Steering False Event Reports: Fabricated Support and Subsequent Evidence Use

Anonymous authors
Paper under double-blind review

Abstract

Activation steering induces false event reports with fabricated support. In 96 blinded generations from 16 events, two independent raters find fabricated support in 0.72 and 0.69 of steered responses, versus at most 0.06 in controls (κ=0.667). Present-minus-absent residual directions raise false-report probability by 0.726--0.999 over norm-matched random controls across three Qwen2.5 checkpoints; unrelated-event directions have equivalent graded effects. These experiments link steered reports to generated supporting content in the tested settings.

Keywords: activation steering, language models, causal interventions, event reporting, confabulation

TL;DR: Event-nonspecific activation steering induces false reports and fabricated support in the tested settings.

1 Introduction

A model can read a record that rules out an event and still affirm that the event occurred. That answer becomes more consequential when the model supplies supporting details or uses it as a draft for later verification. Activation steering lets us change the report while retaining the same supplied record, then measure these consequences separately.

We study deterministic fictional-event contexts with known stated/absent labels. Three questions organize the experiments: does a direction need to come from the queried event, does a changed report come with changed content choices or fabricated support, and does the prior report affect later use of the record? Matched-norm random directions, unrelated-event directions and separate content readouts distinguish these possibilities.

The study makes two contributions:

1. **Steered false reports acquire fabricated support.** Two human raters identify fabricated support in roughly 70% of steered generations in a 96-response, 16-event cohort. The joint coherent-support outcome also increases relative to random controls. Directions fitted on development records induce false reports on disjoint records, and own- and unrelated-event directions have equivalent graded effects. Qwen and Granite also meet the joint report/content criterion on 256 new events. Crossed-label, taxonomy, dose and format controls specify the scope of these effects.
2. **A prior draft changes the benefit of consulting evidence.** In the paired two-pass analysis, Qwen's log-specific correction falls by 0.187 after a steered first-pass draft. The second pass receives the draft text and no activation intervention. SmolLM2 responds differently to the same record manipulation.

Appendix F lists the cohorts, fitting data, intervention sites and readouts.

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.

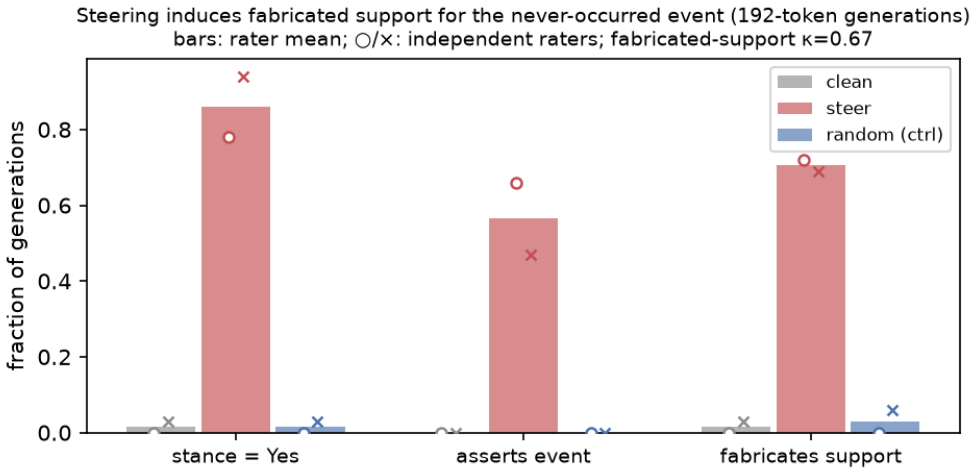


Figure 1. False reports extend to fabricated supporting content. Blinded ratings of 96 generations from 16 events measure affirmative stance, event assertion and fabricated support. Bars show the mean of two independent raters and markers show each rater; fabricated-support inter-rater $\kappa = 0.67$. Steering induces assertion and fabricated support for the absent event. Control rates are zero for Rater 1 and at most 0.06 for Rater 2.

2 Related Work

Verbal reports and their reference. People can supply reasons for behavior they did not have (Nisbett & Wilson, 1977) and defend choices they never made (Johansson et al., 2005); confabulation also occurs after neurological insult (Hirstein, 2005). In language models, chain-of-thought can be systematically unfaithful (Turpin et al., 2023; Lanham et al., 2023; Jacovi & Goldberg, 2020). Our reference label records whether the supplied context states the event. Steering changes the report while holding this stated/absent label fixed.

Relation to human memory studies. Misinformation studies examine later reports about experienced or suggested events (Loftus & Palmer, 1974; Loftus, 2005). Here the reference is an explicitly supplied fictional record. The task measures reports about that record; episodic storage and retrieval would require separate measurements.

Activation steering and model editing. Difference-of-means and contrastive steering move behavior along internal directions (Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023), while ROME and MEMIT edit stored facts (Meng et al., 2022; 2023). ContextFocus uses activation steering to improve contextual faithfulness under knowledge conflict (Anand et al., 2026), whereas broad evaluations show that steering can also induce coherent, cross-task misalignment across model families and scales (Cao et al., 2026). Most directly, Cox et al. (2026) decode answers before chain-of-thought, steer those answer directions to flip decisions, and observe confabulated premises after incorrect flips. Their result connects answer steering to fabricated premises; we extend the behavioral analysis to supplied event records and later record consultation.

Cox et al. examine whether chain-of-thought rationalizes a pre-committed answer. We give the model an external fictional-event record with a known stated/absent label, compare matched and unrelated event directions, directly test whether the unsupported content changes, and obtain blinded human ratings of fabricated support. Our experiment combines a supplied event record with unrelated-event equivalence tests and a separate content readout at each tested checkpoint. The specificity control measures how the report effect transfers between event-derived directions; the taxonomy comparison tests an affirmative direction fitted without event pairs.

Truth directions and the limits of probing. Internal truth-like signals are linearly decodable (Burns et al., 2023; Azaria & Mitchell, 2023; Marks & Tegmark, 2024), but decodability does not establish use (Hewitt & Liang, 2019; Elazar et al., 2021; Geiger et al., 2021). Subspace interventions can even produce the expected behavior by recruiting a dormant parallel pathway rather than the hypoth-

esized natural mechanism (Makelov et al., 2023). Injecting our probe direction therefore establishes sufficiency for this report; natural mediation would require a necessity-preserving intervention.

Hallucination, sycophancy, and self-correction. Hallucination studies measure unsupported outputs (Ji et al., 2023; Farquhar et al., 2024), while sycophancy studies examine prompt-pressure effects (Sharma et al., 2024). Our norm-matched intervention induces false-event reports with blinded ratings of fabricated support. The two-pass comparison additionally tests how the resulting answer text changes the measured benefit of consulting a fixed record during an unsteered verification pass.

3 Methods

3.1 Task: novel-entity event reports

Each item is a fabricated event about an invented person ("Cormac Ulwin traveled to Tallinn"), paired with matched present and absent contexts that differ in whether the event clause is stated, and a shared query suffix ("Did the following happen? … Answer (Yes or No):"). Items are generated deterministically from templates, with disjoint fitting and evaluation pools. SHA-256 hashes identify their contents. The **task truth** is the condition flag, logged independently of the report; the report is P(Yes) over {" Yes", " No"}. Invented entities reduce entity-specific pretraining priors.

3.2 Interventions and evaluation

The wrapper adds forward hooks to Hugging Face causal language models. The direction `v_avg = mean(resid_present) − mean(resid_absent)` is fit on one item pool; effects are measured on absent prompts from the disjoint evaluation pool. Four injections share the norm $\texttt{T} = \beta \cdot \|\texttt{v_avg}\|$: steer, norm-matched random, the event's own (self) direction, and an unrelated event's direction. The primary comparison is steer versus random; self versus unrelated-event tests event specificity with an equivalence analysis away from ceiling. We check task competence with clean present/absent AUROC. Sweeps over sites and doses exclude conditions where the random control itself saturates.

The content experiment fits only on `pilot[0:96]` and evaluates `confirmatory[96:160]` once at the site and dose fixed in advance. After the target-event Yes/No query, an A/B forced choice with target position balanced across items asks the model to select either the unsupported target statement or a concise event entailed by the same context. A model counts as showing the dissociation only if (i) the steer−random report effect is at least +0.20 with a positive 95% lower bound, (ii) hard content accuracy is at least 0.80, and (iii) 90% paired intervals for the steered unsupported-content probability relative to both clean and random lie strictly inside ± 0.10. Results are not pooled across models. Every run logs model revisions, precision, data identifiers, and item-level readouts.

