold from 10 to 40 points. On the **same-subject** family the rates are FT-L 98.2%, FT-M and LoRA 100.0%, ROME 82.9% and MEMIT 64.5%, and the count falls as the threshold rises:

Table 32: Leakage mismatch of the eleven registered reversals by prompt family and threshold (nine under the intersection-union test: 9 at every threshold on the neighbourhood family; 9, 9, 3, 3, 3 and 0 on the same-subject family). At the audit's 20-point default, seven of the eleven are matched on the same-subject family: the three measurement reversals and the four split reversals against ROME. The four that stay mismatched are the split reversals against MEMIT. The audit therefore reports leakage in every family it is given and calls these pairs family-dependent (B.6).

threshold (pp)	10	15	20	25	30	40
mismatched, neighbourhood family	11	11	11	11	11	11
mismatched, same-subject family	11	11	**4**	4	4	**0**

F.5 Leakage on Mistral-7B-v0.1

Experiment 5's registered secondary Mi measured leakage for the two gradient editors of the post hoc Mistral rejection (E.4) over Experiment 2's facts with the same rule: **FT-M leaks on 98.8% of counted neighbourhood prompts and FT-multi on 100.0%**, with 316 of 323 facts at exactly 1 for FT-M and 322 of 323 for FT-multi, and 99.9% and 99.8% on the same-subject family. **Both arms are at the ceiling of the measure, so it cannot resolve a difference between them**, and this pair says nothing about whether their specificity differs; reading saturation as equivalence would be the

error. It marks the limit of the instrument: on a backbone where every gradient editor saturates it, leakage cannot separate them. That rejection is in any case post hoc, involves `FT-multi` (ours), and is removed by its own registered sensitivities (E.4).

G RELEASED-BENCHMARK ANALYSIS

The inputs are HalluEditBench's released per-item results (Huang et al., 2025) and the per-item outcomes its authors released for a KnowEdit run (Zhang et al., 2024), rebuilt from the public releases at pinned commits by `fetch_external_data.py`; the instrument is `external_reanalysis.py`. On HalluEditBench the comparisons are the complete $\binom{7}{2} = 21$ pairs within each of four backbones, 84 in all; the benchmark reports aggregate scores and a ranking, not these contrasts, and no selection enters. A permissive per-probe count, which calls a comparison reversed when any minority-direction probe is individually significant, gives 18 of 84. The reversal test at five strata sharing items is the intersection-union test (B.2); it gives 17 of 84 at .05 and 16 after a per-backbone Bonferroni correction, rejecting the same comparisons as Gail and Simon's test. **Figure 2** shows the HalluEditBench result.

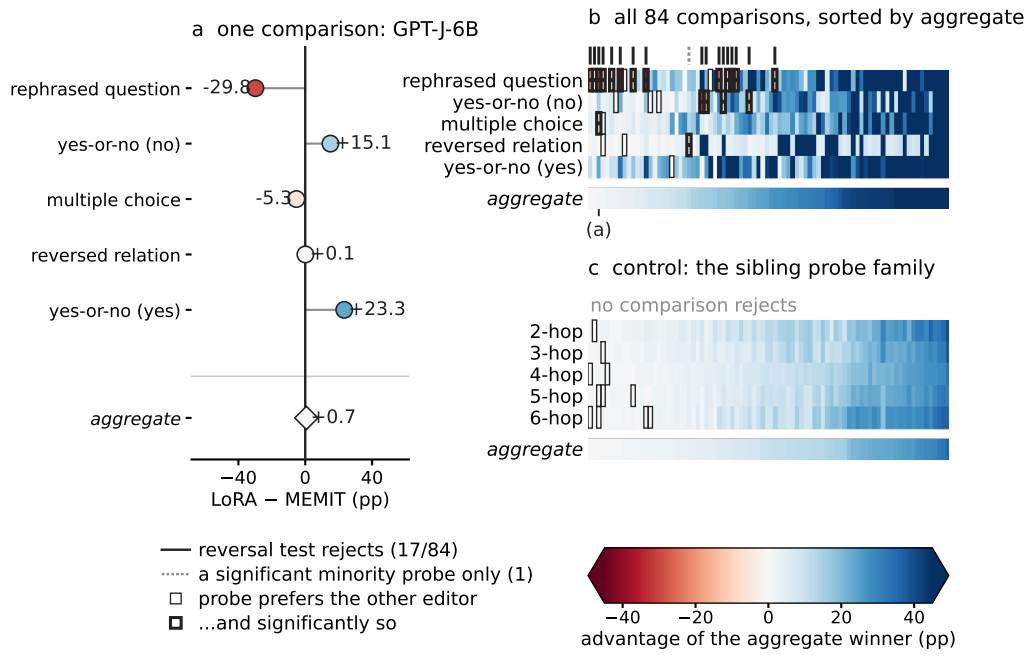

Figure 2: An aggregate ranking can conceal opposite probe effects: HalluEditBench's Generalization Score. Colour is one editor's advantage over another on one probe, in percentage points, oriented so that the *aggregate* favours the named winner: blue probes agree with the aggregate, red probes favour the other editor. Outlined cells are minority-direction probes; a heavy outline with a dot marks those significant at 0.05 on their own paired test. **(a)** One comparison: an aggregate gap of $+0.7$ points sits over probes favouring opposite editors by 29.8 and 23.3. It is the column marked (a) in panel (b). **(b)** All 84 comparisons, sorted by aggregate advantage. Ticks mark the **17** where the reversal test rejects directional agreement at 0.05 (16 after Bonferroni); the dotted tick is the **one** further comparison flagged only by the permissive per-probe count. The red cells concentrate in the open-answer rephrase question. **(c)** The sibling portability family on the same scale: no comparison rejects. The shared scale saturates at ± 45 pp (the largest positive advantage is $+98.0$; no minority-direction cell is clipped); portability post-accuracies are floor-compressed, and the scale-free comparison is in G.2.

G.1 HALLUEDITBENCH: PROVENANCE, QUESTION TYPES AND CHANCE LEVELS

What the release contains. The release is `baixianghuang/HalluEditBench` at commit `0141e32`. Its documentation names three backbones, Llama-2-7b-chat, Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.3, and its data tree shows those three. It also ships a fourth results folder,

gpt_j_6b (179 files against 182 for the others), documented only by a note on "additional results for GPT-J-6B"; its files carry no model metadata, and 4,116 of their 15,968 post-edit rephrase outputs still contain the string "Question:", which the shipped evaluation code cuts from outputs before storing them. The GPT-J results were therefore probably not produced by the shipped code, and this paper treats them as the release's supplementary results rather than as one of the benchmark's own backbones. The in-context method is named ICL in every released file and in the code that runs it; the benchmark's tables relabel it ICE for display. We keep the released name.

How answers are judged. In the public code every question type is answered by greedy generation and scored 1 when the lower-cased output and label are substrings of each other, and otherwise by an LLM judge; multiple-choice questions are answered with a letter. An exactly empty output is a substring of every label and so always scores 1. Of 63,134 post-edit rephrase rows, the 39 exactly empty outputs are all scored correct (GPT-J FT-L 4, Llama-2 ROME 6, Llama-3 FT-L 18, Mistral FT-L 11). The 2,367 whitespace-only outputs, almost all GRACE's, are all scored 0, so a string of whitespace and the empty string are scored oppositely by the same rule. Dropping the 39 cases those rows belong to and re-running the reversal test over all 84 comparisons changes nothing that the paper reports: 17 rejections at .05 and 16 after a per-backbone Bonferroni correction, with the same per-backbone counts (9, 3, 3 and 1). Two things move. Llama-2 MEMIT against ROME, the rejection that a per-backbone correction removes, falls from $p = 0.0079$ to 0.0032 and now clears the corrected threshold under the intersection-union test but not under Gail and Simon's. GPT-J FT-L against GRACE, a non-rejection, becomes untestable, because the excluded cases carried the only variation in one probe's paired differences and its standard error falls to zero. Whether the public code is the code that produced the released results is not verifiable from the release.

(external_reanalysis.json → backbone_strata, gail_simon_structure.)

Table 33: Per backbone, under the reversal test (identical under Gail and Simon's). Magnitudes in percentage points; cnt is the permissive per-probe count, and the reversing question types are counted over those 18 comparisons, which is why they sum to 19 where the 17 Gail and Simon rejections carry 20 (one gen_mc stratum belongs to a rejection the permissive count does not make; §3 quotes the rejection-based figure). The homogeneity p is the Freeman–Halton exact test on the 2×4 table of rejection counts, enumerated over its full support. The pooled rate is not a prevalence.

backbone	test	cnt	med. mag	max mag	med. gap	reversing question types
GPT-J (supplementary)	9/21	9	17.2	43.4	15.9	rephrase 6, no 3, mc 1
Mistral-7B	3/21	4	21.7	39.7	15.0	rephrase 3, reversed 1
Llama-3-8B	3/21	3	11.1	15.3	3.7	rephrase 2, no 1
Llama-2-7B	2/21	2	4.5	5.7	0.5	rephrase 2
all four	**17/84**	**18**				exact homogeneity $p = 0.043$

Excluding the supplementary backbone. gpt_j supplies half the reversals, the largest single magnitude and the worked example, and G.2 diagnoses a model-level pathology on it. Over the three documented backbones the count falls from 17 to 8, a little over a half, while the rate falls by about a third, from 20.2% to 12.7%: 8 of 63 at .05 and 7 after correction, with reversing question types rephrase 7, no 1 and reversed 1; the magnitudes do not fall, since Mistral's median of 21.7 pp is the largest of the four.

Leave one question type out. Each row re-runs the reversal test over all 84 comparisons with one question type removed; its multiplier is the number of strata remaining.

Table 34: Leave-one-question-type-out on HalluEditBench. With gen_yes removed, Gail and Simon's test gives 16 at .05 and the intersection-union test 17; every other row is identical under both.

question type removed	types remaining	significant at .05	after Bonferroni
none	5	17 / 84	16
gen_mc	4	17 / 84	16
gen_reversed	4	17 / 84	16
gen_yes	4	17 / 84	15
gen_no	4	13 / 84	12

question type removed	types remaining	significant at .05	after Bonferroni
gen_rephrase	4	**5 / 84**	**5**

Which rows the substring rule decides. The pipeline tries a bidirectional substring match first and falls back to an LLM judge, and the release records neither which rule fired nor the judge's calls, so a natural worry is that the open-answer rephrase question reaches the judge far more often than the closed-answer ones and is therefore scored by a different instrument. It does not. Applying the public substring rule to all 63,134 post-edit rows of each type and comparing with the recorded score, the rule reproduces **99.99%** of yes rows, 99.99% of no, **99.8%** of rephrase, 99.7% of reversed and **91.0%** of multiple-choice rows. The judge can therefore account for at most two rephrase rows in a thousand, and the type the rule explains least is multiple choice, which is the opposite of the direction the worry predicts. The residual is not a count of judge calls: it is whatever the published rule does not reproduce, which also covers any normalisation the shipped code applies. The rule must be applied without stripping, since the release scores whitespace-only outputs 0 and exactly-empty ones 1; stripping collapses those two cases and mis-scores every row of either kind, and the instrument's self-test pins all four.

Open-answer and closed-answer questions. Removing `gen_rephrase` removes two-thirds of the result. Because every question type is generated and judged the same way, this is not a difference of scoring rule. It is a difference of answer space: rephrased and reversed questions have open answers, while yes, no and multiple-choice questions have closed ones with chance levels of 0.5 and 0.25, and under a bidirectional substring rule even those chance levels are approximate. The likelihood-ranked factorial of Sections 4 and 5 does not score anything this way, so it does not explain HalluEditBench's concentration in the rephrase question. The five rejections that remain without it are:

Table 35: The five rejections that remain without the rephrase question.

backbone	comparison	reversing question type (effect, pp)
`gpt_j`	ICL vs LoRA / MEMIT / ROME	`gen_no` (−34.7, −18.2, −11.6)
`gpt_j`	LoRA vs MEMIT	`gen_mc` (−5.3)
`llama-3`	FT-L vs GRACE	`gen_no` (−9.0)

Three are `gpt_j`/`gen_no`, the cell G.2 diagnoses, and all three compare against ICL, a prompting baseline; a fourth is against GRACE, whose generalization score on that backbone is 7.9%, a collapsed arm. The residual after the rephrase exclusion is therefore **one comparison**, and because `gen_reversed` is open-answer as well that residual is not controlled for answer space either.

Chance levels. HalluEditBench records no chance level. Re-running the audit with nominal chance levels of 0.25 for multiple choice and 0.5 for the yes and no questions leaves every verdict unchanged (16 corrected failures, 67 magnitude heterogeneous, 1 with no detected heterogeneity) and flags every one of the 16 corrected failures for an arm within 0.05 of chance on some question type; with chance levels of zero, four carry no flag (B.6). HalluEditBench therefore supports a magnitude result in raw points, not a probe-attributable one. KnowEdit is not affected by this confound: its three portability probes share one scoring rule, and it returns **12 of 42**, 8 after Bonferroni over the family of 42 (9 over each cell's own family of 21), with its sibling family at 0 of 21.

G.2 What one averaged score is made of

Section 3's worked pair, in full, from `post_hoc_analyses.json` (`halluedit_format_collapse`). The backbone is Mistral-7B-Instruct-v0.3, one of the benchmark's three documented ones, and the cases are the 2,233 on which both editors are scored on all five question types.

Table 36: The released per-item accuracies for one pair, by question type, with the unedited model's level on the same rows. The pre-edit column is a single number per row because both arms start from the same model; it differs between the two arms by at most 0.05 points anywhere, which the instrument's self-test asserts.

question type	answer space	ROME	LoRA	difference	both, pre-edit
`gen_yes`	closed	78.46	14.55	+63.91	56.2
`gen_no`	closed	43.13	12.05	+31.08	29.6
`gen_mc`	closed	77.30	55.40	+21.90	61.1
`gen_reversed`	open	9.85	4.08	+5.78	10.5
`gen_rephrase`	**open**	**68.03**	**89.21**	**−21.18**	15.8
Generalization Score		**55.35**	**35.06**	**+20.30**	

The reversal is certified: Gail and Simon's test gives p below machine precision and the dependence-robust intersection-union test $p = 1.0 \times 10^{-77}$, against a per-backbone threshold of $0.05/21 = 2.4 \times 10^{-3}$.

What the closed-answer columns measure. After LoRA's edit the model does not answer a yes/no question in form: **13.2%** of its yes and no outputs begin with "yes" or "no", against **93.4%** for ROME, from the same unedited model. Its scores of 14.6 and 12.1 are far below the 50% a coin would earn, which is what a model that is not emitting the answer tokens at all looks like, and not what a model that knows the fact and answers wrongly looks like. Three of the five question types therefore contribute a measurement of output form.

What the open-answer column measures. LoRA leads on the rephrase question three ways, and the three are not the same measurement. As **the release scores it**, 89.21 against 68.03, which is the −21.18 the table above prints and the figure quoted in §1. Re-applying the release's **own bidirectional substring rule** to the stored outputs gives 89.2 against 67.5; the half-point difference on ROME is the same disagreement G.1 measures, where that rule reproduces 99.8% of rephrase rows rather than all of them, and it is reported here rather than smoothed away. Under the **first-span match** of E.1, which credits an answer given first and then repeated but not a label buried in degenerate text, 73.2 against 58.4; this one cannot be taken from the release's scores at all, because it needs the outputs. The gap narrows, and the ordering does not change, so this side is not an artefact of the scoring rule. It is worth stating plainly because the opposite result would have been convenient for us: our own factorial found that an overlap score can manufacture a generation deficit (E.1), and here the same family of checks leaves the effect standing.

