Evaluating Agent-Proposed Weight Edits for False Abstention:
A Controlled Case Study in Qwen2.5-3B-Instruct

Qamil Mirza

College of Computing, Data Science, and Society, University of California, Berkeley

qamilmirza@berkeley.edu

Preliminary Technical Report, November 8, 2026

Abstract

Language models sometimes abstain from answering questions even when the necessary information is present in their context. In Qwen2.5-3B-Instruct, such false abstentions often follow a generated note containing the correct answer. Simple interventions can recover these answers but risk increasing unsupported responses. We test whether weight edits proposed by an agent improve this trade-off, and whether the agent's explanations of them hold up.

A Claude agent proposed three edits, which we evaluated using separate selection and held-out confirmation data under a pre-specified protocol. A rank-1 edit to the final layer's attention output projection met the repair criterion, improving utility by 0.117 (95% CI 0.096 to 0.138). It recovered 82% of false abstentions while increasing unsupported answering from 3.5% to 4.2% and damaging 3% of initially correct answers (base-weighted rates). The edit achieved a better trade-off than three simpler interventions under Holm-corrected tests, but superiority over option removal was not established, and the comparison with logit bias was inconclusive.

A separate two-head edit showed that a layer-33 component is sufficient, though not necessary, to release answers, but the agent's account of that component was not established. Follow-up tests supported both edits' specificity relative to random controls, but logit bias achieved similar observed utility, and most newly released answers on natural questions were incorrect. The evidence comes from one model and one agent run.

1 Introduction

In the example of fig. 1, Qwen2.5-3B-Instruct copies the answer into its own note and then replies "unknown". We call any refusal where the context supports an answer a *false abstention*, as opposed to an *appropriate abstention*, where the context really lacks the answer. Copying the answer and then refusing is common. On our confirmation set, the unedited model named the correct answer in its own note before abstaining in 232 of 261 false abstentions. The failure is hard to fix because making a model answer more readily also makes it answer when the evidence is insufficient. A useful intervention has to recover correct answers while keeping appropriate abstention.

Language-model agents can inspect a network and propose targeted weight modifications [13]. Requiring the agent to state its hypothesis makes its explanation testable. An edit can work on the examples the agent explored and fail elsewhere, and it can work while the agent's explanation of it is wrong. We asked a Claude Opus 5.5 agent to propose weight edits for this failure, and it submitted three, each with a hypothesis and predictions (section 4). Training-free weight edits that fix a behaviour have been shown before [12, 13], so our contribution is the evaluation rather than the edit. We ask whether an agent's fix beats simpler interventions under sealed, pre-specified tests, and whether its explanation survives independent testing. In this controlled setting (one model, one task family, one agent run) we contribute:

Memory	`user, 2023/03/01 (Wed) 10:00: Ben lives in Manchester.` `user, 2023/03/02 (Thu) 10:00: Ben works as a manager.`
Question	`(asked on 2023/03/03 (Fri) 12:00) Where does Ben live?`
Unedited model	`Note: user mentioned that Ben lives in Manchester on 2023/03/01 (Wed) 10:00.` `Answer: unknown`
Edited model	`Note: user mentioned that Ben lives in Manchester on 2023/03/01 (Wed) 10:00.` `Answer: Manchester`

Figure 1: A false abstention by Qwen2.5-3B-Instruct, verbatim from stored outputs (case `E-DE01_ANSa_H0`, with the full prompt in section A). The edited output is the agent's layer-35 edit at dose 1 with the unedited note held fixed. The example comes from the split the agent studied, so it illustrates the failure and is not evidence for the repair.

1. **An evaluation protocol** that separates agent exploration, intervention selection and sealed held-out confirmation, and tests repair and explanation independently (section 5).
2. **Evidence for selective repair.** An agent-proposed rank-1 edit improves false-abstention behaviour on held-out data and beats three of five simpler interventions on the recovery–unsupported-answer trade-off, although its advantage over logit bias is unresolved (section 6).
3. **Specificity without a confirmed mechanism.** Both agent edits beat matched random controls, yet neither the agent's mechanistic account nor superiority over an output-level bias is established, and natural-question tests expose limits to generalisation (sections 7 and 8).

2 Related work

Agents and weight edits. Weight-space programming showed that an agent can edit weights without gradient-based optimisation, reducing reasoning loops in Qwen3.5-0.8B [13]. Steer2Edit converts steering directions into rank-1 head and neuron edits [12], and interpretability agents such as MAIA test hypotheses about components [10]. Unlike weight-space programming, our target has abstentions worth keeping, and our confirmation is sealed and pre-specified.