3.3 Models

We evaluate Qwen2.5-Instruct at 0.5B, 1.5B, and 3B. We also test SmolLM2-1.7B-Instruct (Llama lineage) and Granite-3.1-2B-instruct (IBM lineage). The content test evaluates Qwen2.5-1.5B and Granite-3.1-2B. We explore Qwen2.5-7B/14B using a site-and-dose sweep and a separate evaluation on 64 new events (Appendix D). All models have open weights, and all interventions use neutral fictional content.

Analysis design. We select sites, doses and directions on fitting or development data before the report, content-holdout and crossed-task evaluations. The site-and-dose sweeps and free-text observations are exploratory. Additional analyses developed after examining the initial results comprise the joint human-rating outcome, two-pass interaction and factorial draft control, new-event and larger-model evaluations, and dose, taxonomy, format and fitting-pool analyses, including the stratified human audit. Their intervals describe the stated cohorts and fitted interventions.

4 Results

4.1 Clean reports discriminate the supplied context

Clean present-versus-absent P(Yes) is 0.499/0.098 (0.5B), 0.918/0.181 (1.5B), and 0.921/0.000 (3B), giving context-condition AUROC of 0.979, 1.000, and 1.000. Before steering, all three Qwen readouts discriminate whether the supplied context states the event.

4.2 Residual-stream steering flips the report

At sites chosen on the fitting split, steering increases false reports in all three Qwen models on 96 disjoint evaluation events.

Model	clean absent	steer	random	steer − random [95% CI]
0.5B	0.098	0.817	0.091	**+0.726 [+0.709, +0.742]**
1.5B	0.181	1.000	0.126	**+0.874 [+0.851, +0.896]**
3B	0.000	1.000	0.001	**+0.999 [+0.999, +1.000]**

Event-derived directions increase false reporting by 0.726--0.999 relative to norm-matched random injection. The random controls remain near their clean-absent baselines. Appendix D reports the 7B and 14B results on 64 new events.

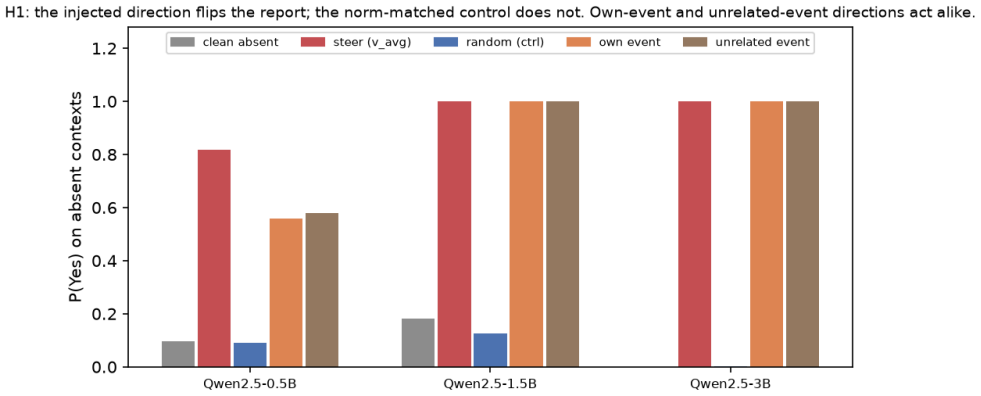


Figure 2. Mean P(Yes) on absent contexts per condition. The steering direction flips the report, the norm-matched control stays near the clean baseline, and the event's own (self) and unrelated-event injections have similar yes/no effects.

4.3 Event-nonspecific reports and content readouts

On the yes/no report, self − unrelated is −0.019 [−0.091, +0.048] at 0.5B, the only case away from ceiling; at 1.5B and 3B every content injection saturates and the contrast is uninformative. A graded readout (a verbalized 0–9 likelihood, taken as an expected value over digit tokens) removes the ceiling: the primary effect holds in graded units at every setting (+0.145 to +0.744), while |self − unrelated| is at most 0.0463. Against the ±0.10 margin fixed in advance, self − unrelated is statistically equivalent to zero at every Qwen evaluation setting; equivalence also holds at ±0.05 in five of six settings. Table F3 reports each setting. The graded comparison establishes event-nonspecific report steering across the tested directions.

A forced-choice experiment measures content identity separately on 64 unseen items per model. Let P_{false} be the probability assigned to the unsupported target rather than the event entailed by the context, Δ_R the report-level steer-minus-random effect, Δ_{C0} the P_{false} steer-minus-clean effect, and Δ_{CR} the corresponding steer-minus-random effect.

Model	Δ_R [95% CI]	Δ_{C0} [90% CI]	Δ_{CR} [90% CI]	dissociation
Qwen2.5-1.5B	**+0.877** **[+0.850, +0.900]**	+0.085 [+0.064, +0.107]	+0.076 [+0.057, +0.096]	no
Granite-3.1-2B	**+0.987** **[+0.979, +0.992]**	+0.002 [+0.001, +0.003]	−0.003 [−0.005, −0.001]	yes

Raw P_{false} values for clean / steer / random are 0.001 / 0.086 / 0.010 for Qwen and 0.008 / 0.010 / 0.013 for Granite. Clean hard-choice accuracy is 1.00 in both models, and both report criteria are met. Granite's two content intervals lie inside the ±0.10 margin fixed in advance: the direction induces a false event report without changing this content-identity readout by a meaningful amount. Qwen's hard choices also remain correct. Its +0.085 shift relative to clean has a 90% upper bound of +0.107, exceeding the +0.10 equivalence margin; its contrast with random meets the equivalence criterion. Only Granite meets the joint criterion in this cohort. The two separate equivalence decisions do not test a checkpoint interaction.

An additional evaluation uses 256 new entities within the same eight templates, with fixed Qwen layer 25/all-token and Granite layer 36/all-token injection at beta 4. Both checkpoints meet the joint report/content criterion. Qwen's content shift relative to clean is +0.0587 [90% CI +0.0416, +0.0759] and relative to the mean of five random directions is +0.0475 [+0.0324, +0.0628]; the report shift is +0.8088. Granite's content contrasts are -0.0002 and -0.0223, with both intervals inside ±0.10. Clean content accuracy is 0.9688 for Qwen and 1.0000 for Granite. Appendix D gives the full results for this cohort.

4.4 Effects across sites and doses

The sweep shows a broad mid-to-late steering band across injection sites and doses. For Qwen-1.5B, 17 of 22 readable conditions support the effect, peaking at +0.854 (post-block layer 25, β = 4); saturated conditions, where the random control reaches 0.90, sit next to the strongest ones and are excluded as unreadable. For SmolLM2-1.7B, the pre-selected site estimate is +0.041, while the sweep peaks at +0.715 (layer 14, β = 16, with the control at 0.192) and 19 of 24 readable conditions support the effect. Effect sizes vary across sites even after excluding saturated controls (Figure 3).

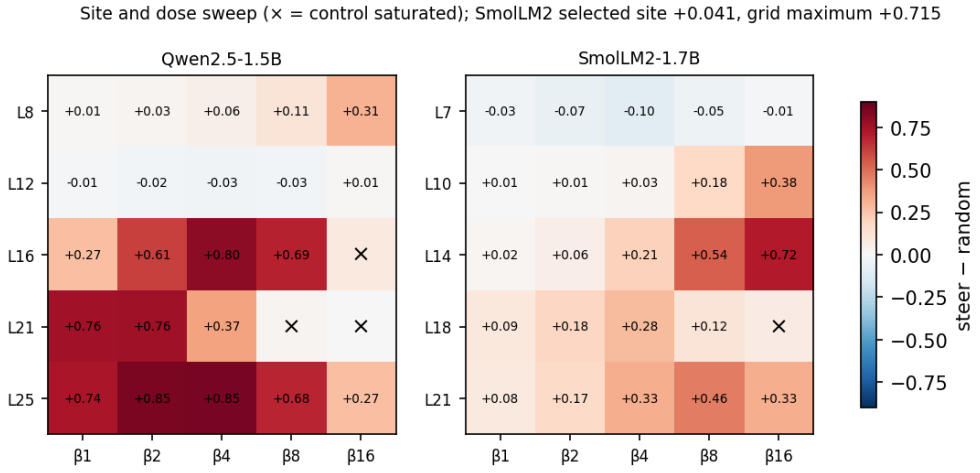


Figure 3. The steer − random contrast over sites and doses; × marks conditions where the random control saturates (≥ 0.90) and the contrast is unreadable. SmolLM2's pre-selected site has a smaller effect than most readable conditions in its sweep.