What the pair does and does not license. It licenses this: on these data the Generalization Score's ordering of these two editors is produced by three question types on which one edited model has stopped answering in the required form, while on the question type that asks the edited fact in other words the ordering is the other way. It does not license a claim that LoRA is the better editor. Its restatement of the fact is credited here, and its damage elsewhere is exactly what a locality or fluency measure is for; the benchmark reports both, and the averaged Generalization Score is where the two become one number.

G.3 Two diagnostics of the released data

The no-probe on GPT-J. `gen_no` is HalluEditBench's only generalization question whose correct answer is "no". Supplied the fact in context, GPT-J moves the wrong way on it (accuracy 38.2 to 13.9, P(yes) on no-items 61.8 to 86.1) while the other three backbones gain 26 to 43 points; on `gen_yes` it also moves the wrong way. The accurate description is an intervention-induced loss of polarity discrimination on that backbone, a property of the model rather than of the question, and three of the five rejections that survive the rephrase exclusion sit in that cell.

The negative control. HalluEditBench's **portability** family is the same file, four models, seven methods, 84 pairs, five probes and different base rates. The reversal test rejects 17 times on generalization and **never** on portability, so the pipeline does not fire indiscriminately. It does not by itself show that the two families differ in kind: portability's median span is 5.8 points against generalization's 49.1,

roughly half the raw span ratio is base-rate arithmetic (forcing one constant log-odds effect gives a null median ratio of 1.06 against the observed 1.92), and its cells sit near the floor (0.067–0.130 pre-edit). On raw sign disagreement the comparison is 24 pairs against 8.

Both diagnostics are reported in full, with their tables, in `SUPPLEMENTARY_S6_BENCHMARK_DIAGNOSTICS.md`.

G.4 KnowEdit in full

KnowEdit's release covers 7 editing methods (FT, FT-L, IKE, LoRA, MEMIT, ROME, SERAC) on **one** base model (Llama-2-7b-chat), with the benchmark's own probe types and per-item outcomes. It scores **all three** portability probes (`Subject_Aliasing`, `Logical_Generalization`, `reasoning`) by one fractional token-match rule, so neither the response format nor the scoring rule changes within the family.

Table 37: The two benchmarks side by side. Relative to each family's own median span the ordering of magnitudes reverses (0.15 `zsre`, 0.37 `recent`, 0.31 HalluEditBench), so no amplification factor is claimed.

benchmark	scoring across probes	reversal test	median reversal
KnowEdit portability	one (fractional token match)	12 / 42	4.4–6.9 pp
HalluEditBench generalization	one judging rule; open- and closed-answer questions	17 / 84	15.3 pp

Table 38: KnowEdit per cell; `cases` is the range of cases per probe type.

KnowEdit cell	cases	pairs	span	gap	ratio	rev. test	$\geq$1 opposite
`zsre` · portability	263–771	21	28.8	12.6	2.60	8/21	11/21
`recent` · portability	353–860	21	18.8	15.7	1.19	4/21	6/21
`recent` · locality (control)	842–1172	21	7.4	14.1	0.80	0/21	1/21

Pairing (M3). Each probe type is paired across methods on the cases where every method has that probe; different probe types use different cases. We use per-probe pairing because it matches the HalluEditBench analysis; requiring each case to carry every probe type would leave `recent` 108 cases and `zsre` none. On `zsre` the probe types' case sets are disjoint (263 and 771), so none of its reversals can be attributed to probe type. On `recent` two probes share 353 cases and 108 carry all three. At the applied threshold the minimum detectable effect on those 108 cases is 14.5 to 14.7 points on the aliasing stratum and 6.2 to 7.1 on the reasoning stratum, estimated from the release's own standard errors; certification needs both, so the joint power against an 8.2-point reversal is about 0.17 (B.5) and a null there says little. On the 108 fully crossed cases the audit finds two corrected reversals with no flag (B.6). `counterfact` is excluded structurally: no probe type reaches 30 cases shared by all seven methods, and the instrument emits the exclusion as a `skipped` record with its reason.

The answer-identity split on `recent` (post hoc). KnowEdit's `reasoning` probe on `recent` pools, as on the item set of Sections 4 and 5, prompts whose gold answer is the edit target with genuine consequences. Classifying each reasoning prompt by whether its ground truth equals the case's `target_new` after case folding, 54 of the 108 crossed cases carry a target-as-gold prompt and 88 carry a true-consequence one (a case can carry both).

Table 39: The two `recent` reversals with the reasoning probe restricted by answer identity, intersection-union test at the per-cell family of 21 (threshold $0.05/21$). Effects in percentage points; the minimum detectable effect (MDE) is for the reasoning stratum at the applied threshold. Computed by `post_hoc_analyses.py` from the release.

pair	reasoning prompts	n	alias	logical	reasoning	p	MDE
MEMIT vs FT-L	pooled	108	$+18.6$	-1.3	-13.6	6.7×10^{-7}	7.1
MEMIT vs FT-L	true consequence	88	$+21.0$	-1.2	-14.8	$\mathbf{7.5 \times 10^{-7}}$	7.4
MEMIT vs FT-L	target as gold	54	$+10.0$	$+0.5$	-20.3	0.063	17.3
ROME vs FT-L	pooled	108	$+16.5$	-0.6	-14.2	8.1×10^{-6}	6.2
ROME vs FT-L	true consequence	88	$+18.6$	-0.3	-14.5	$\mathbf{9.8 \times 10^{-6}}$	6.8
ROME vs FT-L	target as gold	54	$+7.4$	$+0.2$	-21.0	0.18	16.9

Both reversals persist on true-consequence prompts, so label pooling does not explain them. On target-as-gold prompts FT-L's reasoning advantage is larger still, and those subsets are too small to certify anything (MDE 17 points). The pattern is the opposite of the gradient editors' on GPT-J: here the fine-tuning editor is the one ahead on consequences and behind on aliases. The release scores off-target prompts by post-versus-pre agreement rather than as a target-against-gold ranking, so leakage cannot be measured, and these two remain the released-benchmark reversals our checks leave unexplained.

Partial credit for token overlap, and the strictest reading of a true consequence. KnowEdit's portability metric is a token-level accuracy, so a consequence whose gold answer shares a word with the edit target (`Canada` against `O Canada`, `Scotland` against `Flower of Scotland`) can be credited to an editor that merely raises the target. Of the 8,556 true-consequence reasoning golds on `recent`, **1,208 (14.1%) share at least one word with `target_new`**. Dropping every one of them, which leaves 87 of the 88 crossed cases with a usable true-consequence stratum, is the strictest reading the release allows, and both surviving reversals hold on it: MEMIT against FT-L at $p_{\mathrm{IU}} = 9.1 \times 10^{-7}$ ($+21.1$ on aliasing against -14.4) and ROME against FT-L at 1.2×10^{-5} ($+18.6$ against -14.1), both larger on aliasing than on the unrestricted subset. Token overlap is therefore not what these two are made of either, and they remain the released-benchmark reversals no check here explains.

G.5 A THIRD BENCHMARK

Three benchmarks qualified in a budget-limited search; `editing-attack` (Chen et al., 2026) was analysed with the identical statistic and ships in the artifact. Across its 30 method pairs (three editors, five instruction-tuned backbones, three probe types from EasyEdit's taxonomy, 100 items per cell) it returns **1 significant reversal**, with 3 pairs showing any opposite-signed probe. That near-null is **underpowered**: its minimum detectable effect is 15.1 points, above the 4.4–6.9-point reversals KnowEdit shows, and its `locality` and `yes_no` fields are empty in every file. It is evidence neither that the phenomenon fails on instruction-tuned models nor that it holds. A fourth candidate, `behavior-edit`, was rejected before analysis because each file carries a single probe and its "probe classes" are directory-level groupings. `SUPPLEMENTARY_S6_BENCHMARK_-DIAGNOSTICS.md` gives the detail.

H REGISTRATIONS, DEVIATIONS AND PROVENANCE

H.1 THE STATUS OF EVERY REPORTED RESULT

Each row is a result the main text reports. **Registered**: design, family and rule fixed before any output of that experiment existed. **Registered amendment**: added to a registration before the relevant

contrast or outcome was computed. **Post hoc**: designed after the data existed. **Observational**: a retrospective re-analysis of released outputs. H.4 describes what "registered" does and does not establish here.

Table 40: Status of every reported result.

result	where	status
released-benchmark counts and pairing (17 of 84, 12 of 42, the two crossed `recent` reversals)	§3, G, B.6	**observational**; statistic and comparison sets fixed before the counts were read, no pair selected
leave-one-question-type-out and chance levels on HalluEditBench; the answer-identity split on KnowEdit `recent`; GPT-J provenance and empty outputs	§3, G.1, G.4	**post hoc**
Experiments 1 and 2 and the extension: nulls, formats, families, MDEs	C	**registered**; pooling across commits is deviation D10
Experiment 3: populations, family of 40, cell-mean restriction, 3 of 20 and 0 of 20, partial support	§4.2, D	**registered**; the restriction's constants were chosen after Experiments 1 and 2, and the arm-level observation preceded the registration (H.4)
arm-level restriction and the sixteen-setting margin grid	D.4, E.3	**registered** sensitivities
answer-identity split, 0 to 8	§5.2, E.3	**registered amendment** (D2), before any contrast; computed after the primary verdict (D5); hypothesis from overlapping data
held-out facts; flip weights; generation rescored by first-span match and first-span F1; the constrained-choice margin and its dominance check; same-cell effects and the subset $\times$ stratum test; continuation lengths and repetition	§4.2, §5.1, §5.2, D.3, E.1 to E.3	**post hoc**; the generation scores were chosen after a first look at the stored continuations, the dominance rule before it was run
unpairing, random and systematic	§5.3, E.5	**post hoc**
effect sizes, reversal bounds, format mixing and screen analyses on the earlier pools	E.4, E.6	**post hoc**
300-fact seed sweep and the τ breaking points	C.4	**post hoc**
leakage test and `SPECIFICITY` verdict	§5.1, F	**registered** after Experiment 3's outcome; amendments A1 to A10 before any locality row existed
MEMIT leak rate; leakage of the split reversals; the same-subject threshold sweep	§5.1, F.4	**post hoc**, descriptive
Experiment 5: verdict, arms, twelve comparisons; secondary C (calibration), S (seeds), U (baselines), Mi (Mistral)	§5.1, §4.2, J, F.5	**registered**; amendment A9 recorded after the constrained arm's block was queued (H.2)
calibrated arm levels; the efficacy-matched arm under calibration and first-span scoring; the dose comparisons	§5.1, Figure 1, J.4, J.5	**post hoc**
Experiment 4: verdict, pool, arm levels, leak rates	§6, K	**registered**; backbone substituted by amendment M1 before any output
Experiment 6: the two published configurations that ran, their arm levels, leak rates and twelve comparisons; `AdaLoRA`'s divergence	§5.1, §8, M	**registered**; the AdaLoRA rank reading amended in §7 of that registration after it proved unrunnable and before any block existed
Experiment 7: teacher-forced levels and the twelve comparisons of family T; the two fidelity arms and the eight of family F	§3, §5.1, N	**registered**

result	where	status
Experiment 8: L1 to L5 on Llama-2-7b-chat, the arm levels, the six comparisons of family L and the six of family S, `AdaLoRA`'s divergence and the screen's retention	§3, §5.1, §8, O	**registered**
the attribution flags carried by Experiment 8's reversal, read from the audit instrument	O.4	**post hoc**
the sweep of five published papers' own tables	I.4	**post hoc**, observational
flag base rates and the planted check; relation-clustered standard errors; the substring-rule share; the audit of a published table	B.7, E.6, G.1, I.3	**post hoc**
the audit tool's outputs (verdicts, flags, flip weights, bounds); the HELM demonstration	§6, B.6, L	written after the experiments; the HELM run is **post hoc**
the test policy of B.2 and every intersection-union count	throughout	**post hoc** re-analysis; Gail and Simon's test is the registered one

Where the choice of test matters. At two strata the intersection-union p is exactly twice Gail and Simon's, so the two agree except where a p falls between the thresholds. Every headline count is identical under both: HalluEditBench 17 at .05 and 16 after a per-backbone correction, KnowEdit 12 and 8 with two crossed `recent` failures, Experiments 1 and 2 and the extension 0, Experiment 3's measurement contrast 3 and its pooled semantic contrast 0, Experiment 5's two corrected reversals (one under the arm-level rule) and Experiment 4's non-diagnostic verdict. Four counts differ: the answer-identity split, 6 under the intersection-union test against 8 registered; the split on the 278 held-out facts, 3 against 4; unpairing by answerability, 266 pair-splits against 394 (and by relation group, 0 against 5); and secondary C's classification, 5 vanish and 2 attenuate under both, with 3 further comparisons not replicating raw under the registered test. Secondary S classifies eleven reversals at the full families, of which nine are seed-robust and the two LoRA splits under constrained choice are not; the intersection-union test does not certify those two, so under it all nine of its reversals are seed-robust. Counts at the resized families the arm-level restriction keeps (22, 13 and 9) appear only here: 1 arm-level measurement reversal and 2 in Experiment 5.

H.2 Deviations

Labels are each registration's own. Every deviation entered its registration before the contrast it could affect, except where the table says otherwise; dated records are in the registrations and `supplementary_README.md`.

