Conversational Relevance in the Conjunction Fallacy Is Represented Independently in Language Models

Jinil Kim
Departments of Linguistics, Psychology, and Computer Science and Engineering
Seoul National University
haba6030@snu.ac.kr

Jaejun Ryu*
KAIST AI
2fasturtle@kaist.ac.kr

Abstract

The conjunction fallacy is the judgment that a conjunction is more probable than one of its conjuncts. On one account, a preceding context makes one conjunct relevant, and that conjunct gains probability for that reason alone, independently of its posterior probability and of its confirmation by the evidence. Human experiments rest this independence on a null result, and their probability and confirmation ratings came from participants other than those who made the judgments. Here we show that six open-weight language models reproduce the effect, carry a relevance component that is separable from probability and confirmation, and shift the fallacy when that component is edited with little change in those two ratings. On a 15-item main bank and a 165-item model-written bank, context raised the probability assigned to the relevant conjunct relative to the irrelevant one in all 12 model–task cells. A linear probe decoded relevance from held-out activations at AUC 0.70 and 0.80, and decoding survived removal of the probability and confirmation directions. Adding the residual relevance direction induced the fallacy in two of the six models, and ablating it eliminated the fallacy in two others. Under those edits, the probability and confirmation ratings stayed within an equivalence margin fixed before the manipulation check in 8 of their 12 model–task–strength cells. This study shows that language models can reproduce a human heuristic judgment through discourse relevance, which is decodable from their activations, separable from the evidential ratings, and shifts the judgment when added or removed. Language models can thereby give pragmatics a falsifiable test of the judgment and of its independence from the evidential ratings.

1 Introduction

The conjunction fallacy is the tendency to rank a conjunction above one of its conjuncts, which probability theory forbids. People judge *Linda is a bank teller and a feminist* more probable than *Linda is a bank teller* [1]. Three accounts compete over what produces the ranking: a conversational inference that reads the bare conjunct as excluding the other [2], the strength with which the evidence *confirms* the added conjunct [3], or the relevance of that conjunct to the context [4].

On the relevance account, the added conjunct gains probability because the context makes it the conjunct at issue, rather than because the evidence bears on it. Chung et al. [4] normed each conjunct for how probable it is after the evidence (posterior probability, ρ) and for how much the evidence supports it (confirmation, C). They then asked whether the context moved those two ratings. The context moved the conjunction ranking and left the ratings unchanged, a pair of results they read as

*Work done while at the Department of Linguistics, Seoul National University.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Linguistic Principles for Foundation Models (LP4FM).

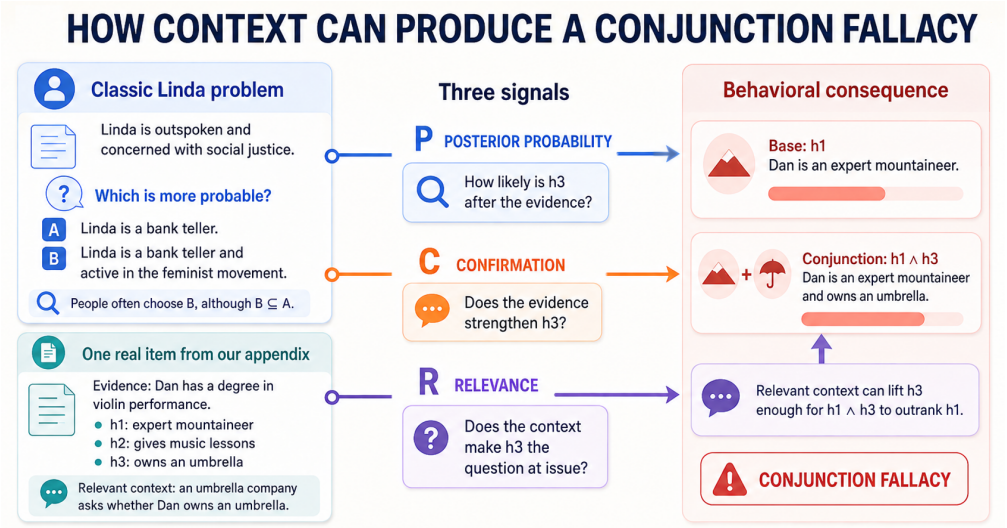


Figure 1: The Linda problem introduces the conjunction fallacy, and the Dan card shows one actual appendix item. The three colored tracks distinguish how probable h_3 is after the evidence (P), how strongly the evidence supports it (C), and whether the conversation makes it the proposition at issue (R). Across all tested model–task cells, relevant context raises h_3; the paper asks whether R can drive that preference beyond the measured P and C signals.

independence. The ratings, however, came from a separate pool of raters, and the independence rests on a null.

A language model removes that obstacle. One model ranks the conjuncts and answers the norming questions, and the residual stream that produced both admits decoding and editing. Prior work on these models has asked whether the fallacy is present rather than what produces it (Section 2). We varied relevance by changing which organization the context paragraph names. We asked the same model whether each conjunct is probable and whether the evidence strengthens it, and read the log-odds of each answer as its two ratings. From the same forward pass we decoded a relevance direction, then injected and ablated it to test whether the fallacy follows that direction.

Here, independence has an operational meaning: relevance remains linearly decodable after removing directions fitted to the model's prompted posterior-probability and confirmation readouts, and selected edits shift CF while those readouts remain within an equivalence margin. This does not establish independence from every possible evidential signal in the model.

Contributions. We report three tests in six open-weight language models. **Behavior.** The context effect appears in 12/12 cells of each item bank, and the evidential ratings neither confound nor carry it ($\rho^2 \leq 0.10$). **Representation.** A linear probe reads relevance from held-out items at AUC 0.704 in 12/12 cells, and decoding survives removal of the posterior-probability and confirmation directions. **Intervention.** Adding the residual relevance direction induces the fallacy in two models and ablating it eliminates the fallacy in two others, with the ratings inside a margin fixed before the manipulation check. The evidential edit meets that check in 0/13 cells, so the causal claim holds conditionally. This extends direction editing to a discourse-structural variable.

2 Related work

Three accounts of the conjunction fallacy. The accounts of the conjunction fallacy predict the same ranking and differ in what they hold responsible for it. Tversky and Kahneman [1] attributed the ranking to representativeness. Politzer and Noveck [2] and Dulany and Hilton [5] argued that participants read the bare conjunct as excluding the added one, so that the ranking follows a conversational inference. Crupi et al. [6] and Tentori et al. [3] showed that inductive confirmation of the added conjunct predicts the fallacy better than its posterior probability. Chung et al. [4] varied whether the added conjunct answers the question the context raises and found an effect of relevance beyond what confirmation explained. Each of these results is behavioral and comes from human

participants, and in each design the evidential norms come from a rater pool separate from the one that made the judgment. The present design takes the judgment and both norms from one system.

Reasoning biases and pragmatics in language models. Language-model work on the conjunction fallacy has measured whether the bias is present and how far it depends on surface cues. Binz and Schulz [7] and Macmillan-Scott and Musolesi [8] report the fallacy in GPT-3 and in later chat models, Echterhoff et al. [9] test debiasing prompts on decision tasks, and Wang et al. [10] score the representativeness heuristic directly. Jiang et al. [11] show that a model's success on the Linda problem collapses when the name is swapped, which separates lexical cues from logic. Pragmatic inference has entered this literature with scalar implicature as the dependent variable [12, 13], and syntax with acceptability judgments over minimal pairs [14]. These studies read the bias from prompted outputs and leave open what produces it. Hu and Levy [15] show that prompted judgments diverge from the model's own probabilities, which is why we score log probabilities and read the representation that produced them.

Reading and editing linear directions. The linear representation hypothesis [16] licenses reading a concept as a direction in the residual stream. Directions have been extracted and edited for refusal [17], for sentiment and other behaviors by activation addition [18, 19], and for truth [20]. The variables in that line are properties of a proposition or of a policy, whereas relevance is a relation between a proposition and the discourse. We extend direction editing to that relation, and each edit is also subject to a manipulation check: the edited direction must leave the measured evidential norms inside a equivalence margin fixed before the manipulation check.

3 Method

3.1 Items, conditions, and models

Each item pairs evidence about a person with two conjuncts. The evidence confirms h_2. The context makes h_3 relevant, and the evidence is silent on it. A frame is one of two wrappers around the prompt (Appendix C). A prefix is the introduction that the frame wraps, and only the introduction varies across conditions. Prefix `N` gives the evidence alone, `R` adds the context paragraph, and `I` keeps a paragraph of the same length that is relevant to neither conjunct (Appendix E). `R` against `I` isolates relevance, where `R` against `N` confounds it with the presence of a paragraph. The `conj` arm conjoins each conjunct with a shared h_1 and the `bare` arm presents them standalone. Each option is scored on its own continuation, since forced-choice rankings change with the scoring method [15, 21]. The frames and the wordings are in Appendix C. The **main bank** holds 15 items: four adapted from Chung et al. [4], three built on the human norming materials of Tentori et al. [3], and eight written by Claude under the same item specification. Screening norms for the eight generated items were set by that specification; the ρ and C readouts reported here were subsequently measured in each tested model. A **second bank** holds 165 model-written items (Appendices B and E). Six open-weight models of 1.5 to 4.3B parameters run every analysis, and every comparison between banks uses the same six (Appendix L).

3.2 Behavioral scores and evidential norms

Write $s_w(x \mid \pi)$ for the length-normalized log probability of continuation x after prefix π under wording w:

$$s_w(x \mid \pi) = \frac{1}{|x|} \sum_{t=1}^{|x|} \log p\big(x_t \mid \pi, w, x_{<t}\big). \tag{1}$$

The gain of conjunct i over the shared h_1 is

$$g_{i,w}(\pi) = s_w(x_i \mid \pi) - s_w(h_1 \mid \pi), \qquad i \in \{2, 3\}, \tag{2}$$

where x_i is $h_1 \wedge h_i$ in the conjunction arm and h_i in the bare arm. Averaging over the wording set W gives

$$\Delta_i^{A|B} = \frac{1}{|W|} \sum_{w} \big[g_{i,w}(A) - g_{i,w}(B)\big], \qquad \mathrm{DoD}^{A|B} = \Delta_3^{A|B} - \Delta_2^{A|B}. \tag{3}$$

A positive DoD means that the context raises the relevance conjunct more than the other. $\mathrm{DoD}^{R|I}$ is the minimal-pair relevance effect, and $\mathrm{DoD}^{R|N} = \mathrm{DoD}^{R|I} + \mathrm{DoD}^{I|N}$ item by item.