4.5 A decoded direction is sufficient to steer the report

A direction trained only to discriminate whether the context states the event also steers the report. At 1.5B (layer 23), P(Yes) moves from a clean 0.181 to 0.933 under probe steering, against 1.000 for the class-mean direction and 0.200 for random. Probe − random is +0.734 [+0.707, +0.760], 92% of the corresponding class-mean effect, despite a cosine of only 0.51 between the two directions. On a dis-

joint 64-event set with directions re-fit, probe − random is +0.630 [+0.612, +0.647], against +0.826 for the class-mean direction. Decodability alone is less selective: probe AUROC is approximately 1.0 at every layer. Thus, the probe direction is sufficient to steer the report at the tested site.

4.6 Steered generations contain fabricated support

We steer each decoding step on absent contexts and obtain blinded ratings of fabricated support. At moderate doses ($\beta \approx 1$–2) the stance flips to "Yes," and the model often invents a justification, sometimes quoting the exculpatory context as proof (for example, "…states that he 'stayed near home and rested for a full week.' This indicates that he indeed traveled to Tallinn…"). The norm-matched random control instead answers "No" with reasoning consistent with the context. At $\beta = 4$, outputs often degenerate into repetition, and fewer outputs contain a selected content word from the queried event. The lexical marker tracks event-word inclusion; human ratings below measure fabricated support and coherence separately. At 3B and $\beta = 1$, the hedge rate is 0.31: some outputs explicitly note the contradiction inside a steered "Yes."

Blinded human ratings provide the primary fabricated-support measure. The supplement includes the rubric and rating tool. Two independent raters scored all 96 blinded generations against the hidden condition key. For Rater 1, fabricated support is 0.72 [0.53, 0.88] under steering and 0.00 in both controls (paired steer − random +0.72 [0.53, 0.88]). For Rater 2, the corresponding rate is 0.69 [0.53, 0.81], versus 0.03 clean and 0.06 random (steer − clean +0.66 [0.50, 0.78]; steer − random +0.62 [0.47, 0.75]). Intervals use an event-cluster bootstrap because each event is shown under multiple conditions and models. Inter-rater agreement on fabricated support is Cohen's κ = 0.667 (raw agreement 0.875; positive agreement 0.750; negative agreement 0.917); stance κ = 0.826 and assertion κ = 0.728.

A blinded LLM judge gives a steering rate of 0.75 [0.56, 0.91], with κ = 0.972 against Rater 1 and 0.699 against Rater 2. On the fixed 192-token generations, steering coherence is 0.91/0.78 for the two raters versus 1.00 in both controls. The joint outcome identifies coherent fabricated support in 20/32 and 18/32 steered responses, versus 0/32 and 2/32 random-control responses. Appendix F gives the full denominators and event-cluster intervals.

A separate stratified audit of the 480-response dose/taxonomy extension adds 211 R1 ratings and 100 overlapping R2 ratings. R1 estimates definite fabricated support at 51.22% for event steering, 49.84% for taxonomy-Yes steering and 19.13% for random control; coherent fabricated support is 40.24%, 35.85% and 19.13%. The event-minus-random fabricated-support difference is +32.09 points [12-comparison-adjusted CI 12.93, 51.26]. All responses and indeterminate labels remain in the denominators (Tables F4--F5).

4.7 Additional checkpoints and baseline competence

SmolLM2 and Granite differ in baseline task competence and fixed-site effects. SmolLM2-1.7B has marginal baseline discrimination (AUROC 0.667 versus Qwen's ≥ 0.979), which limits interpretation of its intervention effect; its exploratory site-and-dose sweep finds a steer − random effect of +0.715 and a small self − unrelated contrast (+0.033 [−0.039, +0.103]). Granite-3.1-2B has discrimination AUROC 1.000, a primary effect of +0.865 [+0.828, +0.902] at the pre-selected site, and a site-and-dose sweep with 19 of 24 readable conditions, peaking at +0.988 with the control at 0.012. The fixed-site Granite result extends the report effect beyond Qwen. These checkpoints do not isolate architecture from training, scale or baseline competence.

4.8 Event-derived steering in a crossed A/B readout

An independent evaluation tests Qwen2.5-1.5B at layer 25 and β=1 on 256 new record units. Each unit crosses an event and a matched static badge attribute with supported/denied truth, positive/negated statements, and both A/B-to-Yes/No mappings. The direction, five norm-matched random controls, chat prompt and dose are fixed from development. The design yields 32,768 trial records at one checkpoint. Clean accuracy is 100% in 15 crossed cells and 87.5% in the denied-event, negated-statement, A=Yes cell.

The primary estimand is the mean signed change in positive-fact probability, averaged equally over truth, polarity and mapping, for event steering versus the mean random control. All three adjusted intervals lie within the prespecified ± 0.05 bound (Table 3). Secondary contrasts show a different response coordinate: P(A) increases by 0.0373 for events and 0.0368 for attributes, while mean P(Yes) decreases by 0.0114 and 0.0144. The small net semantic-score change coexists with an output-letter preference. The bound applies to the average signed contrast and permits opposing cell-level changes.

Table 3. Primary crossed-task contrasts. The three comparisons share a Bonferroni-adjusted family; intervals use a paired record-unit bootstrap within eight fixed templates. They condition on this checkpoint, prompt, dose and fitted directions. Appendix C gives definitions, secondary intervals and all cell-level results. Comparisons with the Yes/No experiment also involve changes in prompts, sites and doses.

Contrast	Estimate	Adjusted 95% CI	± 0.05 bound
Event semantic score	-0.00002	[-0.00014, +0.00011]	within
Attribute semantic score	+0.00288	[+0.00281, +0.00295]	within
Event minus attribute	-0.00290	[-0.00303, -0.00276]	within

4.9 The prior draft changes log-specific correction

The two-pass experiment has 48 shared items per checkpoint at β=4 and β=6. Event steering or a norm-matched random direction supplies a binary draft. The uninjected second pass receives that draft and query with no log, the complete log, or the log minus its target entry. Log benefit is P(Yes) without the log minus P(Yes) with it. The comparison pairs observations within records.

Table 4. Random and Steered columns give log benefits. The difference is random-draft minus steered-draft benefit. Positive values indicate less correction after a steered draft. Intervals use 50,000 item-bootstrap draws and Bonferroni adjustment over all four contrasts; doses reuse the same 48 items. Qwen's two rows repeat the same drafts and readouts.

Checkpoint	β	Random	Steered	Difference [95% CI]
Qwen2.5-1.5B	4	+0.197	+0.009	+0.187 [+0.126, +0.248]
Qwen2.5-1.5B	6	+0.197	+0.009	+0.187 [+0.126, +0.248]
SmolLM2-1.7B	4	-0.126	-0.111	-0.015 [-0.027, -0.006]
SmolLM2-1.7B	6	-0.130	-0.111	-0.019 [-0.030, -0.009]

A factorial control assigns no prior draft, Yes, or No to every record. Qwen's log benefit is 0.366, 0.009 and 0.229, respectively. No-prior-minus-Yes is +0.357 [six-contrast-adjusted CI 0.275, 0.435]; No-minus-Yes is +0.221 [0.157, 0.282]. Both remain positive under template sensitivity. SmolLM2 has negative log benefits in all three arms and median Yes/No token mass 4.48×10^{-5}, so its normalized scores concern a low-mass readout. Appendix A reports the complete grid and separates draft value from removal of the prior-query/answer prefix.

4.10 Taxonomic directions, weaker doses and paired formats

An affirmative direction fit on static noun/category pairs shifts absent-event reports in both Qwen and Granite. At beta 4 its event-report effect is nearly the same as event-derived steering. The comparison changes with dose: at beta 0.5, event-minus-taxonomy is -0.0070 for Qwen and +0.1233 for Granite, while both high-dose curves approach saturation (Figure 4). We evaluate both Yes/No orders, A/B mappings, graded reports and content choices, recording the total probability of each readout's labels. Small graded-label mass further limits interpreting normalized digit scores as spontaneous confidence.