Abstention, release and selectivity. Safety refusal can be mediated by a single direction [1] or by individual heads [15], and a related direction tracks answerability from context [4]. A commit–abstain circuit links attention heads and MLP components to the decision to answer or abstain, and a policy trained on its activations reduces false abstentions [9]. Abstention can follow the mere presence of an "unknown" option [5], and suppressed answers can often be released locally [6]. Selective fixes include weight-folded false-refusal ablation [14], gated residual modules [3] and SAE-feature steering [2], but none is agent-proposed. Outside language models, Modgil [8] used pre-registered once-only tests to separate decodable from causally usable information, and we separate repair from mechanism in the same way.

3 Experimental setup

Model and prompt. We use Qwen2.5-3B-Instruct (36 layers indexed 0–35, residual width 2048) with fp32 weights, greedy decoding and at most 160 new tokens. Each prompt holds a short dated

Table 1: Data splits. E and S use generator seed 51 with disjoint cases, and H and T-city use seeds 52 and 53. Rows from earlier experiments were excluded by prompt hash.

Split	Bases/task	Access	Purpose
E (exploration)	48	Agent, then analysts after the agent finished	Agent exploration and screening
S (selection)	32	Analysts	Candidate and dose selection
H (confirmation)	48	Sealed, generated and read once after F2	Confirmatory evaluation
T-city (transfer)	24	Sealed, generated and read once after F2	Descriptive transfer to unseen cities

memory and a question. The model writes a brief note on the relevant facts, then `Answer: <short phrase>`, and may reply "unknown" if the memory lacks the answer (section A). We treat the note as a prompted summary, not a faithful reasoning trace.

Cases and splits. Synthetic memories cover *direct extraction* (DE), *relational composition* (RC) and *temporal updating* (TU). Each base case yields six answerable questions (two per difficulty level, with more distractors at higher levels), four unanswerable ones (three replace the fact with related but insufficient information, one changes the queried person) and three counterfactuals that change the evidence to another city. Bases are partitioned into four splits (table 1): exploration (E), selection (S), confirmation (H) and transfer to unseen cities (T-city). No base, person–evidence–city key or memory text crosses splits. The agent saw only E.

Grading, populations and utility. A rule-based grader, frozen before the agent saw E, reads only the text after the final `Answer:` and counts only asserted or suggested cities (section A). The unedited model, B0, fixes the populations used for every arm: *false abstentions* (FA, answerable rows on which B0 abstains), *baseline-correct* (C), *unanswerable* (UNA) and *counterfactual* (CF) rows. We report *recovery* (FA answered with the gold city), *unsupported* answering (UNA given an answer), *damage* (C no longer correct), *CF-follow* (CF answered with the changed city) and the *degenerate* rate (share of rows with degenerate output or no answer line). Utility U scores a correct answer $+1$, an abstention 0, a mixed answer (city plus abstention phrase) 0 on answerable and -1 on unanswerable rows, and any other error -1. Answerable and unanswerable rows each carry half of a base's weight, and CF rows are reported separately. All rates are *base-weighted* (computed per base, then averaged over bases).

Comparators. Each dose was selected on S from a fixed grid, then frozen. **B1**: logit bias on tokens beginning with "unknown" after the answer marker (selected -8 from -1, -2, -4, -8, -16 and $-\infty$). **B2**: appended instruction "Do not reply unknown." **B3**: removal of the sentences permitting "unknown". **B4**: a previously built content-free steering direction added at layer 29 from the answer marker on (selected $\alpha = 1$ of four). **B5**: directional orthogonalisation after Arditi et al. [1], projecting a mean-difference abstain-vs-answer direction estimated on E out of all attention and MLP output weights (selected $\beta = 1$ from 0.25, 0.5, 0.75 and 1). **B6**, five random edits matched in size, is used only in the pathway tests.

Table 2: The agent's submitted edits. Hypotheses are paraphrased from the stored submissions and are claims to be tested.

#	Proposed edit	Agent's hypothesis
1	**Two-head ablation (repair).** Zero the `o_proj` columns of heads L32h2 and L33h10 (≈0.52M weights).	The heads add a roughly evidence-independent bias toward "unknown", and removing it recovers false abstentions while preserving warranted abstention.
2	**Rank-1 update (repair).** Modify the layer-35 `o_proj` with two 2048-d vectors (≈4.19M entries).	The last attention layer carries an "answer available" signal, and reading it and writing toward city tokens releases correct answers.
3	**Two-head amplification (diagnostic).** Double the same columns.	If the heads supply a default abstention bias, doubling them turns correct answers into "unknown".

