Hidden Commitment: When Language Models Silently Pick a Side and How Steering Can Surface It

Samuel Dawit Assefa [1] **Jae Won Cho** [2]

Abstract

Language-model answers are often treated as a visible record of the internal decision that produced them. We test this assumption in conflicting-context question answering, where two supplied documents support incompatible answers. A model can process the disagreement and still return a one-sided answer, even under prompts that ask it to disclose conflict. In Qwen 0.8B, hidden source commitment is captured by a residual source-choice direction whose sign tracks which document the answer uses. Sparse middle-layer MLP writers feed the same coordinate, and steering the coordinate changes the chosen source. A transfer run on Gemma 3 1B preserves the source-choice readout and sparse-write bridge under LLM-audited labels. Conflict disclosure follows a different pattern, appearing weaker, later-layer, and only partially steerable. These results support a practical monitoring strategy for retrieval systems that detects silent source commitment with residual readouts and pairs conflict-surfacing prompts with disclosure steering when the model has selected a side under disagreement.

1. Introduction

Transparent AI systems should make consequential model choices visible to users, especially when those choices concern conflicting evidence. This problem has become more important as language models are paired with retrieved or supplied documents in retrieval-augmented generation and assistant systems, a setting introduced by RAG and now surveyed as a standard way to ground LLM output in external evidence (Lewis et al., 2020; Gao et al., 2024). External context can improve factual coverage, but it does not guarantee faithful evidence use, since models can ignore relevant context depending on position, generate fluent statements unsupported by evidence, and produce explanations that misrepresent the computation behind an answer (Liu et al., 2024; Ji et al., 2023; Turpin et al., 2023).

Conflicting evidence makes this transparency problem sharper. Knowledge-conflict surveys and benchmarks distinguish cases where retrieved context conflicts with parametric memory from cases where retrieved documents conflict with each other. FaithEval and WikiContradict show that models can fail to stay faithful when the supplied context is inconsistent or counterfactual (Xu et al., 2024; Ming et al., 2025; Hou et al., 2024). The risk is more than an incorrect answer. In civic, medical, legal, and public-information settings, users may need to know when an answer came from one contested source rather than from settled evidence (Lee et al., 2023; Magesh et al., 2025). A model that silently resolves disagreement can make an internal source-selection decision while presenting the result as ordinary factual output. The user sees the selected answer, but not the conflict that shaped it or the source the model effectively privileged.

We study this failure mode with a controlled proxy from FaithEval-style inter-context conflict evaluation (Ming et al., 2025; Salesforce AI Research, 2024). Each item contains a question and two supplied documents, d_1 and d_2, that fill the same factual slot with incompatible answers. This setup separates events that are conflated in deployed retrieval systems, including registering that the documents disagree, selecting one source for generation, and deciding whether to disclose the disagreement to the user. Our hypothesis is that source selection and conflict disclosure both become encoded in the residual stream, but as separable internal states.

We use Qwen 0.8B as the primary mechanistic case study and repeat the core source-choice analyses on Gemma 3 1B as a transfer check (Qwen Team, 2026a; Gemma Team, 2025). Qwen supplies the full chain of evidence: residual readout, sparse MLP write bridge, bidirectional source steering, and prompt-matched disclosure steering. Gemma changes the model family while staying in the same small-model regime, letting us ask whether the source-choice

[1]KAIST, School of Computing, Daejeon, South Korea. Email: <samuelassefa@kaist.ac.kr> [2]Konkuk University, School of Electrical and Electronics Engineering, Seoul, South Korea. Correspondence to: Jae Won Cho <chojw@konkuk.ac.kr>.

Accepted to the ICML 2026 Trustworthy AI for Good (AI4GOOD) Workshop, Seoul, South Korea, 2026.