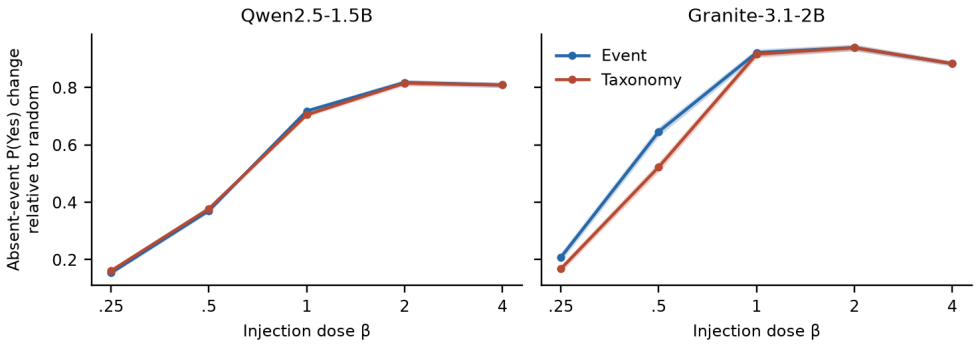


Figure 4. Event and taxonomic steering minus the event-matched mean random control, on the same 256 absent events under the original Yes-first mapping. Bands are pointwise 95% paired-event intervals. Other mappings remain in the complete data view.

Three renderings of 128 shared scenarios show substantial weak-dose format dependence. For Qwen at beta 0.25, event-minus-random is 0.1713 in prose, 0.0577 in dialogue and 0.0933 in transaction logs. Inference pairs all three formats within each of the 128 scenarios. Eight fixed resamples of each original fitting pool assess orientation sensitivity; two new fixed pools per checkpoint assess additional material sensitivity. The Qwen analysis is descriptive and reports its lower taxonomy fit accuracy (Appendix E.3). Appendix E reports all formats, doses and fits. Separate 7B/14B tests on 64 new events also meet the joint report/content criterion (Appendix D).

5 Discussion and limitations

A false report can acquire fabricated support. Human raters identify fabricated support in approximately 70% of steered generations, with a measurable coherence cost. Separately, the graded own/unrelated equivalence test shows that the binary report effect transfers across event-derived directions. Generated support and event specificity therefore require distinct readouts.

A prior answer changes the effect of consulting a record. The paired two-pass interaction quantifies the loss of log-specific correction in Qwen. The second pass is unsteered, and the retained draft is the only first-pass output it receives. Alternative drafts therefore change verification even when the supplied evidence remains fixed. In SmolLM2, consulting the log increases false affirmation under both draft conditions.

Report changes can coexist with stable content choices. Both checkpoints meet the joint report/content criterion on 256 new events: the report shifts while both content-choice intervals stay inside the specified bounds. Section 4.3 gives the original cohort and its intervals. The crossed A/B study identifies a separate response-coordinate effect, with output-letter preference alongside small average signed semantic changes.

Limitations. The independent evaluations cover open-weight checkpoints up to 3B; results at 7B/14B come from the exploratory new-event tests. The forced-choice task measures one content distinction, and the crossed-label average permits opposing cell-level changes. Fixed prompts and fitted directions leave the internal mechanism of the report effect open. The measured outcomes are reports and content choices about fictional contexts. Relating these outcomes to memory, belief or introspection requires additional measures.

The original 96 human-rated generations contain 16 unique events, so their intervals resample events. The additional stratified audit also spans 16 events; its finite-pool intervals account for stratified response sampling, with a separate event-cluster sensitivity analysis. Fabricated-support inter-rater agreement is $\kappa = 0.494$ on the 100-response overlap. Section 4.6 and Appendix F report each rater, including indeterminate labels and coherence costs. SmolLM2's weaker baseline discrimination limits its intervention comparison. The two-pass analysis covers 48 records in six fixed templates; the Qwen doses repeat second-pass inputs. The factorial extension reuses those records with assigned Yes/No or no-prior drafts. Its intervals condition on the checkpoints, and SmolLM2's binary token mass is small.

Appendix E reports sensitivity to dose, format and fitting pool. Graded-label mass accompanies normalized digit scores. Taking the worst contrast against five random draws leaves the strongest readable sweep effects positive (+0.683 Qwen-1.5B, +0.898 Granite, +0.236 SmolLM2). Transfer across prompts, event families and post-training changes remains to be evaluated.

6 Reproducibility

The anonymous artifacts include deterministic event sets, model revisions, fitting and evaluation splits, intervention sites and doses, trial readouts, analysis code and the blinded human-rating instrument. The release includes all 32,768 crossed-label trials and 192 two-pass model/dose/item records, and the factorial draft control adds 864 paired readouts with exact prompts. The on-disk records remain unchanged during intervention. The additional 480-response audit releases the original R1/R2 labels, sampling weights and a CPU replay of Table F4, separately from the original 96-response cohort.

Related anonymous submissions study candidate-selection dependence in verification and answer/query allocation. They share infrastructure but use distinct data, fitted objects, interventions and results; the relationships are disclosed through the submission system.

AI use statement

Generative AI tools were used to critique and refine hypotheses and experimental methodology; suggest experimental parameters; implement and edit analysis and experiment code; support statistical auditing and result interpretation; generate plotting code; search for and format literature; and draft and edit manuscript text. Human ratings supply the reported human endpoints; model judges are identified as study instruments and sampling aids. All AI-assisted analyses are checked against trial-level artifacts and automated tests. The authors take responsibility for the final content.

Ethics statement

This study analyzes open-weight language models and public or template-generated data. The reported rating task concerns fictional events and generated text. Aggregate outcomes use anonymous rater labels. No sensitive personal data are released.

Reproducibility statement

The appendix and anonymous supplement specify data partitions, model revisions, precision, seeds, intervention sites, evaluation criteria, and trial identifiers. The released test suite covers event generation and offsets, metrics, holdout separation, rating aggregation, and content-readout prompts.

References

Nikhil Anand, Shwetha Somasundaram, Anirudh Phukan, Apoorv Saxena, and Koyel Mukherjee. ContextFocus: Activation steering for contextual faithfulness in large language models, 2026. URL `https://arxiv.org/abs/2601.04131`.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Qi Cao, Jian Lou, Meiting Liu, Wenjie Feng, Dan Li, See-Kiong Ng, and Anh Tuan Luu. Activation steering induces emergent misalignment: A more comprehensive evaluation, 2026. URL `https://arxiv.org/abs/2606.08682`.

Kyle Cox, Darius Kianersi, and Adrià Garriga-Alonso. Post-hoc reasoning in chain of thought: Decoding and steering pre-committed answers, 2026. URL `https://arxiv.org/abs/2603.01437`.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019.

William Hirstein. *Brain Fiction: Self-Deception and the Riddle of Confabulation*. MIT Press, 2005.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023. doi: 10.1145/3571730.

Petter Johansson, Lars Hall, Sverker Sikström, and Andreas Olsson. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116–119, 2005. doi: 10.1126/science.1111709.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Elizabeth F. Loftus. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4):361–366, 2005. doi: 10.1101/lm.94705.

Elizabeth F. Loftus and John C. Palmer. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589, 1974. doi: 10.1016/S0022-5371(74)80011-3.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching, 2023. URL https://arxiv.org/abs/2311.17030.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in LLM representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977. doi: 10.1037/0033-295X.84.3.231.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023.

A Log consultation and draft-dependent verification

We compare prompt framing, within-pass log access, injection site and subsequent verification. Section 4.9 reports the paired two-pass interaction.

(1) Consulted-log framing. This comparison changes both prompt framing and the direction fitted for each framing, so its effect combines these two changes.

(2) First-person episodic agent with log ablation. For a content-matched first-person episodic agent, at the only effective injection site the logged condition changes as much as the bare one (+0.825 versus +0.778), and zeroing the model's read of its own log does not restore the original report.

(3) Earlier injection. Earlier injections leave more downstream layers to process the log but produce small steering effects (maximum bare effect 0.055). At these sites, neither the bare nor logged condition shows a report flip.