4 What the agent proposed

A fresh Claude Opus 5.5 API session had read-only access to E and none to the project's records or other splits. It could run the model, inspect activations, apply edits and analyse in Python, but only with closed-form operations (means, covariances, projections, SVD, single linear solves), not gradient or iterative fitting. A candidate could change at most four components in total, each being one layer's `o_proj`, `down_proj` or one of its two layer norms, by rank-≤ 4 updates or sparse edits of up to 512 whole rows or columns. Other projections, embeddings, the final norm and "unknown"-token weights were off limits. Before testing, the agent had to name the component, explain its function, predict effects on every population and the component's activity in the unedited model, and declare an endpoint and a scalar variable f representing the proposed function.

For edit 1 the agent predicted about 85% recovery, at least 97% of correct answers kept and related-distractor unsupported answering at most 4%. For edit 2 it predicted 75–80% recovery, 92–97% kept and at most 3%, and for edit 3 it predicted that 30–40% of correct answers would turn to "unknown". Predictions did not determine selection.

The rank-1 proposal. Edit 2 updates the layer-35 output projection $W_O \in \mathbb{R}^{2048\times 2048}$ as $W_O \leftarrow W_O + s\,c\,d\,(W_O^\top w)^\top$, adding $\Delta h = s\,c\,(w^\top W_O a)\,d$ to the residual stream for input a. The read direction w comes from one closed-form solve on E meant to separate answerable from unanswerable rows. The write direction d is the unit difference between the mean city unembedding and the "unknown" unembedding. The agent fixed $c = 20$, and s was selected on S. The update is rank 1 but touches nearly every entry of the dense matrix.

5 Evaluation design

Two branches, each free to choose its own edit and dose, ask separate questions, so success on one does not establish the other. The *repair* branch asks whether an edit fixes false abstention at acceptable cost, and the *pathway* branch asks whether manipulating the named component produces the predicted behaviour for the predicted reason.

Screening on E and selection on S. After the agent finished, every submission was replayed on E. Repair entry required a utility gain over B0 with one-sided 97.5% bootstrap lower bound above zero and at least 20 false-abstention bases in DE or RC. Pathway entry required an effect of

at least 0.05 in the declared direction, bound beyond zero, and 20 bases. Edits 1 and 2 entered both branches (each with a utility gain of +0.122, and with recovery of 0.862 and 0.739 respectively), while edit 3 entered the pathway branch only. These E results are exploratory. On S, repair chose each candidate's dose by utility and kept the best candidate (ties to fewer weights, then the weaker dose): submission 2 at $s = 2$, utility 0.429 against 0.402 for submission 1 and 0.303 for B0. This was the top of the frozen grid $\{0.5, 1, 1.5, 2\}$, with utility still rising. The pathway branch maximised the declared effect under a degenerate-output limit and chose submission 1 at dose 1.5, with layer 33 as primary component because reverting it on S removed more of the effect than reverting layer 32.

Pre-specification. Here pre-specification means the design was committed to Git before the data it governed existed. It was frozen in three stages: F0 (data, grader, comparators, edit rules) and F1 (bounds, run procedure, cost limits) before agent exploration, and F2 (candidates, doses, comparator settings and analysis code) before any H or T-city output. The harness checked F2 before reading held-out data. There was no external registration. H and T-city were each generated and read once, and no candidate was added afterwards (section B).

Repair criterion and trade-off tests. An edit repairs on H if the one-sided 97.5% lower bound of its utility gain over B0 exceeds zero, damage is at most 0.05, the degenerate rate at most 0.02 and CF-follow at least 0.90 (damage and CF-follow on point estimates). Against each of B1–B5 we tested two orientations: **O1**, more recovery (superiority) with unsupported answering at most 0.05 higher (non-inferiority), and **O2**, fewer unsupported answers with recovery at most 0.05 lower. Each is an intersection–union test (p is the larger component p), Holm-corrected across all ten, and a comparator counts as beaten if either orientation passes. Equivalence with B1 was tested separately by TOST (margins ±0.05). Pathway tests M1–M3 form their own Holm family. Uncertainty comes from a paired bootstrap clustered by base and stratified by task ($B = 10{,}000$), and intervals are 95% unless stated.

6 Results

All results here come from the sealed split H (144 bases and 1,872 rows, including 261 false abstentions, 593 baseline-correct and 576 unanswerable rows), except the transfer check.

The selected edit meets the repair criterion. Utility rose from 0.317 to 0.435, a gain of +0.117 (95% CI 0.096 to 0.138). The edit gave the gold city on 212 of 261 false abstentions (recovery 0.816), and unsupported answers rose from 20 to 24 of 576 (0.035 to 0.042). Damage was 0.030, under the limit on the point estimate but with upper bound 0.055. The edit broke 11 of 593 correct answers, which is 0.019 pooled. The base-weighted rate is higher because the broken rows fall mostly in bases with few correct rows. B1 and B4 broke none. Degenerate outputs were 0.008, and CF-follow 0.917 (threshold 0.90, unedited 0.632).

