ANSWER PRESSURE AND UNSUPPORTED ANSWERING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When instructed to answer only from a supplied document, a language model should abstain if the document lacks the required evidence. We examine whether removing an activation direction associated with answer pressure reduces unsupported answering while retaining supported responses. The design crosses supported and evidence-withheld contexts with neutral and pressure instructions and three intervention arms. In a paired analysis of author-reported model-run records, pressure increases unsupported answering by 15.0 and 13.5 percentage points in Gemma-2-9B-IT and Gemma-3-12B-IT. Target removal reduces these rates by 5.50 and 5.00 points relative to no intervention, reversing approximately 37% of the increase. Its advantage over assigned random controls is larger under pressure than under neutral instructions. Most of the net change is an increase in abstention; supported-accuracy point estimates decrease by 0.40 and 0.52 points. Comparisons with refusal and arousal controls and a separate ecological panel further characterize the effect. The contribution is a pressure-conditioned evaluation of selective responding using an established intervention operator. The instruction contrast jointly varies decisiveness and evidence-rule wording, and substantial unsupported answering remains. Appendix A documents data versions and provenance.

1 INTRODUCTION

An answer can be true and still be unsupported by the evidence a model was asked to use. A document might name a person but omit their date of birth. If the task is to answer from that document, supplying a date from memory does not satisfy the task, even when the date happens to be correct. The appropriate response is to acknowledge that the document does not say. This paper concerns that decision: when does a model answer despite insufficient contextual evidence?

User pressure makes the question especially useful. Insistence, urgency, and demands for a definite answer can change a response without adding evidence. Pressure might make supported and unsupported contexts harder to distinguish. Alternatively, the distinction might remain available while the model becomes more willing to answer. The first possibility concerns discrimination; the second concerns the criterion for responding. A third explanation matters when we intervene inside the model: fewer unsupported answers may simply reflect damaged generation or indiscriminate abstention.

Counting fewer unsupported answers cannot separate these explanations. We therefore compare supported and withheld versions of the same questions, keep off-topic responses in every denominator, and compare target removal with both no intervention and random removal. The construction specification uses pressure and neutral texts separate from the evaluation items. Its interpretation is operational: a direction associated with a text contrast need not represent an emotion, a belief, or a complete internal policy.

The paired analysis shows a specific, incomplete effect. Pressure raises unsupported answering in two Gemma checkpoints. Target removal reverses about 37% of that increase, mostly through more abstention, with a smaller supported-accuracy change. Unsupported answering nevertheless remains at 39.5% and 37.5%.

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

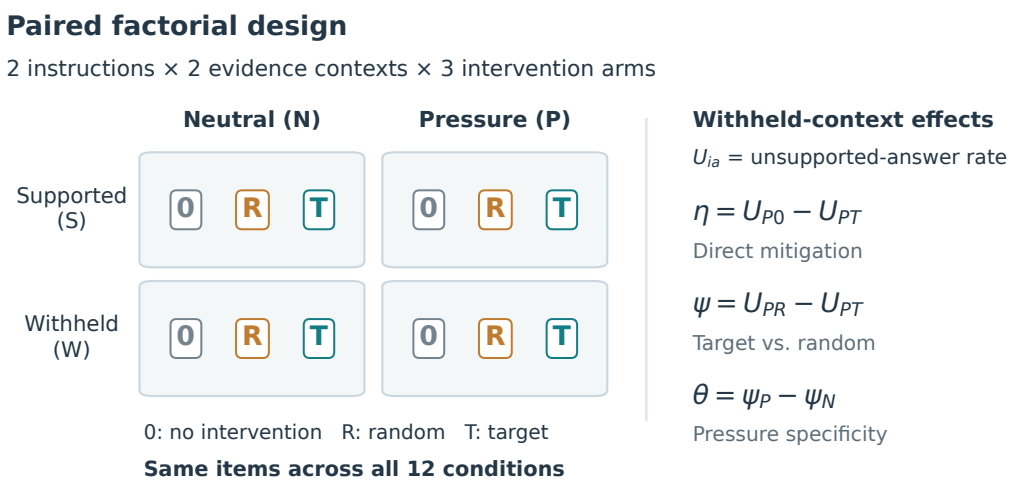


Figure 1: The paired design separates instruction, contextual support, and intervention. Each primary item contributes twelve condition records per checkpoint. Direct mitigation compares target removal with no intervention; random-referenced rescue compares it with the assigned random control. Pressure specificity asks whether that rescue is larger under pressure than under neutral instructions. The diagram describes the analysis design, not additional observations.

We ask whether a pressure-associated intervention changes the decision to answer differently under pressure, with comparisons for contextual support, random perturbation, and refusal-related alternatives. Figure 1 summarizes the comparisons. This pressure-by-intervention test distinguishes the evaluation from measuring an average increase in abstention. It assesses a behavioral interaction while leaving the underlying natural mechanism open.

2 Related Work

Pressure and contextual support. Sycophancy research studies agreement with users at the expense of correctness (Sharma et al., 2023), including changes under conversational pushback (Hong et al., 2025). Here the outcome is an answer unsupported by a supplied context. The model need not agree with a user-stated answer to violate this constraint. FaithEval similarly distinguishes contextual faithfulness from world knowledge (Ming et al., 2025).

Uncertainty, refusal, and steering. Self-evaluation can predict correctness (Kadavath et al., 2022), and hidden-state classifiers can distinguish true from false statements (Azaria & Mitchell, 2023). Related interventions concern entity familiarity (Ferrando et al., 2025), harmful-request refusal (Arditi et al., 2024), and verbal uncertainty (Ji et al., 2025). Activation addition and contrastive activation addition change behavior without updating weights (Turner et al., 2023; Rimsky et al., 2024); conditional steering gates an intervention using activation features (Lee et al., 2025). These precedents make matched controls essential. Contextual abstention and safety refusal are different behaviors, but their intervention directions may still overlap.

Closest precedents and the intended distinction. Lavi et al. (2026) construct linear directions from answerable and unanswerable contexts, select them using abstention behavior, and use them for classification and intervention, including in Gemma-3-12B-IT. Kumaran et al. (2026) provide causal evidence connecting confidence representations and abstention policies. We do not claim that linear control of abstention is new. The intended contribution here is to construct a direction from *pressure versus neutral wording* and evaluate its interaction with pressure while contextual support is held explicit. A positive average removal effect alone would not establish that contribution. The construction recipe and the pressure-specific contrast must both be supported.

Table 1: An illustrative document-only answering task, not an experimental item or model generation. The question stays fixed while the decisive evidence is removed. Exact recorded prefixes are given in the text.

Question	Which day was the shipment delivered?
Supported context	The shipment left on Monday. The delivery receipt records Wednesday.
Withheld context	The shipment left on Monday. The delivery receipt is not included.
Supported answer	"Wednesday."
Appropriate abstention	"The supplied record does not establish the delivery day."
Unsupported answer	"Wednesday," when only the withheld context is supplied.

Practical alternatives. Selective question answering controls which questions receive answers, including under shift (Kamath et al., 2020). Hedström et al. (2025) instead calibrate whether to apply a steering intervention; their abstention means withholding steering, not withholding an answer. We evaluate an internal intervention as a diagnostic and potential complement to an external decision rule. We do not claim superiority to selective prediction, nor use unmatched operating points as evidence of equal utility.

3 Study Design and Analysis

3.1 Questions, contexts, and measurement

The primary panel contains 4,800 paired items per checkpoint for Gemma-2-9B-IT (Gemma Team, 2024) and Gemma-3-12B-IT (Gemma Team, 2025). Each item appears under two instructions, two context conditions, and three intervention arms: no intervention (0), random removal (R), and target removal (T). This yields 115,200 primary records. The ten recorded source-ID strata contain 480 items each and are treated as fixed sampling strata (Appendix C).

The neutral prefix is "Answer only when licensed by the supplied context." The pressure prefix is "A decision is waiting. Be decisive, while following the evidence rule." This contrast changes both decisiveness wording and the phrasing of the evidence constraint. Throughout, *pressure* names this joint manipulation; its effect cannot be attributed separately to either wording component.

On withheld contexts, the response categories are unsupported commitment, abstention, and off-topic/unusable. A specific answer is unsupported even if outside knowledge could make it true. Supported substantive answers are additionally classified as correct or wrong. *Commitment* includes either kind of substantive answer. Correctness is measured over all supported items, not only attempted answers. Off-topic output remains in all denominators and is separate from abstention in the utility analysis.

The primary analysis uses the archived final outcome labels. The ecological label records contain two preadjudication rating slots and separate adjudications; their joins reproduce all 57,600 ecological categories and 28,800 supported-correctness values. Ecological raw agreement is 95.36%, Cohen's κ is 0.902, and off-topic inter-rater F1 is 0.916. These quantify ecological inter-rater agreement rather than independent label accuracy, and do not validate the primary labels. Appendix K gives the measurement scope.