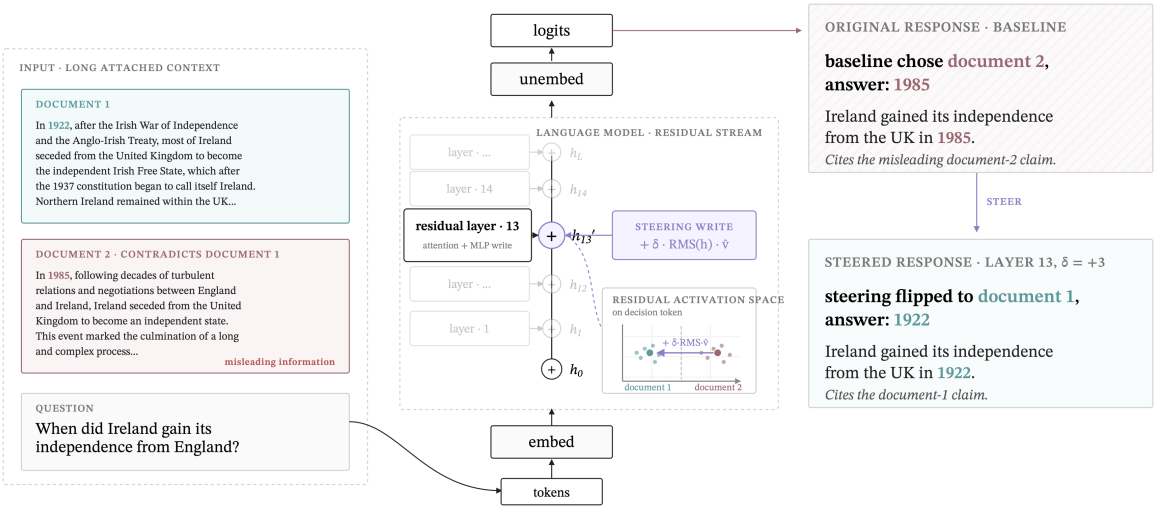


Figure 1. Source-selection steering under conflicting context. A question is paired with two long documents that support incompatible answers. Document 1 states that Ireland gained independence in 1922, while Document 2 states 1985. The baseline model gives a one-sided answer from Document 2. At generation time, we add a learned source-selection direction to the layer-13 residual stream on the decision token, shifting the activation from the Document 2 side of the doc-choice axis toward the Document 1 side. The steered model then answers from Document 1.

coordinate and sparse-write bridge survive outside Qwen. We keep the stronger disclosure-steering claim scoped to Qwen.

Our contributions are:

- We show that competition between two document-supported continuations collapses into a residual source-choice coordinate whose sign tracks the source used in the answer.
- We identify sparse middle-layer MLP writers that feed this coordinate, giving a component-level bridge between localized neurons and the downstream residual state.
- We show that source selection and conflict disclosure separate — the model can internally commit to one source while leaving the conflict invisible in the final answer, and disclosure steering recovers only part of this gap.

These results give retrieval systems a monitoring target beyond output text alone: source commitment without corresponding conflict disclosure.

2. Related Work

Context faithfulness and knowledge conflict. Retrieval-augmented generation gives language models access to external evidence at inference time, which can improve factual coverage and update answers beyond the model's parameters (Lewis et al., 2020; Gao et al., 2024). The same setup creates a conflict-management problem when retrieved documents disagree with one another or with parametric memory. FaithEval evaluates context faithfulness under unanswerable, inconsistent, and counterfactual contexts. WikiContradict focuses on real inter-context contradictions drawn from Wikipedia (Ming et al., 2025; Hou et al., 2024). Knowledge-conflict surveys distinguish context-memory, inter-context, and intra-memory conflicts (Xu et al., 2024). Our experiments sit in the inter-context case, where both candidate answers are supplied in the prompt, so the question is which source the model commits to and whether it discloses the disagreement.

Closest mechanistic work studies how models represent and intervene on conflicts between retrieved context and parametric memory. Zhao et al. analyze residual-stream conflict signals and show that source-use patterns emerge in intermediate and later layers. Their follow-up SpARE method uses pretrained sparse autoencoders to steer whether models use contextual or parametric knowledge (Zhao et al., 2024; 2025). Li et al. argue that context and memory are superposed in attention heads and introduce JuICE, a test-time attention intervention for steering toward either source (Li et al., 2025). Our setting differs in two ways. The competing sources are two supplied documents rather than context versus memory, and the main transparency claim includes the gap between internal source choice and user-facing conflict disclosure.