Comparison with simpler interventions. The edit's trade-off beat the added instruction (B2, O2, Holm-adjusted $p = 0.005$), steering (B4, O2, $p = 0.033$) and orthogonalisation (B5, O1, $p = 0.007$): fewer unsupported answers without lost recovery against B2 and B4, more recovery without added unsupported answers against B5 (table 3, fig. 2).

Logit bias: inconclusive. B1 is the most informative comparator, as the simplest intervention acting directly on the output. Point estimates favour the edit (recovery 0.816 vs 0.793, unsupported

Table 3: Sealed confirmation results on H (base-weighted). The last column gives whether the edit's trade-off beats the comparator, with the orientation that passed and its Holm-adjusted p across ten tests. Values were recounted independently from per-row outputs.

	Recov. ↑	Unsupp. ↓	Damage ↓	CF-follow ↑	U ↑	Edit better? (p_{Holm})
B0 unedited	0	0.035	0	0.632	0.317	—
Selected edit (rank-1, L35, s=2)	**0.816**	**0.042**	**0.030**	**0.917**	**0.435**	—
B1 logit bias (−8)	0.793	0.095	0	0.919	0.406	inconclusive
B2 "Do not reply unknown"	0.721	0.083	0.046	0.882	0.389	yes (O2, 0.005)
B3 "unknown" option removed	0.977	0.682	0.021	0.975	0.130	not established
B4 layer-29 steering ($\alpha = 1$)	0.786	0.078	0	0.921	0.418	yes (O2, 0.033)
B5 orthogonalisation ($\beta = 1$)	0.678	0.031	0.099	0.838	0.401	yes (O1, 0.007)

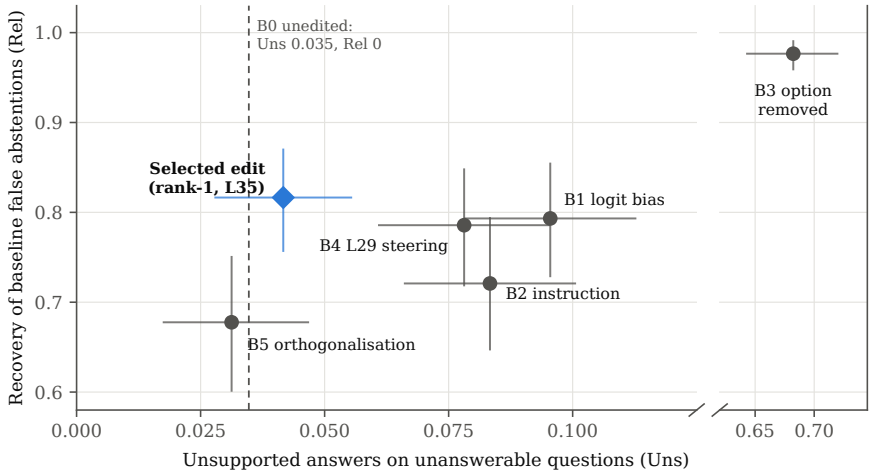


Figure 2: Recovery against unsupported answering on H, where up and left is better. Bars are descriptive marginal 95% bootstrap intervals computed afterwards from per-row outputs, whereas the confirmatory comparisons in table 3 use paired differences with Holm correction. Dashed line: unedited unsupported rate. The horizontal axis is broken to show B3.

0.042 vs 0.095). The unsupported difference is clear (−0.054, 95% CI −0.069 to −0.038), but the recovery difference is not (+0.023, 95% CI −0.043 to +0.088). O2's non-inferiority component gave $p = 0.015$, not surviving Holm correction (adjusted $p = 0.092$), and TOST did not show equivalence ($p = 0.21$ for recovery, 0.70 for unsupported). Neither superiority nor equivalence is shown.

Option removal: not established. B3 recovered 0.977 but answered 0.682 of unanswerable questions. The edit recovered less (−0.160) with far fewer unsupported answers (−0.641), so O1 fails on recovery and O2 on the 0.05 margin.

Descriptive checks. On the 232 false abstentions whose note names the gold city, the edit and B1 recovered at similar rates (0.864 and 0.881). On T-city (unseen cities, read once), utility was 0.422 against 0.303 for B0, recovery 0.817, unsupported 0.062 against 0.059, damage 0.031 and

CF-follow 0.912, with comparator utilities in the same order as on H (B4 0.415, B1 0.412, B5 0.409, B2 0.397, B3 0.107).