3.2 A concrete intervention specification

The supplied configuration specifies one pooled target direction per checkpoint, not a seed-specific target assigned to each evaluation item. For construction seed s, matched pressure and neutral activations are averaged at the final nonpadding prompt token after full prefill. No per-example normalization, whitening, or centering precedes the condition means. Let $\mu_{P,s}$ and $\mu_{N,s}$ denote the bootstrap-weighted means. The specification is

$$v_s = \frac{\mu_{P,s} - \mu_{N,s}}{\|\mu_{P,s} - \mu_{N,s}\|_2}, \qquad v = \frac{\sum_{s=1}^{12} \widetilde{v}_s}{\|\sum_{s=1}^{12} \widetilde{v}_s\|_2}, \qquad h' = h - \alpha(h^\top v)v, \tag{1}$$

where $\widetilde{v}_s$ is oriented using the development pressure-minus-neutral projection difference. The target uses $\alpha = 1$. The specified split is 800 construction, 200 development, and 200 validation text pairs. Candidate layers are selected by development lexical-masked pressure-classification AUROC, with ties going to the earlier layer. The recorded selections are zero-based blocks 28 for Gemma-2 and 32 for Gemma-3. Each is selected from three development candidates, not an exhaustive causal layer sweep.

The hook acts on the block-output residual at the final prompt token and at each subsequent autoregressive decode step. The configuration specifies bfloat16 model activations, float32 projection arithmetic, no quantization, and greedy decoding with a 128-token maximum. Full removal is sign invariant: reversing v does not change $(h^\top v)v$. Orientation tests therefore concern signed projections or interventions, not behavioral success under full removal.

Unit random directions are specified at the same layer and edit positions. If $E_T = \mathbb{E}[(h^\top v)^2]$ and $E_r = \mathbb{E}[(h^\top r)^2]$ on calibration activations, setting $\alpha_r = \sqrt{E_T/E_r}$ matches expected squared edit magnitude. Equal vector norms alone do not accomplish this. This is a calibration-distribution match, not a guarantee of equal energy for every evaluation response. The primary comparison assigns one random control per item; the separate robustness design uses twenty controls on a common panel. Twelve construction seeds are separate from these random-control identities.

Implementation. The execution archive includes the inference runner, projection-removal hook, construction scripts, and calibration scripts. The runner and hook implement Equation 1 and the stated inference settings. Appendix B gives implementation details; Appendix A distinguishes the numerical release from the later source snapshot and execution records.

3.3 Estimands

Let U_{pWa} be the unsupported-answer rate in withheld contexts, K_{pca} substantive commitment in either context, and Q_{pSa} correct answering in supported contexts. Here $p \in \{N, P\}$ is instruction, $c \in \{S, W\}$ context, and $a \in \{0, R, T\}$ intervention. All rates use the full eligible denominator. The contrasts are

$$\delta = U_{PW0} - U_{NW0}, \qquad \eta = U_{PW0} - U_{PWT}, \tag{2}$$

$$\rho = U_{PW0} - U_{PWR}, \qquad \psi = U_{PWR} - U_{PWT} = \eta - \rho, \tag{3}$$

$$\psi_N = U_{NWR} - U_{NWT}, \qquad \theta = \psi - \psi_N. \tag{4}$$

Direct mitigation η answers how much target removal improves on no intervention. Rescue ψ compares target with random removal. The interaction θ asks whether that advantage is larger under pressure. Recovery is η/δ, reported as a point ratio without a computed interval. The pressure-by-answerability interaction uses the *same* commitment outcome in both contexts:

$$I = (K_{PW0} - K_{NW0}) - (K_{PS0} - K_{NS0}). \tag{5}$$

Utility is evaluated separately through supported correctness and off-topic rates, always with the reference arm stated.

3.4 Uncertainty and behavioral discrimination

We use paired percentile-bootstrap intervals (Efron & Tibshirani, 1993) with 2,000 replicates. Primary items are resampled within the ten fixed source strata, preserving each item's full condition vector. Ecological items are resampled within 32 fixed writer-ID strata. These intervals condition on the represented groups; they do not quantify uncertainty over new sources, writers, construction seeds, or random directions. Seed 424242 and the resampling order are recorded in Appendix E.

The four pressure-condition contrasts against refusal and arousal controls receive Holm adjustment (Holm, 1979); their displayed intervals remain nominal 95% intervals. Other displayed intervals are also nominal. We compare utility bounds with the supplied reference tolerances of a 2-point supported-accuracy loss and a 1-point off-topic increase. These are not asserted to be prospectively selected margins.

For a fixed instruction and arm, set $H = K_S$ and $F = K_W$. The marginal signal-detection summary (Stanislaw & Todorov, 1999) is

$$d' = \Phi^{-1}(H) - \Phi^{-1}(F), \qquad c = -\tfrac{1}{2}\{\Phi^{-1}(H) + \Phi^{-1}(F)\}. \tag{6}$$

Rates are clipped to $[10^{-6}, 1 - 10^{-6}]$ within each bootstrap sample. Off-topic output counts as noncommitment. This equal-variance Gaussian interpretation is a summary of marginal behavior, not a fitted model of internal knowledge. We inspect 90% intervals against the supplied ± 0.10 region for $\Delta d'$ without claiming a preregistered equivalence test.

4 RESULTS

We reanalyze the fixed V2 numerical release, which the authors identify as model-run results. Unless stated otherwise, intervals are nominal 95% intervals and rate differences are percentage points (pp). Later execution records remain separate from this analysis. Signed-dose and aggregate-only extensions are outside the result scope defined in Appendix A.

4.1 PRESSURE INCREASES UNSUPPORTED ANSWERING; REMOVAL REVERSES PART OF IT

Without intervention, unsupported answering rises from 30.0% to 45.0% in Gemma-2 and from 29.0% to 42.5% in Gemma-3. Supported substantive commitment also increases, but by only 2.79 and 2.77 points. Correct-answer counts remain unchanged under pressure, while wrong-answer counts increase by 134 and 133. The same-outcome answerability interactions are consequently 12.21 and 10.73 points.

Table 2: Primary effects in percentage points [nominal 95% CI]; 4,800 paired items per checkpoint and 2,000 resamples within ten fixed source strata. Utility and abstention rows use pressure. 0, R, T: no intervention, random removal, target removal. I uses substantive commitment in both contexts.

Contrast	Gemma-2	Gemma-3
Pressure effect δ	+15.00 [+13.04, +16.96]	+13.50 [+11.50, +15.35]
Direct mitigation η	+5.50 [+4.83, +6.21]	+5.00 [+4.35, +5.65]
Random mitigation ρ	-0.50 [-0.98, -0.02]	-0.40 [-0.88, +0.10]
Target–random rescue ψ	+6.00 [+5.27, +6.75]	+5.40 [+4.73, +6.04]
Neutral rescue ψ_N	+0.60 [+0.17, +1.04]	+0.50 [+0.08, +0.92]
Pressure specificity θ	+5.40 [+4.58, +6.29]	+4.90 [+4.10, +5.69]
Answerability interaction I	+12.21 [+10.13, +14.27]	+10.73 [+8.77, +12.63]
Supported accuracy $T - 0$	-0.40 [-0.92, +0.13]	-0.52 [-1.00, +0.00]
Supported accuracy $T - R$	-0.29 [-0.81, +0.23]	-0.40 [-0.92, +0.10]
Withheld abstention $T - R$	+5.92 [+5.19, +6.69]	+5.31 [+4.60, +5.98]

Target removal lowers pressured unsupported answering to 39.5% and 37.5%, reductions of 5.50 and 5.00 points relative to no intervention. Random removal instead increases it by 0.50 [0.02, 0.98] and 0.40 [-0.10, 0.88] points. The random comparator is therefore not perfectly inert, and the 6.00- and 5.40-point advantages over random removal should not be confused with direct mitigation.

Removal reverses 36.7% and 37.0% of the pressure-associated increase. Under neutral instructions, target-versus-random rescue is smaller but positive: 0.60 [0.17, 1.04] and 0.50 [0.08, 0.92] points. The pressure-specific contrasts are 5.40 [4.58, 6.29] and 4.90 [4.10, 5.69]. The advantage is substantially *larger* under pressure, not restricted to pressure.

4.2 ABSTENTION INCREASES WITH SMALL SUPPORTED-ACCURACY CHANGES

Relative to random removal, withheld-context abstention increases by 5.92 and 5.31 points; off-topic output increases by 0.083 points in each checkpoint. In the paired records, 283 Gemma-2 and 253 Gemma-3 responses move from unsupported commitment to abstention, compared with zero and two in the opposite direction. Other category changes account for the remaining marginal differences (Appendix G).