Residual readouts and activation steering. The transformer-circuits view treats each token position as a residual vector repeatedly updated by attention and MLP blocks (Elhage et al., 2021). Logit Lens and Tuned Lens use these intermediate residual states as approximate prediction trajectories, asking what the model is ready to predict before the final layer (nostalgebraist, 2020; Belrose et al., 2023). Representation-engineering work extends the same idea from measurement to intervention: directions

fitted from contrastive activations can be added, removed, or projected out to change model behavior (Li et al., 2023; Rimsky et al., 2024; Marks & Tegmark, 2024). Refusal directions provide a close methodological precedent for treating a behavior as low-dimensional residual geometry (Arditi et al., 2024). We apply this readout-and-steering recipe to source selection under contradictory retrieved context, and extend the readout to conflict surfacing.

MLP writers and transparency gaps. A readable residual direction raises a component-level question: which parts of the model write into it? Feed-forward layers have been studied as sites of factual storage and editing, and transformer-circuits analyses describe both attention and MLP blocks as additive writers to the residual stream (Geva et al., 2021; Meng et al., 2022; Elhage et al., 2021). We adapt the contribution-normalization idea behind CETT (Zhang et al., 2024), originally a layer-level sparsity metric, into a per-neuron contribution score, following its per-neuron use in H-Neurons (Gao et al., 2025). The sparse analysis here tests whether the localized MLP writes point into the same residual source-choice coordinate that we read and steer. The transparency motivation is related to unfaithful-explanation work: surface text can fail to reveal the computation that produced it (Turpin et al., 2023). Our case is silent conflict resolution in a retrieval-style setting rather than generated rationales.

3. Setup and Methodology

Task and labels. Qwen 0.8B is the primary mechanistic model, with Gemma 3 1B used as a transfer check (Qwen Team, 2026a; Gemma Team, 2025). The small-model setting reduces, but does not eliminate, the chance that parametric recall dominates the supplied documents. We therefore also use closed-book recall controls. Each FaithEval item pairs a question with two documents, d_1 and d_2, that answer the same slot with incompatible strings (Salesforce AI Research, 2024; Ming et al., 2025). Because the benchmark order couples d_1 with the original answer and d_2 with the conflicting answer, we analyze *document choice* rather than truth, and use order swaps plus closed-book recall controls to reduce truth, position, and memory confounds. Source labels come from answer-alias matching. Conflict-surfacing labels come from an LLM judge that distinguishes one-sided answers, conflict-aware leaning answers, strict unresolved disclosure, weak disclosure, and truncation. Full prompts, model revisions, splits, and judge audits are in Sections B.1, B.4, C and D.

Residual readouts. We capture residuals as inputs to transformer layers. For row i, layer ℓ, and readout position set P_i, the aggregated state is

$$r_i^{(\ell,P)} = |P_i|^{-1} \sum_{t \in P_i} h_{i,t}^{(\ell)}. \tag{1}$$

We evaluate prompt-final, first-response, last-decision, and mean-decision positions. For source choice, $s_i = +1$ denotes a Document 2 answer and $s_i = -1$ denotes a Document 1 answer. The readout direction is the normalized difference between class centroids,

$$\hat{v}_{\text{doc}}^{(\ell,P)} = \frac{\mu_+^{(\ell,P)} - \mu_-^{(\ell,P)}}{\|\mu_+^{(\ell,P)} - \mu_-^{(\ell,P)}\|_2}, \quad \mu_\pm = |\mathcal{D}_\pm|^{-1} \sum_{i \in \mathcal{D}_\pm} r_i. \tag{2}$$

Held-out rows are scored by projection onto $\hat{v}_{\text{doc}}$. This is a class-centroid (mean-difference) direction, not a logistic-regression probe, because we also steer with it rather than only classify. In Qwen, validation selects layer 13 at `mean_decision`, giving validation AUROC 0.9959 and held-out test AUROC 0.9945.