7 What the edits explain

Beyond repair, we ask whether manipulating the named component changes the behaviour (*pathway*) and whether the evidence supports the agent's account of what it computes. These confirmatory tests concern the two-head edit, which the pathway branch selected, not the layer-35 repair. They use a *controlled-prefix* setting: the intervened model is fed the unedited model's own note and generates only the answer, so the endpoint (false abstentions answered with the gold city) cannot move through a changed note, although internal states over the prefix can differ.

Pathway evidence through layer 33. At dose 1.5 the two-head edit is strong and unselective (on S it answered 96% of unanswerable rows), so its role is to locate where release can be produced. *M1:* reverting only its layer-33 part reduced recovery by 0.293 (95% CI 0.225 to 0.368), and the full edit recovered 0.717 (95% CI 0.664 to 0.770) more than the mean of five random edits of the same entries (B6). *M2:* adding the edit's layer-33 output change, at the answer position only, to the unedited model recovered 0.808 (95% CI 0.744 to 0.870), while clamping that output back in the edited model removed only 0.193 (95% CI 0.138 to 0.253). Both passed (Holm-adjusted $p < 0.001$): the layer-33 change is sufficient but not necessary, with the rest of the release presumably travelling another route, perhaps the layer-32 head (untested). Because B6 randomises the same entries, M1 does not show that the two heads are unusual. Section 8 tests that with random head pairs.

The account of L33h10 was not established. The agent said L33h10 writes an evidence-independent "unknown" prior, with f the layer-33 attention output at the answer position projected onto that head's mean output direction. M3 passed only if all three parts did. (a) All four activity predictions passed: removing evidence lowered f (-1.08, 95% CI -1.54 to -0.62), and changing the city (-0.16), adding distractors ($+1.03$, 95% CI 0.52 to 1.55) and swapping absent for unrelated evidence (-0.04) stayed within frozen margins of 5.2 to 6.6, although the distractor effect excludes zero and exceeds the agent's own bound of 0.5. (b) Setting f in the unedited model, compared with the logit bias whose effect on S was closest, gave an answerable-vs-unanswerable interaction of 0.0005 (95% CI -0.025 to 0.027; $p = 0.49$), so setting f acted like a generic push away from "unknown". (c) Adding f to a frozen predictor of false abstention changed AUROC by -4.5×10^{-5} ($p = 0.82$). M3 therefore failed (Holm-adjusted $p = 0.82$). The failure counts little against the account: the predictor's baseline AUROC was already 0.998, and an evidence-independent prior should not separate answerable from unanswerable questions, so a null in (b) is close to what the account predicts.

8 Follow-up: are the edits specific?

After confirmation closed, a follow-up was pre-specified in the same way to ask whether random edits of the same form would work as well and whether the repair carries to natural questions. It cannot change any confirmation verdict.

Design. A new split, N (seed 54, 1,950 rows), is disjoint in keys from E and S but not from H (867 of its 1,350 keyed rows share a key with H). One fifth (N-dev) served calibration only, and the rest (N-held) was generated and read once, after a hashed manifest was committed. A fixed rule drew

Table 4: Follow-up claims on N-held. "Rank" is the target's rank among 21 edits.

Claim	Target	20 random edits	Rank	Result
C1: zeroing the heads recovers false abstentions	0.751	max 0.464, median 0.163	1	Shown
C2: doubling the heads turns correct to "unknown"	0.381	max 0.345, median 0.094	1	Shown (narrowly)
C4: layer-35 read direction matters (utility)	0.403	max 0.300, median −0.069	1	Shown
co-requirement: unsupported answers	0.058	median 0.228	—	met
C5: four activity predictions, layer-35 signal	—	—	—	Shown ($p \approx 10^{-4}$)

150 false-abstention, 150 baseline-correct and 150 unanswerable rows from N-held, so levels are not directly comparable with H. Each target faced 20 random edits of the same form. **C1**: zeroing the two heads (dose 1) must recover more than every random layer-32/33 pair and beat their median by 0.30. **C2**: doubling them must cause more correct-to-"unknown" flips than every doubled random pair. **C4**: the layer-35 edit (dose 2) must beat the utility of every control that keeps its write vector with a random read vector (scaled on N-dev to match its gate and sign-aligned, a design revision recorded before any held-out output), with unsupported answering below their median. **C5**: four activity predictions for submission 2 must hold in the unedited model (three sign tests, one TOST at ±2 N-dev s.d.). C3, the cost of the head ablation, was descriptive. Rank tests cannot give p below $1/21 \approx 0.048$. No family-wise control is claimed, and under exchangeability the chance that some claim passes by luck is at most about 0.19.