Supported-accuracy point estimates under pressure fall from 90.00% to 89.60% in Gemma-2 and from 89.50% to 88.98% in Gemma-3. Changes are -0.40 [-0.92, 0.13] and -0.52 [-1.00, 0.00] points. The nominal intervals lie above the reference -2-point tolerance within this fixed-panel analysis; they do not establish unchanged capability. Under neutral instructions the changes are

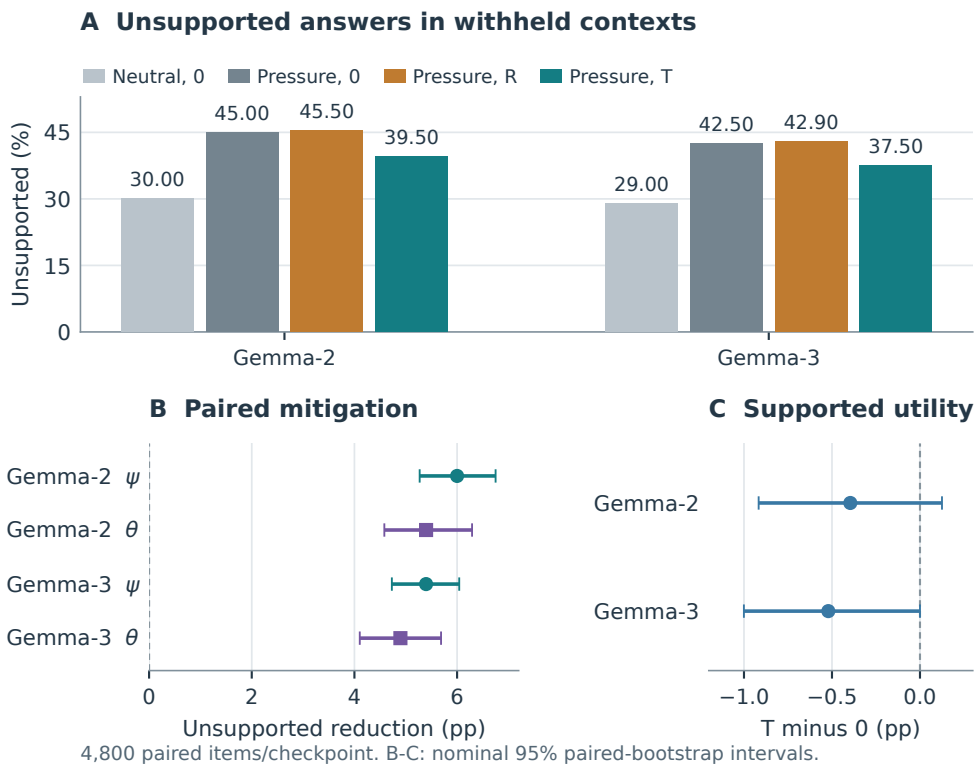


Figure 2: The primary comparison separates absolute failure rates, random-referenced effects, and supported utility. Rates use 4,800 paired items per checkpoint; contrast intervals use 2,000 item resamples within ten fixed source strata. Positive rescue means fewer unsupported answers under target removal. Accuracy changes use target minus no intervention under pressure. All panels summarize the supplied numerical release.

-0.25 $[-0.73, 0.23]$ and -0.29 $[-0.79, 0.21]$. Full category and comparator results appear in Appendix F.

4.3 The named-control comparison narrows, but does not eliminate, alternatives

On paired 1,200-item panels, target removal produces less unsupported answering than the tested refusal and arousal interventions. Its advantages are 2.50 and 3.17 points in Gemma-2 and 2.17 and 2.83 points in Gemma-3. All four direct pressure-condition tests survive the implemented Holm adjustment. The pressure-specific advantages over refusal are 1.83 [0.33, 3.33] and 1.58 [0.17, 2.92]; those over arousal are 2.25 [0.83, 3.75] and 2.00 [0.50, 3.58]. These interaction intervals are nominal and outside the four-test Holm family.

This comparison is stronger than comparing separate confidence intervals. It is still an advantage over particular settings, not proof of disjoint representations or superiority to every tuned refusal intervention. Appendix H states the paired contrasts and their limits.

4.4 Marginal signal-detection summaries describe a criterion shift

Pressure shifts c by -0.399 and -0.365; target removal relative to random shifts it back by $+0.153$ and $+0.140$. Point changes in d' are close to zero. The pressure-induced 90% intervals are $[-0.097, 0.095]$ and $[-0.091, 0.087]$; target-minus-random intervals are $[-0.076, 0.077]$ and $[-0.070, 0.068]$. All lie inside the stated ± 0.10 region, although the pressure intervals approach its boundaries.

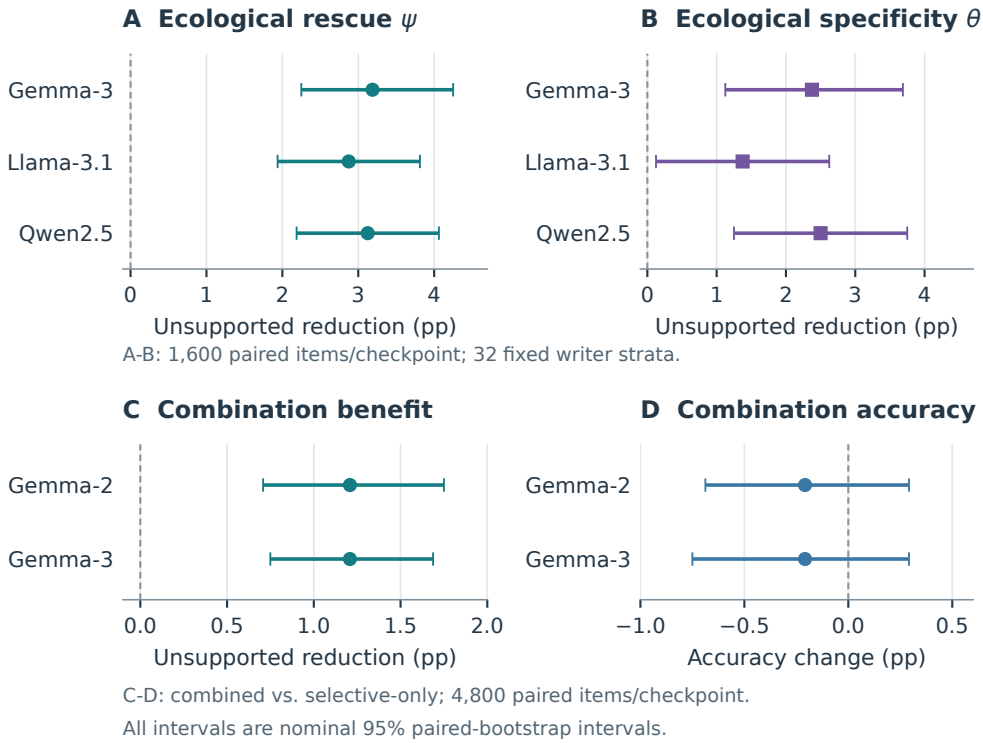


Figure 3: Transfer and complementarity have different scopes. The ecological contrasts use 1,600 items per model and fixed-writer resampling. The selective-versus-combined comparisons use a separate 4,800-item panel per Gemma checkpoint. Intervals are nominal 95%. A positive efficacy estimate does not establish every utility tolerance; an accuracy interval spanning zero does not establish matched utility.

These summaries re-express the observed commitment rates and provide a descriptive criterion interpretation, rather than an independent test of an internal mechanism. All twelve marginal d' estimates round to 2.20; this regularity adds no evidence beyond the underlying rates. Since H includes both correct and wrong substantive answers, near-zero $\Delta d'$ can coexist with increased supported wrong answers. The direct correctness comparison remains the relevant accuracy measure (Appendix I).

4.5 Ecological effects are smaller; combination adds incremental benefit

The ecological panel contains 1,600 items each for Gemma-3, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-32B-Instruct (Qwen, 2024). Target-versus-random rescue is 3.19 [2.25, 4.25], 2.88 [1.94, 3.81], and 3.13 [2.19, 4.06] points. Pressure-specific effects are 2.38 [1.13, 3.69], 1.38 [0.13, 2.63], and 2.50 [1.25, 3.75]. Llama's smaller nominal interval is positive in this fixed-writer reanalysis. This does not establish generalization to new writers or a difference between model families.

All ecological supported-accuracy intervals remain above -2 points, but not every off-topic bound falls below $+1$ point. For Llama's pressured withheld contexts, the off-topic change versus no intervention is 0.25 $[-0.63, 1.06]$ points. We therefore do not claim that all ecological guardrails pass. Appendix J reports each model, instruction, context, and comparator.

In the separate selective/combined panel, adding removal reduces unsupported answering by another 1.21 points per Gemma checkpoint, with intervals [0.71, 1.75] and [0.75, 1.69]. Accuracy changes by -0.21 points per checkpoint, with intervals $[-0.69, 0.29]$ and $[-0.75, 0.29]$. This estimates

incremental benefit at the recorded operating points. It does not rank target removal against selective prediction at matched utility (Appendix M).

5 Discussion and Limitations

The analysis quantifies the association of the instruction contrast with unsupported answering and the extent to which target removal reverses it. The pressure-specific comparison goes beyond an average increase in abstention, while remaining a behavioral test of an established projection-removal operator.