Table 41: Every deviation, the analysis it affects, and its effect on reported conclusions.

deviation	affected analysis	resolution	effect on reported conclusions
Exp. 1–2 D1: backbone GPT-J-6B instead of Llama-2-7b-chat (gated, no access token)	Experiment 1	GPT-J-6B has results in HalluEditBench's supplementary release (G.1); facts and taxonomy unchanged	none on any claim; scope is GPT-J
Exp. 1–2 D2: compute partition changed	Experiments 1–2	runs checkpoint per fact; a truncated run is stamped PARTIAL	none
Exp. 1–2 D3: editing framework not importable; gradient editors written as our own code from its configuration values	Experiments 1–2 and every gradient-editor result	registered extension adds ROME and MEMIT from the framework's files (C.3); gradient editors documented in A.4	"gradient editors" means these implementations

deviation	affected analysis	resolution	effect on reported conclusions
Exp. 1–2 D4–D7: gold position cycled, type-matched distractors, likelihood-ranking scoring, `recent` and `counterfact` pooled; all before any editor ran	item set and scoring (A.1, A.2)	fixed in code with regression tests	the measurement is the one the paper describes
Exp. 1–2 D8: LoRA as plain LoRA at lr 1e-4 with clipping (published AdaLoRA setting diverged on calibration facts)	every LoRA result	frozen before analysed facts; holdout efficacy 0.386 against 0.381 on calibration. Experiment 6 re-ran the published setting with AdaLoRA's own parameterisation: it diverged on 20 of 20 calibration facts (M.1)	LoRA results concern this configuration; no result in this paper describes EasyEdit's published LoRA editor
not registered: FT-L's objective changed from the target at the last prompt position to cross-entropy over prompt and target, beside its clamp	every FT-L result	recorded in A.4 only; no reason is documented anywhere. Experiment 6 runs EasyEdit's own FT-L and ROME's FT+L on the same facts (M)	FT-L results concern this variant; the published configuration repeats twice as often and, unlike ours, reverses against ROME under three of four generation scores (M.3)
not recorded: no EasyEdit commit was pinned when configuration values and the ROME and MEMIT files were copied	A.4, C.3	values match EasyEdit's current configuration files	the "published configuration" is that version
Exp. 1–2 D9: LoRA excluded from Experiment 2 after diverging mid-run	Experiment 2	recorded before any Experiment 2 contrast; pool of three	family of 3 instead of 6
Exp. 1–2 D10: extension pooled across two commits	extension (C.3)	exact FT-L reproduction across the commits; pooling refuses any other difference. The pooling decision preceded that reproduction, and is recorded here rather than in a registration	none on counts; see C.3
Exp. 3 D1: calibration and holdout facts removed	Experiment 3 population	P0 is 533 facts	none
Exp. 3 D2: answer-identity split registered, before any contrast	semantic contrast	family extended to 80 with its own sensitivities	the M1 result (E.3)
Exp. 3 D3 and a scope note: registration-stamp reader repaired; yes/no rationale reworded	analyser	before any contrast; no cell changed	none
Exp. 3 D4: the decision rule's implementation read only half of clause (c); **repaired after the verdict**	Experiment 3 verdict	the repair can only downgrade a verdict; the pre-repair artifact `exp3_-result.PRE_D4.json` ships	no verdict name and no result value changed
Exp. 3 D5: split computed after the primary verdict	M1 result	rule fixes no free parameter and reproduces D2's stated counts	none
Specificity A1–A10: leakage measure replaces free-generation survival; after Experiment 3's outcome, before any locality row	specificity test (F)	registered correlation kept as secondary; the deciding family chosen knowing ROME's same-subject bleed	the registered test is the leakage test; F.4 reports both families

deviation	affected analysis	resolution	effect on reported conclusions
Exp. 4 M1: Qwen2.5-7B-Instruct replaces Llama-3-8B-Instruct (licence), before any output	Experiment 4	same size, published EasyEdit configurations	scope is Qwen
Exp. 4 LoRA: published configuration diverged (20.8% of facts against a 5% cap)	Experiment 4 pool	configuration of Experiments 1–3 used by the registered rule; published block ships	none on the verdict; that block most likely trained other modules than it records (A.4)
Exp. 5 A9: selected arm's values written into the registration after its block was queued	Experiment 5 arm G*	values fixed earlier by the calibration artifact; no contrast computed	none
Exp. 5 rationale: truncation said not to reverse a sign	Experiment 5 primary analysis	holds only under stochastic dominance (§2.3); analysis kept as registered	restricted reading reported beside it (J.3)
Exp. 6 §7: the published `rank: 8` read as AdaLoRA's initial and target rank, which peft cannot run; amended to EasyEdit's own construction, before any block existed	Experiment 6's `AdaLoRA` arm	amendment appended to the registration with its hash chain and the two qualification records that existed; the arm diverged afterwards in any case	none on any computed comparison; the arm produced no block
All experiments: the registered test assumes independent strata	every corrected-reversal count	post hoc test policy (B.2): Gail and Simon's test where valid, the intersection-union test where not, both reported	counts under both in H.1; no claim depends on the choice

H.3 Registered outcomes not reported

None of these could have changed a primary result, and none was dropped after a result was seen.

Table 42: Registered outcomes of Experiment 1 that are not reported, and why.

registered outcome of Experiment 1	status
§5.5 the **three-type secondary** on the 108-fact subset (registered power 0.25; 0.28 with the correct mixing weights of B.2)	**not computable**: the frozen item set carries two semantic types, so the subset is empty; the semantic leave-one-probe-out routed to it goes with it
§5.6 **per-model exact homogeneity** across the controlled experiments	**not reported**: with two experiments and zero rejections in each it returns $p = 1$ by construction
§5.3 the ranking sweep's **best-alternative margin** and **leave-one-probe-out**	**not reported** for the controlled experiments; at two probe types the leave-one-probe-out leaves no contrast
§5.4 the **collapsed-arm** and **per-paraphrase** leave-one-outs	the collapsed-arm exclusion is vacuous (no controlled arm has mean post-accuracy at or below 0.10); the per-paraphrase one is reported in E.4

A replication of Experiment 3 on Mistral-7B-v0.1 was registered and has not been run. It supports no claim in this paper; its registration ships as `SUPPLEMENTARY_S3_REGISTERED_NOT_RUN.md`.

H.4 What the registrations do and do not protect

The record. Each registration is append-only and hashed, and every output row carries the commit, registration hash, qualification artifact, item-set hash and screen hash that produced it; the analyses refuse blocks whose provenance disagrees. The hash chains, commit times and scheduler records are described in `supplementary_README.md`.

What it does not establish. Nine registrations were written within fifteen days by the same authors who ran the experiments, one of them (the Mistral replication of Experiment 3) never run, and every

timestamp is author-controlled: commit dates, hashes and scheduler records, none of them third-party. The record shows the order in which files were committed; it cannot show what the authors had seen. Specifically:

- **Experiment 3's arm-level observation came first.** The observation that three of five arms in the extension's decisive cell sit at ceiling was committed at 11:30 on 2026-09-14, and `PREREG_SYMMETRIC_2026-09-14.md` was first committed at 16:07 the same day, with the earlier commit as its ancestor. The registration named the arm-level restriction as a sensitivity rather than as the primary rule, with that observation already in the record.
- **The decision rule was repaired after the verdict** (D4), in the direction that can only downgrade a verdict.
- **The leakage test was registered after Experiment 3's outcome was known**, and its deciding family was chosen knowing that the other family favoured ROME's leakage (F.1).
- **Experiment 5's selected arm** was written into its registration after its block was queued (A9).
- **The answer-identity split** was proposed from data that share 255 facts with Experiment 3's population (E.3).
- **The audit framing, the test policy of B.2 and every post hoc analysis of H.1** postdate all registrations.

The registrations therefore constrain analysis choices without providing independent confirmatory protection, and H.1 labels every result accordingly.

H.5 Instruments

Every number in the paper is produced by a shipped instrument with a self-test, and `doclint.py` checks the documents against the artifacts those instruments write. `supplementary_README.md` lists the instruments, their self-test counts and the commands that produce each artifact; `AUDIT_LOG.md` records defects found in the instruments and documents during the study and how each was resolved.

I Related work in detail

I.1 The closest works

Table 43: The closest works, point by point.

work	question	what this paper adds relative to it
Meng et al. (2022), CounterFact and its GPT-J table	do edits generalise to paraphrases without spilling onto neighbours?	the ranking consequence of that trade-off: on these data, scoring rules that credit a broad boost order editors oppositely to rules that do not
Zhang et al. (2025), EVOKE	do edited models over-weight the edit target (editing overfit)?	the same direction appears in KnowEdit's own `Reasoning` label, which pools target-as-gold items with true consequences (E.3)
Hoelscher-Obermaier et al. (2023), CounterFact+	does measured specificity depend on how neighbourhood prompts are built?	leakage reported for two prompt families, with the conclusion shown to depend on the family (F.4)
Pohl et al. (2025)	do protocol choices (scoring function, generation length, batch size) change editor rankings?	holds the protocol fixed and varies the probe; adds generation scoring as a further protocol choice that reorders editors (E.1)
Yang et al. (2025)	does teacher forcing at test time inflate edit success?	our free-generation cell is closer to their non-teacher-forced protocol (E.7)

work	question	what this paper adds relative to it
Ethayarajh & Jurafsky (2020); Colombo et al. (2022); Zhang & Hardt (2024)	how should benchmark results be aggregated, and how stable are aggregate rankings?	the per-pair question of whether a composite ranking is also a Pareto ranking over the benchmark's own probe classes
Oh (2026)	do ranking claims match evidence such as cross-task win rates and breakdown points?	the same principle applied to editing benchmarks' probe classes, with checks for why probe classes disagree
Hu et al. (2026)	do pooled root-cause-analysis leaderboards hide system-specific winners?	checks for label pooling, ceiling and chance arms, item sets and target-as-gold credit; controlled experiments in knowledge editing
Zheng & Yang (2026)	are near-tied LLM rankings robust to DIF-guided benchmark recomposition?	benchmark-native probe taxonomies; reversals of several to tens of points
Gail & Simon (1985); Berger (1982); Piantadosi & Gail (1993)	is a treatment effect directionally consistent across subgroups?	used as published (B.2)

Zheng & Yang (2026) also hold scoring fixed and measure sign reversals on paired item-level data, but their strata come from a DIF residual rather than a published taxonomy, the pairs they recompose start from aggregate gaps below one point, which is a property of the pairs they select and not the size of the reversals they then find, and their work is stronger on replication (five benchmarks against our two usable ones, with its own negative control). Hu et al. (2026) audit pooled leaderboards against their benchmarks' own subsystem strata and release a module for it; Appendix L is the same kind of subgroup audit. Pohl et al. (2025) find MEMIT best on CounterFact under the multiple-choice scoring its authors chose and worst under both alternatives. That is **not** the shape M4 describes, and saying so cuts against our own account: MEMIT is the *least* leaky editor in our data (0.3% of neighbourhood prompts on Experiment 5's seed-42 blocks, Table 59; 0.11% on Experiment 3's, F.4), so target-as-gold credit does not predict that it should gain most from multiple choice. Their result is a scoring-protocol reordering, the phenomenon of E.1, and their setting measures no specificity.

The editing literature widens the probe set because one query underdetermines whether an edit took: MQuAKE adds multi-hop queries (Zhong et al., 2023), RippleEdits adds logical-generalisation, compositionality, subject-aliasing and forgetfulness probes (Cohen et al., 2024), and later work adds long-form and unstructured settings. Four adjacent strands are not engaged here: short-form against long-form evaluation and robustness to rephrasing, for which this paper cites nothing; the pitfalls literature, of which Li et al. (2024b) is an empirical study rather than a survey; and forgetting under sequential editing, where Gupta et al. (2024) measure the loss across a sequence and Ma et al. (2025) restrain it. Nothing in this paper edits sequentially, so none of these results is tested here; they bound what a single-edit factorial can be asked to say.

Not contributed by this paper: Proposition 1, which is the Pareto-dominance fact for positive linear scalarisation; the reversal test, which is Gail and Simon's test and its intersection-union form; the leakage measure, which is a variant of existing specificity scores (I.2); the checks M1 to M3, which are subgroup cancellation, ceiling and floor compression and confounding of condition with items; the observation that probe types differ, stated in the audited benchmarks' own text; and the general finding that benchmark composition changes rankings (Wang et al., 2024d; Sprague et al., 2025; Oh, 2026; Gilda & Gilda, 2026).

I.2 Leakage and existing specificity metrics

CounterFact's Neighborhood Score is a **post-edit** comparison: on prompts about other subjects sharing the relation, does the model still assign more probability to the original object than to the edit target (Meng et al., 2022). KnowEdit's locality score asks whether the post-edit answer still *agrees with the pre-edit answer*. CounterFact+ adds a component that changes how neighbourhood prompts are presented and a KL-divergence specificity metric (Hoelscher-Obermaier et al., 2023), and EVOKE measures editing overfit, the edited model's tendency to favour the edit target on prompts where it is not the answer (Zhang et al., 2025). The leakage measure asks nearly the same question as the Neighborhood Score and EVOKE's overfit measure. The differences are narrow: leakage applies

the A4 pre-edit floor, keeping only prompts the unedited model already ranked correctly; it scores under the **same length-normalised rule constrained choice uses**, so it is commensurable with the advantage it is meant to explain; and it compares the target against **the prompt's own gold** rather than against the fact's original object, which the item set does not carry (A.1), so a true Neighborhood Score cannot be computed here. Its limits are the same either way: it is a property of one scoring rule, it saturates on editors that raise the target everywhere (F.3, F.5), and it depends on how the off-target prompts are chosen (F.4).

I.3 A PUBLISHED STATE-OF-THE-ART CLAIM, AUDITED FROM ITS OWN TABLE

A fair objection to this paper is that nobody asserts Pareto dominance, so the cross-probe reading is a straw man (§2.1 concedes the first half). The objection is answerable from published tables alone, without per-item data, and this subsection answers it once, on a claim the introduction quotes.

PMET reports "state-of-the-art performance on both the COUNTERFACT and zsRE datasets" (Li et al., 2024a). Its Table 1 gives, for GPT-J-6B under 10,000 CounterFact edits, the three probes its composite is built from, with 95% confidence intervals, as printed:

Table 44: PMET's own Table 1, GPT-J-6B, 10,000 CounterFact edits. Numbers as printed; the Score column is the harmonic mean of the other three, which we recomputed as 86.13 and 85.80.

probe	PMET	MEMIT	difference
Efficacy	99.5 (0.1)	98.9 (0.2)	$+0.6$
Generalization	92.8 (0.4)	88.6 (0.5)	$+4.2$
Specificity	71.4 (0.5)	73.7 (0.5)	$\mathbf{-2.3}$
Score (harmonic mean of the three)	**86.2**	**85.8**	$+0.4$

The composite lead is **0.4 points**, and the per-probe effects disagree in direction. Each of the three gaps exceeds the sum of the two printed half-widths, which is the conservative comparison because it treats two arms measured on the same items as independent, so the minority-direction effect is not noise on the paper's own reported precision. Under equal linear weights the difference is $+0.83$ points, and the ranking ties once specificity carries **0.51** of the weight with the other two left equal, or **0.68** if the two agreeing probes are collapsed into one block, which is the form the two-stratum formula of §2.1 takes. An earlier version of this subsection printed 0.32 here and called it the weight on specificity; 0.32 is the weight the *agreeing* block keeps at the tie, its complement on that collapsed scale. `published_table_audit.py` computes both and its self-test pins them apart. On zsRE, its other dataset, the same pair runs $+0.2$, $+0.9$ and $+0.1$, so there PMET does lead on every probe.

Three things follow, and we claim no more than these. First, the cross-probe reading is **not** a straw man: an abstract-level claim about two datasets holds across probes on one of them and fails on the other, and a reader cannot tell which from the composite. Second, the claimant says so itself, in a sentence a reader of the abstract does not see: "With the exception of being slightly inferior to MEMIT in terms of specificity, PMET outperforms all baselines in all other metrics." The information is in the paper; the ranking language is what loses it. Third, this is a **sign analysis, not a certification**: the release carries no per-item outcomes, so no reversal test can be run, the intervals are the paper's own, and PMET's composite is a harmonic mean, so Proposition 1's positive-linear-weighting form does not apply to it literally. What does apply is the dominance question, which is scale-free.

This is one worked example, chosen because the introduction quotes the claim, and it is not a survey of the literature. The same arithmetic on ROME's GPT-J table, which §7 already quotes, gives the same shape: unconstrained fine-tuning at 96.6 on paraphrases against a neighbourhood score of 10.3, where its constrained variant is at 47.9 and 78.6.

I.4 THE SAME AUDIT OVER FIVE PAPERS' OWN TABLES

I.3 answers the straw-man objection on one claim. A reader may still ask whether that claim is unusual, and the answer needs more than one table. `published_table_audit.py` (26 checks) applies the same reading to **nine tables from five papers** (Li et al., 2024a; Fang et al., 2025; Wang

et al., 2024a; Gu et al., 2026; Huang et al., 2026), chosen because each prints one row per method and one column per probe, which is what the audit needs and what a per-item release is not.