The same model also supplies two evidential ratings, which we call the **norms**. Posterior probability ρ is read from the log-odds of `yes` against `no`. Confirmation C is read from `strengthen` against `weaken`:

$$\begin{aligned} \rho_h(\pi) &= \log p(\texttt{yes} \mid \pi, q_\rho) - \log p(\texttt{no} \mid \pi, q_\rho), \\ C_h(\pi) &= \log p(\texttt{strengthen} \mid \pi, q_C) - \log p(\texttt{weaken} \mid \pi, q_C). \end{aligned} \tag{4}$$

Every behavioral judgment is thus paired item by item with norms from the same model and prefix. The 165-item bank and every separation and intervention analysis score the full continuations `yes`/`no` and `strengthen`/`weaken`. The 15-item behavioral norms were scored on the first answer token after a chat template (Appendix C).

3.3 Inference and association tests

A cell is one model by one task. For item-level values $d_1, \ldots, d_n$, we use the two-sided sign-flipping permutation test

$$p = 2^{-n} \sum_{\epsilon \in \{\pm 1\}^n} \mathbf{1}\left[\left| \tfrac{1}{n} \sum_i \epsilon_i d_i \right| \geq \left| \tfrac{1}{n} \sum_i d_i \right| \right]. \tag{5}$$

The test is exact at 15 items (2^{15} sign assignments) and Monte Carlo with 200,000 draws at 165. We Holm-correct across the two tasks within a model [22]. For cross-model summaries, we first average the two tasks within model and then bootstrap the six model means. The model is the resampling unit because the two tasks correlate at $\rho \approx .96$ over items.

Mediation would require the DoD to covary with the context-induced change in a norm. Confounding would require it to covary with the norm level. For posterior probability, the two item-level regressors are

$$u_i = \big[\rho_{h_3}(R) - \rho_{h_3}(N)\big]_i - \big[\rho_{h_2}(R) - \rho_{h_2}(N)\big]_i, \qquad v_i = \overline{\rho_{h_3}}_{,i} - \overline{\rho_{h_2}}_{,i}, \tag{6}$$

with C treated identically and the bar denoting the mean over the two prefixes. We correlate (Spearman) DoD with u_i for mediation and with v_i for confounding.

3.4 Decoding relevance and separating the norms

Let $a_\ell(i, \pi, x) \in \mathbb{R}^d$ be the layer-ℓ residual stream averaged over the tokens of x. For training items T, the relevance direction holds the conjunct fixed and varies only the prefix:

$$v_R^{(\ell)}(T) = \mathrm{unit}\left(\frac{1}{|T|} \sum_{i \in T} \big[a_\ell(i, R, h_3) - a_\ell(i, I, h_3) \big] \right), \qquad G_c = \mathrm{AUC}(h_3) - \mathrm{AUC}(c). \tag{7}$$

The layer is selected inside each held-out fold. A cell decodes only when its held-out AUC exceeds chance and the 95th percentiles of 1,000 random directions and 200 label permutations. The control c in G_c is h_2, h_1, or the two conjuncts pooled. Specificity gaps are tested by resampling items paired across the target and controls.

Frisch–Waugh–Lovell residualization [23, 24] removes the two norm columns before fitting and scoring. With target y, controls Z, and $x_i = a(i, R, h_3) - a(i, I, h_3)$,

$$\tilde{y} = y - Z\hat{\beta}_y, \qquad \tilde{x}_i = x_i - Z_i\hat{\beta}_x, \qquad v_{R|PC} = \mathrm{unit}(\tilde{X}^\top \tilde{y}), \tag{8}$$

where $y = \mathbf{1}$ for the relevance target and $Z = [\Delta\rho, \Delta C]$. The held-out item is scored by $\tilde{x}_{\mathrm{held}} \cdot v_{R|PC}$. We write $R \mid PC$ for this decoding analysis, $PC \mid R$ for the reverse analysis, and $R_{|PC}$ for the direction used in an edit.

3.5 Conditional intervention

The intervention outcome is the conjunction-fallacy score under the relevance prefix:

$$\mathrm{CF}_i = s(h_1 \wedge h_3 \mid R)_i - s(h_1 \mid R)_i, \qquad \Delta\mathrm{CF}_i = \mathrm{CF}_i^{\text{edit}} - \mathrm{CF}_i^{\text{base}}. \tag{9}$$

Injection targets items with $\mathrm{CF}^{\text{base}} \leq 0$. Ablation targets items with $\mathrm{CF}^{\text{base}} > 0$. Stage assignment uses one baseline frame and evaluation the other, then swaps the frames and averages within item. This prevents selection and outcome measurement from using the same observation.

Both edits act at the layer selected inside the held-out fold. Injection begins at the context paragraph, and ablation removes a fraction of the component already present:

$$a'_t = a_t + \alpha\,\sigma\,\hat{v}\,\mathbf{1}[t \geq t_{\text{intro}}], \qquad a' = a - \lambda\,\langle a, \hat{v}\rangle \hat{v}, \tag{10}$$

where σ is the norm of the position-averaged across-item SD of the training activations, $\alpha \in \{0.5, 1, 2\}$, and $\lambda \in \{0.5, 1\}$.

We edit two residual directions. For the intervention, x_i and the two norm columns are the difference-in-differences contrasts $[a(i,R,h_3) - a(i,I,h_3)] - [a(i,R,h_2) - a(i,I,h_2)]$ and the matching norm differences, so that the edited direction targets the same contrast as the behavioral DoD. Let $\hat{v}_P$ and $\hat{v}_C$ be the fitted norm directions, B an orthonormal basis for their span, and $\hat{v}_R = \mathrm{unit}(\bar{x})$. The shared norm scalar is $\mathrm{pc}_i = (z(P_i) + o\,z(C_i))/\sqrt{2}$, where o is the sign of the training correlation, and $v_{PC} = \tilde{X}^\top \widetilde{\mathrm{pc}}$. The edited directions are

$$R_{|PC} = \mathrm{unit}\big(\bar{x} - BB^\top \bar{x}\big), \qquad PC_{|R} = \mathrm{unit}\big(v_{PC} - \hat{v}_R \hat{v}_R^\top v_{PC}\big). \tag{11}$$

When the constructs share variance, the projection assigns that variance to one side by construction. Random and label-shuffled directions receive the same constraint.

3.6 Manipulation check and decision gates

With the identical hook and fold, we re-ask both norming questions. For $M \in \{\rho, C\}$, the edit-induced change is

$$\begin{aligned} \mathrm{DID}_M &= (M_{R,h_3} - M_{I,h_3}) - (M_{R,h_2} - M_{I,h_2}), \\ \Delta\mathrm{DID}_M &= \mathrm{DID}_M^{\text{edit}} - \mathrm{DID}_M^{\text{base}}. \end{aligned} \tag{12}$$

For $R_{|PC}$, both norms must remain inside a fixed equivalence margin. For $PC_{|R}$, both must shift in the predicted direction. The margin and the two intersection–union decisions are

$$\delta_M = 0.25 \times \mathrm{median}_i |\mathrm{DID}_M^{\text{base}}|, \qquad p^{\text{eq}} = \max_M p_M^{\text{TOST}}(\delta_M), \qquad p^{\text{sh}} = \max_M p_M^{\text{one-sided}}, \tag{13}$$

using two one-sided equivalence tests [25] and an intersection–union combination [26].

Five gates determine whether an intervention cell is read: representation, the predicted sign of $\Delta\mathrm{CF}$, specificity against random and shuffled controls, preservation of the base score, and fluency. The base-collapse guard requires $\max_i |s'(h_1 \mid \pi)_i - s(h_1 \mid \pi)_i| \leq 2\,\mathrm{SD}_i[s(h_1 \mid \pi)_i]$. The fluency guard permits less than a 10% WikiText perplexity increase. Appendix H gives the complete grid and cell counts.

4 Results

4.1 Behavioral effect and evidential norms

Context increases the gain of the relevance conjunct more than that of the other, in every model and outside conjunction. $\mathrm{DoD}^{R|N}$ was positive and Holm-significant in 12/12 cells of both banks, with a mean of 0.313 at 15 items and 0.356 at 165, and it held in the worst leave-one-item-out fold in 12/12 cells. The bare arm gave the larger DoD in 12/12 cells of both banks (Figure 2a,b). The effect therefore reaches a relevant proposition presented on its own. The two arms also differ in length, which this comparison leaves uncontrolled. $\mathrm{DoD}^{R|I}$, the minimal pair that isolates relevance, was Holm-significant in 12/12 cells of both banks, while the irrelevant paragraph alone, $\mathrm{DoD}^{I|N}$, reached significance in 0/12 cells at 15 items and in 12/12 at 165. The relevance term carried 58 to 85% of the full context effect at 15 items and 33 to 56% at 165. The relevance contrast is therefore the part of the context effect that both banks share.