Both edits beat random edits. All four claims passed (table 4). The pair is also unusual in its cost (C3, descriptive), since its ablation gave the highest unsupported rate of all 21 (0.075, against a random maximum of 0.071 and 0.031 unedited). Single heads released less than the pair (L32h2 0.486, L33h10 0.606), and random pairs from layers 30–35 at most 0.306. Random-read controls for the layer-35 edit either answered unanswerable questions (one: recovery 0.94, unsupported 0.88) or broke generation (two were degenerate on 0.91 and 1.00 of rows), and random *write* vectors reached utility at most 0.148. C5 concerns the agent's variable for submission 2, the layer-35 attention output at the answer position projected onto the read direction w. In the unedited model, removing evidence lowered it (−3.77, 95% CI −3.89 to −3.65), adding distractors lowered it (−1.12, 95% CI −1.38 to −0.86) and unrelated evidence raised it (+0.22, 95% CI 0.13 to 0.32). The agent had stated −3.8, −0.76 and +0.24 for these three effects. Changing the city left it within tolerance (−0.03, 90% CI −0.13 to 0.08, tolerance ±1.12). No predictor or separating-intervention test was run for submission 2, so C5 does not establish its account.

Direct logit attribution (descriptive) in the unedited model, at the answer position on false abstentions, shows L32h2 and L33h10 pushing toward "unknown" over the gold city's first token by +7.6 and +10.5 logits, against +1.4 and +0.5 for the two layer-35 heads the agent named. The layer-35 edit's write at dose 2 contributes −27.6 on false abstentions, −34.8 on correct rows and −0.5 on unanswerable rows. These are attributions and do not by themselves establish a mechanism.

Comparison with logit bias. In a descriptive comparison on the same sample, the frozen B1 bias, which was not retuned, had recovery 0.698, unsupported answering 0.075, damage 0 and utility

0.406. The layer-35 edit had utility 0.403 and the two-head ablation 0.414. B1's unsupported rate exceeded the layer-35 edit's (0.058), but utility was about the same.

Natural questions: most released answers are wrong. On 126 LongMemEval development questions (descriptive), the unedited model refused 28, 26 with the answer in view, and answered 35 correctly. The two-head ablation released 20 of the 26 in-view refusals, only 8 with the gold, and the layer-35 edit released 19, 7 with the gold. Both raised correct answers from 35 to 45, so some gains came from questions other than the in-view refusals. They damaged 2 and 3 of the 35, and answered 4 of the 12 questions that warrant abstention (unedited: 6).

9 Discussion and limitations

What the repair establishes. With the agent kept away from selection and confirmation data, a closed-form agent-proposed rank-1 edit met a criterion fixed in advance and beat three of five simpler fixes on corrected tests. That shows such an edit exists for one model and task family, but it says nothing about how often agent edits succeed. The repair also damaged correct answers that the logit bias and steering left intact (section 6). On natural questions most released answers were wrong, so synthetic recovery rates should not be read as accuracy gains.

Comparison with an output bias. Both edits beat random edits of the same form, yet a logit bias against "unknown", which changes nothing inside the model, was not clearly beaten on H and matched both edits' utility on N. The simplest explanation is that the agent found internally selective ways to do roughly what an output bias does. The layer-35 edit is close to one by construction: its write direction d lies in output-vocabulary space and is added at the last layer, so it acts much like a logit bias gated by the read projection. The attribution in section 8 fits this (descriptive), with a write near zero on unanswerable rows, which would explain the lower unsupported rate than the ungated B1. A logit bias gated on the same projection is the direct test, and we did not run it. Whether the edit's lower unsupported rate on both H and N (0.042 against 0.095, and 0.058 against 0.075) is a real advantage needs a larger, B1-targeted comparison.

Behaviour and explanation. For the two-head edit, the pathway tests passed and the heads proved unusual, but the account was not established. For the layer-35 edit, activity predictions held but the stronger tests were never run. Activity predictions show how a variable responds to input, not that the model uses it as claimed. Edits and explanations should be tested separately and with comparable strength.

Is the failure partly prompt-induced? The prompt gives two different conditions for abstaining: "If the memory does not contain the answer, reply: unknown" and "Reply unknown only if the memory has nothing at all on the topic" (section A). With both sentences removed, the unedited model answered 145 of the 150 sampled false-abstention rows correctly, leaving almost nothing for the edits to repair. The option still does real work, since removing it (B3) produced unsupported answers on 0.682 of unanswerable rows. The main open threat is that the false abstentions depend on the inconsistent wording, so that the edit repairs a prompt-induced bias rather than a general failure. A rerun with one consistent abstention condition would settle this.