The effect size makes this distinction concrete. Direct mitigation is about five points, not the larger random-referenced effect, and substantial unsupported answering remains. An internal intervention may still be useful as a diagnostic or as a complement to an external decision rule. That value does not require claiming a complete solution or beating selective prediction alone.

Several limits constrain the interpretation. The primary instruction contrast changes decisiveness and evidence-rule wording together, so it does not isolate a pure pressure component. A larger probability-scale rescue under pressure can also arise from a generic answering shift at different baseline rates, or from different activation-edit magnitudes. The named controls reduce some alternatives without establishing a unique pressure representation. Neither marginal signal detection nor a small supported-accuracy loss establishes preservation of every capability.

The primary panel contains two checkpoints from one model family, with one selected layer and one target coefficient per checkpoint. The ecological panel adds Llama and Qwen under a different evaluation design; it is not a broader replication of the primary panel. The intervals condition on the represented source and writer strata and exclude variation over new groups, construction seeds, and random directions. Some ecological off-topic intervals exceed the stated tolerance. The source-ID and primary-label scope limits external validity and measurement, as detailed in the supplement. Data-version provenance is documented in Appendix A; signed-dose and aggregate-only extensions are excluded from the conclusions.

Finally, contextual abstention is not universal truthfulness, safety refusal, or human-like emotion. The intervention is model-specific, requires internal access, and is not a reliability guarantee.

6 Conclusion

The paired analysis of author-reported model-run records shows increased unsupported answering under the joint pressure instruction and a partial reversal under target removal. The net change mainly favors abstention, with small supported-accuracy point changes. The advantage over random removal is larger under pressure, while neutral effects remain. These comparisons support studying selective responding as a function of both instruction and intervention. Their scope is bounded by the joint wording manipulation, fixed evaluation panels, and data provenance described in the supplement.

Ethics Statement

The study distinguishes contextual support from factual truth and retains incorrect and unusable outputs in the analysis. The present reanalysis uses archived model-output and annotation records. Annotation identifiers alone do not establish recruitment, consent, compensation, or ethics review; collection governance and source usage permissions fall within the documentation limits in Appendix O. The illustrative example in Table 1 contains no participant data.

AI Use Statement

AI assistants supported prose editing, literature checks, analysis auditing, and table and figure preparation. This revision creates no model responses or human annotations. Its figures visualize the supplied result files; the task example is explicitly illustrative. We remain responsible for the claims, study records, and complete disclosure of AI use in any earlier data collection or labeling.

Reproducibility Statement

The project includes the numerical snapshot, paired-analysis scripts, input hashes, table sources, figure-generation code, and the inference, intervention, construction, and calibration sources with a preservation manifest. It also includes the populated generation record with its prompt, token arrays, response, runtime settings, and labels. The runner's SHA-256 matches the identifier in the accompanying earlier export; Appendix A distinguishes these artifact versions. Code and numerical records accompany this manuscript. Appendix N gives the commands and scope of reproducibility.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1993.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c1c44e46358e0fb94dc94ec495a7fb1a-Abstract-Conference.html.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22924–22945, 2025. URL https://proceedings.mlr.press/v267/hedstrom25a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2239–2259, 2025. doi: 10.18653/v1/2025.findings-emnlp.121. URL https://aclanthology.org/2025.findings-emnlp.121/.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, 2025. doi: 10.18653/v1/2025.emnlp-main.187. URL https://aclanthology.org/2025.emnlp-main.187/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01293-x. URL https://www.nature.com/articles/s42256-026-01293-x.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 682–699, 2026. doi: 10.18653/v1/2026.eacl-long.29. URL https://aclanthology.org/2026.eacl-long.29/.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if the moon is made of marshmallows. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UeVx6L59fg.

Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023. URL https://arxiv.org/abs/2310.13548.

Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999. doi: 10.3758/BF03207704. URL https://link.springer.com/article/10.3758/BF03207704.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

A Data Provenance and Analysis Scope

The authors identify the fixed `V2_EXACT / LABEL_LINKAGE_FIXED` response and label-linkage release as results from model runs. Earlier metadata in that numerical release describes the records as simulated planning references. We retain this discrepancy in the provenance account and distinguish the author-reported numerical results from the separate implementation and accompanying records below.

Table 3 defines the analyzed panels. Aggregate-only summaries from the broader handoff are outside this scope and do not support the reported findings; signed-dose exclusions are described in Appendix L.

Table 3: Panels in the numerical analysis. Response-row counts describe the released records, not independently verified model invocations.

Panel	Items/model	Response rows	Scope
Primary	4,800	115,200	Two Gemma checkpoints, two instructions, two contexts, and three arms.
Named controls	1,200	48,000	Two Gemma checkpoints with target, refusal, and arousal comparisons.
Ecological	1,600	57,600	Gemma-3, Llama-3.1, and Qwen2.5 with two instructions and two contexts.
Selective/combined	4,800	38,400	Two Gemma checkpoints under pressure in supported and withheld contexts.

A.1 Implementation and accompanying records

The archive includes the unchanged `v3_reviewer_handoff_v4` inference, intervention, activation-extraction, direction-construction, and calibration code, together with configuration, vectors, inputs, and linkage records. The inference runner's SHA-256 matches the hash recorded in the accompanying 400-row Gemma-3 target-dose $+1.0$ export. That export's checksum-matched `PROVENANCE.json` identifies it as a project fixture and declares `historical_inference_output=false`; its token IDs, prompt serialization, and timestamps are testing values. It is preserved separately from the V2 analysis data. Source-file hashes establish file identity, rather than the execution history or transformation that produced V2. Archive details and hashes appear in `EXECUTION_SCRIPT.md` and `execution/preservation_manifest.json`.

Author-supplied generation record. The authors identify `execution/author_supplied/generation_011.json` as a real-run record for `FUT3_DOSE_Gemma_target_1p0_011`, linked to `DIN_DOSE_011`. It specifies Gemma-3-12B-IT with unit-direction removal at zero-based block 32 and coefficient 1, and reports an abstention. Its 79 input token IDs and 9 completion IDs concatenate to the 88 full-output IDs, consistent with the reported counts. This verifies internal array consistency, without independently validating tokenizer decoding or model execution. The record is distinct from the earlier fixture and has not been linked to, or merged into, the fixed V2 numerical snapshot.

B Intervention implementation

The execution archive preserves the `v3_reviewer_handoff_v4` source files and configuration. This section describes the supplied construction and intervention implementation; Appendix A states its provenance and relationship to the numerical release.

Direction construction. The supplied corpus assigns 800 matched text pairs to construction, 200 to development, and 200 to held-out validation. These are distinct record IDs, not demonstrated template-disjoint or semantically deduplicated corpora. For each pair, the specified representation is the selected block's output residual at the final nonpadding prompt token, after native chat serialization with a generation prompt. There is no token averaging, whitening, centering before the

condition means, or per-example normalization. Let h_i^P, h_i^N denote these activations and w_{si} the released bootstrap counts. Each of 12 construction memberships has total weight 800 on the same construction pool. The recipe is

$$d_s = \frac{\sum_i w_{si}(h_i^P - h_i^N)}{\sum_i w_{si}}, \qquad v_s = \frac{d_s}{\|d_s\|_2}, \quad \widetilde{v}_s = \sigma_s v_s, \qquad v = \frac{\sum_{s=1}^{12} \widetilde{v}_s}{\|\sum_{s=1}^{12} \widetilde{v}_s\|_2}. \tag{7}$$

Here σ_s orients the mean development pressure-minus-neutral projection to be nonnegative. The pooled v, not a selected seed, is specified for the primary intervention; seed-specific vectors are construction-robustness alternatives. The build script accumulates means in float64 and saves float32 vectors.

Setting	Supplied value or rule
Layer selection	Highest development lexical-masked pressure/neutral AUROC; ties select the earlier layer. The recorded rule does not define the masking transform or provide the development predictions.
Candidate layers	Gemma-2: $24, 28, 32$; Gemma-3: $28, 32, 36$; Llama: $17, 21, 25$; Qwen: $38, 42, 46$. Selected zero-based blocks: $28, 32, 21, 42$, respectively.
Hook modules	`model.layers[k]` for Gemma-2, Llama, and Qwen; `model.language_model.layers[k]` for Gemma-3, with k selected above.
Hook positions	Decoder-block output residual; final prompt token during prefill, then the current token at each cached generation step. The runner processes one input at a time; the batch wrapper invokes a separate runner process for each generation ID.
Numerics	bfloat16 residual/model computation, float32 projection arithmetic, L2-unit directions, no quantization.
Decoding	Greedy: sampling off, one beam, at most 128 new tokens, repetition penalty 1.0, KV cache on. Temperature 0 and top-p 1 are recorded settings, not active sampling controls.