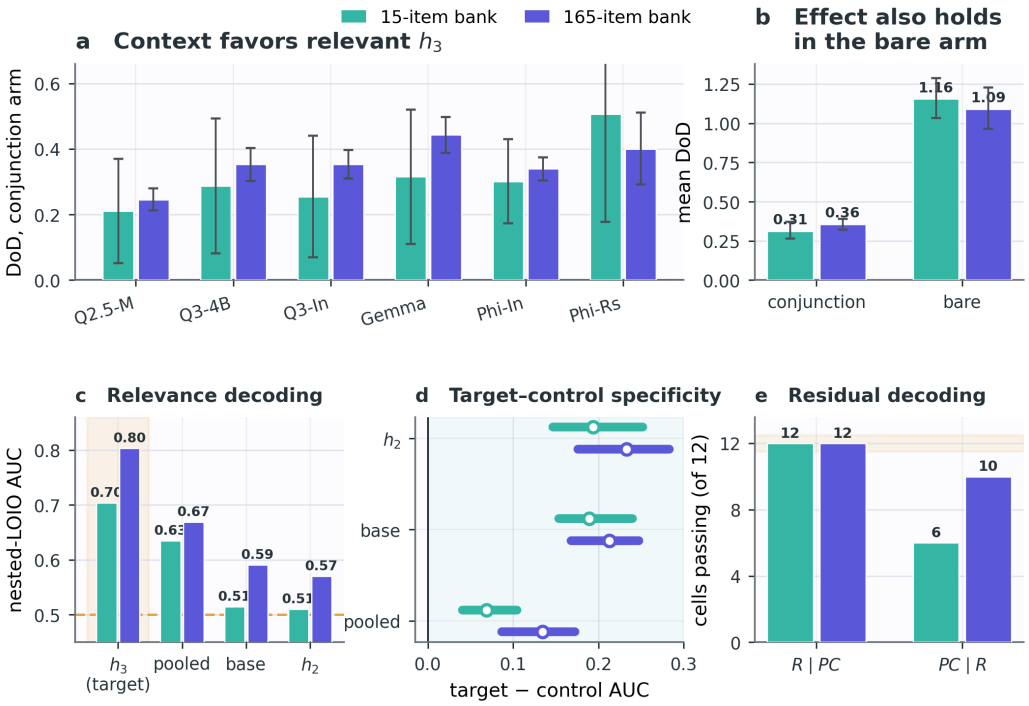


Figure 2: Behavioral and representational summary across six models and two tasks (12 model–task cells per bank). Light marks the 15-item bank and dark the 165 model-written items; intervals are item-bootstrap 95% CIs. (a) A positive difference-in-differences (DoD) means that replacing the no-context prefix with relevant context raises h_3 more than h_2; all cells are positive. (b) The larger bare than conjunction DoD shows that the contextual lift applies to the proposition itself, not only to a conjoined string. (c) Held-out AUC above the dashed chance line shows that the $R - I$ activation contrast identifies whether h_3 is conversationally relevant. (d) Positive target–control AUC gaps localize that contrast to the conjunct made relevant. (e) AUC remains above chance after the measured probability and confirmation directions are removed. Together, the panels show a relevance effect supported across both item banks and a readable relevance component beyond the two evidential ratings.

bank	$\mathrm{DoD}^{R\|I}$	$\mathrm{DoD}^{I\|N}$	$\mathrm{DoD}^{R\|N}$	$R\|I$ share	$I\|N$ Holm
15 items	0.215	0.097	0.312	0.58–0.85	0/12
165 items	0.164	0.192	0.356	0.33–0.56	12/12

Table 1: Where the conjunction-arm context effect comes from. Entries are mean normalized log-probability DoDs across the six models: $R|I$ changes only whether the paragraph makes h_3 relevant, $I|N$ adds a matched but irrelevant paragraph, and their item-wise sum is the full $R|N$ context effect. Positive values mean that h_3 gains more than h_2. "$R|I$ share" reports, across model–task cells, the fraction of the full effect attributable to relevance; the final column counts cells in which the paragraph-only term survives within-model Holm correction. Relevance explains most of the 15-item effect (58–85%) and remains substantial at 165 items (33–56%), while the irrelevant-paragraph term appears only with the larger bank's greater power.

Context shifts the evidential norms, which neither confound nor carry the effect. The 95% intervals excluded zero in 14/24 model by contrast comparisons at 15 items and in 21/24 at 165, and both ρ and C shifted for both conjuncts, with a sign that differed across models. Context therefore moves the ratings here, where it left them unchanged in humans [4]. The remaining question is whether those movements explain the behavioral contrast.

The norm shifts show no stable association with the DoD across banks. The confounding correlations for ρ were -0.034 $[-0.114, +0.040]$ and -0.025 $[-0.091, +0.042]$ in the 15- and 165-item banks.

For C they were $+0.023$ $[-0.132, +0.167]$ and $+0.016$ $[-0.030, +0.063]$. All four intervals include zero. The mediation correlations for ρ were -0.276 $[-0.394, -0.177]$ and $+0.080$ $[+0.023, +0.145]$, while those for C were -0.080 $[-0.243, +0.141]$ and $+0.094$ $[+0.053, +0.141]$. Thus the estimates that exclude zero reverse sign between banks, explain at most $\rho^2 = 0.10$, and replicate in direction in only 1/6 models for ρ and 2/6 for C. Only 1/12 and 2/12 cells were significant on their own. No equivalence margin was fixed for this correlational analysis, so this null is an absence of association rather than equivalence. The intervention check below supplies the stronger test (Appendix I).

4.2 Representational separation

Relevance is decodable from held-out activations, is specific to the conjunct it attaches to, and survives removal of the norms. The probe reached AUC 0.704 in 12/12 cells at 15 items and 0.803 at 165, with a gap over the h_2 control of $+0.194$ $[+0.146, +0.252]$ and $+0.233$ $[+0.176, +0.283]$. After residualization $R \mid PC$ decoded in 12/12 cells of both banks, against 6/12 and 10/12 for $PC \mid R$ (Figure 2c–e, with held-out curves in Figure S1). At 15 items the gate separated target from controls, 12/12 for h_3 against 1/12 for h_2 and 3/12 for h_1. At 165 items it saturated and passed all 48 cells, which is why we state specificity as an AUC gap. The residualization removed 76% of the norm of v_R at 165 items and 67% at 15, and relevance was still decodable in every cell. We read the asymmetry as a power artifact.

The two banks play complementary roles. The 15-item bank supplies the tighter irrelevant prefix and contains the tighter-control stimuli. There the target passes while most controls fail. The 165-item bank supplies narrower intervals and enough items for conditional intervention, but its permutation gate accepts AUCs as low as 0.516. For that bank, the continuous AUC gap is the more informative statistic. Results from both banks support the presence of a linearly decodable relevance signal across these stimulus sets.

4.3 Conditional intervention

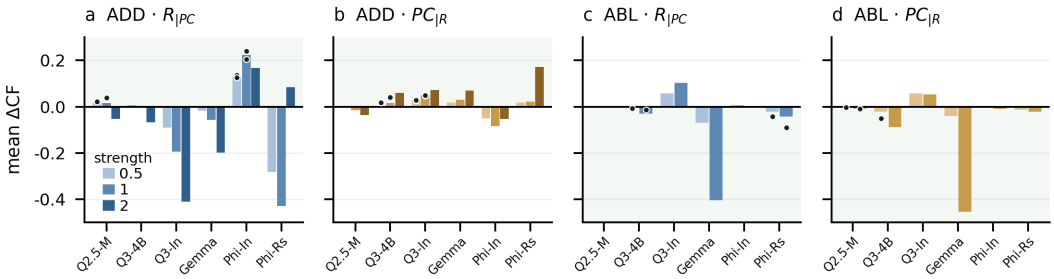


Figure 3: Conditional intervention on the two separated directions. Each bar is the mean change in conjunction-fallacy score, ΔCF, for one model and edit strength, averaged over the two tasks; positive values mean that an edit makes the fallacy more likely, and negative values mean that it suppresses the fallacy. The shaded half-plane marks the preregistered direction of success, and a black dot marks a cell that also passes the representation, matched-control specificity, base-preservation, and fluency gates; an undotted bar is therefore not counted as causal evidence even if its sign is favorable. (a,b) Injection is evaluated only on items lacking the fallacy at baseline and asks whether adding $R_{|PC}$ or $PC_{|R}$ induces it. (c,d) Ablation is evaluated only on baseline-fallacy items and asks whether removing the same component eliminates it. Ten combinations pass, spanning five models, whereas no constraint-matched random or label-shuffled direction passes. Dose response and the required norm checks appear in Figure 4.

The preregistered edit of the single mean direction returned a null. On the 15-item bank we edited v_R at every layer and position, with four refinements and thirteen controls (Appendix F). Ablation and injection each supported the hypothesis in 1 of 10 adjudicated cells, and the manipulation check passed in 0 of 20. Every direction moved the norms alike, by $+2.5$ to $+3.8$ in C, which is a non-specific disturbance. The geometry locates the cause in the sample. The held-out item's own contrast lies at $74.3°$ from the axis averaged over the other fourteen (mean cosine $+0.271$), and

item×model is the largest variance term (20.8% for ablation, 24.4% for injection). A decoder needs only the order of the projections, whereas an edit needs the component itself. That gap motivated the conditional design on the 165-item bank (Appendices G and N).

Editing the two separated directions induces or eliminates the fallacy in a subset of models and in no other cell. Ten model by task by stage by direction combinations passed every gate, with the predicted sign and an effect increasing with strength in 10/10, while the constrained random and shuffled directions passed none (Figures 3 and 4a; per-model counts in Appendix K). The The passing combinations came from five models. No model passed both injection and ablation of the same direction within one task; Qwen3-4B passed both for $PC_{|R}$ across different tasks. No model passed both for $R_{|PC}$. Eight cells shifted against the prediction under $R_{|PC}$ injection, so its 12-cell mean at $\alpha = 1$ was -0.073 against $+0.161$ in the passing cells. At $\alpha = 2$ the fluency guard cleared 0/12 cells, and the correlated tasks leave the 12 cells worth 6 independent ones. In the model–task–strength cells that pass the stated gates, the edit shifts the CF score in the predicted direction. The result is conditional on model, task, direction, and strength.

The passing pattern is heterogeneous rather than universal. $R_{|PC}$ injection passed in Qwen2.5-Math and Phi-4-mini-instruct, while its ablation passed in Qwen3-4B and Phi-4-mini-reasoning. $PC_{|R}$ injection passed in Qwen3-4B and Qwen3-4B-Instruct. Its ablation passed in Qwen2.5-Math and Qwen3-4B. Gemma passed no cell. No model passed both injection and ablation for $R_{|PC}$; the two stages succeeded in different models, so these results do not establish within-model sufficiency and necessity for that direction. Random and label-shuffled directions were projected out of the norm span in the same way as the targets and passed 0 cells, so the passing effects are specific to the fitted direction rather than to the projection.