(4) Unsteered second-pass verification. An agent writes four first-person episodes through an on-disk store, produces a first answer under steering (effect +0.778), and then answers again with steering removed. After a random-control first pass, consulting the store reduces false reporting relative to mere reconsideration (log-specific effect +0.197 [+0.159, +0.229]). After a steered first pass, the log-specific point effect is +0.009 [−0.026, +0.046]; P(Yes) remains 0.64 against a 0.20 random-control first-pass baseline. Removing the topically related exculpatory entry lowers P(Yes) by 0.144, a harmful effect of including that entry in this prompt comparison. No injection is active during the second pass.

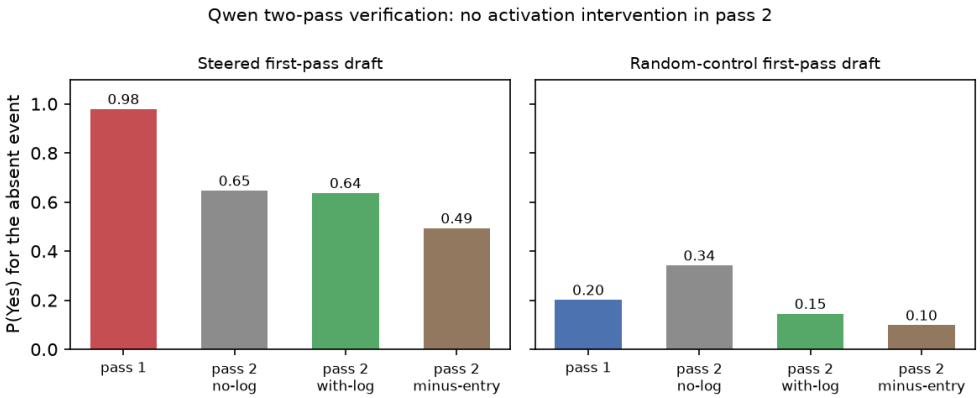


Figure 5. P(Yes) for the false event at pass 1 and in three pass-2 arms. Consulting the store reduces false reporting after random-control drafts (right). After steered drafts (left), the log-specific reduction is smaller; removing the exculpatory entry lowers false affirmation.

In SmolLM2 at β=4, reconsideration raises mean P(Yes) after random-control drafts from 0.49 to 0.76, and consulting the log raises it further to 0.88. The log-specific benefit is -0.126 under random-control drafts and -0.111 under steered drafts. Table 4 compares the paired effects at both doses. These model-specific results distinguish loss of an otherwise beneficial correction in Qwen from harmful log consultation in SmolLM2.

We define the interaction as the random-draft log benefit minus the steered-draft log benefit. Each bootstrap draw resamples the same 48 item identifiers jointly across both checkpoints, both doses and all record views. Source probabilities were saved to five decimal places. Intervals condition on the tested checkpoints and fitted directions. The six-template cluster-bootstrap sensitivity interval is [0.084, 0.262] for Qwen at both doses, [-0.032, -0.003] for SmolLM2 at β=4, and [-0.034, -0.006] at β=6. Omitting any one template leaves Qwen's interaction between +0.149 and +0.215; SmolLM2's point interactions remain negative. Both sensitivity analyses reuse the same records.

A.1 Factorial prior-draft control

The factorial control crosses three draft arms with all three log views on the same 48 records and both checkpoint revisions, yielding 864 readouts. Yes and No differ by the single draft word. The no-prior arm removes the two-line prior-query/prior-answer prefix; the log instructions and final question remain unchanged. All prompts run without activation interventions, in float32 with eager attention. Each item contributes one grid per checkpoint. The original Qwen experiment supplied Yes in all steered drafts and six random-control drafts.

Table A1. Conditional Yes/No probabilities and log benefit in the factorial control. Each row averages all 48 records.

Checkpoint	Draft	No log	Full log	Minus entry	Log benefit
Qwen	No prior	0.567	0.201	0.044	0.366
Qwen	Yes	0.644	0.636	0.494	0.009
Qwen	No	0.301	0.071	0.036	0.229
SmolLM2	No prior	0.584	0.751	0.720	-0.168
SmolLM2	Yes	0.779	0.889	0.884	-0.110
SmolLM2	No	0.729	0.870	0.867	-0.141

Table A2. Direct paired differences in log benefit and entry benefit, where entry benefit is P(Yes) with the target omitted minus P(Yes) with the full log. The six contrasts (three per checkpoint) share a Bonferroni-adjusted 95% family. The template column is a separate pointwise 95% sensitivity interval.

Checkpoint	Contrast	Difference [adjusted CI]	Template sensitivity CI
Qwen	No prior - Yes, log	+0.3571 [+0.2752, +0.4351]	[+0.1899, +0.4842]
Qwen	No - Yes, log	+0.2208 [+0.1566, +0.2818]	[+0.1047, +0.3084]
Qwen	No prior - Yes, entry	-0.0148 [-0.1000, +0.0587]	[-0.1987, +0.1011]
SmolLM2	No prior - Yes, log	-0.0574 [-0.0919, -0.0222]	[-0.1270, +0.0052]
SmolLM2	No - Yes, log	-0.0309 [-0.0418, -0.0209]	[-0.0512, -0.0128]
SmolLM2	No prior - Yes, entry	-0.0263 [-0.0520, -0.0002]	[-0.0797, +0.0265]

The 50,000 bootstrap draws resample records jointly across both checkpoints and all draft/log cells (seed 20260925). A second 50,000-draw analysis resamples the six whole templates (seed 20260926); all six leave-one-template-out values are released. Qwen's no-prior-minus-Yes log difference ranges from 0.298 to 0.389 across template omissions; its No-minus-Yes difference ranges from 0.180 to 0.251. The adjusted entry-benefit difference spans zero. SmolLM2's log differences are negative, while its no-prior and entry contrasts span zero in the template sensitivity. Across the complete grid, Yes/No token mass has median 0.983 and minimum 0.956 for Qwen, versus median 4.48×10^{-5} and minimum 1.15×10^{-6} for SmolLM2. The readout normalizes over those two tokens.

These interventions identify an effect of the supplied draft text and, for the no-prior arm, its surrounding prefix. Every second pass starts from its prompt with no hidden-state carryover. Intervals resample records and condition on the checkpoints and prompt construction.

B Association and repair selectivity

The preceding intervention results are exogenous: we add a direction and observe the report. A follow-up study asks whether the same linear coordinate is associated with errors without an intervention. It uses misleading gist lures: contexts that strongly suggest an event (the adoption paperwork is picked up, the boarding passes are printed) but never assert it. With no injection, models affirm these unsupported events at high rates (false-affirmation 0.677 for Qwen, 0.615 for SmolLM2, 0.969 for Granite).

The prediction is supported in both architectures where it can be evaluated: the clean activation's projection onto the steering direction, at the site fixed in advance, predicts, item by item, which lures are falsely affirmed, with AUROC 0.998 [0.991, 1.000] for Qwen and 0.732 [0.627, 0.830] for SmolLM2, each beating all K=5 random directions under the worst-case rule. The near-identity between projection and report (correlation +0.993 in Qwen) is partly mechanical, however, because a linear output readout is approximately a projection. This coupling limits the projection's interpretation as an independent predictor. Granite produces only three rejections in 96 items, too few to evaluate the binary outcome, so we report only its continuous coupling, +0.863.

The repair interventions reveal a correction/selectivity tradeoff. Parameter-free projection-out raises false affirmation in Qwen: the worst-case contrast is C = -0.223 [-0.274, -0.173] (raw change from clean +0.216), because zeroing the coordinate also removes negative projections associated with 31 correct rejections. Projection-out does little in SmolLM2; in Granite it lowers false affirmation by 0.473 but lowers true affirmation more (selectivity S = -0.213). The same tradeoff appears with counter-steering: counter-steering toward rejection reduces false affirmation beyond matched random subtraction (Qwen +0.413, Granite +0.568) but shows no selectivity in any architecture, reducing true affirmation at least as much. The direction acts as a signed event-evidence coordinate on these readouts: operations that reduce false affirmation also suppress true affirmation. None of the tested operations meets the selective-repair criterion.

A secondary observation is that gist-lure susceptibility dissociates from literal context discrimination. Granite, with task-faithfulness AUROC 1.000 on explicit absence, affirms 97% of lures, whereas Qwen, equally accurate on explicit absence, rejects a third of them. Literal discrimination on this task does not imply resistance to strongly suggested but unstated events.