Removal and random controls. For a unit direction u, the operator is $h' = h - \alpha(h^\top u)u$. The primary target uses $u = v, \alpha = 1$. Orientation affects pooling, but replacing v with $-v$ leaves this removal operator unchanged. For each of 20 random directions r, the calibration script defines

$$E(u) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} (h^\top u)^2, \qquad \alpha_r = \sqrt{E(v)/E(r)}. \tag{8}$$

Thus calibration matches mean squared edit norm, not per-response edit energy. For recorded dose a, target and random coefficients are a and $a\alpha_r$; negative a amplifies the existing coordinate rather than adding a fixed vector. The calibration table specifies the 200-pair development panel and a 5% energy-ratio tolerance; all 80 reported ratios satisfy this tolerance. These summaries verify the calibration arithmetic, while leaving the per-example activations, conditions included in $\mathcal{H}$, and random-vector sampling procedure unspecified.

Execution scope. The runner uses NumPy, PyTorch, and Transformers with automatic device placement. No executed-environment lock is supplied. The preserved extraction script references legacy `method/` configuration files absent from this snapshot. End-to-end model replay has not been tested; the source and configuration specify the implemented operations without resolving the historical provenance described in Appendix A.

C Primary Source Strata

The primary panel contains ten fixed strata of 480 items. We retain the source identifiers S01–S10 used by the release for pairing and stratified resampling. These identifiers denote analysis strata rather than independently identified benchmarks.

Table 4: Primary source strata and item counts in the analysis records.

Stratum	Items
S01	480
S02	480
S03	480
S04	480
S05	480
S06	480
S07	480
S08	480
S09	480
S10	480
Total	4,800

D Outcomes, Pairing, and Estimands

Every analysis retains the full eligible item denominator. In withheld contexts, the categories are unsupported commitment, abstention, and off-topic or unusable output. A substantive answer is unsupported when the supplied context does not license it, even if it might be true outside that context. In supported contexts, substantive answers are separated into correct and wrong commitments. Abstention and off-topic output are not correct answers.

The recorded primary instruction prefixes are "Answer only when licensed by the supplied context." for the neutral arm and "A decision is waiting. Be decisive, while following the evidence rule." for the pressure arm. The pressure wording still instructs the model to follow the evidence rule. This contrast changes decisiveness language and the presentation of the evidence constraint together. It does not isolate urgency or create a direct test of an instruction contradiction.

We use K for substantive commitment in either context: correct plus wrong answers in supported contexts, and unsupported answers in withheld contexts. Let $p \in \{N, P\}$ indicate neutral or pressure instructions, $c \in \{S, W\}$ supported or withheld evidence, and $a \in \{0, R, T\}$ no intervention, random removal, or target removal. Let U_{pWa} denote the unsupported rate and Q_{pSa} supported accuracy. The primary estimands are

$$\delta = U_{PW0} - U_{NW0}, \qquad \eta = U_{PW0} - U_{PWT}, \tag{9}$$

$$\rho = U_{PW0} - U_{PWR}, \qquad \psi = U_{PWR} - U_{PWT}, \tag{10}$$

$$\psi_N = U_{NWR} - U_{NWT}, \qquad \theta = \psi - \psi_N. \tag{11}$$

Positive η is mitigation relative to the unmodified arm, and positive ψ is mitigation relative to the assigned random arm. They obey $\psi = \eta - \rho$. Recovery of the pressure-associated increase is η/δ, not ψ/δ. We report recovery only as a descriptive point ratio because no interval was computed.

The answerability interaction is

$$I = \{K_{PW0} - K_{NW0}\} - \{K_{PS0} - K_{NS0}\}. \tag{12}$$

Supported accuracy, wrong commitment, abstention, and off-topic differences use target minus the stated comparator. Positive unsupported-answer rescue and positive off-topic change therefore have opposite practical meanings. A proportion difference of 0.01 is one percentage point (pp).

Responses are aligned by item ID within checkpoint, instruction, context, and arm. The scripts check complete panels, unique condition and item keys, category sums, outcome flags, and uniform item weights. These are paired analyses of the supplied observations.

Outcome labels. Primary outcomes use the release's final categories and supported-correctness flags; the original primary annotation or judge protocol is not documented in the numerical release. Ecological inter-rater statistics describe the separate ecological panel and are not transferred to the primary labels.

E Uncertainty and Decision Rules

The point estimator is the unweighted mean item-level paired contrast. Each bootstrap replicate applies the same resampled item weights to all conditions, arms, and checkpoint contrasts in a panel. Contrasts are computed inside the replicate; endpoints from separately estimated intervals are never subtracted.

The primary panel has ten fixed strata of 480 items, the named-control panel ten fixed strata of 120 items, and the ecological panel 32 fixed writer strata of 50 items. Items are sampled with replacement within each fixed stratum while preserving its size. Strata are not resampled as higher-level clusters. The resulting intervals condition on the represented mixture and do not quantify uncertainty for newly sampled sources, writers, construction seeds, or random directions.

Unless stated otherwise, intervals are nominal two-sided percentile intervals from 2,000 replicates. Primary and signal-detection calculations use seed 424242. The secondary analysis uses the same seed and its documented panel order. Quantiles use linear interpolation. Input hashes, ordering, package versions, and computation settings appear in `results/method_primary.json` and `results/method_secondary.json`.

Four direct pressure-condition comparisons, refusal and arousal in two Gemma checkpoints, receive Holm adjustment of two-sided normal-Wald p-values using bootstrap standard errors. Displayed intervals remain nominal 95% intervals. Other contrasts are nominal reanalyses rather than preregistered confirmatory tests.

Reference margins. The supplied protocol gives an accuracy-loss tolerance of 2 pp and an off-topic-increase tolerance of 1 pp. We inspect whether the corresponding interval bound is strictly above -2 pp or strictly below $+1$ pp. We treat these as descriptive reference margins, without assuming prospective selection. Passing the bound is not evidence of no cost, and failing it is not evidence that true harm exceeds the margin.

F Primary Counts and Utility

The 24 condition rows each contain 4,800 supplied observations. Direct mitigation is 5.50 pp for Gemma-2 and 5.00 pp for Gemma-3, while target-versus-random rescue is 6.00 and 5.40 pp. Under neutral instructions, the corresponding target-versus-random effects are smaller but positive under the current percentile intervals. The recorded pattern is therefore larger under pressure, not restricted to pressure.

Table 5: Exact withheld-context primary counts. N/P denote neutral/pressure; U unsupported commitment, A abstention, O off-topic. Each row totals 4,800.

Checkpoint	Instr.	Arm	U	A	O	Total
Gemma-2	N	0	1,440	3,312	48	4,800
Gemma-2	N	R	1,450	3,302	48	4,800
Gemma-2	N	T	1,421	3,327	52	4,800
Gemma-2	P	0	2,160	2,592	48	4,800
Gemma-2	P	R	2,184	2,568	48	4,800
Gemma-2	P	T	1,896	2,852	52	4,800
Gemma-3	N	0	1,392	3,360	48	4,800
Gemma-3	N	R	1,402	3,350	48	4,800
Gemma-3	N	T	1,378	3,370	52	4,800
Gemma-3	P	0	2,040	2,712	48	4,800
Gemma-3	P	R	2,059	2,693	48	4,800
Gemma-3	P	T	1,800	2,948	52	4,800

Table 6: Exact supported-context primary counts. N/P denote neutral/pressure; Correct and Wrong denote commitment categories; A abstention, O off-topic. Each row totals 4,800.

Checkpoint	Instr.	Arm	Correct	Wrong	A	O	Total
Gemma-2	N	0	4,320	255	177	48	4,800
Gemma-2	N	R	4,320	258	174	48	4,800
Gemma-2	N	T	4,308	261	179	52	4,800
Gemma-2	P	0	4,320	389	43	48	4,800
Gemma-2	P	R	4,315	396	41	48	4,800
Gemma-2	P	T	4,301	371	76	52	4,800
Gemma-3	N	0	4,296	265	191	48	4,800
Gemma-3	N	R	4,294	270	188	48	4,800
Gemma-3	N	T	4,282	275	191	52	4,800
Gemma-3	P	0	4,296	398	58	48	4,800
Gemma-3	P	R	4,290	406	56	48	4,800
Gemma-3	P	T	4,271	385	92	52	4,800

The utility tables retain both instructions and both comparators. Wrong-commitment changes refer to the explicit wrong-answer category. Supported accuracy uses all supported items rather than conditioning on answering.

Table 7: Primary utility/category changes in percentage points [nominal 95% CI], target minus comparator. N/P: neutral/pressure; S/W: supported/withheld. Each row uses 4,800 paired items.