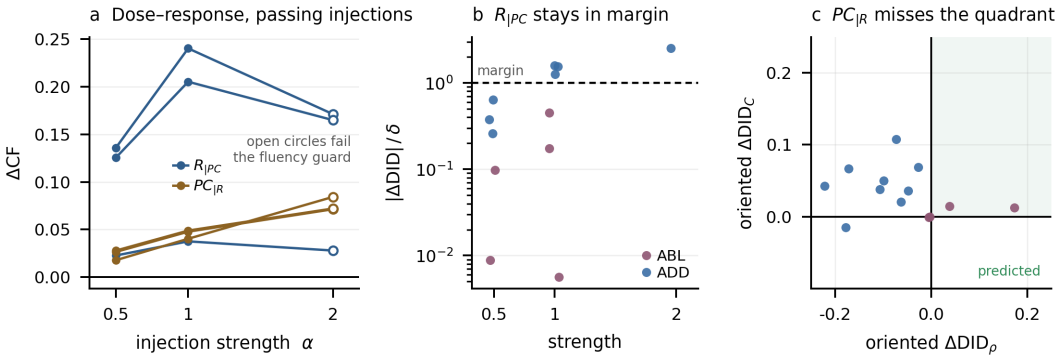


Figure 4: Dose response and evidential-norm checks for Figure 3. (a) ΔCF over injection strength; all six passing combinations move in the predicted direction, and open points fail the fluency guard. (b) Absolute $R_{|PC}$-induced displacements in posterior probability (ρ) and confirmation (C), each divided by its fixed equivalence margin: the lower-left unit square preserves both norms (8/12 cells). (c) Sign-oriented $PC_{|R}$ shifts: a coherent evidential edit belongs in the upper-right quadrant, but none does (0/13), chiefly because ρ moves oppositely. Thus the relevance edit can change the fallacy while preserving the norms; the pooled evidential edit fails its positive-control criterion.

The relevance edit keeps the norms inside the margin fixed before the manipulation check, and the evidential edit moves them in opposite directions. Under $R_{|PC}$, ρ and C were both equivalent in 8/12 strength cells, and under $PC_{|R}$ both shifted as predicted in 0/13 (Figure 4b,c). Under $R_{|PC}$ every ablation cell passed, 5/5 including $\lambda = 1$, and the four cells outside the margin were injections at $\alpha \geq 1$. The constrained controls separated two causes of leakage: at $\alpha = 1$ only the target direction shifted the ratings, consistent with $\cos(v_R, v_P) = .335$, and at $\alpha = 2$ random directions shifted them as well. Under $PC_{|R}$, C shifted as predicted in 6/13 cells while ρ shifted against it in all nine Qwen3 injection cells. The relevance direction therefore changes the fallacy while holding the two ratings, which is what the separation predicts. Control of the evidential direction remains open and requires separate ρ and C edits (Section 5).

5 Discussion and limitations

5.1 What the three analyses jointly support

Taking the judgment and its two ratings from the same model, in each of six models, yields three results. The effect is supported by both banks, the probability and confirmation ratings neither confound nor carry it ($\rho^2 \leq 0.10$), and its decodable component survives residualization. The bare arm has the larger DoD in 12/12 cells of both banks, so the context effect is common to the relevant proposition rather than specific to conjunction. Editing the decodable component changes the fallacy in two models per arm with the predicted sign. At $\alpha \leq 1$ the ratings stay within a fixed quarter of the context effect. In the tested models, the measured relevance signal is linearly separable and conditionally steerable with respect to the measured ho and C readouts: it remains decodable after linear residualization, while $R_{|PC}$ edits shift CF in selected model–task–strength cells.

The independence these results support is independence from the two measured norms. Three measurements carry it. After posterior probability and confirmation are projected out, a relevance direction still decodes held-out items in 12/12 cells of both banks. An edit along that projected direction changes the fallacy in every cell that passes the five gates. Under that edit, both norms remain inside their equivalence margin in 8 of 12 strength cells. The claim is confined to the projected direction. Before projection, v_R shares 76% of its norm with the ρ–C plane at 165 items, so the unprojected relevance and norm directions overlap. The claim is also confined to the passing cells. They cover five of the six models.

5.2 Why the two edits behave differently

The intervention results have a common cause in the geometry of the two directions. At 165 items, the ρ–C plane contains 76% of $\|v_R\|$, and v_R contains 72% of $\|v_{PC}\|$. Orthogonalization must therefore assign substantial shared variance, and it risks deleting relevance or moving the norms with it. Eight of twelve $R_{|PC}$ injections reverse, and $\alpha \geq 1$ increasingly leaves the equivalence margin. Ablation is cleaner, since it eliminates the fallacy while preserving both norms in 5/5 passing strength cells. The pooled $PC_{|R}$ edit instead moves ρ and C oppositely and passes 0/13 positive checks, so separate $\rho_{|R}$ and $C_{|R}$ edits are needed before claiming control of a coherent evidential component.

5.3 Limitations and implications

Three limits bound these claims. The six models form a convenience sample. Three come from the Qwen lineage, all have between 1.5 and 4.3B parameters, and gemma alone ran in fp32 (Appendix L). The results therefore describe these six models. The correlational analysis fixed no equivalence margin, so its null on mediation shows an absence of association rather than equivalence. No model passed both injection and ablation for $R_{|PC}$; the two stages succeeded in different models, so these results do not establish within-model sufficiency and necessity for that direction. Associative priming between words in the context and h_3 remains a possible contributor to the context effect. The present design does not isolate that contribution from conversational relevance.

Within those limits, the results show that a discourse-level variable shapes the behavior of these language models and is present in their representations. For pragmatics, the intervention supplies the test that human data left open for the relevance account of the fallacy [4]. In selected model–task–strength cells, adding or removing the residual relevance direction shifts the CF score. The measured probability and confirmation readouts remain within the stated equivalence margin in 8 of 12 passing strength cells. A claim about discourse structure thus becomes a test that one system can pass or fail. For language-model research, the results locate a pragmatic property, which proposition the discourse makes relevant, as a linear component that a probe reads and an edit moves. These models reproduce a human heuristic judgment. Whether training should preserve that heuristic or correct it as a probabilistic error is now a question that can be put to the responsible component.

References

[1] A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

[2] G. Politzer and I. A. Noveck. Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, 20(2):83–103, 1991.

[3] K. Tentori, V. Crupi, and S. Russo. On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, 142(1):235–255, 2013.

[4] W. Chung, K. Dorst, M. Mandelkern, and S. Mascarenhas. The conjunction fallacy: Confirmation or relevance? *Thinking & Reasoning*, 31(1):82–108, 2025.

[5] D. E. Dulany and D. J. Hilton. Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, 9(1):85–110, 1991.

[6] V. Crupi, B. Fitelson, and K. Tentori. Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, 14(2):182–199, 2008.

[7] M. Binz and E. Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

[8] O. Macmillan-Scott and M. Musolesi. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255, 2024.

[9] J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He. Cognitive bias in decision-making with LLMs. In *Findings of EMNLP*, pages 12640–12653, 2024.

[10] P. Wang, Z. Xiao, H. Chen, and F. L. Oswald. Will the real Linda please stand up...to large language models? Examining the representativeness heuristic in LLMs. In *COLM*, 2024.

[11] B. Jiang, Y. Xie, Z. Hao, X. Wang, T. Mallick, W. J. Su, C. J. Taylor, and D. Roth. A peek into token bias: Large language models are not yet genuine reasoners. In *EMNLP*, pages 4722–4756, 2024.

[12] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *NeurIPS*, pages 20827–20905, 2023.

[13] Y. Cho and S. Kim. Pragmatic inference of scalar implicature by LLMs. In *ACL Student Research Workshop*, pages 10–20, 2024.

[14] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

[15] J. Hu and R. Levy. Prompting is not a substitute for probability measurements in large language models. In *EMNLP*, pages 5040–5060, 2023.

[16] K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, pages 39643–39666, 2024.

[17] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *NeurIPS*, pages 136037–136083, 2024.

[18] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv:2308.10248*, 2023.

[19] N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering Llama 2 via contrastive activation addition. In *ACL*, pages 15504–15522, 2024.

[20] S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *COLM*, 2024.

[21] P. Tsvilodub, H. Wang, S. Grosch, and M. Franke. Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. *arXiv:2403.00998*, 2024.

[22] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[23] R. Frisch and F. V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933.

[24] M. C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

[25] D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

[26] R. L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982.

[27] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, et al. Qwen3 technical report. *arXiv:2505.09388*, 2025.

[28] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv:2409.12122*, 2024.

[29] Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, et al. Gemma 3 technical report. *arXiv:2503.19786*, 2025.

[30] Microsoft, A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *arXiv:2503.01743*, 2025.

[31] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. In *ICLR*, 2017.

Appendix

A Supplementary figures

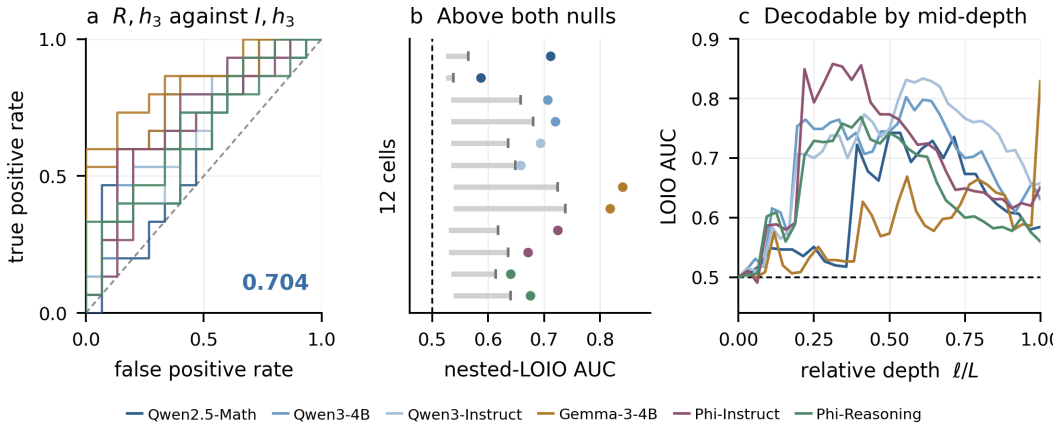


Figure S1: The relevance probe, held out. Twelve model by task cells of the 15-item bank, one color per model. (a) Receiver operating characteristics for R, h_3 against I, h_3 under nested leave-one-item-out, pooled within cell; the mean AUC over the twelve cells is 0.704. (b) Each cell against its own two null distributions, the 95th percentiles of 1,000 random directions and of 200 label permutations; every cell clears both. (c) Held-out AUC as a function of relative depth ℓ/L, averaged over the two tasks. Relevance becomes decodable from roughly mid depth, and the first fifth of the stack stays at chance.