C Crossed event and static-attribute evaluation

Design. We selected Qwen2.5-1.5B and the two-example chat readout on development data and set the procedure using a separate 64-unit matched-control development set before evaluating new records. The model revision is `989aa7980e4cf806f80c7fef2b1adb7bc71aa306`; injection is at `resid_post.25`, on every input token, with β=1. We reuse the event, unrelated-event and five norm-matched random directions fitted on development data. The 256 evaluation records use entity names absent from development data. The records are independent units within eight template families at one checkpoint. Development selection required at least 90% clean accuracy in every crossed cell. On the evaluation set, 15 clean cells have 100% accuracy and one has 87.5%.

Every unit has 2 task types $\times$ 2 truth conditions $\times$ 2 statement polarities $\times$ 2 output mappings $\times$ 8 interventions. Static controls describe eight badge attributes rather than events. Clean and intervened trials use the same text within each crossed cell. The evaluation includes the clean, event, unrelated and five random conditions for every record, using the prompts and doses selected on development data. FP16 inference uses exact-length batches of up to eight. The public protocol pins the model, direction and source hashes.

Readout and estimands. The A and B probabilities sum the bare and single-space token variants and normalize over those two semantic labels; their total vocabulary probability is retained separately. Mapping yields P(Yes), and question polarity yields the probability assigned to the positive event or attribute proposition. This last quantity is the semantic report score. For each unit and task, we average the signed event-minus-mean-random change across both truths, polarities and mappings. The two task means and their difference form the three primary contrasts. Their family-adjusted 95% intervals must lie wholly inside [$-$0.05,+0.05] to meet the prespecified average-effect bound. Opposing trial-level changes can cancel in this average.

Sample size was chosen using development variance, a planned adjusted half-width of 0.05, and a minimum of 256 / maximum of 512 units. The minimum determined the final sample. All 10,000 bootstrap draws resample record units within each of the eight fixed templates, carrying all task and intervention cells together. Intervals are conditional on the fitted directions, templates and checkpoint.

Secondary response coordinates. The four task-by-coordinate comparisons below use a separate Bonferroni family. Positive P(A) changes coexist with negative P(Yes) changes in both tasks. The intervention thus increases preference for the letter A under this protocol.

Table C1. Event direction minus the mean of five fixed random directions, averaged over truth, polarity and label mapping. The supplement includes all 128 cell means, accuracies and median label masses, together with the 32,768 trial outputs. Unrelated-direction contrasts are descriptive secondary analyses. The accompanying CPU script reproduces all scientific summary fields from these records.

Task	Readout	Estimate	Adjusted 95% CI
Event	P(Yes)	-0.01140	[-0.01153, -0.01128]
Event	P(A)	+0.03726	[+0.03698, +0.03756]
Attribute	P(Yes)	-0.01444	[-0.01451, -0.01438]
Attribute	P(A)	+0.03681	[+0.03659, +0.03702]

D New events and additional model scales

These exploratory evaluations use sites, doses and directions set before collecting their evaluation outputs. The first uses 256 new entities/events, balanced within the same eight templates. Qwen2.5-1.5B uses layer 25, Granite-3.1-2B layer 36, all-token injection and beta 4 with FP16 inference. The

larger-model evaluation uses a separate 64-event pool, Qwen2.5-7B layer 21 and 14B layer 43, all-token injection and beta 2 with native BF16. Fitted directions and norm-matched controls are fixed. Bootstrap draws pair conditions within events and stratify by template; the mean of five random directions is formed within an event. Both evaluations use the same eight template families.

Checkpoint	n	Δ_{C0} [90% CI]	Δ_{CR} [90% CI]	Δ_R
Qwen2.5-1.5B	256	0.0587 [0.0416, 0.0759]	0.0475 [0.0324, 0.0628]	0.8088
Granite-3.1-2B	256	-0.0002 [-0.0012, 0.0007]	-0.0223 [-0.0247, -0.0198]	0.8834
Qwen2.5-7B	64	0.0336 [0.0232, 0.0442]	0.0336 [0.0232, 0.0442]	0.9998
Qwen2.5-14B	64	0.0003 [0.0002, 0.0003]	0.0003 [0.0002, 0.0003]	1.0000

For 7B and 14B, the clean and mean-random content probabilities are close enough that the two contrasts round to the same four-decimal values. Both content intervals lie inside ± 0.10 in each row; the report contrast is positive. Section 4.3 reports the separate 64-item cohort for Qwen-1.5B and Granite. The supplement includes all mappings, 95% intervals and individual random-direction comparisons.

E Dose, readout and fitting-material sensitivity

The taxonomic control uses 24 fitting and 24 disjoint development noun/category pairs. Its Yes-minus-No direction is rescaled to the original event-direction norm; the reverse direction is also evaluated. The dose grid is beta in {0.25, 0.5, 1, 2, 4} on the same 256 events. For all 11 readouts, we record label mappings, full-vocabulary token probabilities and total label mass. Intervals below are pointwise, from 10,000 paired event resamples.

Checkpoint	Dose	Event − random	Taxonomy − random	Event − taxonomy [95% CI]
Qwen2.5-1.5B	0.25	0.1537	0.1611	-0.0074 [-0.0077, -0.0071]
Qwen2.5-1.5B	0.5	0.3706	0.3775	-0.0070 [-0.0077, -0.0062]
Qwen2.5-1.5B	1	0.7179	0.7046	0.0133 [0.0130, 0.0137]
Qwen2.5-1.5B	2	0.8175	0.8152	0.0024 [0.0023, 0.0024]
Qwen2.5-1.5B	4	0.8088	0.8088	4.29e-07 [4.05e-07, 4.54e-07]
Granite-3.1-2B	0.25	0.2082	0.1679	0.0403 [0.0382, 0.0425]
Granite-3.1-2B	0.5	0.6458	0.5225	0.1233 [0.1203, 0.1263]
Granite-3.1-2B	1	0.9214	0.9161	0.0054 [0.0050, 0.0057]
Granite-3.1-2B	2	0.9387	0.9387	6.07e-06 [5.91e-06, 6.23e-06]
Granite-3.1-2B	4	0.8833	0.8833	2.06e-06 [2.03e-06, 2.09e-06]

Large-dose convergence of normalized scores is compatible with saturation. At beta 4, Qwen's graded digit-label mass has medians 2.94e-5 for event steering and 2.52e-5 for taxonomic steering; the graded scores are expectations conditional on these digit labels. The fixed-site bridge separately evaluates the original 64 items at Qwen layer 23/last-token and layer 25/all-token sites, and Granite layer 36/all-token. Layer and intervention positions change together in the Qwen contrast. The complete mapping-level summaries of the 28,800 bridge rows accompany the scientific artifact.

E.1 Paired formats at low and high doses

A further bank contains 128 unique scenarios in four semantic families, rendered as prose, dialogue and transaction logs. Repeated formats share the scenario as the resampling unit. The bank crosses every format with beta 0.25, 0.5, 1 and 4. The following table gives original-mapping absent-event contrasts; all families, mappings, paired format differences and intervals remain in the complete data view.