What the instrument does and does not do. It computes nothing about the papers: every value in it is transcribed from the table named in its entry, with the URL and the date it was read, and the instrument derives only differences, directions and weights from those transcriptions. A transcription error is the one defect its self-test cannot catch, so `--sources` prints every source for spot-checking and no table was included that could not be read as text. Bar charts were not estimated: one candidate paper reports its editing results only as rasterised panels and is therefore absent rather than approximated.

Table 45: Each row audits the proposed method against every baseline in that table. A "composite win" is a pair where the proposed method leads the aggregate; the last column counts those where at least one probe runs the other way.

paper	table	baselines	composite wins	with a probe disagreeing
PMET	CounterFact, GPT-J	1	1	**1**
PMET	zsRE, GPT-J	1	1	0
AlphaEdit	CounterFact, GPT-J	7	7	0
WISE	ZsRE, Llama-2, $T{=}1$	8	8	**3**
WISE	ZsRE, Llama-2, $T{=}1000$	8	8	**1**
MoEEdit	CounterFact, Qwen3-30B	4	4	0
MoEEdit	CounterFact, GPT-OSS-20B	4	4	**3**
MoEEdit	ZsRE, GPT-OSS-20B	4	4	**1**
JNO	RippleEdits, GPT-J	9	9	(scoped; see below)

The headline, restricted to the papers' own aggregates. Six of the nine tables print an aggregate column themselves (PMET's harmonic-mean Score, WISE's Avg., MoEEdit's Uti.). Over those, **9 of 29 composite wins have a probe pointing the other way.** The other three tables print no aggregate, and there the composite is ours rather than theirs, so they are excluded from that count even though the per-probe reading is unchanged.

A claim contradicted on the probe it names. MoEEdit's (Gu et al., 2026) abstract states that it "attains state-of-the-art efficacy and generalization while preserving high specificity". On Qwen3-30B-A3B that holds against all four baselines. On **GPT-OSS-20B**, three baselines lead it on generalization while it wins the printed Uti.: `FT` by 14.30 points, `AdaLoRA` by 10.90 and `UnKE` by 0.30, against printed half-widths that make the first two wider than the two intervals summed and the third not. This is the pattern of §1's worked pair in a second literature and on a second construct: the aggregate is won and the probe the claim names is lost.

A paper whose claim survives. AlphaEdit (Fang et al., 2025) leads every one of its seven baselines on all three probes, so its composite ordering is Pareto and nothing here qualifies it. This is reported because an audit that only ever convicts is not measuring anything.

A paper that draws the line itself. JNO's (Huang et al., 2026) abstract claims improvement "on propagation and preservation metrics", which are five of the six probes its table prints, and its own `Improve` row prints -3.0% on the sixth, Reliability, where `SAKE` and `AlphaEdit` lead it. Audited on the probes the claim names, JNO leads on all of them against every baseline; the probe it loses is outside the claim and the paper says so. **That is the practice this paper is asking for**, and it is worth naming as such: the scope is stated, the losing probe is printed, and a reader can see both. The instrument records the claimed subset separately from the full set for exactly this case.

What this sweep does not establish. Five papers are not a sample of the field, and they were not drawn at random: they were chosen for having a readable per-probe table, which selects for papers that report more rather than less. The counts are therefore a demonstration that the reading applies

to published claims at a non-trivial rate, not an estimate of that rate. Nor does any row here carry a significance test: two of the nine tables print no precision at all, and where precision is printed the comparison used is the conservative one of I.3, a gap wider than the two printed half-widths summed, which treats two arms measured on the same items as independent. Of the nine disagreements, **four** are in tables printing no precision at all, one more has a gap narrower than the two half-widths summed (`MoEEdit` against `UnKE`, 0.30 points), and the remaining **four** clear that bar. The table above does not distinguish them; the artifact does, per pair.

J EXPERIMENT 5: LEAKAGE VARIED WITHIN EDITOR FAMILIES

J.1 DESIGN

In Experiment 3 every high-leakage editor is a gradient editor and every low-leakage editor a locate-and-edit editor, so the association between corrected reversals and mismatched leakage is one family difference observed eleven times, and no re-analysis of those blocks can separate leakage from subject keying, edited layer, edit strength or efficacy. Experiment 5 was registered to vary leakage **within** each family. Its registration was frozen before any of its outputs existed and fixes the decision rule and the sentence each outcome supports.

Table 46: The registered 2×2. Held fixed: the backbone (GPT-J-6B), the item set, P0's 533 facts, the probes, formats, scoring rules, test, unit and the leakage measure with its A4 floor. Changed: a constraint on the gradient editor's update, and the key the locate-and-edit editor writes at. L1 is `ROME` with `fact_token: last` rather than `subject_last`, a value the published implementation supports, so the edit is keyed on the final token of the edit prompt rather than on the subject.

	high leakage	low leakage
gradient	`FT-M` at its published configuration (G0)	`FT-M` with a specificity constraint (G*, ours)
locate-and-edit	`ROME` keyed on the prompt's last token (L1, ours)	`ROME` at its published configuration (L0)

The predictions. If a reversal is produced by non-specific credit, corrected reversals appear in leakage-mismatched pairs whichever family the arms come from, including the two **within-family** pairs, and not in matched pairs. If it is produced by something else about the families, they appear in cross-family pairs whether or not leakage is matched, and not within a family.

How G* was chosen, before it ran, on facts that are never analysed. Two candidate constraints were registered, a KL anchor penalty and an L_∞ clamp, over a grid of six λ and five ε values. The rule: among candidates passing all five qualification witnesses, keeping at least 80% of G0's free-generation efficacy and leaking at most 0.30, take the lowest leak rate, ties by higher efficacy. Calibration ran on the twenty calibration facts of the screen's excluded set and the KL anchors come from the twenty holdout facts, so no analysed fact is edited, scored or used as an anchor in the step that chooses the arm. **Exactly one of the thirty candidates satisfied all three conditions** (clamp $\varepsilon = 10^{-4}$: leak 0.111 against G0's 1.000, efficacy 0.588 against a floor of 0.313). Two further settings, each its knob's candidate closest to a leak rate of 0.50, are kept for a descriptive dose curve and enter no family. The timing of amendment A9, which recorded these values, is in H.2.

Test, family and rule. The registered test is Gail and Simon's at $I = 2$ on fact-unit paired differences, with the intersection-union p twice it; both give the same rejections here. The primary family is the four arms' six pairs $\times$ two semantic cells **= twelve comparisons** at Bonferroni $0.05/12$, every one counted, including any that cannot be computed. No informativeness restriction enters the primary analysis; Experiment 3's arm-level rule is reported as a sensitivity (J.3). An arm is high-leakage at a rate ≥ 0.70 and low-leakage at ≤ 0.30; a pair is mismatched at a difference ≥ 0.40 and matched at ≤ 0.20, and indeterminate pairs enter no branch. The 2×2 is **formed** only if G* is low-leakage and L1 is high-leakage on P0. The outcomes are `NON_DIAGNOSTIC` (not formed, fewer than 400 facts, or a failed qualification), `NO_REVERSAL`, `LEAKAGE_TRACKS`, `FAMILY_TRACKS` and `MIXED`, each with the sentence it licenses fixed in advance; if the published pair G0 against L0 does not reject, the verdict is `NO_REVERSAL` whatever the other pairs do. Power was computed before the

run: at $n = 533$ the per-stratum minimum detectable effect is 3.4 to 5.1 points, against the 15-point effects Experiment 3 reports for that pair. A matched pair that does not reject is reported with its minority-direction bound, never as equivalence.

Registered secondaries. (i) **S**, seeded replication of `ROME`, `MEMIT` and `LoRA` at seeds $\{42, 1, 2\}$ (J.7). (ii) **C**, calibrated scoring: every corrected reversal with a constrained-choice stratum recomputed with a domain-conditional pointwise mutual information score (Holtzman et al., 2021), each option's length-normalised log-likelihood minus its log-likelihood under the null stem `"Answer:"`, classified PERSISTS, ATTENUATES or VANISHES (J.4). (iii) **Mi**, `FT-M` and `FT-multi` leakage on Mistral-7B-v0.1 (F.5). (iv) **U**, unedited per-cell baselines (J.6). (v) The dose curve, descriptive (J.5).

J.2 VERDICT

All four arms completed **533 of 533** facts; no fact was dropped for any arm; every arm clears the registered witnesses (max $|dW|$ 1.2×10^{-4} to 4.3×10^{-2}, changed fraction 0.80 to 1.00, mean target-NLL drop 3.53 to 5.13 nats against a floor of 0.5).

Table 47: Leak rates on P0's 533 facts after the A4 floor, and efficacy on the edit prompt. The unedited model scores 0.051 on the same efficacy measure.

arm	what it is	neighbourhood leak	free-generation efficacy
G0	`FT-M`, published	**97.6%**	0.423
G*	`FT-M` + L_∞ clamp 10^{-4} (ours)	**10.5%**	0.336
L0	`ROME`, published	**6.4%**	0.916
L1	`ROME`, `fact_token: last` (ours)	**47.5%**	0.871

The constraint did what it was chosen to do: G* is low-leakage at 10.5%, and it keeps 79.5% of G0's free-generation efficacy on P0, just under the 80% the calibration step required on the twenty calibration facts, where G* was the *more* effective arm (0.588 against 0.391). Keying `ROME` on the last token raised its leakage sevenfold, from 6.4% to 47.5%, but 47.5% is neither the $\geq 70\%$ the registration calls high-leakage nor the $\leq 30\%$ it calls low. **The 2×2 is therefore not formed, and the registered verdict is `NON_DIAGNOSTIC`.** The mechanism branches, and the sentences they would support, are not available. No arm failed qualification, and the population is 133 facts above the registered floor, so the unformed design is the only condition that applied. Figure 3 shows the design that was required beside the one the arms could form.

J.3 THE TWELVE REGISTERED COMPARISONS

Table 48: All twelve comparisons of the registered family, at Bonferroni $0.05/12 = 0.00417$; differences in percentage points at the fact unit, $n = 533$. A p of 0.5, the test's maximum at two strata, means the two strata agree in sign. Two comparisons reject; generation is scored by the registered token F1.

pair	cell	Gail and Simon p	d_{CC}	d_{gen}	leakage	family
L0 vs L1	Subject_Aliasing	3.9×10^{-13}	-9.29	$+21.12$	mismatched	**within**
G0 vs L0	Subject_Aliasing	4.3×10^{-13}	$+16.32$	-16.42	mismatched	cross
G0 vs L0	Reasoning	4.6×10^{-3}	$+5.72$	-4.00	mismatched	cross
G0 vs G*	Reasoning	0.273	$+13.51$	-0.88	mismatched	within
G0 vs L1	Reasoning	0.301	$+0.84$	-5.19	mismatched	cross
G0 vs G*	Subject_Aliasing	0.5	$+28.80$	$+25.64$	mismatched	within
G0 vs L1	Subject_Aliasing	0.5	$+7.04$	$+4.69$	mismatched	cross
G* vs L0	Subject_Aliasing	0.5	-12.48	-42.06	**matched**	cross
G* vs L0	Reasoning	0.5	-7.79	-3.12	**matched**	cross
G* vs L1	Subject_Aliasing	0.5	-21.76	-20.95	indeterminate	cross
G* vs L1	Reasoning	0.5	-12.66	-4.31	indeterminate	cross
L0 vs L1	Reasoning	0.5	-4.88	-1.19	mismatched	within

The two rejections. The first is the **published pair**: `FT-M` against `ROME` on `Subject_Aliasing` reproduces Experiment 3's reversal on a fresh run at $+16.3$ points under constrained choice against

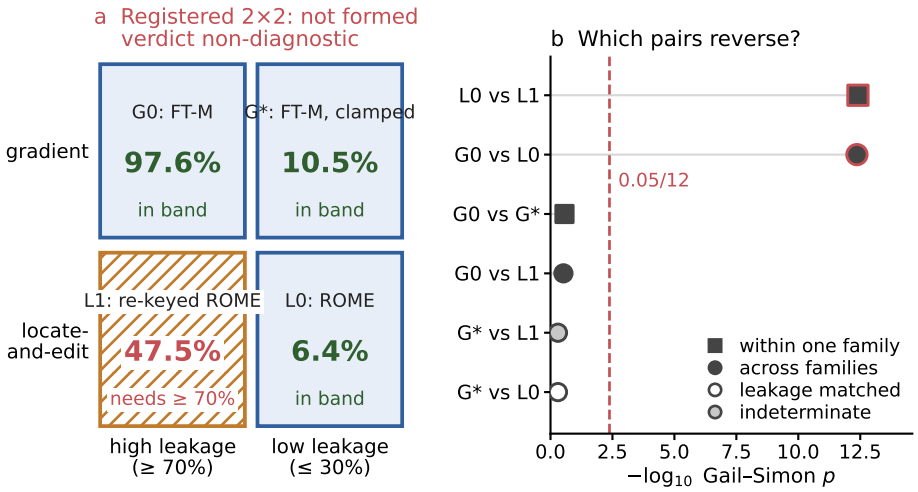

Figure 3: Experiment 5's registered design was not formed. (a) The 2×2 the registration required, with each arm's neighbourhood leakage. The constraint moved FT-M into the low-leakage band ($\leq 30\%$); re-keying ROME on the last token reached 47.5%, below the high-leakage threshold ($\geq 70\%$), so the locate-and-edit high-leakage cell is empty and the registered verdict is non-diagnostic. **(b)** The six editor pairs, each at its strongest of the twelve comparisons, on $-\log_{10}$ of Gail and Simon's p; the dashed line is the primary threshold $0.05/12$. Squares are pairs within one family, circles across families; fill gives the registered leakage class. Two pairs reject, one across families (the published FT-M against ROME) and one within ROME. The within-family gradient pair, mismatched in leakage, does not reject.

-16.4 under free generation. The second is ROME against **ROME keyed on the last token**: the same editor, layer and hyper-parameters, differing only in which token the update is keyed on, and differing in leakage by 41 points. This is a corrected reversal inside one editor family, so editor family alone does not describe when reversals occur. Re-keying an edit on the prompt's last token changes what the edit is about as well as how far it spreads, so this pair separates leakage from *editor family* without separating it from *what was edited*; it does not isolate specificity as the operative property, and no design in this paper does.

The other comparisons. The registered leakage-matched pair, G* against L0 at 10.5% and 6.4%, does not reject in either cell, but G* keeps 0.336 free-generation efficacy against L0's 0.916 and loses both strata by 12 to 42 points, so an arm weaker everywhere could hardly reverse and its non-rejection says little about leakage (J.5). The within-family gradient pair G0 against G*, mismatched in leakage, does not reject ($p = 0.273$ and 0.5), contrary to what a leakage account predicts, although G* is again the weakest arm. The two rejections meet the condition of the LEAKAGE_TRACKS branch, but that branch is available only if the 2×2 is formed, so it is not the verdict.