Limitations.

- *Scope.* One 3B model, synthetic city questions in three templates, one prompt and note format, one agent run. In earlier project work false abstention was much rarer in three other small models, so the problem may be specific to this model.
- *Transfer.* A reworded-instruction split was dropped before the first freeze, and T-city varies only the answer vocabulary, so new instructions and misleading contexts [11] are untested. The edit's write direction is built from city unembeddings, so other answer types, such as dates, occupations or numbers, are untested too.
- *Dose.* The selected dose (and B5's β) sat at the top of its grid, so the edit is one point on a dose curve explored only up to $s = 2$.
- *Agent.* Other sessions or models could propose different edits, and we did not test whether weight access was necessary. In a related auditing setting, a calculation tool mattered more than weight access [7].
- *Comparators.* We did not compare against component-level methods such as Steer2Edit [12], false-refusal ablation [14], gated edits [3] or SAE steering [2].

Reproducibility and cost. Two deviations were recorded at F1. An import bug that crashed the first dry run was fixed, and the API was set to return the agent's summarised reasoning so that it could be recorded. Four reviewed tooling fixes followed (a comparator file missing from run bundles, smoke tests unable to apply full-size edits, a missing key that crashed the first selection run, and a guard exiting without logging). None of the four changed a scoring, selection or decision rule. The agent run cost $3.58 in API usage, and replay on E took about 59 laptop-CPU hours. On one NVIDIA A40, selection cost $4.38 (including $0.09 for the failed attempt), confirmation 4.36 hours and $2.57, and the follow-up 4.38 hours and $2.59, for $9.54 of GPU charges.

10 Conclusion

Under an evaluation that kept exploration, selection and confirmation apart, an AI agent proposed a rank-1 edit that repaired Qwen2.5-3B's habit of answering "unknown" with the answer in its own note. It recovered most false abstentions on sealed data, added few unsupported answers and beat three of five simpler fixes, at the cost of breaking a small share of correct answers. Both of the agent's edits beat random edits of the same form. The study does not show that the agent found something a logit bias cannot do, that it understood why the model fails, or that the edits generalise to natural questions, where most released answers were wrong. The next steps, in order, are a rerun with a consistent abstention prompt, a logit bias gated on the edit's read projection, a causal test of whether the unedited model uses what that projection detects, and answer types other than cities.

Disclosure of AI assistance

The weight edits studied here were proposed by Claude Opus 5.5 (Anthropic), run through the API as described in section 4. AI assistants (Claude and GPT models) also helped design the study, write and review the analysis code, and independently recount the results from per-row outputs. The author reviewed the design, the results and the text, and is responsible for the content.

References

[1] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2406.11717.

[2] Farid Hasanov, Riccardo Renzulli, Bahram Zonooz, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Dr. No: Feature steering for abstention in vision-language models, 2026. arXiv:2611.03862.

[3] Bryce Hinkley and Peyman Najafirad. Residual paving: Diagnosing the routing bottleneck in selective refusal editing, 2026. arXiv:2605.20262.

[4] Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–699. Association for Computational Linguistics, 2026. arXiv:2509.22449.

[5] Zipeng Ling, Shuliang Liu, Yuehao Tang, Junqi Yang, Shenghong Fu, Seonil Son, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. LLM abstention can be a prompt artifact, in addition to genuine uncertainty. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 12612–12633. Association for Computational Linguistics, 2026. arXiv:2507.16199.

[6] Yiqi Liu, Yang Wang, Songxin Wang, Chenghao Xiao, and Chenghua Lin. Broken symmetry in LLM refusal: Answer release is more local than refusal restoration. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 57765–57785. Association for Computational Linguistics, 2026. arXiv:2608.15772.

[7] Zhiyu Liu and Xinyu Yang. Computation, not weight access: Large language models as auditors of clinical prediction networks, 2026. arXiv:2611.07847.

[8] Manvendra Modgil. Readable is not steerable: Pre-registered tests of physical structure in learned fluid representations, 2026. arXiv:2611.05577.

[9] Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, Renqiang Luo, Erik Cambria, and Xiuzhen Zhang. The commit-abstain circuit: Why language models hallucinate instead of abstaining. In *Advances in Neural Information Processing Systems*, 2026. arXiv:2609.32964.

[10] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44293–44321, 2024. arXiv:2404.14394.

[11] Yohanes Andre Setiawan. Prompt-based abstention fails under misleading context: A controlled study of small frozen RAG models, 2026. arXiv:2608.22228.