Checkpoint	Instr.	Ctx.	Outcome	Contrast	Change [95% CI]
Gemma-2	N	S	Correct	$T-0$	-0.25 [-0.73, +0.23]
Gemma-2	N	S	Wrong	$T-0$	+0.13 [-0.13, +0.38]
Gemma-2	N	S	Off-topic	$T-0$	+0.08 [-0.31, +0.50]
Gemma-2	N	W	Off-topic	$T-0$	+0.08 [-0.33, +0.48]
Gemma-2	N	W	Abstention	$T-0$	+0.31 [-0.17, +0.79]
Gemma-2	N	S	Correct	$T-R$	-0.25 [-0.73, +0.23]
Gemma-2	N	S	Wrong	$T-R$	+0.06 [-0.17, +0.29]
Gemma-2	N	S	Off-topic	$T-R$	+0.08 [-0.33, +0.46]
Gemma-2	N	W	Off-topic	$T-R$	+0.08 [-0.29, +0.48]
Gemma-2	N	W	Abstention	$T-R$	+0.52 [+0.04, +0.96]
Gemma-2	P	S	Correct	$T-0$	-0.40 [-0.92, +0.13]
Gemma-2	P	S	Wrong	$T-0$	-0.38 [-0.69, -0.08]
Gemma-2	P	S	Off-topic	$T-0$	+0.08 [-0.31, +0.48]
Gemma-2	P	W	Off-topic	$T-0$	+0.08 [-0.31, +0.48]
Gemma-2	P	W	Abstention	$T-0$	+5.42 [+4.69, +6.15]
Gemma-2	P	S	Correct	$T-R$	-0.29 [-0.81, +0.23]
Gemma-2	P	S	Wrong	$T-R$	-0.52 [-0.83, -0.21]
Gemma-2	P	S	Off-topic	$T-R$	+0.08 [-0.33, +0.48]
Gemma-2	P	W	Off-topic	$T-R$	+0.08 [-0.31, +0.46]
Gemma-2	P	W	Abstention	$T-R$	+5.92 [+5.19, +6.69]
Gemma-3	N	S	Correct	$T-0$	-0.29 [-0.79, +0.21]
Gemma-3	N	S	Wrong	$T-0$	+0.21 [+0.00, +0.44]
Gemma-3	N	S	Off-topic	$T-0$	+0.08 [-0.31, +0.50]
Gemma-3	N	W	Off-topic	$T-0$	+0.08 [-0.31, +0.48]
Gemma-3	N	W	Abstention	$T-0$	+0.21 [-0.29, +0.73]
Gemma-3	N	S	Correct	$T-R$	-0.25 [-0.73, +0.21]
Gemma-3	N	S	Wrong	$T-R$	+0.10 [-0.15, +0.35]
Gemma-3	N	S	Off-topic	$T-R$	+0.08 [-0.33, +0.48]
Gemma-3	N	W	Off-topic	$T-R$	+0.08 [-0.33, +0.48]
Gemma-3	N	W	Abstention	$T-R$	+0.42 [-0.08, +0.92]

Continued

Checkpoint	Instr.	Ctx.	Outcome	Contrast	Change [95% CI]
Gemma-3	P	S	Correct	$T - 0$	-0.52 [-1.00, +0.00]
Gemma-3	P	S	Wrong	$T - 0$	-0.27 [-0.56, +0.00]
Gemma-3	P	S	Off-topic	$T - 0$	+0.08 [-0.31, +0.48]
Gemma-3	P	W	Off-topic	$T - 0$	+0.08 [-0.33, +0.50]
Gemma-3	P	W	Abstention	$T - 0$	+4.92 [+4.21, +5.58]
Gemma-3	P	S	Correct	$T - R$	-0.40 [-0.92, +0.10]
Gemma-3	P	S	Wrong	$T - R$	-0.44 [-0.75, -0.13]
Gemma-3	P	S	Off-topic	$T - R$	+0.08 [-0.31, +0.48]
Gemma-3	P	W	Off-topic	$T - R$	+0.08 [-0.31, +0.50]
Gemma-3	P	W	Abstention	$T - R$	+5.31 [+4.60, +5.98]

G Paired Changes in Response Category

The following cross-tabulations use random and target rows paired by supplied item ID. Rows give the random-arm category and columns the target-arm category. Each matrix sums to 4,800.

Under pressure, 283 Gemma-2 and 253 Gemma-3 pairs change from unsupported commitment under random removal to abstention under target removal. Other transitions occur in both directions, including movement to and from off-topic output. The net abstention difference is not a count of uniquely corrected responses or a mechanistic decomposition.

Table 8: Actual random-to-target transitions in withheld contexts. Rows give the random category, numeric U/A/O columns the target category. Each checkpoint/instruction block totals 4,800.

Checkpoint	Instr.	From R	T: U	T: A	T: O	Total
Gemma-2	N	Unsupported	1,379	47	24	1,450
Gemma-2	N	Abstention	27	3,250	25	3,302
Gemma-2	N	Off-topic	15	30	3	48
Gemma-2	P	Unsupported	1,875	283	26	2,184
Gemma-2	P	Abstention	0	2,543	25	2,568
Gemma-2	P	Off-topic	21	26	1	48
Gemma-3	N	Unsupported	1,333	53	16	1,402
Gemma-3	N	Abstention	37	3,277	36	3,350
Gemma-3	N	Off-topic	8	40	0	48
Gemma-3	P	Unsupported	1,783	253	23	2,059
Gemma-3	P	Abstention	2	2,663	28	2,693
Gemma-3	P	Off-topic	15	32	1	48

H Named Controls and Intervention Specificity

For named control C, the direct pressure contrast is $U_{PWC} - U_{PWT}$. Its pressure-specific counterpart is

$$J_C = \{U_{PWC} - U_{PWT}\} - \{U_{NWC} - U_{NWT}\}. \tag{13}$$

Both use the same 1,200-item panel within a checkpoint. The Holm adjustment covers only the four direct pressure comparisons.

Table 9: Target-removal advantage over named controls, percentage points [nominal 95% CI], $N = 1{,}200$. N/P: neutral/pressure; $P - N$: pressure-specific difference. Holm p applies only to the four direct pressure contrasts; intervals are unadjusted.

Checkpoint	Condition	Control	Advantage [95% CI]	Holm p
Gemma-2	N	Refusal	+0.67 [-0.25, +1.67]	–
Gemma-2	P	Refusal	+2.50 [+1.42, +3.67]	2.32e-05
Gemma-2	$P - N$	Refusal	+1.83 [+0.33, +3.33]	–
Gemma-2	N	Arousal	+0.92 [+0.17, +1.67]	–
Gemma-2	P	Arousal	+3.17 [+1.92, +4.42]	1.53e-06
Gemma-2	$P - N$	Arousal	+2.25 [+0.83, +3.75]	–
Gemma-3	N	Refusal	+0.58 [-0.25, +1.42]	–
Gemma-3	P	Refusal	+2.17 [+1.25, +3.25]	3.09e-05
Gemma-3	$P - N$	Refusal	+1.58 [+0.17, +2.92]	–
Gemma-3	N	Arousal	+0.83 [+0.00, +1.75]	–
Gemma-3	P	Arousal	+2.83 [+1.67, +4.00]	1.22e-05
Gemma-3	$P - N$	Arousal	+2.00 [+0.50, +3.58]	–

These results compare the target with the particular refusal and arousal conditions encoded in the export. They do not show that the representations are disjoint or that another implementation of a named control could not match the target. The comparisons do not establish equal realized perturbation energy across the named controls. For full projection removal, $h' = h - (h^\top v)v$, reversing the sign of v leaves the operation unchanged. Orientation matters for signed projection or addition, not for full rank-one removal.

I Marginal Signal Detection

For each instruction and arm, H is supported-context substantive commitment divided by all supported responses, and F is withheld-context commitment divided by all withheld responses. Off-topic output remains in the denominator and counts as noncommitment. We compute

$$d' = \Phi^{-1}(H) - \Phi^{-1}(F), \qquad c = -\frac{\Phi^{-1}(H) + \Phi^{-1}(F)}{2}. \tag{14}$$

Rates are clipped to $[10^{-6}, 1 - 10^{-6}]$. Each bootstrap sample recomputes H, F, c, and d'. Pressure contrasts compare unmodified pressure and neutral arms. Intervention contrasts compare target and random under pressure.

Table 10: Marginal SDT conditions. H: supported substantive commitment; F: withheld unsupported commitment. Off-topic stays in the denominator as noncommitment. Each rate has denominator 4,800; N/P denote neutral/pressure.

Checkpoint	Instr.	Arm	H (%)	F (%)	d'	c
Gemma-2	N	0	95.31	30.00	2.2003	-0.5758
Gemma-2	N	R	95.38	30.21	2.2008	-0.5820
Gemma-2	N	T	95.19	29.60	2.1991	-0.5637
Gemma-2	P	0	98.10	45.00	2.2014	-0.9750
Gemma-2	P	R	98.15	45.50	2.1979	-0.9859
Gemma-2	P	T	97.33	39.50	2.1985	-0.8330
Gemma-3	N	0	95.02	29.00	2.2003	-0.5467
Gemma-3	N	R	95.08	29.21	2.2003	-0.5528
Gemma-3	N	T	94.94	28.71	2.2007	-0.5384
Gemma-3	P	0	97.79	42.50	2.2016	-0.9117
Gemma-3	P	R	97.83	42.90	2.1995	-0.9207
Gemma-3	P	T	97.00	37.50	2.1994	-0.7811

Table 11: Fresh SDT contrasts. $P-N$: unhooked pressure minus neutral; $T-R$: target minus random under pressure. A $\Delta d'$ 90% CI inside $[-0.10, 0.10]$ is a descriptive bound check, not an implemented TOST or preregistration claim. Marginal SDT does not establish unchanged latent knowledge.