The registered sensitivity. Experiment 3's arm-level rule keeps **9 of the 12** comparisons; the three it drops are those in which FT-M sits at 0.998 under constrained choice in Subject_Aliasing, the same near-ceiling cell that carried Experiment 3's measurement rejections. At the registered family of twelve, **one** of them still rejects: ROME against ROME keyed on the last token in Subject_Aliasing ($p = 3.9 \times 10^{-13}$); FT-M against ROME in Reasoning ($p = 4.6 \times 10^{-3}$) rejects only at the family of 9 the restriction keeps, which B.3 does not use. The within-family reversal therefore does not depend on a saturated arm.

J.4 Secondary C: rescoring against a null prompt

Each corrected reversal was recomputed with the constrained-choice score replaced by the domain-conditional PMI score, leaving free generation unchanged, and classified by a rule fixed before the run: PERSISTS (opposite signs, still rejecting at the reversal's own registered threshold), ATTENUATES (opposite signs, no rejection) or VANISHES (signs agree).

Table 49: Secondary C on Experiment 5's two rejections. Percentage points at the fact unit, $n = 533$.

comparison	raw d_{CC} / d_{gen}	calibrated	calibrated p	class
G0 vs L0 @ `Subject_Aliasing`	$+16.32$ / -16.42	-3.38 / -16.42	0.5	**VANISHES**
L0 vs L1 @ `Subject_Aliasing`	-9.29 / $+21.12$	-4.78 / $+21.12$	1.4×10^{-5}	**PERSISTS**

The published pair's constrained-choice advantage falls from $+16.3$ points to -3.4 and then agrees in sign with free generation, so no reversal remains. The within-family reversal persists. The rescoring is one operationalisation of credit that does not depend on the question, it decides nothing in any primary analysis by registration, and a single alternative scoring rule does not establish a mechanism.

What calibration does to arm levels (post hoc). A rescoring that lowered every edited arm in proportion to its edit strength would remove any advantage automatically. The calibrated levels show something narrower:

Table 50: Constrained-choice accuracy before and after null-prompt calibration, on Experiment 5's blocks for P0's 533 facts. Leak is the neighbourhood leak rate; leak rates are from Experiment 3's locality run for FT-L, LoRA and MEMIT and from Experiment 5 otherwise.

arm	leak	restatement, raw → calibrated	consequence, raw → calibrated
unedited		.366 → .498	.251 → .584
ROME (L0)	6.4%	.835 → .902	.448 → .718
ROME, last token (L1)	47.5%	.928 → .950	.497 → .694
MEMIT	0.1%	.715 → .810	.389 → .665
FT-L	90.2%	.992 → .863	.492 → .536
FT-M (G0)	97.6%	.998 → .869	.506 → .454
LoRA	99.9%	.995 → .560	.541 → .309
FT-M, clamp 3×10^{-4} (efficacy-matched)	31.1%	.917 → .799	.455 → .674
FT-M, KL $\lambda = 30$	47.9%	.917 → .811	.474 → .667
FT-M, clamp 10^{-4} (G*)	10.5%	.710 → .734	.371 → .629

On restatement, calibration raises the unedited model and every locate-and-edit arm, including ROME re-keyed to 47.5% leakage, and lowers every gradient arm except the weakest, G*: by 11 to 13 points for FT-L, FT-M and the two intermediate FT-M arms at 31.1% and 47.9% leakage, and by 44 points for LoRA. It therefore does not penalise strong edits in general, but on these arms it tracks editor family at least as closely as leakage. It is not an accuracy scale anchored at chance, however: the unedited model scores 0.584 on the consequence probe after calibration, against a chance level of 0.25, and LoRA's restatement score falls to 0.560, only 6 points above the unedited model's 0.498, which is the overshoot visible in the LoRA row below.

Why the unedited model gains 33 points, measured rather than assumed. A reader asked whether that gain is knowledge or an artefact of how the options were built; it is the latter, and the size of it can be read off the stored log-likelihoods. Both terms of the score are length-normalised in the same way (the mean log-probability per option token, A.2), so the difference is not a length effect. On the 1,066 unedited consequence items the gold option is **1.21 nats less likely than the average distractor under the null stem alone** (-8.66 against -7.44), while under the question it leads by only 0.29 nats. The distractors are other facts' answers for the same relation, which are commoner strings than a counterfactual consequence, so the raw rule charges the gold a prior handicap and the calibrated rule removes it. On restatement the same quantities are -0.17 and $+0.93$ nats, which is why calibration moves the unedited model by 13 points there and 33 points here. The calibrated score therefore measures the question's evidence for the gold net of the option's unconditional plausibility; it does not measure whether the model knows the consequence, and a rise in it is not evidence of knowledge. We report it as a re-expression of leakage, and the main text says so.

Experiment 3's reversals. Secondary C covers every corrected reversal of the registered analysis that involves a constrained-choice stratum. Experiment 3's blocks carry no calibrated score, because

the null-stem pass did not exist when they ran; Experiment 5 re-ran all five editors on the same 533 facts, frozen item set and scoring rules, so the recomputation runs on a **fresh run of the same design**. The raw recomputation is reported beside the calibrated one, because a reversal that does not replicate raw on these blocks cannot be said to vanish under rescoring.

Table 51: Secondary C on the ten reversals of the registered analysis that involve a constrained-choice stratum, recomputed on Experiment 5's replication blocks, $n = 533$ (355 in the true-consequence subset). Percentage points at the fact unit. The measurement rows are in `Subject_Aliasing` unless named, and the split rows in the true-consequence subset. For a measurement comparison the two strata are constrained choice and free generation; for an answer-identity split they are the two semantic sub-labels, both under constrained choice. Each row is classified against its own registered family, $0.05/40$ for the measurement contrast and $0.05/80$ for the split; the last row is the arm-level reversal at the resized family of 22, which is not a corrected reversal at the full family (B.3). "Not replicated" means the raw comparison did not reject on these blocks, so no rescoring claim is made about it. Figure 4 plots the same ten rescorings.

reversal	raw s1 / s2	p raw	calib. s1 / s2	p calibrated	class
FT-L vs ROME (meas.)	$+15.76$ / -16.00	2.6×10^{-13}	-3.94 / -16.00	0.5	**VANISHES**
FT-M vs ROME (meas.)	$+16.32$ / -16.42	4.3×10^{-13}	-3.38 / -16.42	0.5	**VANISHES**
LoRA vs ROME (meas.)	$+16.07$ / -14.44	1.7×10^{-10}	-34.21 / -14.44	0.5	**VANISHES**
FT-L vs ROME (split)	$+15.63$ / -11.69	4.8×10^{-10}	-3.80 / -30.00	0.5	**VANISHES**
FT-L vs MEMIT (split)	$+26.76$ / -11.27	2.8×10^{-9}	$+5.35$ / -30.00	6.6×10^{-3}	ATTENUATES
FT-M vs MEMIT (split)	$+26.90$ / -10.56	6.3×10^{-9}	$+6.48$ / -41.83	1.1×10^{-3}	ATTENUATES
FT-M vs ROME (split)	$+15.77$ / -10.99	8.6×10^{-9}	-2.68 / -41.83	0.5	**VANISHES**
LoRA vs ROME (split)	$+15.77$ / -5.49	6.8×10^{-3}	-32.82 / -48.59	0.5	not replicated
LoRA vs MEMIT (split)	$+26.90$ / -5.07	9.9×10^{-3}	-23.66 / -48.59	0.5	not replicated
FT-M vs ROME (meas., `Reasoning`)	$+5.72$ / -4.00	4.6×10^{-3}	-26.36 / -4.00	0.5	not replicated

Intervals behind VANISHES. A sign call on two point estimates does not say whether the sign is resolved. Each calibrated difference carries a simultaneous interval at its own row's threshold (`C_calibrated_exp3[*].calibrated_ci_pp`): for the three measurement reversals the constrained-choice difference falls to -3.9 $[-8.9, +1.0]$, -3.4 $[-8.2, +1.4]$ and -34.2 $[-41.6, -26.9]$ points. In two of the three, VANISHES therefore means the advantage is **no longer distinguishable from zero**, not that it has turned negative; only for `LoRA` against `ROME` is the calibrated difference resolved and negative in both strata.

Both rescorings together (post hoc). With constrained choice calibrated and generation scored by first-span match, FT-L and FT-M against ROME keep opposite signs without rejecting (-3.9 against $+7.9$, and -3.4 against $+21.9$; $p = 0.005$ and 0.011). LoRA against ROME reverses in the other direction (-34.2 under calibrated choice against $+23.7$ under first-span match, $p < 10^{-15}$): ROME leads under the calibrated score and LoRA under the first-span one. That reversal is a product of calibration's overshoot for LoRA meeting a generation score that credits LoRA's first answer, and it is reported as such rather than as a finding about the editors.

Summary. Of the seven reversals at the full families, five vanish, two attenuate, and none persists; three further comparisons did not replicate raw. **All three primary measurement reversals vanish.** What this establishes is that the constrained-choice advantage these reversals are built from does not survive removing question-independent credit under this operationalisation, which the leakage measure (F.3) also measures; it does not establish that such credit is their unique cause.

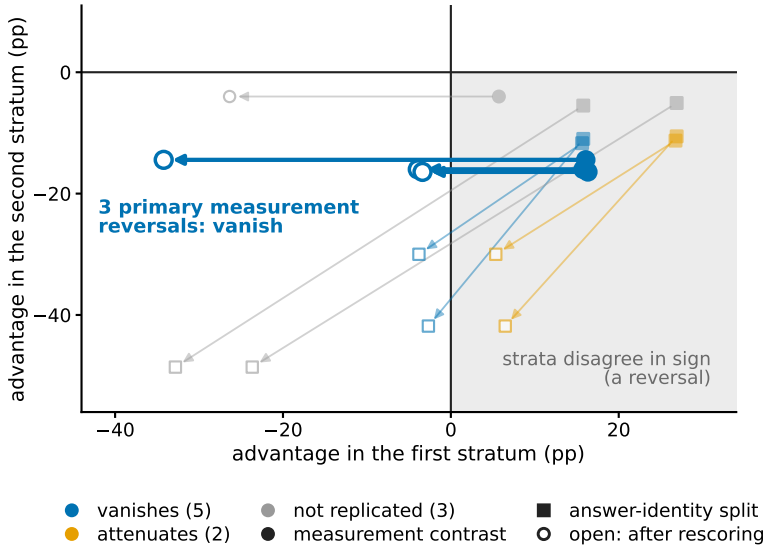

Figure 4: Rescoring against a null prompt, for all ten reversals it covers. Each arrow runs from a comparison's raw position (filled) to its position after constrained choice is rescored against the null prompt (open). The axes are the comparison's two strata: the two response formats for a measurement contrast, the two semantic sub-labels for an answer-identity split. Shaded: the strata disagree in sign, so a point there is a reversal. The three primary measurement reversals (large, dark arrows) all leave the shaded region. Of the other seven, two leave it, two stay without rejecting (attenuate) and three did not replicate raw (grey). The rescoring is one operationalisation of question-independent credit: this is a classification, not a demonstration of cause.

J.5 The dose arms and the efficacy-matched comparison (post hoc)

Two further settings were kept because their calibration leak rate was closest to 0.50 on each knob; their levels on P0 are below.

Table 52: The dose curve on P0's 533 facts. Efficacy is trimmed token F1 on the edit prompt against the edit target (A.2). No test enters a registered family for these arms, and the registration forbids reporting them as a primary result.

arm	neighbourhood leak	free-generation efficacy
G0, `FT-M` published	97.6%	0.423
`FT-M-KL`, $\lambda = 30$	47.9%	0.689
`FT-M-clamp`, $\varepsilon = 3 \times 10^{-4}$	31.1%	**0.919**
G*, `FT-M-clamp`, $\varepsilon = 10^{-4}$	10.5%	0.336

Leakage falls monotonically as the clamp tightens. Efficacy does not trade off against it monotonically: the intermediate clamp reaches 31.1% leakage **and** 0.919 efficacy, more than twice G0's 0.423, while the tighter clamp the selection rule chose reaches 10.5% at an efficacy of 0.336. The registered rule chose the arm that is worse on efficacy on the analysed population, because it took the lowest leak rate among candidates clearing an efficacy floor measured on twenty calibration facts, where this arm scored 0.588 against G0's 0.391; twenty facts is a small set on which to fix a floor. The arm was not re-selected after seeing P0, which the registration forbids.

The efficacy-matched comparison. The registered matched pair is confounded with edit strength (J.3). The dose curve contains an arm matched to L0 on efficacy: `FT-M-clamp` at $\varepsilon = 3 \times 10^{-4}$, whose efficacy is **0.919** against L0's **0.916**.

Table 53: Each gradient arm against L0 (ROME, leak 6.4%, efficacy 0.916) in Subject_Aliasing, on P0's 533 facts, post hoc and in no registered family; Bonferroni over the twelve comparisons of the dose analysis is $0.05/12$. Differences in percentage points, constrained choice / generation.

gradient arm	leak	efficacy	raw choice / token F1	raw choice / first-span match	calibrated choice / token F1
G*, clamp 10^{-4}	10.5%	0.336	-12.48 / -42.06, $p = 0.5$	-12.48 / -41.37, $p = 0.5$	-16.89 / -42.06, $p = 0.5$
clamp 3×10^{-4}	31.1%	**0.919**	$+8.16$ / -24.88, $p = 2.1 \times 10^{-9}$	$+8.16$ / -24.02, $p = 2.1 \times 10^{-9}$	-10.32 / -24.88, $p = 0.5$
KL $\lambda = 30$	47.9%	0.689	$+8.16$ / -13.50, $p = 1.2 \times 10^{-7}$	$+8.16$ / -4.78, $p = 0.020$	-9.10 / -13.50, $p = 0.5$
G0, FT-M published	97.6%	0.423	$+16.32$ / -16.42, $p = 4.3 \times 10^{-13}$	$+16.32$ / $+21.86$, $p = 0.5$	-3.38 / -16.42, $p = 0.5$

At matched efficacy the reversal is present under the registered scores: absent at 10.5% leakage, present at 31.1%, 47.9% and 97.6%. Neither of its arms is within 0.05 of ceiling or chance, and it **survives first-span scoring** ($+8.2$ against -24.0): the efficacy-matched arm's generation deficit is not repetition. It **does not survive calibration**, which turns its constrained-choice advantage into a deficit of 10.3 points. Three further limits apply. The evidence does not strengthen as leakage rises. Efficacy is matched to L0 at one arm only and varies across the rest (0.336, 0.919, 0.689, 0.423), so the series is not a leakage dose-response, and the single arm that does not reverse, G*, is also the weakest. And every comparison here is a gradient arm against ROME, so the table does not bear on whether a reversal needs two different editor families. The two dose arms share a d_{CC} of $+8.16$ because both score exactly 0.916510 on Subject_Aliasing under constrained choice, against L0's 0.834897. The efficacy-matched arm's leakage gap against L0 is 24.7 points, which the registered bands call **indeterminate**, so this comparison would have entered no branch even had the 2×2 formed.

ROME against re-keyed ROME persists under every combination: raw or calibrated constrained choice against token F1, first-span match or first-span F1 (p from 3.9×10^{-13} to 1.4×10^{-5}). A registered series of efficacy-matched arms spanning leakage within one editor, which would test the leakage account directly, was not run (§8).