[12] Chung-En Sun, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Steer2Edit: From activation steering to component-level editing. In *Advances in Neural Information Processing Systems*, 2026. arXiv:2602.09870.

[13] Rikiya Takehi, Jayden Teoh, Alex Zhang, and Omar Khattab. Sparks of weight-space programming: LLMs can natively read & write neural network weights. https://rikiyat.github.io/blog/weight-space-programming/, November 2026. Blog post, MIT CSAIL, 2 November 2026.

[14] Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *International Conference on Learning Representations*, 2025. arXiv:2410.03415.

[15] Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025. arXiv:2410.13708.

A Methods in detail

Prompt. Below is the user turn for case `E-DE01_ANSa_H0`, which sits inside the default chat template (system turn: "You are Qwen, created by Alibaba Cloud. You are a helpful assistant."). Only the memory and question vary. B2 appends "Do not reply unknown.", and B3 deletes the two sentences mentioning "unknown".

```
Below is your memory of a long conversation, followed by a question about it.  Answer with a short phrase
that states the fact asked for.  If the memory does not contain the answer, reply:  unknown
Each memory line shows who said it and when.  First write a brief note of at most 80 words:  name the facts
that bear on the question in a few words each (do not copy whole lines), then do any counting, adding or
date arithmetic.  Give dates as absolute dates (for example 7 May 2023), never relative ones such as
'yesterday' or 'last year'.  If the question asks for advice or a suggestion, answer with a brief
suggestion that uses what the memory says about the user.  Reply unknown only if the memory has nothing at
all on the topic.  End with one line of the form 'Answer:  <short phrase>'.

## Memory
user, 2023/03/01 (Wed) 10:00:  Ben lives in Manchester.
user, 2023/03/02 (Thu) 10:00:  Ben works as a manager.

## Current input
Question:  (asked on 2023/03/03 (Fri) 12:00) Where does Ben live?

Note:
```

Grading and utility. A missing `Answer:` marker counts as degenerate. Asserted ("York", "lives in York") and suggested ("possibly York") mentions are answers, but negated, doubted ("I don't know whether York") and quoted ones are not. With no answer mention, an abstention phrase gives the abstention class. A single city is graded gold, wrong or other. A city with an abstention phrase is mixed, and two or more distinct cities are conflicting. A separate frozen rule decides whether a note is answer-bearing. Utility uses error penalty $\lambda = 1$ (primary) or 2 (secondary).

Settings and edits. In the *total-effect* setting (repair) the model writes note and answer under the intervention. In the controlled-prefix setting (pathway) the model receives the unedited model's stored tokens up to and including the answer marker and generates at most 16 new tokens. B6 draws Gaussian changes on the same entries, matched in Frobenius norm column by column (seeds 1–5).

Inference. The paired percentile bootstrap resamples bases within task ($B = 10{,}000$, fixed seed). For a one-sided null $\theta \le \theta_0$, $p = (1 + \#\{b : \theta_b^* \le \theta_0\})/(B_{\text{valid}} + 1)$. Undefined replicates are excluded and counted, and a statistic with more than 1% failed draws gets no decision. If all valid draws equal the point estimate, $p = 1/(B_{\text{valid}} + 1)$ on the alternative side, else 1. These are approximate bootstrap tests. The 0.05 margins, 0.90 CF-follow threshold and 20-base minimum are conventions fixed before the data they govern.

B Artefacts

Commit identifiers refer to the project's Git repository, which is not yet public. The generator, grader and analysis code, the redacted agent transcript and submissions, the per-row outputs and the figure scripts are held but not yet released. The proposal revisions are `d9e67b7`, `1be9281` and `15ec3ee`, with F0 at `d29263a`, F1 at `1a37a72` (built at `c9ca458`) and F2 at `02bb2f6` (design hash `d02d9f9c...`). The tooling fixes are `b6e3d8e`, `4bf49ad`, `c880239` and `bf22ac3`. The confirmation report `h_report.json` (sha256 `51bbbd2e...`) was committed at `321a4f1`. The model revision is `aa8e7253...`, and the cases are 5,928 rows in 456 bases (sha256 `5eeb7c40...`). The edit recipes are `c64ea090781450cf`, `8364bf135ac59a83` and `87224655b7a5361e`, and edited weights will be released only if the model licence allows. For the follow-up, the proposal revisions are `b70e7c5`, `3a01186` and `e3b3018/ea88023`, the build is `9ff3d61` and the sign rule is `1dce499`. Its manifest (`b2b14bac...`, 86 arms) was committed at `a177536`, the N split has sha256 `651c930c...`, and the report (`report.json`, `f1ba6dad...`) was committed at `7b9e271`. An independent recount of all 87 arms (86 plus B0) reproduced it.