Steering and sparse writes. For residual steering, we add the saved direction at selected token positions,

$$\tilde{h}_t^{(\ell)} = h_t^{(\ell)} + \delta \, \mathrm{RMS}(h_t^{(\ell)}) \, \hat{v}_{\text{doc}}. \tag{3}$$

Because the Qwen direction is fitted on the input to layer 13, the hook edits the output of layer 12 so that layer 13 receives the steered residual. For sparse MLP analysis, CETT gives each MLP neuron a contribution-normalized activation score. It combines the neuron's activation with the norm of its down-projection vector, then normalizes by the total MLP output norm at that token. High-CETT neurons are therefore not just active but also make a comparatively large contribution to the MLP write into the residual stream. We use these CETT features to localize neurons whose answer-span contributions predict source choice, then project each selected neuron's actual output write onto the source-choice direction,

$$m_{i,j,t}^{(\ell)} = z_{i,j,t}^{(\ell)} \left\langle W_{\text{down},:,j}^{(\ell)}, \hat{v}_{\text{doc}} \right\rangle. \tag{4}$$

Aggregating these signed writes tests whether the localized MLP set feeds the same residual coordinate that is readable and steerable downstream.

Conflict-surfacing directions. For surfacing, the contrast is conflict-aware disclosure versus plain one-sided answers under a prompt that explicitly asks the model to describe document disagreement before answering. The fitting recipe matches the source-choice readout, but the label is disclosure rather than answered source. The selected Qwen disclosure direction appears at layer 17 and `pre_response_last`. Steering is evaluated only on prompt-matched failures, where the prompt asked for disclosure but the model still answered one-sidedly.

4. Experimental Findings

4.1. Source selection becomes a residual state

4.1.1. Residual readout reveals source commitment

We first ask whether hidden source commitment is present in the residual stream while the answer is being formed. The learned direction behaves like a selected-source coordinate across shortcut checks. In Qwen, validation selects the layer-13 `mean_decision` state, and the direction is near ceiling on held-out rows (validation AUROC 0.996; pooled test AUROC 0.994). Under document-order swaps, the sign follows the document the model answers from rather than the original position or the truthful side (original/swap AUROC 0.993/0.997). Rows where answer strings are lexically similar give the same interpretation: both the residual projection and sparse CETT score follow the selected document role rather than the visible answer string alone.

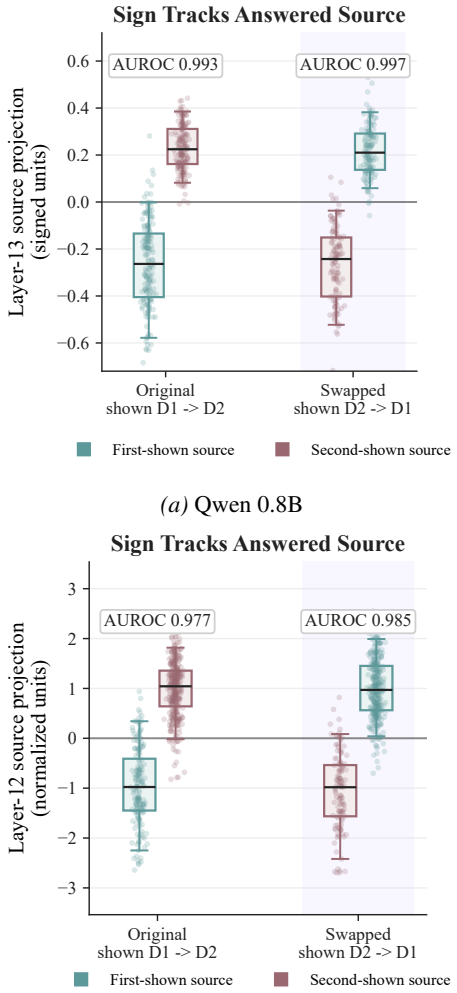
Gemma preserves the qualitative geometry with an independently selected answer-span direction (clean held-out AUROC 0.980 on 907 clean doc-position rows; original/swap AUROC 0.977/0.985). The exact layer and aggregation differ: Qwen peaks at layer 13 with `mean_decision`, whereas Gemma peaks at layer 12 with `last_decision`. The shared pattern is answer-span timing. Prompt-final and first-response probes are much weaker in Qwen, while the clean readout appears when the model is about to produce and commit to one of the two supported answers.

4.1.2. Steering the source-choice state changes the answered source

A linear readout becomes more informative when intervening on it changes behavior in the predicted direction. We steer the Qwen layer-13 source-selection axis over the final prompt tokens and beginning of the decision span, using an orthogonalized random residual unit as the control.