B The 165-item bank

A second bank of 165 model-written items was screened as in Appendix E and run on the same six models. This bank tests whether the behavioral and representational results depend on hand-built stimuli, and it carries the conditional intervention of Section 4, whose per-item stage assignment needs more items than the main bank holds. Its irrelevant prefix is a weaker control than the main bank's, so it supports the behavioral result across a different stimulus set.

The DoD is positive and Holm-significant in 12/12 cells at a mean of 0.356, against 0.313 in the main bank, and the unconjoined arm is larger in 12/12 cells (1.091 against 0.356). Context shifts the norms in 21 of 24 comparisons. Both confounding contrasts contain zero. The mediation contrasts exclude zero with the opposite sign to the main bank, yet every interval overlaps its counterpart and models paired across banks agree in sign in 1 of 6 cases for ρ and 2 of 6 for C, so no difference between banks is established. What excludes zero here is consistency rather than size, with ρ^2 under 1%.

The probe reaches a held-out AUC of 0.803 in 12/12 cells. The permutation gate saturates at its resolution floor and passes all 48 cells, down to AUC 0.516, which is why the main text states specificity as a gap. The gap against h_2 is $+0.233$ $[+0.176, +0.283]$ and against h_1 $+0.213$ $[+0.168, +0.248]$. Residualization leaves relevance decodable in 12/12 cells and the reverse in 10/12, against 6/12 in the main bank. The projection that removes the ρ–C plane deletes 76% of the norm of v_R here and 67% in the main bank; this vector-level reading is post hoc (Appendix N).

C Prompt wordings

Every prompt is one of two frames wrapped around one question. Nothing else varies between conditions; the introduction is substituted verbatim.

Frames.

```
F1:  Consider the following text:
{intro}
{question}

F2:  {intro}
{question}
```

The judgment. Each option is scored on its own continuation, so the model never sees a list and no option occupies a position. Six wordings per task are averaged; the first of each set is Chung's.

```
TWO-STATEMENT TASK
Which of the two statements is more probable?  / Which is more
probable?  / Of the two statements below, which is more likely to
be true?  / Choose the more probable statement.  / Which of the
following two is more probable?  / Answer with the more probable
statement.

THREE-STATEMENT TASK
Rank the three statements from most to least probable.  / Which of
the three statements is most probable?  / Which statement is most
probable?  / Which is most likely to be true?  / Rank the following
by probability.  The most probable is:  / Answer with the most
probable statement.
```

Probability norm (ρ). Three wordings, scored on `" yes"` versus `" no"`:

```
Is it probable that {hyp}?  Answer yes or no.
Would you say that {hyp}?  Answer yes or no.
Based on the text above, is it likely that {hyp}?  Answer yes or no.
```

Confirmation norm (C). Three wordings. The 15-item behavioral norms score the first answer token, `" strength"` versus `" weak"`, after a chat template. The 165-item bank and the separation and intervention analyses score the full continuations `" strengthen"` versus `" weaken"` on the raw prompt. The evidence is already in the introduction, so the question refers to it rather than repeating it, since repetition would introduce lexical priming.

```
Consider the hypothesis that {hyp}.
Does the evidence described in the text strengthen or weaken that
hypothesis?

Consider the hypothesis that {hyp}.
Given the evidence described above, does that hypothesis strengthen
or weaken?

Consider the hypothesis that {hyp}.
How does the evidence described above affect that hypothesis --
strengthen or weaken?
```

Two frames $\times$ three wordings gives the six variants averaged within item; answer candidates are token-length matched, and a run halts otherwise.

D The fifteen items

Each block gives the two `ORG` fillers that separate R from I, the evidence, the three conjuncts, the question the context raises, and the norms used to screen the item. Norms are measured in human participants for the four Chung items, taken from Tentori et al. [3] for the three edited ones, and declared by the slot specification for the eight generated ones. The context paragraph is the same sentence frame throughout, with only `ORG` and the target's name substituted:

> ***[consultant]*** *is a consultant doing research for* ***[ORG]****, trying to discover new target groups in Europe for the company to market to. She calls a randomly selected person,* ***[target]****, and starts asking* ***[target]*** *questions. She finds out that* ***[evidence]***

`chung-1` *Chung* `violin`

ORG in R	an umbrella company
ORG in I	a dairy company
evidence e	Dan has a degree in violin performance.
h_1	Dan is an expert mountaineer.
h_2 (confirmation)	gives music lessons
h_3 (relevance)	owns an umbrella
question raised	*Does Dan own an umbrella, or does he not?*
norms	P_{h_2}=0.35 P_{h_3}=0.67 C_{h_2}=5.6 C_{h_3}=-0.1

chung-2 *Chung* swiss-man

ORG in R	a car company
ORG in I	a mattress company
evidence e	Noah is a Swiss man.
h_1	Noah likes making Italian desserts.
h_2 (confirmation)	can ski
h_3 (relevance)	has a driving license
question raised	*Does Noah have a driving license, or does he not?*
norms	P_{h_2}=0.68 P_{h_3}=0.83 C_{h_2}=4.7 C_{h_3}=-0.6

chung-3 *Chung* italian-undergrad

ORG in R	an American travel agency
ORG in I	a Belgian chocolate maker
evidence e	Carlo is an Italian undergraduate student.
h_1	Carlo has red hair.
h_2 (confirmation)	studied abroad in Barcelona in 2017
h_3 (relevance)	spent his summer holidays in America in 2017
question raised	*Did Carlo spend his summer holidays in America in 2017, or did he not?*
norms	P_{h_2}=0.16 P_{h_3}=0.12 C_{h_2}=3.9 C_{h_3}=-0.4

chung-4 *Chung* swedish-woman

ORG in R	a shampoo company
ORG in I	a bicycle company
evidence e	Alice is a Swedish woman.
h_1	Alice studies in Italy.
h_2 (confirmation)	works as a model
h_3 (relevance)	has brown hair
question raised	*Does Alice have brown hair, or does she not?*
norms	P_{h_2}=0.19 P_{h_3}=0.25 C_{h_2}=2.6 C_{h_3}=-4.1

ling-fix-1 *ling-fix* lingfix-russian

ORG in R	a chain that surveys who can collect an order without a car
ORG in I	a chain that surveys who prefers to shop before breakfast
evidence e	Katya is a Russian woman.
h_1	Katya lives in New York.
h_2 (confirmation)	is an interpreter
h_3 (relevance)	is a cyclist
question raised	*Is Katya a cyclist, or is she not?*
norms	P_{h_2}=0.3 P_{h_3}=0.25 C_{h_2}=4.0 C_{h_3}=0.0

ling-fix-2 *ling-fix* lingfix-american

ORG in R	a maker of padlocks choosing which size to produce
ORG in I	a maker of postcards choosing which scenes to print
evidence e	Jack is an American man.
h_1	Jack speaks Italian fluently.
h_2 (confirmation)	has visited twenty states
h_3 (relevance)	has a garden shed
question raised	*Does Jack have a garden shed, or does he not?*
norms	P_{h_2}=0.45 P_{h_3}=0.35 C_{h_2}=4.5 C_{h_3}=0.0

ling-fix-4 *ling-fix* lingfix-surgeon

`ORG` in R	a maker of precision tools choosing which grip to produce
`ORG` in I	a maker of wall calendars choosing which months to feature
evidence e	V. works as a surgeon.
h_1	V. plays in a brass band.
h_2 (confirmation)	is a man
h_3 (relevance)	is right-handed
question raised	*Is V. right-handed, or are they not?*
norms	P_{h_2}=0.77 P_{h_3}=0.78 C_{h_2}=2.4 C_{h_3}=0.2

`llm-1` *llmMade* `astronomer-gloves`

`ORG` in R	a brand testing who stays warm outdoors after midnight
`ORG` in I	a brand testing who reads printed newspapers after midnight
evidence e	Petra observes variable stars from her garden.
h_1	Petra is a qualified radiographer.
h_2 (confirmation)	logs eclipse timings
h_3 (relevance)	owns thermal gloves
question raised	*Does Petra own thermal gloves, or does she not?*
norms	P_{h_2}=0.45 P_{h_3}=0.6 C_{h_2}=6.0 C_{h_3}=0.3

`llm-2` *llmMade* `chess-stall`

`ORG` in R	a firm that surveys who is equipped to host a street stall
`ORG` in I	a firm that surveys who sorts household recycling correctly
evidence e	Nadia is a competitive chess player.
h_1	Nadia is a licensed pharmacist.
h_2 (confirmation)	studies endgame positions
h_3 (relevance)	owns a folding table
question raised	*Does Nadia own a folding table, or does she not?*
norms	P_{h_2}=0.35 P_{h_3}=0.6 C_{h_2}=6.0 C_{h_3}=0.0

`llm-3` *llmMade* `chorister-lisbon`

`ORG` in R	a firm that surveys who has experience of Portuguese payroll rules
`ORG` in I	a firm that surveys who has experience of local planning appeals
evidence e	Emil holds a diploma in early music.
h_1	Emil is a licensed drone operator.
h_2 (confirmation)	sang in a cathedral choir
h_3 (relevance)	worked a season in Lisbon
question raised	*Did Emil work a season in Lisbon, or did he not?*
norms	P_{h_2}=0.5 P_{h_3}=0.15 C_{h_2}=6.0 C_{h_3}=0.0