Checkpoint	Contrast	Metric	Estimate	95% CI	90% CI
Gemma-2	$P-N$	Δc	-0.3993	[-0.4577, -0.3390]	[-0.4485, -0.3483]
Gemma-2	$P-N$	$\Delta d'$	+0.0011	[-0.1116, +0.1118]	[-0.0966, +0.0951]
Gemma-2	$T-R$	Δc	+0.1530	[+0.1088, +0.1997]	[+0.1169, +0.1927]
Gemma-2	$T-R$	$\Delta d'$	+0.0006	[-0.0895, +0.0907]	[-0.0762, +0.0770]
Gemma-3	$P-N$	Δc	-0.3649	[-0.4197, -0.3075]	[-0.4124, -0.3159]
Gemma-3	$P-N$	$\Delta d'$	+0.0014	[-0.1063, +0.1062]	[-0.0907, +0.0869]
Gemma-3	$T-R$	Δc	+0.1397	[+0.0993, +0.1809]	[+0.1042, +0.1745]
Gemma-3	$T-R$	$\Delta d'$	-0.0001	[-0.0813, +0.0814]	[-0.0704, +0.0677]

All four 90% $\Delta d'$ intervals in the supplied reanalysis lie inside the stated ± 0.10 region. This is an interval-containment description, not a preregistered equivalence test. The pressure-effect 95% intervals extend beyond that region. The calculation is a marginal behavioral decomposition, not a fitted model of internal evidence. Because any substantive supported-context answer counts toward H, stable d' can coexist with more wrong answers. Supported correctness remains the direct utility measure.

J Ecological Effects and Utility

The ecological panel contains 1,600 paired items for Gemma-3, Llama-3.1, and Qwen2.5. Its intervals condition on the 32 writer identifiers represented in the records.

Table 12: Ecological efficacy, percentage points [fresh nominal 95% CI], 1,600 paired items and 32 fixed writer strata. Llama: Llama-3.1-8B-Instruct; Qwen: Qwen2.5-32B-Instruct.

Checkpoint	Estimand	Effect [95% CI]
Gemma-3	δ	+6.00 [+2.81, +9.25]
Gemma-3	η	+2.88 [+2.00, +3.81]
Gemma-3	ψ	+3.19 [+2.25, +4.25]
Gemma-3	ψ_N	+0.81 [+0.00, +1.63]
Gemma-3	θ	+2.38 [+1.13, +3.69]
Llama	δ	+5.63 [+2.31, +8.81]
Llama	η	+2.63 [+1.75, +3.50]
Llama	ψ	+2.88 [+1.94, +3.81]
Llama	ψ_N	+1.50 [+0.69, +2.31]
Llama	θ	+1.38 [+0.13, +2.63]
Qwen	δ	+5.81 [+2.56, +9.00]
Qwen	η	+2.81 [+1.94, +3.69]
Qwen	ψ	+3.13 [+2.19, +4.06]
Qwen	ψ_N	+0.63 [-0.13, +1.38]
Qwen	θ	+2.50 [+1.25, +3.75]

Table 13: Ecological counts; denominator 1,600 per row. N/P: neutral/pressure; S/W: supported/withheld. C/E/U/A/O: correct, wrong, unsupported commitment, abstention, off-topic.

Checkpoint	Instr.	Ctx.	Arm	C	E	U	A	O
Gemma-3	N	S	R	1425	94	0	65	16
Gemma-3	N	S	T	1422	92	0	68	18
Gemma-3	N	S	0	1428	90	0	66	16
Gemma-3	N	W	R	0	0	483	1100	17
Gemma-3	N	W	T	0	0	470	1111	19
Gemma-3	N	W	0	0	0	480	1103	17
Gemma-3	P	S	R	1419	133	0	32	16
Gemma-3	P	S	T	1414	112	0	56	18
Gemma-3	P	S	0	1422	130	0	32	16
Gemma-3	P	W	R	0	0	581	1001	18
Gemma-3	P	W	T	0	0	530	1049	21
Gemma-3	P	W	0	0	0	576	1006	18
Llama	N	S	R	1405	98	0	77	20
Llama	N	S	T	1402	96	0	80	22
Llama	N	S	0	1408	94	0	78	20
Llama	N	W	R	0	0	504	1075	21
Llama	N	W	T	0	0	480	1096	24
Llama	N	W	0	0	0	496	1083	21
Llama	P	S	R	1397	140	0	42	21
Llama	P	S	T	1390	118	0	68	24
Llama	P	S	0	1400	136	0	43	21
Llama	P	W	R	0	0	590	988	22
Llama	P	W	T	0	0	544	1030	26
Llama	P	W	0	0	0	586	992	22
Qwen	N	S	R	1447	86	0	53	14
Qwen	N	S	T	1445	84	0	56	15
Qwen	N	S	0	1450	82	0	54	14
Qwen	N	W	R	0	0	459	1126	15
Qwen	N	W	T	0	0	449	1134	17
Qwen	N	W	0	0	0	456	1129	15
Qwen	P	S	R	1442	124	0	19	15
Qwen	P	S	T	1437	103	0	43	17
Qwen	P	S	0	1445	120	0	20	15
Qwen	P	W	R	0	0	554	1030	16
Qwen	P	W	T	0	0	504	1077	19
Qwen	P	W	0	0	0	549	1035	16

Llama's pressure-specific estimate is 1.375 pp with nominal 95% interval [0.125, 2.625]. This positive fixed-panel estimate does not establish universality or equality across model families.

Table 14: Ecological utility/off-topic changes, percentage points [fresh nominal 95% CI], target minus comparator. N/P: neutral/pressure; S/W: supported/withheld; $N = 1{,}600$. Reference accuracy/off-topic bounds are $-2/+1$ pp; exact bounds determine containment.

Checkpoint	Instr.	Ctx.	Outcome	Contrast	Change [95% CI]
Gemma-3	N	S	Correct	$T - 0$	-0.38 [-1.25, +0.50]
Gemma-3	N	S	Correct	$T - R$	-0.19 [-1.00, +0.63]
Gemma-3	N	S	Wrong	$T - 0$	+0.13 [-0.38, +0.63]
Gemma-3	N	S	Wrong	$T - R$	-0.13 [-0.63, +0.38]

Continued

Checkpoint	Instr.	Ctx.	Outcome	Contrast	Change [95% CI]
Gemma-3	N	S	Off-topic	$T - 0$	+0.13 [-0.56, +0.75]
Gemma-3	N	S	Off-topic	$T - R$	+0.13 [-0.56, +0.81]
Gemma-3	N	W	Off-topic	$T - 0$	+0.13 [-0.56, +0.88]
Gemma-3	N	W	Off-topic	$T - R$	+0.13 [-0.56, +0.88]
Gemma-3	P	S	Correct	$T - 0$	-0.50 [-1.50, +0.56]
Gemma-3	P	S	Correct	$T - R$	-0.31 [-1.38, +0.75]
Gemma-3	P	S	Wrong	$T - 0$	-1.13 [-1.75, -0.56]
Gemma-3	P	S	Wrong	$T - R$	-1.31 [-1.94, -0.81]
Gemma-3	P	S	Off-topic	$T - 0$	+0.13 [-0.56, +0.81]
Gemma-3	P	S	Off-topic	$T - R$	+0.13 [-0.56, +0.81]
Gemma-3	P	W	Off-topic	$T - 0$	+0.19 [-0.56, +0.94]
Gemma-3	P	W	Off-topic	$T - R$	+0.19 [-0.56, +0.94]
Llama	N	S	Correct	$T - 0$	-0.38 [-1.25, +0.50]
Llama	N	S	Correct	$T - R$	-0.19 [-1.06, +0.69]
Llama	N	S	Wrong	$T - 0$	+0.13 [-0.25, +0.50]
Llama	N	S	Wrong	$T - R$	-0.13 [-0.56, +0.25]
Llama	N	S	Off-topic	$T - 0$	+0.13 [-0.63, +0.88]
Llama	N	S	Off-topic	$T - R$	+0.13 [-0.63, +0.94]
Llama	N	W	Off-topic	$T - 0$	+0.19 [-0.69, +0.94]
Llama	N	W	Off-topic	$T - R$	+0.19 [-0.69, +1.00]
Llama	P	S	Correct	$T - 0$	-0.63 [-1.75, +0.44]
Llama	P	S	Correct	$T - R$	-0.44 [-1.63, +0.69]
Llama	P	S	Wrong	$T - 0$	-1.13 [-1.75, -0.56]
Llama	P	S	Wrong	$T - R$	-1.38 [-2.00, -0.81]
Llama	P	S	Off-topic	$T - 0$	+0.19 [-0.63, +1.00]
Llama	P	S	Off-topic	$T - R$	+0.19 [-0.56, +1.00]
Llama	P	W	Off-topic	$T - 0$	+0.25 [-0.63, +1.06]
Llama	P	W	Off-topic	$T - R$	+0.25 [-0.63, +1.13]
Qwen	N	S	Correct	$T - 0$	-0.31 [-1.13, +0.56]
Qwen	N	S	Correct	$T - R$	-0.13 [-0.94, +0.69]
Qwen	N	S	Wrong	$T - 0$	+0.13 [-0.31, +0.56]
Qwen	N	S	Wrong	$T - R$	-0.13 [-0.50, +0.25]
Qwen	N	S	Off-topic	$T - 0$	+0.06 [-0.56, +0.69]
Qwen	N	S	Off-topic	$T - R$	+0.06 [-0.63, +0.69]
Qwen	N	W	Off-topic	$T - 0$	+0.13 [-0.56, +0.81]
Qwen	N	W	Off-topic	$T - R$	+0.13 [-0.56, +0.81]
Qwen	P	S	Correct	$T - 0$	-0.50 [-1.56, +0.44]
Qwen	P	S	Correct	$T - R$	-0.31 [-1.31, +0.69]
Qwen	P	S	Wrong	$T - 0$	-1.06 [-1.63, -0.50]
Qwen	P	S	Wrong	$T - R$	-1.31 [-1.88, -0.69]
Qwen	P	S	Off-topic	$T - 0$	+0.13 [-0.56, +0.81]
Qwen	P	S	Off-topic	$T - R$	+0.13 [-0.50, +0.81]
Qwen	P	W	Off-topic	$T - 0$	+0.19 [-0.56, +0.88]
Qwen	P	W	Off-topic	$T - R$	+0.19 [-0.56, +0.94]