J.6 Secondary U: unedited baselines

Table 54: Secondary U, the unedited model's mean in each analysed cell, generation by the registered token F1. Chance is 0.25 under constrained choice and 0 under free generation.

population	aliasing CC	aliasing gen	reasoning CC	reasoning gen
P0 (GPT-J-6B, 533)	0.366	0.085	0.251	0.126
P0 / true_consequence (355)	0.354	0.076	**0.230**	0.144
P0 / same_answer (178)	0.390	0.102	0.295	0.090
Mistral-7B (357)	0.464	0.120	0.444	0.160
Qwen2.5-7B-Instruct (533)	0.407	0.070	0.345	0.142

On P0 the near-ceiling Subject_Aliasing constrained-choice arms are a real edit effect, from 0.366 unedited to 0.995 and above. The Reasoning constrained-choice cell is at chance whichever population is used: 0.251 over all 533 facts, and **0.230** over the 355 true-consequence facts on which the split's arms are measured. At that matching population the gradient arms (0.249–0.293) sit 1.9 to 6.3 points above an unedited baseline that is itself below the 0.25 chance level, so the cell carries no graded effect either way, which is why the arm-level restriction excludes it. On Qwen, at matching populations, the same cell is 0.345 unedited against FT-M's 0.463 on all 532 analysed facts, and the 0.989 that makes the semantic comparisons ceiling/chance-adjacent is on the 178 same_answer facts, so the saturation there is an edit effect on a subset rather than a set of trivially answerable items.

J.7 SECONDARY S: SEED REPLICATION

ROME, MEMIT and LoRA, the three editors whose stochastic component is unseeded in Experiment 3, were re-run on all 533 facts at seeds $\{42, 1, 2\}$; FT−L and FT−M have no seed to vary and contribute one block each. They are not bit-identical across runs: against Experiment 3's blocks **97 of 6,396** item rows differ for FT−L and **71 of 6,396** for FT−M, 1.5% and 1.1%, concentrated in the two free-form formats (50 and 23 in generation, 42 and 44 in yes/no, 5 and 4 in constrained choice). FT−L − FT−M on Subject_Aliasing under free generation is therefore $+0.42$ points on Experiment 5's blocks against $+0.55$ on Experiment 3's. "Deterministic" in the registration means no seeded stochastic component, not reproducible to the item, and this run-to-run variability is small beside every effect the paper reports. A reversal is SEED-ROBUST if it rejects at its own registered family in at least two of the three seeds with the same sign pattern.

Table 55: Secondary S on P0's 533 facts, Gail and Simon's test. Nine of the eleven corrected reversals at the full families are SEED-ROBUST; the twelfth row is not a corrected reversal at the full family (B.3).

reversals (registered analysis)	seeds rejecting	verdict
all **3** measurement reversals (FT−L, FT−M, LoRA against ROME in Subject_Aliasing)	3 of 3	SEED-ROBUST
4 split reversals at constrained choice (FT−L/ROME, FT−L/MEMIT, FT−M/MEMIT, FT−M/ROME)	3 of 3	SEED-ROBUST
2 split reversals at generation (FT−L/MEMIT, same-answer LoRA/ROME)	2 of 3	SEED-ROBUST
2 split reversals at constrained choice (LoRA/ROME, LoRA/MEMIT)	0 of 3	SEED-FRAGILE
FT−M against ROME in Reasoning, arm-level restriction at the resized family of 22	1 of 3	SEED-FRAGILE

The two seed-fragile splits keep **opposite signs at every seed**; what moves is the p-value's position relative to a nearby threshold (0.0007 to 0.0099 against 6.25×10^{-4}). They are exactly the two split reversals that the intersection-union test does not certify, so under that test all nine of its reversals are seed-robust, and a statement of that kind reflects the choice of test as much as the stability of the effects.

The criterion is lenient. Seed 42 reproduces Experiment 3's own template draw, so for a ROME or MEMIT pair a SEED-ROBUST call at two of three needs only **one** new seed to agree with the original run. Requiring both *new* seeds would move one generation split, same-answer LoRA against ROME, into the fragile column and leave eight of eleven robust; FT−L against MEMIT at generation rejects at both new seeds and not at seed 42.

The run-level component is not identified at this n. Across all twelve arm × cell combinations the between-seed variance of the cell means never exceeds the within-cell variance divided by n, so τ^2 is bounded at zero. The between-seed standard deviation of the cell means is at most **0.90** points (median 0.42), consistent with the **0.79**-point estimate of the earlier 300-fact sweep (C.4) but not separable from sampling noise here. The sensitivity that imposes a run-level error term (C.4) remains the way the paper prices it.

K EXPERIMENT 4: A PROSPECTIVE TEST OF THE AUDIT ON QWEN2.5-7B-INSTRUCT

The controlled evidence stands on base models from 2021 and 2023, with the factorial and the specificity test on GPT-J-6B alone. Experiment 4 asks whether the audit's checks behave the same way on a contemporary instruction-tuned backbone. Its registration was frozen before any output of any backbone existed; its amendment M4 (amendment labels M1 to M4 here are the registration's own and unrelated to the attribution checks) records that the analysis instrument implementing its sections 8 to 11 was written, self-tested and hashed while the jobs were queued and before any output existed.

Backbone and pool. The registered backbone was `Meta-Llama-3-8B-Instruct`, whose licence must be accepted by a person; amendment M1, written before any output existed, substitutes **`Qwen2.5-7B-Instruct`**: ungated, contemporary, of the registered size, with published EasyEdit configurations for `FT`, `LoRA` and `ROME`. `MEMIT` and `AlphaEdit` are **excluded**, because their published second-moment statistics are not distributed for this backbone and section 5 forbids computing our own. The pool is three editors, three pairs, six comparisons per pair, a family of **18**, Bonferroni $0.05/18$.

LoRA's configuration. Amendment M2 ran `LoRA` at both EasyEdit's published Qwen configuration (rank 8, alpha 32, dropout 0.1, `q_proj` and `v_proj`, 60 steps, lr 5×10^{-3}) and the configuration Experiments 1–3 used, with the published one deciding unless it failed section 6's divergence rule. It failed it: the published arm's loss rose on **111 of 533 facts (20.8%)**, against a cap of 5%, with the same convergence witness that excluded `LoRA` from Experiment 2 as deviation D9, and it diverged at the same published learning rate on GPT-J. Our configuration failed on **1** fact (0.19%). The pool therefore uses ours, recorded as a deviation (H.2); the published block ships beside it, and no arm was re-tuned. That block records the published modules, but the code that produced it most likely trained `up_proj` and `down_proj` instead (A.4), so its divergence rate is not evidence about EasyEdit's published configuration. The analysed population is the **532** facts every pooled arm edited, above the registered floor of 300.

The verdict is `NON_DIAGNOSTIC`, because of ceiling and chance arms. Section 11's first clause removes a run in which more than half of either contrast's comparisons are uninformative, in this paper's terms ceiling/chance-adjacent. Nine of the twelve semantic comparisons are, and so are three of the six measurement ones. The arm levels show why, and they fail at **both** ends:

Table 56: Arm means on Qwen2.5-7B-Instruct. The two contrasts run on different populations: the measurement contrast on all 532 analysed facts, the semantic contrast inside each answer-identity subset, so the same arm has three levels in the `Reasoning` cell. Bold marks an arm outside $[\text{chance} + 0.05,\ 0.95]$, the informativeness window fixed in advance. Chance is 0.25 under constrained choice and 0 under free generation.

arm	population	`S_A` CC	`Reasoning` CC	`S_A` gen	`Reasoning` gen
`FT-M`	all 532, measurement	**0.995**	0.463	0.377	0.129
`LoRA` (ours)	all 532	**0.996**	0.507	0.466	0.168
`ROME`	all 532	0.730	0.449	0.289	0.189
`FT-M`	`same_answer`, 178	**0.992**	**0.989**	0.359	0.343
`LoRA` (ours)	`same_answer`, 178	**0.994**	**0.992**	0.461	0.422
`ROME`	`same_answer`, 178	0.716	0.458	0.321	0.119
`FT-M`	`true_cons.`, 354	**0.997**	**0.199**	0.386	**0.021**
`LoRA` (ours)	`true_cons.`, 354	**0.997**	**0.263**	0.469	**0.041**
`ROME`	`true_cons.`, 354	0.737	0.445	0.273	0.223

Both gradient arms sit within half a point of the ceiling on `Subject_Aliasing` under constrained choice, which makes all three measurement comparisons in that cell ceiling/chance-adjacent; the three in `Reasoning` are scored, because on the pooled population both arms sit at 0.46 to 0.51 there. Inside the answer-identity subsets the same two arms reach 0.989 to 0.992 on `Reasoning` under constrained choice and fall to 2–4% on it under free generation, which makes nine of the twelve semantic comparisons ceiling/chance-adjacent. A stratum difference measured between a saturated arm and a collapsed one is bounded by the truncation rather than by edit quality, and the registered rule declines to read a ranking from it. **No measurement comparison rejects** (Gail and Simon's $p = 0.5$, 0.284 and 0.125; the intersection-union p is twice each). The per-stratum minimum detectable effect is 6.8 points at $\alpha = .05$ and, at the thresholds actually applied, **8.8 points** under Gail and Simon's test and **9.3** under the intersection-union test (B.5). Both exceed the registered cap of 8, so the run was underpowered as well as non-diagnostic; the verdict does not change.

The registered sentence for this outcome. *On a contemporary instruction-tuned backbone with this pool the registered audit could not be applied, because more than half of the semantic comparisons have an arm at ceiling or at chance; this experiment neither supports nor contradicts the paper's findings.*

What this means for the paper's claims. The audit behaved as designed, returning non-diagnostic rather than a verdict built on saturated arms, and the GPT-J pattern is absent rather than merely

unreadable: ROME does not beat the gradient arms on Subject_Aliasing under free generation (0.289 against 0.377 and 0.466). The reversals of §4 and §5 should not be assumed to hold for contemporary instruction-tuned models.

Reported whatever the verdict (section 11). Per-editor leakage on the neighbourhood family has the same structure as on GPT-J: FT-M **100.0%**, LoRA (ours) **97.8%**, ROME **0.29%**, with FT-M at a per-fact rate of exactly 1 on 99.8% of facts and ROME at exactly 0 on 74.8%. The simultaneous minority-direction bound over every pair leaves reversals of up to **8.8** points unexcluded, so the null here is not an equivalence. Two of the three pairs have opposite-signed strata, which is one ranking event rather than two independent ones. No corrected reversal exists to classify under the calibrated constrained-choice score (the registration's secondary RQ5, the same null-stem rule as Experiment 5's secondary C), and the pooled-label and cell-mean counts are in exp4_result.json with every exclusion and every lost fact.

L THE AUDIT OUTSIDE KNOWLEDGE EDITING: HELM MMLU

The checks are properties of an evaluation, not of knowledge editing, and the audit takes long-format per-item outcomes from any benchmark. This section runs it, unchanged, on an ordinary language-model leaderboard (post hoc). It shows the instrument running on another benchmark and what that benchmark's design can resolve; it is not evidence about MMLU rankings.

The data. HELM's classic suite at release **v0.3.0** (Liang et al., 2023), which publishes a per-instance record for every run. We take MMLU (Hendrycks et al., 2021) as HELM runs it: the five subjects it covers with these settings (abstract_algebra, college_chemistry, computer_security, econometrics, us_foreign_policy), **567 instances**, about 100 per subject, four-way multiple choice, so the chance level is 0.25 and is written into every row. fetch_general_eval.py (15 checks) downloads **140 pinned runs** and records each file's storage generation number, serving URL and sha256 in a manifest; the build refuses if any hash moves, if a run is absent, if a prediction carries no exact_match, or if a score leaves $[0, 1]$. Here the **systems** are language models and the **item** is a benchmark instance. Two families of **probe** strata are scored on the same items: three HELM perturbations (identity, a typo-and-synonym mix, a dialect rewrite) and three answer-format adaptations (joint multiple choice, separate, separate-calibrated). The five published **subjects** are content subgroups, disjoint by construction, declared to the tool as such (--strata-are subgroup).

Table 57: probe_consistency_general.json (tool version 5.0), 330 pair verdicts on HELM classic v0.3.0, with verdicts on all prespecified strata (B.3): qualitative failure (QF), magnitude heterogeneous (MH), no detected heterogeneity (NDH). Subjects are tested as independent subgroups with Gail and Simon's test, the perturbation and answer-format strata with the intersection-union test. Families are every pair in the section: 45, 15, 45 and 45 per subject. Each section's command and input sha256 are in the artifact, which also retains a subject_as_probe section in which subjects are treated as probe strata and flagged unpaired.

analysis	strata	pairs	QF	MH	NDH	flagged M2	median / largest bound
perturbations	3 perturbations	45	**0**	9	36	30	0.0 / 9.3 pp
answer formats	3 adaptations	15	**0**	5	10	14	5.8 / 8.1 pp
subjects as subgroups	5 subjects	45	**0**	28	17	44	16.6 / 29.5 pp
perturbations within each subject	3 perturbations	225	**0**	12	213	170	10.6 / 33.3 pp

No corrected qualitative failure appears anywhere. The smallest reversal-test p is 0.055 in the within-subject analysis and 0.072 across subjects, both against a Bonferroni threshold of 0.0011, and 0.35 across answer formats against 0.0033.

What the design can resolve. With about 100 items per subject and a family of 45 pairs, a subject-level reversal must reach a one-sided threshold of $0.05/90$ in a stratum of 100 binary items to be certified or excluded. Across subjects, the median minority-direction bound is **16.6 points** and the largest **29.5** (mpt-30b over gpt-3.5-turbo-0613 in college_chemistry); no pair excludes a subject-level reversal outright. A leaderboard gap of a few points is therefore compatible

with a double-digit reversal in a named subject, which describes the resolution of five 100-item subjects under a 45-pair correction, not MMLU. Of the 28 magnitude-heterogeneous pairs across subjects, 27 remain heterogeneous on the log-odds scale and one is additive only. 23 of the 45 pairs show a raw sign disagreement across subjects: `llama-2-70b` over `pythia-12b` is $+66.7$ points on `us_foreign_policy` and -0.9 on `abstract_algebra`.

Attribution flags. Arms near the 0.25 chance level dominate: 258 of the 330 verdicts carry an M2 flag, 44 of 45 across subjects and 170 of 225 within them, most often for `abstract_algebra`. The largest raw format reversal in the demonstration is one of them: `bloom` against `opt-175b` runs -2.35 points under joint multiple choice and $+6.35$ under separate-calibrated scoring **on the identical 567 items**, but `bloom`'s joint mean is within 0.05 of chance, the pair shows no detected heterogeneity on all three strata, and the diagnostic on the two remaining strata does not reject. **18 of 45** pairs in the perturbation analysis and **2 of 15** in the answer-format analysis are label-dependent: the verdict moves when the subject label is split out, which with 170 flagged within-subject verdicts mostly records thin strata near chance.

Limits. The perturbation strata are a crossed subset rather than fully paired (`identity` and the typo mix cover all 567 instances, the dialect rewrite 502), so those tests run on the 502 crossed items with each stratum's minimum detectable effect recorded; two further perturbations fire on 4 and 1 instances and are filtered out by a recorded flag. HELM classic v0.3.0 predates current frontier models and covers 5 of MMLU's 57 subjects with these settings.