`llm-4` *llmMade* `pastry-van`

`ORG` in R	a firm that surveys who is able to deliver its own stock
`ORG` in I	a firm that surveys who is able to recite its slogan
evidence e	Bruno trained at a pastry school in Lyon.
h_1	Bruno is a volunteer coastguard.
h_2 (confirmation)	can temper chocolate
h_3 (relevance)	can drive a van
question raised	*Can Bruno drive a van, or can he not?*
norms	P_{h_2}=0.6 P_{h_3}=0.7 C_{h_2}=6.5 C_{h_3}=0.0

`llm-5` *llmMade* `rower-donor`

`ORG` in R	a service that needs to know who could give at its next session
`ORG` in I	a service that needs to know who could steward at its next session
evidence e	Corinne trains with a national rowing squad.
h_1	Corinne is a piano tuner.
h_2 (confirmation)	is a registered athlete
h_3 (relevance)	is a blood donor
question raised	*Is Corinne a blood donor, or is she not?*
norms	P_{h_2}=0.8 P_{h_3}=0.25 C_{h_2}=7.0 C_{h_3}=0.0

`llm-6` *llmMade* `vet-shellfish`

`ORG` in R	a caterer deciding which dishes are safe to hand round
`ORG` in I	a caterer deciding which napkins are cheap to hand round
evidence e	Daniela runs a small-animal veterinary practice.
h_1	Daniela is a magistrate.
h_2 (confirmation)	is trained in animal surgery
h_3 (relevance)	is allergic to shellfish
question raised	*Is Daniela allergic to shellfish, or is she not?*
norms	P_{h_2}=0.75 P_{h_3}=0.1 C_{h_2}=7.0 C_{h_3}=0.0

`llm-7` *llmMade* `wildlife-move`

`ORG` in R	a firm that surveys who has changed address recently
`ORG` in I	a firm that surveys who has changed energy supplier recently
evidence e	Farida shoots for a wildlife magazine.
h_1	Farida is a school governor.
h_2 (confirmation)	photographed nesting ospreys
h_3 (relevance)	moved house last spring
question raised	*Did Farida move house last spring, or did she not?*
norms	P_{h_2}=0.55 P_{h_3}=0.18 C_{h_2}=6.5 C_{h_3}=0.0

`llm-8` *llmMade* `yacht-trumpet`

`ORG` in R	a society recruiting for its brass ensemble
`ORG` in I	a society recruiting for its weekly quiz league
evidence e	Anders crews on a racing yacht.
h_1	Anders is a hospital porter.
h_2 (confirmation)	can read a compass
h_3 (relevance)	can play the trumpet
question raised	*Can Anders play the trumpet, or can he not?*
norms	P_{h_2}=0.55 P_{h_3}=0.12 C_{h_2}=6.0 C_{h_3}=0.0

E Item construction and screening

Four items are Chung's originals; their `ORG` slots were not separated in the published materials, so we factored them out and chose an irrelevant filler for each under four rules: it must answer neither conjunct, share no content word with the item, keep the paragraph's length within a few words, and stay plausible as a market-research client. Three further items edit that frame. The remaining eight were generated by Claude Opus (versions 4.6 and 5) against a slot specification and screened by the same family under an LLM-as-a-judge protocol that checks, for each candidate, that the evidence confirms h_2 and not h_3, that the context makes h_3 the answer to the question it raises, and that the two conjuncts are matched for length and register. Generation prompts, the judge rubric, and the accepted and rejected candidates are in `data/llm-build/`.

Norms P and C in Appendix D were fixed before any model was run and are used only to screen items, never as a covariate in the reported tests. No human participants took part in this study.

F Preregistered intervention grid and controls

Edits. Ablation removes the direction, $a' = a - \lambda P_k a$, with $\lambda \in \{0.25, 0.5, 1.0\}$; the preregistered cell is every layer and position at $\lambda = 1$, and we also run a variant restricted to the fold-selected layer so that ablation and injection match in layer budget. Injection adds $a' = a + \alpha\sigma\hat{v}$ at the selected layer from the context paragraph onward, with $\alpha \in \{0.5, 1, 2, 4\}$ and σ the across-item SD norm of the training activations.

Directions. Alongside the relevance direction we run three refinements and thirteen controls. The refinements are its projection onto the top two principal components, a two-dimensional item-adaptive push, and its component orthogonal to v_P and v_C. The controls are the same recipe on h_2 and on the bare option, a pooled variant, v_P, v_C, their span, a random one-dimensional direction, eight random two-dimensional subspaces, and a label-shuffled contrast. Ablation is applied to all twenty; injection to seventeen, since v_P, v_C and their span are pushed by a separate preregistered arm.

Guards. A cell is read only if it clears the base-collapse guard and the WikiText fluency guard, each evaluated at the worst of the fifteen folds.

Manipulation check. With the hook in place we re-ask the norming questions, giving $\Delta\rho$ and ΔC for each of h_2 and h_3. We require all four confidence intervals to contain zero for the cell to count as having moved relevance rather than evidence.

G The preregistered intervention

The preregistered grid edited the single mean direction v_R at every layer and position on the 15-item bank, with four refinements and thirteen controls (Appendix F). It returned a null. Ablation and injection each supported the hypothesis in 1 of 10 adjudicated cells, every direction moved the contrast by a similar amount, and the manipulation check passed in 0 of 20 cells. The geometry below accounts for the null and motivated the conditional design of Section 3 (Appendix N).

The edit disturbs the norms as much as the contrast. With the hook in place and the norming questions re-asked, all four intervals contain zero in 0 of 20 cells. The size of the shift is close to constant across directions: ablating v_{R,h_3}, the h_2 control, the bare control, v_P, v_C or their span each shifts C by $+2.5$ to $+3.8$. The edit therefore produces a non-specific disturbance rather than a change confined to relevance, and because the check gates both hypotheses, the signs below carry no reading in terms of relevance.

The behavioral signs also run counter to the hypotheses. Layer-matched ablation *raises* the DoD ($+0.009$) and injection lowers it in all four arms (-0.019 to -0.032), and the only directions that move as predicted are two controls. The three refinements give the same pattern, and the preregistered ablation cell clears the guards in 0 of 40 cells. The geometry accounts for this pattern, and it implicates the sample rather than the method.

Why fifteen items support decoding but not editing. Fifteen items suffice to decode a direction but not to edit one, because the direction is a mean of fifteen item contrasts that disagree. Holding out one item and averaging the other fourteen gives an axis whose cosine with the held-out item's own contrast averages $+0.271$, that is $74.3°$, and the cosine is negative in 14 of 150 folds. A decoder tolerates this, since it needs only the *order* of the projections onto the axis. An edit requires more, since the component it can move is a fifth of the contrast and the remainder stays as it was.

Higher-dimensional subspaces recover little more of the contrast. We swept the subspace dimension k from 1 to 14 under the same leave-one-item-out protocol. At $k = 14$ the basis spans the full training set, the maximum this design allows, and captures 0.289 of the held-out contrast against 0.178 at $k = 1$. Specificity falls with it: the advantage over a random subspace of the same size drops from $451\times$ to $49\times$, and the push direction stays within cosine 0.705 of the one-dimensional one. The limit is set by the disagreement between items rather than by the number of dimensions.

k	capture	random	ratio	
1	0.178	0.00039	$451\times$	current arms A, B, D
2	0.212	0.00097	$219\times$	current arm C
5	0.242	0.00218	$111\times$	
14	0.289	0.00592	$49\times$	full span of the training items

Table 2: Held-out capture as a function of subspace dimension, averaged over ten model–wording cells.

The variance is concentrated in the items. Decomposing the ablation effect over model, arm, dose and item, the largest term is item $\times$ model (20.8% for ablation, 24.4% for injection), roughly twice the item main effect and two orders of magnitude above the arm main effect (0.04–0.11%). Which item responds is a property of the item in a given model rather than of the item alone, so averaging fifteen of them leaves the noise the method assumes it removes.

Decoding and capture are dissociated. Across cells, decoding AUC and capture are uncorrelated (Spearman $\rho = +0.20$, $p = 0.58$): the cell with the lowest capture (0.073) decodes at 0.711, and the one with the highest (0.279) at 0.671. A component too small to change behavior under an edit is still signed consistently enough to order the items, which is what the decoder requires.

H The conditional intervention

Design. Items: the 165-item bank. Models: the six of Table 7. Cells: model $\times$ task wording. The CF outcome compares two length-normalized continuation scores. It is not a probability over nested events and need not satisfy the conjunction inequality. A positive CF is therefore a conjunction-style continuation preference, not by itself a demonstrated probabilistic violation. Stage per item from the baseline CF on one frame, evaluated on the other, cross-swapped and averaged within item. Directions $R_{|PC}$ and $PC_{|R}$ as in Section 3, with random and label-shuffled directions under the same orthogonality constraint. Injection $\alpha \in \{0.5, 1, 2\}$; ablation $\lambda \in \{0.5, 1\}$, both at the fold-selected layer. Gates: representation gate; one-sided item sign-flip on ΔCF (injection > 0, ablation < 0); paired specificity against random and shuffled controls; base-collapse guard (≤ 2 item SD); WikiText fluency guard ($< 10\%$ perplexity increase over 1,000 tokens, all 165 folds for the target, 15 diagnostic folds for controls). Raw p-values are reported without Holm. Items per stage and cell range from 45 to 131.

stage	direction	strength	cells passed	mean ΔCF, all	mean ΔCF, passing	fluency
ABL	$PC_{\|R}$	0.5	2/12	-0.005	-0.027	20/24
ABL	$PC_{\|R}$	1	1/12	-0.088	-0.009	16/24
ABL	$R_{\|PC}$	0.5	2/12	-0.005	-0.025	20/24
ABL	$R_{\|PC}$	1	2/12	-0.061	-0.052	16/24
ADD	$PC_{\|R}$	0.5	3/12	$+0.001$	$+0.024$	21/24
ADD	$PC_{\|R}$	1	3/12	$+0.003$	$+0.046$	17/24
ADD	$PC_{\|R}$	2	0/12	$+0.047$	—	0/24
ADD	$R_{\|PC}$	0.5	3/12	-0.039	$+0.095$	21/24
ADD	$R_{\|PC}$	1	3/12	-0.073	$+0.161$	17/24
ADD	$R_{\|PC}$	2	0/12	-0.079	—	0/24