Checkpoint	Format	Dose	Event vs random [95% CI]	Taxonomy vs random [95% CI]
qwen	prose	0.25	0.1713 [0.1691, 0.1735]	0.1862 [0.1841, 0.1884]
qwen	prose	0.5	0.4224 [0.4194, 0.4253]	0.4433 [0.4408, 0.4459]
qwen	prose	1	0.7823 [0.7809, 0.7837]	0.7790 [0.7774, 0.7806]
qwen	prose	4	0.8824 [0.8799, 0.8850]	0.8824 [0.8799, 0.8850]
qwen	dialogue	0.25	0.0577 [0.0549, 0.0606]	0.0685 [0.0655, 0.0717]
qwen	dialogue	0.5	0.2072 [0.2008, 0.2139]	0.2471 [0.2408, 0.2537]
qwen	dialogue	1	0.7464 [0.7422, 0.7505]	0.7829 [0.7800, 0.7857]
qwen	dialogue	4	0.9730 [0.9714, 0.9745]	0.9730 [0.9714, 0.9745]
qwen	log	0.25	0.0933 [0.0895, 0.0972]	0.1091 [0.1050, 0.1132]
qwen	log	0.5	0.2745 [0.2675, 0.2813]	0.3163 [0.3098, 0.3228]
qwen	log	1	0.7387 [0.7349, 0.7423]	0.7717 [0.7687, 0.7745]
qwen	log	4	0.9423 [0.9391, 0.9455]	0.9423 [0.9391, 0.9455]
granite	prose	0.25	0.0551 [0.0524, 0.0579]	0.0431 [0.0410, 0.0452]
granite	prose	0.5	0.4498 [0.4392, 0.4604]	0.3013 [0.2928, 0.3099]
granite	prose	1	0.9941 [0.9939, 0.9942]	0.9768 [0.9762, 0.9774]
granite	prose	4	0.9631 [0.9624, 0.9638]	0.9631 [0.9624, 0.9638]
granite	dialogue	0.25	0.0087 [0.0081, 0.0093]	0.0073 [0.0069, 0.0078]
granite	dialogue	0.5	0.1045 [0.0991, 0.1101]	0.0680 [0.0647, 0.0715]
granite	dialogue	1	0.9706 [0.9694, 0.9719]	0.8854 [0.8815, 0.8894]
granite	dialogue	4	0.9813 [0.9808, 0.9818]	0.9813 [0.9808, 0.9818]
granite	log	0.25	0.0134 [0.0119, 0.0149]	0.0111 [0.0100, 0.0122]
granite	log	0.5	0.1301 [0.1213, 0.1392]	0.0854 [0.0797, 0.0915]
granite	log	1	0.9706 [0.9695, 0.9717]	0.8897 [0.8863, 0.8931]
granite	log	4	0.9777 [0.9768, 0.9786]	0.9777 [0.9768, 0.9786]

E.2 Fitting-pool sensitivity

We draw eight stratified bootstrap samples from the original fitting pool and evaluate each at beta 1 and 4 in both checkpoints. Holding the original intervention norm fixed isolates sensitivity to direction orientation within this pool. Two additional fixed pools for Granite each contain 96 event and 24 taxonomy pairs with new fitting names and nouns, retaining the original templates and categories.

Both pools are evaluated at beta 0.25 and 1 on the same 256 events. Intervals condition on each fit; the range summarizes variation between the two specified pools.

Checkpoint	Dose	Original-pool resamples	Event range	Taxonomy range
qwen	1	8	[0.7167, 0.7186]	[0.7004, 0.7065]
qwen	4	8	[0.8088, 0.8088]	[0.8088, 0.8088]
granite	1	8	[0.9213, 0.9215]	[0.9149, 0.9168]
granite	4	8	[0.8833, 0.8833]	[0.8833, 0.8833]

The two new Granite fitting pools give:

Pool	Dose	Event vs random [95% CI]	Taxonomy vs random [95% CI]
1	0.25	0.2160 [0.2050, 0.2270]	0.1674 [0.1584, 0.1764]
2	0.25	0.2146 [0.2037, 0.2255]	0.1679 [0.1589, 0.1769]
1	1	0.9216 [0.9116, 0.9314]	0.9160 [0.9061, 0.9257]
2	1	0.9216 [0.9115, 0.9313]	0.9163 [0.9064, 0.9259]

E.3 Descriptive Qwen fitting-material sensitivity

A descriptive sensitivity analysis evaluates Qwen2.5-1.5B at layer 25 with two fixed fitting pools, each containing 96 event pairs and 24 taxonomy pairs. We evaluate both doses (0.25 and 1), both pools and all 11 readouts on the same 256 events. The original intervention norm and matched controls are held fixed. For taxonomy fitting, the positive-claim cell in pool 2 has accuracy 21/24, below the 0.90 competence criterion. We include both pools to describe sensitivity to fitting materials, reporting their accuracy alongside intervention effects. The 33,792 new measurements are repeated readouts of 256 events. Pointwise 95% intervals use 10,000 paired event resamples, stratified by template, conditional on each fixed fit.

Pool	Claim truth	Correct / 24	Accuracy	Median label mass
1	true	23	0.9583	0.9833
1	false	24	1.0000	0.9808
2	true	21	0.8750	0.9822
2	false	24	1.0000	0.9776

Absent-event Yes/No reports, with Yes first:

Pool	Dose	Event vs random [95% CI]	Taxonomy vs random [95% CI]
1	0.25	0.1537 [0.1507, 0.1567]	0.1587 [0.1558, 0.1616]
2	0.25	0.1543 [0.1513, 0.1574]	0.1575 [0.1546, 0.1603]
1	1	0.7187 [0.7143, 0.7231]	0.7015 [0.6971, 0.7058]
2	1	0.7182 [0.7138, 0.7226]	0.6971 [0.6928, 0.7014]

At dose 1, the event-minus-random report contrast is 0.7182--0.7187 across the two pools, compared with 0.6971--0.7015 for taxonomy-derived steering. Readout dependence remains: taxonomy-minus-random is negative for the normalized graded score at dose 1, despite its positive binary-report shift. Graded scores have very small digit-label mass and describe probabilities conditional on those ten labels.

Pool	Dose	Graded taxonomy vs random [95% CI]	Taxonomy digit mass
1	0.25	0.0028 [0.0027, 0.0030]	6.71e-07
2	0.25	0.0024 [0.0022, 0.0025]	6.70e-07
1	1	-0.0040 [-0.0045, -0.0035]	1.21e-06
2	1	-0.0061 [-0.0066, -0.0055]	1.20e-06

The supplement reports every mapping, present/absent report cell, content-choice cell, reverse taxonomy direction and paired comparison with the original fit.

F Protocol map and joint human-rating outcome

Tables F1 and F2 map the cohorts supporting the main results. Q denotes Qwen2.5-Instruct, G denotes Granite-3.1-2B-instruct, and S denotes SmolLM2-1.7B-Instruct. Layer numbers refer to `resid_post` hook indices; last/all specifies the injected token positions. Every stated dose is relative to the fitted direction norm. Tables F1--F2 describe the 96-response cohort; Table F4 reports the separate dose/taxonomy cohort.

Table F1. Fit/evaluation counts are per checkpoint; shared items are counted once when determining the sampling unit. Report, content and crossed-task evaluations use directions fixed before their evaluation outputs. Sections 4.4--4.5 describe the site-and-dose sweeps and probe tests.

Study / section	Checkpoint	Fit / evaluation units	Layer / positions / beta	Output
Fixed report / 4.2	Q 0.5B	96 / 96 events	13 / all / 4	Yes/No
Fixed report / 4.2	Q 1.5B	96 / 96 events	23 / last / 4	Yes/No
Fixed report / 4.2	Q 3B	96 / 96 events	30 / all / 2	Yes/No
Fixed report / 4.7	S 1.7B	96 / 96 events	16 / all / 4	Yes/No
Fixed report / 4.7	G 2B	96 / 96 events	28 / all / 4	Yes/No
Content holdout / 4.3	Q 1.5B; G 2B	96 / 64 events	25; 36 / all / 4	Yes/No + A/B content
Human ratings / 4.6	Q 1.5B; Q 3B	96 / 16 shared events	23 / last / 2; 30 / all / 1	Free text, 192 tokens
Crossed tasks / 4.8	Q 1.5B	Fixed vectors / 256 units	25 / all / 1	Two-example chat A/B
Two-pass / 4.9	Q 1.5B; S 1.7B	48 / 48 records	17; 14 / last / 4, 6	Binary draft; Yes/No
New events / App. D	Q 1.5B; G 2B	96 / 256 events	25; 36 / all / 4	Yes/No + A/B content
Larger models / App. D	Q 7B; Q 14B	96 / 64 events	21; 43 / all / 2	Yes/No + A/B content

The first five rows fit present-minus-absent directions on fitting contexts and evaluate the first 96 test events. The graded Qwen tests reuse those 96 events and hook sites at the original dose and beta 1, fitting directions on graded training prompts. Fixed-report, graded and original content-holdout comparisons each use one fixed norm-matched random direction per setting; the crossed-task and new-event studies use five. The content holdout uses the next 64 events, `confirmatory[96:160]`, and the original 96-item fit. It compares a target-event report with a separate target-versus-supported-content choice.

The human-rating cohort contains 16 of the original evaluation events, crossed with two checkpoints and clean/steer/random conditions: 96 generations, 32 per condition, each rated twice. Directions are fitted separately on 96 fitting prompts requesting explanations and applied during generation. The joint outcome below combines the support and coherence labels from this exploratory free-text cohort.