M EXPERIMENT 6: THE PUBLISHED CONFIGURATIONS

Three of the five editors are our code, and two depart from the configuration they are named after (A.4). Two results depend on how those arms behave: the generation deficit of E.1, which token F1 charges partly to continuations dominated by one repeated word, and the target-as-gold account of F, which rests on gradient arms placing the edit target above an unrelated subject's own answer on nearly every neighbourhood prompt. If repetition and saturation belong to our arms rather than to the methods, both describe editors nobody runs. Experiment 6 runs the published configurations unchanged on the same 533 facts.

It is registered in `PREREG_EXPERIMENT6_FAITHFUL_2026-09-18.md`, frozen with its hash chain before any arm was fitted, and analysed once by `analyze_exp6.py`, written against that registration before a block existed. Everything but the three configurations is inherited verbatim: backbone and revision, the frozen item set, the population P0, the probes, both formats and their scoring rules, the null-prompt pass, the leakage measure with its A4 floor, the fact unit and the reversal test. Reference arms are Experiment 5's seed-42 blocks, so every comparison here is across two runs of one design and the run-level component of C.4 applies to all of them.

M.1 THE REGISTERED FAMILY, AND THE ARM THAT COULD NOT RUN

Table 58: Experiment 6's three arms and the published configuration each one restores.

arm	configuration	source
`FT-L-EasyEdit`	layer 21, 25 steps, lr 5×10^{-4}, objective `prompt_last`, no constraint	EasyEdit's GPT-J FT configuration
`FT+L`	layer 0, 25 steps, lr 5×10^{-4}, L_∞ $\varepsilon = 5 \times 10^{-5}$, stop at loss 0.03	ROME's constrained fine-tuning baseline
`AdaLoRA`	rank 8, α 32, dropout 0.1, 50 steps, lr 5×10^{-3}, no clipping	EasyEdit's published GPT-J LoRA configuration

The registered family is **twelve comparisons**: three arms $\times$ two reference editors (`ROME`, `MEMIT`) $\times$ two semantic cells, on the measurement contrast, at Bonferroni $0.05/12$. Generation is scored by the registered token F1; first-span match, first-span F1 and alias containment are registered sensitivities.

`AdaLoRA` produced no block. Each arm must first pass the five witnesses of A.3 on the twenty calibration facts, which are disjoint from P0. At EasyEdit's published learning rate the arm raised

the runtime convergence witness on **20 of 20** of them; on the first, the edit loss moved 9.44, 7.06, then 572.0. The registration forbids re-tuning a diverging arm and bars a failing one from running, so its four comparisons are recorded as **not computed**, inside the family of twelve and at the same threshold, rather than the family being rescoped to eight with a looser one.

Two things follow, and only two. Deviation D8's stated reason is **correct**: the paper's `LoRA` arm runs at a fiftieth of the published rate because that rate diverges here, and it diverges with AdaLoRA's own parameterisation, not only with the plain-LoRA substitute; Experiment 4 saw the same on the other backbone, where the published Qwen LoRA configuration diverged on 20.8% of facts. And this is a limit, not a licence: **no result in this paper describes EasyEdit's published LoRA editor**, and Experiment 6 cannot supply one. Whether that configuration diverges inside EasyEdit's own code, dtype and loop is not a question this pipeline can answer.

The configuration had to be corrected before it could run at all, recorded as §7 of the registration. The frozen text read the published `rank: 8` as AdaLoRA's initial *and* target rank, and that reading is unrunnable anywhere: with the budgets equal, peft's rank allocator asks for the zeroth smallest importance score and raises. EasyEdit's own LoRA implementation settles what its configuration file left open: it passes neither bound, leaving peft's defaults of 12 and 8, and its training loop never calls the allocator, so the published path never reallocates rank. The amendment was appended, with the qualification records that existed, before any block or any score on P0.

The other two arms qualified with no divergence and completed all 533 facts. `FT-L-EasyEdit` moved 99.3% of the edited tensor by a mean 8.95×10^{-3} and improved 95% of calibration facts; `FT+L`, whose constraint is two orders of magnitude tighter, moved 87.2% by 6.10×10^{-5} and improved every one. On the analysed population their edit-prompt efficacy separates them sharply: `FT+L` restates the target on the edit prompt itself at F1 **0.827** (target NLL 0.257), the highest of any arm here, and `FT-L-EasyEdit` at **0.205** (NLL 0.912). The analysed set is the intersection over all arms, **532** of the 533 facts; the one exclusion is a fact Experiment 5's `LoRA` block does not cover, not a fact either new arm failed.

M.2 What the two arms do

Table 59: Repetition, continuation length, constrained choice on the restatement (r.) and consequence (c.) probes, and the neighbourhood and same-subject leak rates under the A2 measure with the A4 floor. Repetition and length are the restatement cell.

arm	repeats	words	choice r.	choice c.	leak N	leak L
`FT-L-EasyEdit`	0.636	7.32	0.998	0.515	0.953	0.999
`FT+L`	0.000	6.08	0.654	0.376	0.037	0.254
`FT-L` (ours)	0.308	7.07	0.993	0.492	0.904	0.981
`FT-M` (ours)	0.588	7.68	0.998	0.506	0.976	1.000
`LoRA` (ours)	0.576	7.43	0.995	0.541	0.998	0.999
`ROME`	0.000	4.03	0.835	0.448	0.064	0.829
`MEMIT`	0.000	4.93	0.715	0.389	0.003	0.642

Table 60: Free generation on the restatement probe under all four scores, Experiment 6's two arms beside Experiment 5's seed-42 reference blocks.

arm	token F1	first-span EM	first-span F1	contained
`FT-L-EasyEdit`	0.254	0.360	0.577	0.391
`FT+L`	0.184	0.128	0.183	0.141
`FT-L` (ours)	0.425	0.608	0.690	0.687
`FT-M` (ours)	0.421	0.748	0.854	0.757
`LoRA` (ours)	0.440	0.765	0.871	0.799
`ROME`	0.585	0.529	0.590	0.553
`MEMIT`	0.381	0.312	0.382	0.327

O1, repetition: mixed, so the registration licenses no summary sentence, and the split is the informative part. EasyEdit's own FT-L repeats on **0.636** of restatement continuations, more than

twice our FT-L's 0.308 and above FT-M's 0.588, so the repetition token F1 penalises is not an artefact of the objective we substituted; the published objective repeats more. FT+L, also a gradient editor, repeats on **0.000**, like ROME and MEMIT, while writing continuations as long as theirs (6.08 words). Repetition tracks how far the weights move, not the editor family: the unconstrained arm repeats, the arm clamped at 5×10^{-5} does not.

O2 and O3: the target-as-gold account applies to the published FT-L and not to FT+L. FT-L-EasyEdit saturates the constrained-choice restatement probe at **0.998**, inside the 0.992 to 0.998 band of our gradient arms and far above ROME's 0.835, and it leaks **high** in both prompt families (0.953 and 0.999). FT+L does neither: **0.654** on the same probe, below MEMIT, and leak rates of **0.037** and **0.254**, the lowest of any gradient arm here and, on neighbourhood prompts, below ROME's. The two arms sit at opposite ends of both measures, and their reversal counts follow the same order.

FT+L is therefore the gradient, low-leakage editor the pool of Experiment 3 does not contain, and the one Experiment 5 was registered to supply and did not (J). It shows **no** qualitative interaction anywhere: against ROME and MEMIT, in both cells, under all four generation scores, its differences share a sign. That is consistent with leakage rather than editor family accompanying the reversals, but it is weak evidence and stays a flag: FT+L trails both reference editors on both probes by 1.3 to 41 points, so this is a null from an arm that is worse everywhere, not a matched contrast. Its profile is unusual on its own terms, the best edit-prompt efficacy of any arm (0.827) with the lowest leakage, which is what an L_∞ clamp of 5×10^{-5} at layer 0 buys: the edit holds on the prompt it was fitted to and does not move the neighbourhood, or the paraphrases.

M.3 O4, THE MEASUREMENT CONTRAST, AND WHERE THE SCORING DEPENDENCE STOPS

Of the eight computable comparisons, **four reject under the registered token F1** at $0.05/12$ under Gail and Simon's test, the intersection-union test agreeing on all four; all four are FT-L-EasyEdit, which leads both reference editors under constrained choice and trails them under generation. Under the three registered sensitivities the count falls to **one, one and zero**.

Table 61: The four FT-L-EasyEdit comparisons, in percentage points: constrained choice against each generation score. ‡ marks a rejection at $0.05/12$ under both tests. FT+L's four comparisons run -1.31 to -18.11 points under constrained choice and share that sign under every generation score, so none of them reject. The four AdaLoRA comparisons are not computed and are counted in the family of twelve.

comparison	choice	token F1	span EM	span F1	contained
vs ROME, restatement	$+16.32$	-33.15 ‡	-16.89 ‡	-1.29	-16.14 ‡
vs ROME, consequence	$+6.66$	-9.20 ‡	$+0.28$	$+2.50$	-0.75
vs MEMIT, restatement	$+28.33$	-12.77 ‡	$+4.78$	$+19.50$	$+6.38$
vs MEMIT, consequence	$+12.57$	-7.23 ‡	$+4.13$	$+4.10$	$+1.59$

Three of the four behave as §5.1 describes: the reversal is present under token F1 and gone under every alternative, which is what a repetition artefact looks like. **The fourth does not.** FT-L-EasyEdit against ROME on restatement reverses under token F1 ($p < 10^{-29}$ under both tests), under first-span exact match (-16.89 points, $p = 2.9 \times 10^{-13}$) and under alias containment (-16.14, $p = 1.9 \times 10^{-12}$); only first-span F1 closes the generation gap, to -1.29 points, and does not reject. The pattern is coherent with what the arm writes: its first span overlaps the accepted alias in tokens often enough for partial credit to erase the gap, while reproducing the alias exactly, or containing it anywhere in the trimmed continuation, it does **17 and 16 points less often than ROME**, at a constrained-choice advantage of 16 points in the other direction.

The choice side behaves oppositely, and that is worth stating beside it. Null-prompt calibration removes FT-L-EasyEdit's constrained-choice advantage over ROME completely: $+16.3$ points raw becomes -15.7 calibrated, a larger fall than any of our own arms shows (-3.9, -3.4, -34.2 for FT-L, FT-M and LoRA). FT+L, whose leakage is the lowest of any gradient arm, moves the other

way, from -18.1 to -16.5. So the two halves of §5.1 do not travel together: the **choice-side** account reproduces at the published configuration, and the **generation-side** one does not.

So the scoping of §5.1 is what this experiment changes. That result was measured on our arms, and on our arms it stands; at EasyEdit's published FT-L configuration, one cross-probe reversal against ROME survives two of the three alternative scores at $p \approx 10^{-13}$. **Whether the registered reversals are a scoring artefact is itself configuration-dependent**, and neither the paper's earlier "they vanish under rescoring" nor its opposite can be stated of published editors in general. Four qualifications hold this to its size: it is one arm on one backbone, one item set and one prompt family; the comparison is across two runs of one design (C.4); the generation ranking still moves by 32 points across the four scoring rules, so no score identifies it; and the arm that produces it is the one that saturates the other probe and leaks on 95% of neighbourhood prompts, which is the confound F reports as a flag and does not resolve.

O5, divergence. Neither arm that ran diverged: 533 of 533 facts completed for both, a rate of 0.000. The divergence in this experiment is `AdaLoRA`'s, at qualification, reported in M.1.

N EXPERIMENT 7: THE BENCHMARK'S OWN METRIC, AND TWO RESTORED CHOICES

Three questions the paper could not answer are answerable from one sweep, and Experiment 7 runs it. It is registered in `PREREG_EXPERIMENT7_FIDELITY_2026-09-18.md`, frozen with its hash chain before any arm was fitted, and analysed once by `analyze_exp7.py` (11 checks), written against that registration before a block existed.

(a) Does the metric produce the mirror? Section 3's two surviving reversals have `FT-L` behind on aliasing and ahead on reasoning; our arms show the opposite, and three axes could explain it, the backbone, the implementations and the metric. KnowEdit scores portability by **teacher-forced token accuracy**; the factorial scores a free continuation by trimmed token F1. The metric axis can be tested without a single new edit, by scoring the same edited models the way the benchmark does. M has already narrowed the second axis, since the published FT-L behaves like our arms.

(b) Are repetition and leakage properties of an over-trained arm? A.4 records that our gradient arms have no early stopping, while EasyEdit leaves its loop once the loss falls below 10^{-2}. `FT-M` carries Figure 1, Experiment 5 and the target-as-gold account, and M.2 showed that repetition tracks how far the weights move, which is what a stopping rule changes.

(c) Is ROME's profile the published algorithm's? Every ROME number here comes from `mom2_adjustment: false`, EasyEdit's identity-covariance variant; ROME's own GPT-J results use the C^{-1}-whitened update. ROME is the low-leakage, non-repeating reference arm in every controlled comparison, so if that profile is an artefact of the shortcut, the comparisons inherit it.

N.1 WHAT RUNS, AND THE TWO REGISTERED FAMILIES

`run_exp5.py --teacher-forced` adds, for every generation item it already scores, KnowEdit's own metric as EasyEdit's `test_prediction_acc` computes it: concatenate prompt and target, take the argmax at every position of one forward pass, compare positions $[\ell - 1, -1)$ of the argmax with the target tokens at $[\ell, \cdot)$, and average over target tokens. It is a second *reading* of one edit, not a second edit: the registered score of that item is written by the same call and is untouched. The arms are re-run rather than rescored from Experiment 3's stored text, because a teacher-forced score needs a forward pass through the edited model and cannot be recovered from a stored continuation; the configuration, facts and seed are those of the arm it is matched to. Two values are recorded, the item's first gold string (primary) and the best accepted alias (a sensitivity), because our item set carries aliases the benchmark does not.

Two arms restore one published choice each, with everything else identical to the arm they are matched to: `FT-M-EasyEditStop` is `FT-M` under EasyEdit's stopping rule, which skips the update once the loss is below 10^{-2} and then leaves, *before* the step rather than after it; and `ROME-mom2` is `ROME` with `mom2_adjustment: true` and the precomputed Wikipedia second moments for the edited layer.

The registered families are **T**, twelve comparisons on the measurement contrast (three gradient arms against ROME and MEMIT in two semantic cells) at $0.05/12$, and **F**, eight comparisons for the two fidelity arms at $0.05/8$. Teacher-forced accuracy against the first gold is primary and the best-alias reading is a sensitivity, not a second family. Every comparison is counted, including any that cannot be computed.

What this cannot show. It restores two of the departures A.4 records and not all of them: the arms still run in bfloat16 with FP32 masters, inside our loop rather than EasyEdit's, and LoRA remains plain LoRA because the published configuration diverges here (M.1). Teacher-forced accuracy is computed on our items and paraphrases, not on KnowEdit's own portability records, so it tests the metric axis and not the item axis. A null on (a) would not show that the release and the factorial agree; it would show that this metric does not account for their disagreement on these facts.

What the provenance rows say, and one thing they must not be read as saying. Experiments 6 and 7 run through the Experiment 5 runner, so every block's provenance row stamps Experiment 5's registration and calls the experiment "5". The registration a block belongs to is the one named in its output directory and in this appendix, and the hash chains of PREREG_EXPERIMENT6_HASHES.json and PREREG_EXPERIMENT7_HASHES.json are the record that matters. ROME-mom2's block is a re-run: a first attempt left a header that did not describe the run that followed it, so the block was discarded and the arm run again from scratch rather than repaired, and the runner refuses a resume whose stored provenance differs from the current one (run_exp5.py, 35 checks). Every number below comes from the re-run.