Table 3: Grid summary. *cells passed* counts cells passing every gate. The all-cell mean and the passing-cell mean can differ in sign. *fluency* is the number of the 24 stage cells (12 target, 12 control) under the 10% perplexity threshold.

model	stage	direction	strengths	tasks	ΔCF range	PPL rise
Qwen2.5-Math-1.5B	ABL	$PC_{\|R}$	0.5, 1	2	-0.009 to -0.004	0.1–0.2%
Qwen3-4B	ABL	$PC_{\|R}$	0.5	2	-0.050	2.2%
Phi-4-mini-reasoning	ABL	$R_{\|PC}$	0.5, 1	3	-0.091 to -0.043	0.6–1.6%
Qwen3-4B	ABL	$R_{\|PC}$	0.5, 1	3	-0.013 to -0.007	0.8–14%
Qwen3-4B	ADD	$PC_{\|R}$	0.5, 1	3	$+0.018$ to $+0.040$	3.3–20%
Qwen3-4B-Instruct-2507	ADD	$PC_{\|R}$	0.5, 1	2, 3	$+0.026$ to $+0.049$	0.4–16%
Phi-4-mini-instruct	ADD	$R_{\|PC}$	0.5, 1	2, 3	$+0.125$ to $+0.240$	0.7–38%
Qwen2.5-Math-1.5B	ADD	$R_{\|PC}$	0.5, 1	3	$+0.023$ to $+0.038$	1.1–5.8%

Table 4: The ten passing combinations, grouped by model, stage and direction. All ten carry the predicted sign and are dose-monotone. PPL rise spans the strengths run, including $\alpha = 2$, where it exceeds the guard. Gemma passes no cell: specificity holds in 4 of 20 target cells, and ablation at $\lambda = 1$ raises perplexity 2,718-fold.

I Norm shifts and the evidential path

Context shifts the norms in 14 of 24 model by contrast comparisons at 15 items and 21 of 24 at 165. The confounding correlations between the DoD and the norm level are -0.034 $[-0.114, +0.040]$ for ρ and $+0.023$ $[-0.132, +0.167]$ for C at 15 items, and -0.025 $[-0.091, +0.042]$ and $+0.016$ $[-0.030, +0.063]$ at 165; all four include zero. The mediation correlations, between the DoD and the context-induced change in the norm, are -0.276 $[-0.394, -0.177]$ for ρ and -0.080

model	T	stage	direction	str.	ΔDID_ρ	p_ρ	ΔDID_C	p_C	pass
Qwen2.5-Math-1.5B	2	ABL	$PC_{\|R}$	0.5	$+0.003$	.921	$+0.000$	.685	no
Qwen2.5-Math-1.5B	2	ABL	$PC_{\|R}$	1	$+0.005$	.924	$+0.001$	.689	no
Qwen3-4B	2	ABL	$PC_{\|R}$	0.5	-0.038	.002	-0.015	.051	no
Qwen3-4B	2	ABL	$PC_{\|R}$	1†	-0.173	$< .001$	-0.012	.215	no
Qwen3-4B	3	ADD	$PC_{\|R}$	0.5	-0.063	1.00	$+0.021$	$< .001$	no
Qwen3-4B	3	ADD	$PC_{\|R}$	1	-0.107	1.00	$+0.038$	$< .001$	no
Qwen3-4B	3	ADD	$PC_{\|R}$	2†	-0.172	1.00	$+0.067$	$< .001$	no
Qwen3-4B-Instruct-2507	2	ADD	$PC_{\|R}$	0.5	-0.027	.904	$+0.069$	.010	no
Qwen3-4B-Instruct-2507	2	ADD	$PC_{\|R}$	1	-0.073	.957	$+0.107$	.038	no
Qwen3-4B-Instruct-2507	2	ADD	$PC_{\|R}$	2†	-0.222	.993	$+0.043$	.344	no
Qwen3-4B-Instruct-2507	3	ADD	$PC_{\|R}$	0.5	-0.048	.998	$+0.036$	.026	no
Qwen3-4B-Instruct-2507	3	ADD	$PC_{\|R}$	1	-0.099	.997	$+0.050$	.109	no
Qwen3-4B-Instruct-2507	3	ADD	$PC_{\|R}$	2†	-0.177	.985	-0.015	.571	no
Qwen2.5-Math-1.5B	3	ADD	$R_{\|PC}$	0.5	-0.016	$< .001$	$+0.004$	.001	yes
Qwen2.5-Math-1.5B	3	ADD	$R_{\|PC}$	1	-0.040	.993	$+0.006$	.210	no
Qwen3-4B	2	ABL	$R_{\|PC}$	1†	$+0.000$	$< .001$	$+0.053$	$< .001$	yes
Qwen3-4B	3	ABL	$R_{\|PC}$	0.5	$+0.002$	$< .001$	$+0.000$	$< .001$	yes
Qwen3-4B	3	ABL	$R_{\|PC}$	1	$+0.000$	$< .001$	$+0.001$	$< .001$	yes
Phi-4-mini-instruct	2	ADD	$R_{\|PC}$	0.5	-0.010	$< .001$	-0.038	$< .001$	yes
Phi-4-mini-instruct	2	ADD	$R_{\|PC}$	1	-0.071	.001	-0.157	.996	no
Phi-4-mini-instruct	2	ADD	$R_{\|PC}$	2†	-0.279	.941	-0.255	1.00	no
Phi-4-mini-instruct	3	ADD	$R_{\|PC}$	0.5	-0.028	$< .001$	-0.026	$< .001$	yes
Phi-4-mini-instruct	3	ADD	$R_{\|PC}$	1	-0.095	.016	-0.128	.903	no
Phi-4-mini-reasoning	3	ABL	$R_{\|PC}$	0.5	-0.014	$< .001$	$+0.005$	$< .001$	yes
Phi-4-mini-reasoning	3	ABL	$R_{\|PC}$	1	-0.026	$< .001$	$+0.010$	$< .001$	yes

Table 5: Manipulation check, all 25 strength cells. For $PC_{|R}$ the test is oriented superiority of both norms; for $R_{|PC}$ it is two one-sided equivalence tests of both norms with margin $\delta_M = 0.25 \times \text{median}_i|\text{DID}_M^{\text{base}}|$. Joint $p = \max(p_\rho, p_C)$. † cells enter through the post-hoc fluency reading. Margins: Qwen2.5-Math $\delta_\rho = 0.025$, $\delta_C = 0.009$; Qwen3-4B 0.218, 0.118; Qwen3-4B-Instruct 0.285, 0.257.

$[-0.243, +0.141]$ for C at 15 items, and $+0.080$ $[+0.023, +0.145]$ and $+0.094$ $[+0.053, +0.141]$ at 165. Cell-level $p < .05$ holds in 1 of 12 cells at 15 items and 2 of 12 at 165, and the mean ρ^2 is 0.098 and 0.013. Models paired across banks agree in sign in 1 of 6 cases for ρ and 2 of 6 for C.

The unconjoined arm gives a larger DoD than the conjunction arm in 12/12 cells of both banks, 1.156 against 0.313 at 15 items and 1.091 against 0.356 at 165.

J Collinearity and the two edited directions

Collinearity constrains what the edits can show. At 165 items the plane of ρ and C contained 76% of the norm of v_R, and v_R contained 72% of the norm of v_{PC}. Orthogonalization assigns this shared variance to one construct by construction. Removing too much eliminates relevance-specific variance that is aligned with the norms, and the injected vector then carries only part of the causal component. Removing too little retains shared variance in the edited direction, and the edit then shifts the other construct. The injection results fit the first error: eight of twelve cells shifted against the prediction under $R_{|PC}$ injection, and injection above $\alpha = 1$ leaked into the norms. The ablation results do not fit the second error. Ablating $R_{|PC}$ eliminated the fallacy while the norms remained within margin in 5/5 cells, at $\lambda = 1$ as well as 0.5. The relevance component that survives residualization is therefore also the component whose removal changes behavior without changing the norms. That strengthens the separation from a correlational reading to an ablation-based one, at $\alpha \leq 1$ and in the models where the ablation passed.

The evidential arm is weaker by construction, so its outcome is less informative. $PC_{|R}$ combined ρ and C into one shared scalar before the direction was fit. The two norms were not identified separately, and no control fixed their loadings on the edited direction. Under that edit ρ shifted against the prediction in both injection and ablation. A fixed linear loading does not predict this, and the pooled scalar leaves it undetermined. A component-wise arm that edits $\rho_{|R}$ and $C_{|R}$ separately would resolve this, and it remains to be run.

K Per-model summary

Table 6 lists, for each of the six models, how far it passed the four layers of Section 4: behavior, decoding, separation, and intervention. The column means equal the cross-model values reported in the main text.

model	$\mathrm{DoD}^{R\|N}$		AUC h_3		$R\|PC$	$PC\|R$	passing edits
	15	165	15	165	15 / 165	15 / 165	stage·direction
Qwen3-4B	0.287	0.354	0.713	0.832	2/2 · 2/2	2/2 · 2/2	ABL $R_{\|PC}$ · ABL $PC_{\|R}$ · ADD $PC_{\|R}$
Qwen3-4B-Instruct	0.255	0.354	0.676	0.865	2/2 · 2/2	0/2 · 2/2	ADD $PC_{\|R}$
Qwen2.5-Math-1.5B	0.210	0.246	0.649	0.665	2/2 · 2/2	0/2 · 2/2	ABL $PC_{\|R}$ · ADD $R_{\|PC}$
Gemma-3-4B-IT	0.315	0.443	0.829	0.862	2/2 · 2/2	2/2 · 2/2	none (specificity 16/20 fail)
Phi-4-mini-instruct	0.301	0.340	0.698	0.813	2/2 · 2/2	2/2 · 2/2	ADD $R_{\|PC}$
Phi-4-mini-reasoning	0.506	0.400	0.658	0.781	2/2 · 2/2	0/2 · 0/2	ABL $R_{\|PC}$
mean / total	0.312	0.356	0.704	0.803	12/12 · 12/12	6/12 · 10/12	10 combinations · 5 models

Table 6: Per-model summary across the two banks. $\mathrm{DoD}^{R|N}$ is the conjunction-arm context effect averaged over the two tasks, and its 15-item mean (0.312) differs from the main-text value (0.313) by rounding. AUC is held-out decoding of h_3. $R|PC$ and $PC|R$ count the wordings that remain decodable after residualization. Passing edits are the model, stage, and direction combinations of Table 4 that clear every gate; ADD is injection and ABL is ablation. Qwen3-4B-Instruct abbreviates Qwen3-4B-Instruct-2507.