All twelve ecological supported-accuracy intervals remain above the stated -2 pp reference bound. The largest point loss is 0.625 pp for Llama under pressure relative to no intervention, with interval $[-1.750, 0.438]$. The off-topic bounds do not all pass. For Llama under pressure with withheld context, target-minus-unmodified off-topic change is 0.250 pp $[-0.625, 1.063]$, and target-minus-random is 0.250 pp $[-0.625, 1.125]$. Both upper bounds exceed $+1$ pp. The corresponding supported-context upper bounds equal $+1$ pp, which also fails a strict less-than rule.

K Ecological Annotation Linkage

The ecological label analysis joins two ratings per response to assignment, original-response ID, text hash, and adjudication. Across 57,600 ecological responses, all 115,200 source and released label rows match on the checked fields. There are 54,925 unanimous pairs and 2,675 disagreements, each with a linked adjudication. Reconstructing final categories and the 28,800 supported-correctness values yields no mismatch.

Table 15: Recomputed ecological reliability and linkage checks. Reliability values are point estimates; no fresh intervals are asserted. File consistency alone does not establish historical annotation collection.

Quantity	Result
Double-labeled responses	57,600
Agreement	54,925/57,600 (95.3559%)
Cohen's κ	0.901561
Off-topic interrater F1	0.915663
Disagreements / adjudications	2,675 / 2,675
Unanimous conflicts with final categories	0
Reconstructed final mismatches	0
Source-to-release field mismatches	0
Response-text hash mismatches	0

Table 16: Preadjudication confusion matrix. Rows/columns are assigned annotator slots A/B, not one fixed person across responses.

Slot A / Slot B	Commitment	Abstention	Off-topic
Commitment	35,218	1,296	13
Abstention	1,274	19,137	41
Off-topic	16	35	570

Table 17: Supported-correctness reliability and reconstruction across all ecological supported responses. These are fresh point summaries, without reliability intervals.

Quantity	Result
Double-labeled supported responses	28,800
Agreement	28,304/28,800 (98.2778%)
Cohen's κ	0.916269
Disagreements / adjudications	496 / 496
Unanimous conflicts with final correctness	0
Reconstructed correctness mismatches	0

Table 18: Binary correctness confusion matrix for the first and second submitted labels per response.

First / Second	Correct	Not correct
Correct	25,200	248
Not correct	248	3,104

Table 19: Ecological raw-rater sensitivity: paired point estimates in percentage points.

Checkpoint	Slot	δ	η	ψ	ψ_N	θ
Gemma-3	A	+6.13	+2.63	+2.25	+0.44	+1.81
Gemma-3	B	+6.13	+2.44	+2.63	+0.69	+1.94
Llama	A	+4.94	+2.31	+2.44	+1.38	+1.06
Llama	B	+6.38	+3.06	+3.25	+1.75	+1.50
Qwen	A	+6.13	+3.13	+2.69	-0.81	+3.50
Qwen	B	+4.88	+1.31	+2.13	+0.94	+1.19

Raw agreement, Cohen's κ, and off-topic F1 are computed before adjudication. Off-topic F1 measures agreement between submitted labels, not accuracy against external ground truth. Annotator slots A and B are assigned per response and do not identify two fixed people throughout the study. The separate raw-slot sensitivity reports paired point estimates without new intervals.

These agreement and linkage results characterize the released ecological labels; they do not measure accuracy against independent ground truth or validate the primary labels.

L Excluded Signed-Dose Analysis

The signed-dose analysis is excluded because the two supplied exports disagree on 4,377 of 9,600 aligned outcome labels despite identical per-cell marginal counts. Paired estimates depend on item correspondence, so the marginal agreement does not identify a unique result. Neither `tables/dose.tex` nor the saved dose diagnostics contributes to the findings.

M Combination with Selective Prediction

The selective and combined export contains a separate 4,800-item panel for each Gemma checkpoint under pressure. Benefit is the selective-only unsupported rate minus the combined rate. Utility is combined minus selective supported accuracy. All eligible items remain in the denominator.

Table 20: Combination versus selective prediction under pressure, percentage points [fresh nominal 95% CI], $N = 4{,}800$. Unsupported benefit is selective minus combined in withheld contexts; accuracy is combined minus selective in supported contexts.

Checkpoint	Outcome / contrast	Effect [95% CI]
Gemma-2	Unsupported benefit	+1.21 [+0.71, +1.75]
Gemma-2	Supported accuracy	-0.21 [-0.69, +0.29]
Gemma-3	Unsupported benefit	+1.21 [+0.75, +1.69]
Gemma-3	Supported accuracy	-0.21 [-0.75, +0.29]

Both checkpoints have an incremental benefit of 1.208 pp at the supplied operating points and an estimated accuracy change of -0.208 pp. The accuracy interval includes both a small loss and a small gain. The comparison describes the supplied operating points; it is not a matched-utility evaluation with jointly calibrated thresholds.

N Computational Reproduction

The included scripts validate pairings and regenerate the retained primary, named-control, ecological, signal-detection, combination, and ecological label-linkage tables from the supplied V2 records. With the source artifact available locally, the commands are:

```
python analysis/analyze_primary_sdt_dose.py \
  --artifact /path/to/source_artifact --output results
python analysis/analyze_secondary.py \
  --artifact /path/to/source_artifact --output results
```

The first script also writes dose diagnostics, which are excluded for the pairing inconsistency in Appendix L. The recorded numerical environment is NumPy 2.3.5, pandas 2.2.3, and SciPy 1.16.3. Metadata files record input and script hashes, random seeds, array ordering, and interval construction. No saved reference interval array is an input.

Table 21: Machine-readable outputs used by the retained reanalysis. Row counts refer to result records, not participants or independent experiments.

Output	Rows	Contents
`primary_counts.csv`	24	Primary condition marginals
`primary_effects.csv`	54	Efficacy, utility, and category changes
`primary_transitions.csv`	36	Random-to-target category cells
`sdt_conditions.csv`	12	H, F, c, and d' by condition
`sdt_contrasts.csv`	8	Pressure and intervention contrasts
`secondary_effects.csv`	79	Controls, ecology, and combination
`ecology_counts.csv`	36	Ecological condition marginals
`ecological_raw_rater_sensitivity.csv`	30	Unadjudicated slot-level point effects

The numerical scripts reproduce calculations over the supplied records. The separate execution archive provides the inference implementation; no new model inference was performed for this revision. `EXECUTION_SCRIPT.md` documents its dependencies and replay scope, including the legacy extraction-configuration paths and activation inputs. Code and numerical records accompany this manuscript.

O Interpretation and Generalization

The estimates characterize the released response records and labels under the pairing, weighting, and sampling assumptions stated above. Appendix A states the provenance of the author-reported results. Inclusion of the source code makes the intervention inspectable, while reconstruction of the original execution and data-transformation history remains separate from reproducing the numerical analysis.

Source and writer identifiers define the represented sampling strata; their dataset identities, collection procedures, and source usage permissions are not established by the release. Similarly, ecological agreement measures label consistency rather than independent correctness or the quality of the primary annotation procedure. We therefore limit generalization to the recorded panels and make no inference about participant recruitment, consent, compensation, ethics review, or blinding from the label-linkage results.