N.2 What the benchmark's own metric says

Every arm completed; only LoRA stopped on a fact, b090614064e2, the same fact its Experiment 5 block stopped on, a rate of 1 in 533. The analysed set is the intersection, **532** facts. FT-M-EasyEditStop passed the five witnesses before it ran (F4).

Table 62: The restatement probe under KnowEdit's own teacher-forced token accuracy (t-forced), its best-alias reading, the three text scores of E.1 and constrained choice, on the 532 analysed facts.

arm	t-forced	best alias	token F1	first-span	contained	choice
FT-L	0.877	0.891	0.425	0.608	0.687	0.993
FT-M	0.962	0.968	0.421	0.748	0.757	0.998
LoRA	0.963	0.970	0.439	0.760	0.797	0.995
FT-L-EasyEdit	0.755	0.771	0.254	0.360	0.391	0.998
FT+L	0.515	0.592	0.184	0.128	0.141	0.654
FT-M-EasyEditStop	0.872	0.881	0.509	0.565	0.596	0.984
ROME	0.759	0.787	0.585	0.529	0.553	0.835
ROME-mom2	0.712	0.753	0.535	0.483	0.504	0.781
MEMIT	0.597	0.661	0.381	0.312	0.327	0.715
unedited	0.316	0.449	0.085	0.025	0.029	0.366

T1: the mirror is not the metric. The release has FT-L **behind** on aliasing and **ahead** on reasoning. Under the benchmark's own metric our gradient arms show neither half of that: they lead both reference editors on **both** probes, by $+11.8$ to $+36.7$ points on restatement and $+1.2$ to $+7.7$ on the consequence probe. Not one of the six pairs is release-shaped, so the rule assigns its second sentence, quoted from the registration as frozen: "the mirror is not explained by the metric, and the remaining axes are the backbone and the implementations". In this paper's own terms, **the mirror fails no check tied to the metric**, and the backbone and the implementations remain. This is the answer that does not connect the two halves of the paper, and it is the one the data give. What it does remove is a candidate: the disagreement between section 3 and section 5 is not a matter of teacher forcing against free generation on these facts.

T2: the benchmark's own metric agrees with the three alternatives, not with token F1. Of the twelve comparisons, **3 reject under token F1** and **0 under teacher-forced accuracy**, as under first-span match, first-span F1 and containment. Under the benchmark's metric the gradient arms' advantage has the same sign on both strata (choice $+15.8$ to $+28.3$, generation $+11.8$ to $+36.7$), so there is no qualitative interaction to certify. The registered reversal is therefore specific to trimmed

token F1 among five scores, one of which is the rule the audited benchmark actually applies. E.1's four scores were ours and shared one idea; this one is not ours and does not share it, which is what the earlier sections could not say.

F1: EasyEdit's stopping rule matters, and our `FT-M` is over-trained. Restoring it, with everything else identical, moves the arm a long way:

Table 63: `FT-M` under EasyEdit's own stopping rule. The registered band for "matters" is 0.20 on repetition or on either leak rate; two of the three clear it.

quantity	FT-M	FT-M-EasyEditStop	change
repeated restatement continuations	0.588	**0.197**	−0.39
mean continuation length (words)	7.68	5.48	−2.20
neighbourhood leak rate	0.976	**0.651**	−0.32
same-subject leak rate	1.000	0.912	−0.09
constrained-choice restatement	0.998	0.984	−0.01
edit-prompt efficacy (token F1)	0.423	**0.659**	+0.24

Repetition falls by two thirds and neighbourhood leakage by a third, while the edit gets **better** on the prompt it was fitted to. So the reading that our gradient arms are over-trained is correct, and E.1's repetition figures describe an arm trained past the published stopping point. Two things do not change. The arm still **saturates** the constrained-choice restatement probe at 0.984, above the 0.95 band, so the ceiling M2 flags is not an artefact of over-training; and its leak rate of 0.651 is no longer "high" by the locality registration's bands but is ten times ROME's 0.064, so the target-as-gold account of F applies to it in direction if not in degree. Under token F1 it still reverses against ROME on restatement ($+14.9$ choice against -7.6 generation, $p = 1.8 \times 10^{-4}$ at $0.05/8$), and under the other four scores, teacher-forced included, it does not.

F2: ROME's profile is not an artefact of EasyEdit's covariance shortcut. With `mom2_adjustment: true` and the published second moments, ROME repeats on 0.003 of continuations against 0.000, leaks on **0.006** of neighbourhood prompts against 0.064, and scores slightly lower everywhere (choice 0.781 against 0.835, teacher-forced 0.712 against 0.759). The whitened update is a little weaker on these facts and has the same shape, so the low-leakage, non-repeating reference arm the controlled comparisons rest on is the published algorithm's behaviour and not the shortcut's. Both of its comparisons against our gradient arms reverse under token F1 and under nothing else, exactly as `ROME` does.

F3 and F4. Of family F's eight comparisons, **3 reject under token F1** and 0 under each of the other four scores: `FT-M-EasyEditStop` against ROME, and `ROME-mom2` against `FT-M` and `LoRA`. No arm diverged except `LoRA`'s single fact, and the qualification record is in the artifact. Every number here is a field of `exp7_result.json`.

What Experiment 7 changes in the paper. Section 5.1 gains a score that is not ours and agrees with ours; section 3's mirror loses one of its three candidate explanations; and A.4's list of departures loses two entries, one of which mattered and one of which did not. What it does not do is make the factorial a mechanism study for the release's two surviving reversals: after this, the backbone and the implementations remain, and testing those needs a Llama-2-7b-chat run that this paper does not have.

O EXPERIMENT 8: THE AUDITED RELEASE'S OWN BACKBONE

Section 3 leaves two reversals in KnowEdit's `recent` cell that fail no check available in the release, and names three axes that could explain why our factorial shows their mirror: the metric, the implementations and the backbone. N ruled out the metric on the GPT-J facts; M characterised the implementations. This experiment varies the third. It is registered in `PREREG_EXPERIMENT8_LLAMA_2026-09-18.md`, frozen with its hash chain before any arm was fitted, and analysed once by `analyze_exp8.py` (7 checks), written against that registration before a block existed.

O.1 WHAT RUNS

`NousResearch/Llama-2-7b-chat-hf` at commit `351844e75ed0bcbbe...`, in bfloat16 with FP32 master copies. This is a mirror of a gated repository, and the reason is recorded in the registration: `meta-llama/Llama-2-7b-chat-hf` returns HTTP 401 without an accepted licence, which is deviation D1's original blocker. It is not a substitution of a different model; the weight shards and tokenizer are byte-identical to Meta's by the sha256 the Hugging Face API reports for both repositories, and the comparison is reproducible from the two public metadata endpoints.

The items are the frozen 573-fact set of Experiment 3 (`450ea4aa19e7e21e...`), unchanged, and the analysed population is the registered 533 facts, so every comparison with Experiments 6 and 7 is on identical items. Each arm is EasyEdit's published Llama-2 configuration, read from `hparams/*/llama-7b.yaml`.

Two arms do not run, and neither absence is a choice. `AdaLoRA` at the published rate diverged on **all twenty** calibration facts, exactly as it did on GPT-J, so it produced no block; the registration named it the arm at risk and forbids re-tuning it, and it is recorded as failing. `MEMIT` is absent because its Llama-2 configuration sets `mom2_adjustment: true`, which needs second-moment statistics for layers 4 to 8 that are not published for this backbone and are not in our cache; computing them means a full English Wikipedia download and 100,000 forward passes per layer. **One of the two release reversals this experiment bears on is MEMIT against FT-L, and it therefore cannot be addressed at all.** The other, ROME against FT-L, can. The three arms that ran (`FT-L-EasyEdit`, `FT-M`, `ROME`, beside `UNEDITED`) qualified on the twenty calibration facts and diverged on none of the 533 analysed facts.

O.2 L1: THE RELEASE'S DIRECTION IS NOT RECOVERED

The registered test asks, for each gradient arm against `ROME` under KnowEdit's own teacher-forced accuracy, whether the pair of differences (aliasing, reasoning) takes the release's direction, which is the gradient arm behind on aliasing and ahead on reasoning. Neither does: `FT-L-EasyEdit` is behind `ROME` on both (-9.7 and -3.5 points) and `FT-M` is ahead on both ($+35.1$ and $+9.5$). The registered sentence this licenses is that **the backbone does not account for the mirror either, and on these items no axis we can vary reproduces the release's direction.** Among the three axes considered here, the implementations remain.

O.3 L2: ONE REVERSAL, AND THE SCORE UNDER WHICH IT APPEARS

Family L is six comparisons at $0.05/6$, of which four are computable without `AdaLoRA`. Under four of the five generation scores none reverses. Under the fifth, **KnowEdit's own teacher-forced accuracy**, one does: `FT-L-EasyEdit` against `ROME` on `Subject_Aliasing`, ahead by **12.20** points under constrained choice and behind by **9.74** under generation, rejecting under both tests ($p = 5.6 \times 10^{-17}$ Gail and Simon, intersection-union also rejecting at $0.05/6$). This is the first reversal in this paper on the audited release's own backbone, at the published configuration, under the published metric.

Table 64: `FT-L-EasyEdit` against `ROME` on `Subject_Aliasing`, in points, under the registered Gail and Simon test and the dependence-robust intersection-union test, at the family threshold $0.05/6 = 0.0083$. The constrained-choice side is one number because only the generation score changes down the rows. Two of the four non-rejecting scores carry a negative generation sign without clearing the threshold, so the row that separates them is not the only one pointing that way.

generation score	choice	gen.	p	IUT p	rev.
trimmed token F1	$+12.20$	$+5.59$	0.50	1.00	no
first-span match	$+12.20$	-2.25	0.089	0.178	no
first-span F1	$+12.20$	$+8.17$	0.50	1.00	no
alias containment	$+12.20$	-3.10	0.044	0.088	no
teacher-forced accuracy	$+12.20$	-9.74	5.6e-17	1.0e-16	**yes**

This is the converse of the GPT-J result, and it is the reason the experiment is worth its compute. In the factorial, the registered reversal appears under trimmed token F1 and is absent under the

other four scores, teacher-forced accuracy included (E.1, N.2). Here the arrangement is reversed: teacher-forced accuracy is the only one of the five under which the reversal appears, and token F1 is not. Neither backbone licenses the reading that one scoring rule is the sound one and the others are not. What the two together establish is narrower and harder: **whether a cross-format reversal appears depends jointly on the scoring rule and on the model and configuration, and no scoring rule tested here resolves the disagreement on both backbones.**

The semantic contrast, family S, is six comparisons at $0.05/6$ with four computable, and **none** reverses under any score.

O.4 What the checks say about this reversal

The flags are not restated in prose here; the strata are handed to `probe_consistency_report.py`, the audit instrument the paper ships, and read back from its record by `post_hoc_analyses.exp8_audit_flags` (post hoc†, since the registration fixes L1 to L5 and not this). Its verdict is `QUALITATIVE FAILURE` with the cross-probe claim **REFUTED**; the composite flips at a constrained-choice weight of 0.444 and the aggregate winner holds 55.4% of the weight simplex.

- **M2 does not fire.** No arm is within 0.05 of ceiling or of its probe's chance level in either stratum (choice 0.837 and 0.715 against a chance level of 0.25; generation 0.497 and 0.594 against 0). The instrument reports no ceiling/chance-adjacent stratum.
- **M4 does fire, on one family of two.** Leakage differs by 24.0 points on same-subject prompts (`FT-L-EasyEdit` 61.5%, `ROME` 37.5%), which is `MISMATCHED` at the 20-point threshold, and by 10.8 points on neighbourhood prompts (11.5% against 0.7%), which is `MATCHED`. **This reversal is therefore not flag-free, and the paper does not claim it is.**
- **The calibrated rescoring points the other way.** Null-prompt calibration is the paper's re-expression of leakage, and on GPT-J it takes EasyEdit's own `FT-L` from $+16.3$ to -15.7 against `ROME` (M). Here it does not remove the choice-side advantage: it **widens** it, from $+12.20$ to $+14.45$. So the flag says look and the diagnostic that would express its mechanism does not support the artefact reading. B.7 already reports that M2 carries no information about which pairs reverse; this is the same lesson for M4 on a single pair, and it is why the paper reports flags as flags.
- **Repetition does not account for the generation side.** EasyEdit's `FT-L` repeats on 63.6% of restatement continuations on GPT-J (M.2) and on **3.6%** here, with `ROME` at 0.0% on both. The mechanism that makes a trimmed overlap score read a correct answer as a failure is close to absent on this backbone, which is also why token F1 does not produce the reversal here.

The arm is weak on the edit prompt, and that travels with the result. `FT-L-EasyEdit` reaches an edit-prompt efficacy F1 of **0.345** on this backbone, against `ROME`'s 0.906 and `FT-M`'s 0.825, so it installs the edit far less reliably than the arm it is compared with. That is a real limit on how much the pair can be asked to carry. It is not, however, a floor comparison: on the reversing cell both arms are above the unedited model in both strata (choice 0.837 and 0.715 against 0.420; teacher-forced accuracy 0.497 and 0.594 against 0.356), so the disagreement is between two arms that both moved the model, in opposite directions on the two formats.

O.5 L3, L4 and L5

L3, the profile does not transfer. The registered rule is that the GPT-J profile transfers if every gradient arm is saturated at ≥ 0.95 on constrained-choice restatement and leaks at ≥ 0.70 on neighbourhood prompts. Neither arm satisfies both: `FT-L-EasyEdit` is at 0.837 and leaks 0.115, `FT-M` is at 0.970 and leaks 0.509. The saturated, heavily leaking gradient arm that carries Figure 1 and the target-as-gold account on GPT-J does not appear on this backbone.

L4, qualification and divergence. `FT-L-EasyEdit` and `FT-M` qualified with 0 of 20 calibration facts diverging; `AdaLoRA` diverged on 20 of 20 and is reported as failing, not re-tuned. Across the 533 analysed facts no arm that ran diverged on any fact. The sham and cross-editor witnesses both passed.

L5, the screen. The answerability screen is re-run on this backbone because a screen is model-specific, and it is **reported, not used as a gate**: the analysed population is the registered 533 either way. On Llama-2-7b-chat **14** of 573 facts pass the screen's rule, against **15** of 573 on GPT-J. The concern the registration named, that an instruction-tuned chat model might answer these rendered formats differently enough to carry the comparison, is not supported: the two retention rates are within one fact of each other.

O.6 What this experiment cannot show

It runs single edits, one item set, one prompt family and three editing arms, and **MEMIT is absent**, so it bears on one of the release's two surviving reversals and not the other. It uses the published *configurations* inside our pipeline, not EasyEdit's code, exactly as Experiment 6 did, and inherits that experiment's limits. The items are our renderings of KnowEdit's questions, not KnowEdit's own portability records, so L1's null does not show that the release and this run disagree about the same items; it shows that this backbone, with these configurations, on these items, does not produce the release's direction. **"The backbone does not account for the mirror" is a statement about this controlled comparison, not about backbone effects in general.** Nothing here bears on sequential or batch editing.