L Models, assets, and compute

model	params	prec.	license	passing cells
Qwen3-4B [27]	4.0B	fp16	Apache-2.0	3
Qwen3-4B-Instruct-2507 [27]	4.0B	fp16	Apache-2.0	2
Qwen2.5-Math-1.5B [28]	1.5B	fp16	Apache-2.0	2
Gemma-3-4B-IT [29]	4.3B	fp32	Gemma	0
Phi-4-mini-instruct [30]	3.8B	fp16	MIT	2
Phi-4-mini-reasoning [30]	3.8B	fp16	MIT	1

Table 7: The six models. All six enter the behavioral, decoding and conditional intervention analyses. *passing cells* counts the model's task by stage by direction combinations that pass every gate of the conditional intervention (Table 4). Gemma is run fp32 because fp16 produced degenerate logits; the other five are fp16, so precision is not held constant across the sample.

Code and data. Both item banks, every prompt, the run and analysis scripts, the derived CSVs behind every number, and six executed notebooks that recompute each reported value from those CSVs are at `https://github.com/haba6030/cf-relevance-release` and in the supplementary material. Raw activations and per-item edit logs are omitted for size; the scripts that regenerate them are included.

Other assets. The fluency guard scores WikiText-103 [31]. Items `chung-1` to `chung-4` adapt the published materials of Chung et al. [4], and the screening norms for the three edited items come from Tentori et al. [3]. Weights are loaded through `transformers` 4.57.1 on `torch` 2.9.1 (CUDA 12.8). Licenses are the strings the model cards carry. Runs pin the revisions `1cfa9a7` (Qwen3-4B), `cdbee75` (Instruct-2507), `4a83ca6` (Qwen2.5-Math), `093f9f3` (Gemma-3), `cfbefac` (Phi-4-mini-instruct) and `0e3b1e2` (Phi-4-mini-reasoning). Items `llm-1` to `llm-8` and the 165-item bank were written and screened with Claude Opus 4.6 and 5.

Compute. Every run used one NVIDIA TITAN RTX (24 GB). The preregistered intervention grid took 56 GPU-hours summed over five models, at peak allocation 18.8 GB. The conditional intervention took 11 GPU-hours over six models and the manipulation check 6.5 GPU-hours, on four GPUs in parallel. Development and abandoned configurations, including the retired decoder labeling and the forced-choice readout, consumed several times the reported total.

M Broader impacts

The study measures a reasoning bias rather than improving a capability, and the released material is 180 short text items. Two consequences follow. A direction that predicts how relevant a statement is to a discourse could be used to steer a model's probability judgments without changing its stated evidence, which is a manipulation surface in any deployment where a model reports likelihoods. Our edits move the fallacy only conditionally and at low strength, so the finding bears on the diagnosis rather than on any deployed system. Separately, behavior on a human paradigm is evidence about the model rather than about human cognition. We report the pattern in these six models and confine the claim to them.

N Deviations from preregistration

Four analyses departed from the preregistration. Each change was made after the data were seen, and we report it in that order together with its reason.

1. **Decoder labeling.** The preregistration specified a one-versus-three rank AUC over the four prefix–conjunct cells. That construction pools the prefix contrast with the conjunct contrast in one ranking, so its label mixes relevance with the identity of the conjunct. We replaced it with the conjunct-fixed binary contrast (R, h_3 against I, h_3), which isolates the prefix manipulation and matches the behavioral minimal pair. The retired construction gave mean AUC 0.559 (6/12 cells); the replacement gives 0.704 (12/12).
2. **Separation contrast.** The residualization analysis was preregistered on the h_3 contrast $R - I$, and the reported separation counts use that contrast. The intervention directions use the difference-in-differences contrast $(R - I)_{h_3} - (R - I)_{h_2}$, because the edit targets the behavioral DoD and the difference removes item-level components shared by the two conjuncts. The vector-level removal fractions (76% and 67% of the norm of v_R) come from this post-hoc reading.
3. **Conditional intervention.** The design of Section 3, with the fallacy itself as outcome, per-item stage assignment, cross-frame evaluation, FWL-fitted directions and constraint-matched controls, was fixed after the preregistered grid returned its null and was run on the 165-item bank. The grid's null traces to disagreement between the fifteen item contrasts (Appendix G); the conditional design addresses it by scoring the fallacy directly, by assigning each item to the stage its baseline supports, and by using enough items to shrink the item by model variance term. Cross-frame evaluation keeps selection and outcome on different observations. We report it as the main intervention and the preregistered grid in Appendix G. The conditional intervention is therefore exploratory. Its per-cell p-values are reported without a multiple-comparison adjustment. A passing-cell mean is selected partly by a significance gate in the predicted direction; it describes selected cells, not an unbiased estimate of an effect across the full grid.
4. **Manipulation check under the conditional edit.** The equivalence margin ($0.25\times$ the median absolute context-induced DID, per model and norm), the two one-sided tests for $R_{|PC}$, the oriented superiority tests for $PC_{|R}$ and the intersection–union combination were fixed before the check ran and after the behavioral edit results were seen. The margin depends only on the baseline context effect, which the edit outcomes do not enter, and a margin of 0.5 was fixed as a sensitivity setting.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 4 claim three results in six models: a behavioral relevance effect in both item banks, a linear representation of relevance that survives residualizing the two evidential ratings out, and an edit of that representation that shifts the CF score in selected model–task–strength cells. The third claim is stated as conditional on model and direction rather than as a general causal result.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 names the limits in two groups. The causal reading is bounded because injection and ablation do not pair within a model for $R_{|PC}$, the evidential edit moves the two ratings in opposite directions, and the statement holds only at $\alpha \leq 1$, which collinearity partly explains (Appendix J). The sample and the design are bounded because three of the six models are Qwen descendants, the two banks provide supporting agreement across different stimuli, the behavioral nulls are absences of evidence rather than equivalence results, and the conditional design followed the preregistered grid's null (Appendix N). Appendix G gives the preregistered grid, its guards, and the geometry behind its null.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper contains no theorems or formal results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 gives the design, every estimator as an equation, and the statistical procedure. Appendix C gives every prompt verbatim, Appendix D all fifteen items of the main bank, Appendix B the 165-item bank, Appendices H and F the conditional and the preregistered intervention, and Appendix L the model versions, precisions, and library versions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository (Appendix L; also attached as supplementary material) contains both item banks, every prompt wording, the GPU run scripts and queue files for behavior, norming, activation extraction, the conditional intervention and the manipulation check, the analysis scripts, and six executed notebooks that recompute every number, table and figure of the paper from the released CSVs without a GPU, each checked against the printed value. Raw activations (13 GB) and per-item edit logs are omitted for size; the scripts that regenerate them are included with exact commands and pinned versions (`torch` 2.9.1, `transformers` 4.57.1). All fifteen items are printed in Appendix D and every prompt in Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 gives the three prefixes, the two task arms, the scoring rule, the two norming questions, the decoding and residualization recipes, and the two edited directions with their doses. Appendix H gives the conditional intervention's design, its five gates and its full grid, and Appendix F the preregistered grid. Layer selection sits inside the fold, so no reported number uses a layer chosen with the item it scores.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Within a cell the unit is the item, $n = 15$ in the main bank and $n = 165$ in the second, and decisions use a sign-flipping permutation test, two-sided at $\alpha = .05$ and exact over all 2^{15} assignments in the main bank. Holm correction is applied across the two task formats. Across cells the resampling unit is the model, because the two formats correlate at $\rho \approx .96$; figures show bootstrap 95% confidence intervals, and the text reports sign counts and the worst leave-one-item-out p. The manipulation check in Section 3.6 adds a fixed equivalence margin, with two one-sided tests combined by intersection-union.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix L reports one NVIDIA TITAN RTX (24 GB) per run at peak allocation 18.8 GB, 56 GPU-hours for the preregistered intervention grid over five models, 11 GPU-hours for the conditional intervention over six models and 6.5 GPU-hours for its manipulation check, run on four GPUs in parallel. It also states that development and abandoned configurations consumed several times the reported total.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses public open-weight models and text stimuli written by the authors or adapted from published materials, and involves no human participants.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix M names the manipulation surface that a relevance direction would open, and states that behavior on a human paradigm is evidence about the model rather than about human cognition.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: No model or dataset is released beyond the two stimulus banks, which are 180 short text items about fictional people and pose no misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix L names every model with its citation, parameter count, precision, and license, and credits WikiText-103, the published materials of Chung et al. [4], and the screening norms of Tentori et al. [3].

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Both stimulus banks are new. Appendix D prints all fifteen items of the main bank in full, Appendix E documents the slot specification, the generation and screening protocol and the provenance of each item's norms, and Appendix B describes the 165-item bank. Both banks are in the released repository.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human participants took part. The screening norms are published values from Chung et al. [4] and Tentori et al. [3] or values declared by the slot specification, as stated in Appendix E.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human participants took part, so no review board approval was required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Eight of the fifteen items of the main bank and all 165 items of the second bank were generated and screened with Claude Opus 4.6 and 5 under an LLM-as-a-judge protocol, which is a non-standard component of the stimulus construction. Appendices E and L declare this; the generation prompts and judge rubric are in `data/llm-build/`.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.