The crossed-task study reuses fixed development vectors after a separate 64-unit matched-control development study. Its 256 new record units carry both tasks, truth values, polarities, label mappings and eight interventions. The two-pass study instead fits on 48 fitting records containing episodes and evaluates 48 separate records, each containing four stored episodes. Only its first pass is steered; the

second receives the retained draft and a record view. Comparisons between these cohorts combine differences in prompts, sites, doses and readouts.

Appendix E supplies the sensitivity protocols: 24 taxonomic fitting pairs with 24 disjoint development pairs; five doses on the 256-event cohort; a fixed-site readout bridge on the original 64-item content holdout; three formats of 128 shared scenarios; and repeated fitting pools. Their repeated conditions remain paired within the event or scenario. Human ratings are reported separately below.

Table F2. Coherent fabricated support requires both binary labels to equal one for the same response and rater. Denominators include all responses in each condition, including incoherent responses. Rates under steering are 0.625 and 0.5625. The 2,000-draw bootstrap resamples the 16 event identifiers, carrying both checkpoints and all conditions together (seed 0). Intervals are pointwise and descriptive. We estimate each rater's outcome separately from the same 96 generations. Zero-width intervals reflect the observed all-zero cells in this small sample. The audit of the additional 480-response pool is separate from these estimates.

Rater	Clean, n/N [95% CI]	Steered, n/N [95% CI]	Random, n/N [95% CI]	Steer minus random [95% CI]
1	0/32 [0.00, 0.00]	20/32 [0.44, 0.81]	0/32 [0.00, 0.00]	+0.625 [+0.44, +0.81]
2	1/32 [0.00, 0.09]	18/32 [0.41, 0.72]	2/32 [0.00, 0.16]	+0.500 [+0.31, +0.69]

Table F3. Graded report contrasts on 96 events per setting. Scores normalize the expected 0–9 digit over nine, so the 0.10 margin corresponds to 0.9 rating points. The paired 2,000-draw bootstrap uses seed 0. A 90% interval strictly inside the margin implements the two one-sided equivalence decision at level 0.05; report-effect intervals are 95%. These are per-setting decisions, conditional on the fitted directions. Granite uses layer 36/all-token injection. Digit-label mass is rounded to three decimals, so values displayed as zero can have small positive probability. The separate content-choice criterion bounds absolute probability movement and can permit large relative changes near a probability floor.

Checkpoint	Beta	Steer minus random [95% CI]	Self minus unrelated [90% CI]	Equivalent at 0.10 / 0.05
Q 0.5B	4	+0.5714 [+0.5529, +0.5891]	-0.0463 [-0.0842, -0.0068]	yes / no
Q 0.5B	1	+0.2849 [+0.2644, +0.3053]	-0.0198 [-0.0327, -0.0070]	yes / yes
Q 1.5B	4	+0.4930 [+0.4681, +0.5166]	-0.0060 [-0.0163, +0.0040]	yes / yes
Q 1.5B	1	+0.1447 [+0.1368, +0.1527]	-0.0034 [-0.0092, +0.0019]	yes / yes
Q 3B	2	+0.7440 [+0.7104, +0.7781]	-0.0019 [-0.0085, +0.0046]	yes / yes
Q 3B	1	+0.6580 [+0.6241, +0.6907]	+0.0012 [-0.0135, +0.0166]	yes / yes
G 2B	4	+0.3271 [+0.3121, +0.3432]	-0.0180 [-0.0408, +0.0052]	yes / yes
G 2B	2	+0.1583 [+0.1542, +0.1622]	-0.0068 [-0.0195, +0.0065]	yes / yes

F.1 Stratified human audit of the dose/taxonomy extension

The extension crosses Qwen2.5-1.5B (layer 25/all tokens) and Granite-3.1-2B (layer 36/all tokens), five doses (0.25, 0.5, 1, 2, 4), 16 shared events and three conditions, yielding 480 responses. R1 scores 211: all 91 responses in the model-judge disagreement stratum and 120 sampled from the 389 agreement responses. Sampling is stratified by agreement, generator and condition. R2 scores a nested 100-response overlap. Rates use inverse inclusion weights; R2 uses the product of the initial and conditional overlap probabilities. We analyze each rater separately and include U/unclear as a distinct label.

Table F4. Design-weighted rates and pointwise 95% sampling intervals for each 160-response condition pool. Coherent fabricated support requires both labels to equal one; all responses remain in the denominator. Taylor ratio variances use first- and pairwise inclusion probabilities for R1 sampling and the nested R2 sample, including sampling without replacement. R2 indeterminate rates remain point estimates because two observed cells contain no U labels. The analysis includes both raters and all label states.

Measure	Event	Taxonomy-Yes	Random
R1 sampled responses	71	64	76
R1 definite Yes (%)	67.11 [57.58, 76.65]	62.52 [51.65, 73.40]	5.92 [1.38, 10.46]
R1 definite fabricated support (%)	51.22 [40.54, 61.90]	49.84 [38.76, 60.91]	19.13 [11.65, 26.60]
R1 fabricated support U (%)	15.42 [8.07, 22.76]	15.21 [7.46, 22.97]	11.61 [6.25, 16.96]
R1 coherent (%)	72.35 [64.04, 80.66]	68.11 [57.90, 78.33]	89.02 [83.66, 94.37]
R1 coherent fabricated support (%)	40.24 [29.72, 50.75]	35.85 [25.32, 46.39]	19.13 [11.65, 26.60]
R2 sampled responses	35	35	30
R2 definite fabricated support (%)	59.83 [42.39, 77.28]	52.92 [36.05, 69.80]	8.93 [0.00, 19.23]
R2 fabricated support U (%)	0.00	0.00	1.71

Table F5. Condition-minus-random contrasts in percentage points. Both interval columns adjust the same 12 comparisons (two raters, two contrasts and three outcomes). Finite-pool intervals condition on the 480 recorded responses. The additional sensitivity treats the 16 events as exchangeable clusters, carries all models, doses and conditions together, and retains both response-sampling stages. It uses a t_{15} reference; it describes variation across these synthetic events under that assumption. All intervals are approximate Taylor/Wald intervals.

Rater / contrast / outcome	Difference [finite-pool CI]	Event sensitivity CI
R1 / Event / support	32.09 [12.93, 51.26]	[14.14, 50.05]
R1 / Event / coherent support	21.11 [2.14, 40.08]	[0.97, 41.25]
R1 / Event / coherent	-16.67 [-31.20, -2.14]	[-38.90, 5.57]
R1 / Taxonomy / support	30.71 [11.07, 50.35]	[4.87, 56.55]
R1 / Taxonomy / coherent support	16.73 [-2.27, 35.72]	[-7.12, 40.57]
R1 / Taxonomy / coherent	-20.90 [-37.86, -3.95]	[-41.20, -0.60]
R2 / Event / support	50.90 [20.92, 80.89]	[21.70, 80.11]
R2 / Event / coherent support	22.39 [-5.10, 49.89]	[-5.52, 50.30]
R2 / Event / coherent	-30.36 [-59.57, -1.14]	[-64.14, 3.42]
R2 / Taxonomy / support	43.99 [14.76, 73.23]	[12.52, 75.47]
R2 / Taxonomy / coherent support	24.16 [-3.51, 51.83]	[-6.13, 54.45]
R2 / Taxonomy / coherent	-17.51 [-44.64, 9.62]	[-50.16, 15.15]

The nested-sample-weighted fabricated-support kappa for the original pool is 0.516 [pointwise 95% CI 0.392, 0.640]; the corresponding agreement is 71.21% [62.38, 80.05]. These population-weighted estimates and the raw overlap statistics below have different estimands.

On the 100-response overlap, raw agreement/Cohen kappa are 0.83/0.705 for stance, 0.70/0.494 for fabricated support, 0.93/0.835 for coherence and 0.43/0.065 for event assertion. Agreement is lowest for event assertion. Both raters' adjusted support contrasts exclude zero under both uncertainty definitions. The event-minus-random coherent-support interval excludes zero for R1 and spans zero for R2; the event coherence contrast spans zero in both event-sensitivity intervals. Similar event and taxonomy point estimates leave their equivalence unresolved. These estimates describe the 480-response cohort; Table F2 describes the separate 96-response cohort.