The steering subset contains 100 balanced held-out qids under both original and swapped order, giving 200 rows. Baseline answers use doc2 in $103/200 = 51.5\%$ of rows. Positive steering makes doc2 win more often: at $\delta = +3$, the doc2 rate among decided rows rises to 74.7%, with 41 doc1-to-doc2 flips and one reverse flip. Negative steering makes doc1 win more often: at $\delta = -3$, the doc2 rate falls to 32.2%, with 35 doc2-to-doc1 flips and one reverse flip. The matched random direction stays near baseline, with doc2 rates between 51.0% and 52.3%.

Gemma steering transfers in a weaker but diagnostic form. Response-only steering of the layer-12 source-choice direction produces far more answered-source flips than a stat-matched random unit ($28/960$ versus $2/960$), with almost all localized flips following the predicted sign ($27/28$; Fisher $p = 3.65 \times 10^{-7}$). Because the flip rate is low and wording changes also increase, we treat Gemma steering as low-rate causal support rather than a Qwen-strength behavioral replication. Together, the Qwen and Gemma results show that the readout is not only predictive — by the relevant middle layer, the two document-supported continuations have compressed into a residual coordinate that can influence which source wins.



(a) Qwen 0.8B

(b) Gemma 3 1B

Figure 2. Residual sign tracks answered source in both models. Colors encode presentation order, so the separation cannot be reduced to a first-document or truth readout.

Steering which source wins leaves a separate transparency question, namely whether the model will tell the user that the two sources were in conflict.

4.2. Sparse MLP writers feed the selection state

A residual readout could still be a passive trace of the answer tokens. We therefore localize candidate writers behind the readout. CETT first identifies MLP neurons whose answer-span activations predict source choice. The write-projection

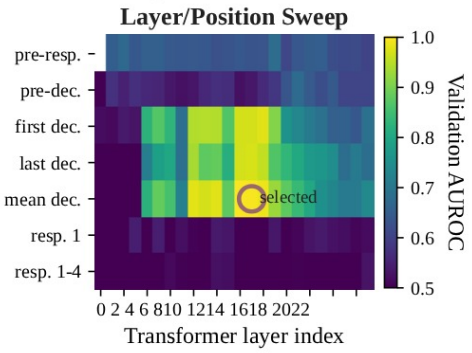
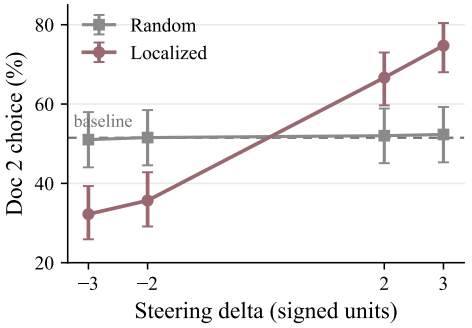


Figure 3. The source-choice readout is answer-span aligned. The strongest validation AUROC appears at decision-token positions and peaks at `mean_decision`. The layer/position sweep shows the same band of high readout performance across middle layers near the answer span, with the selected layer-13 direction marked.



Representative row. How old was Peyton Manning in 2015?	
Documents	d_1 says 39; d_2 says 34.
Baseline	"Peyton Manning turned 39 in the 2015 off-season." (d_1)
Local +2	"Peyton Manning turned 34 in the 2015 off-season." (d_2)
Random +2	"Peyton Manning turned 39 in the 2015 off-season." (d_1)

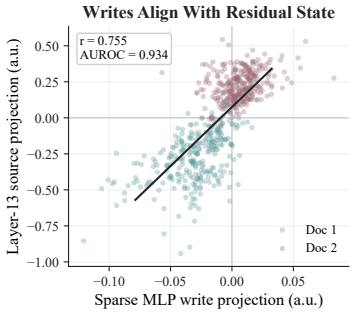
Figure 4. Residual steering of the source-selection axis. The curve aggregates original and swapped held-out rows. Error bars are Wilson 95% confidence intervals over decided rows. The table shows a representative source flip.

test then asks whether those neurons' actual residual updates point into the learned source-choice direction.

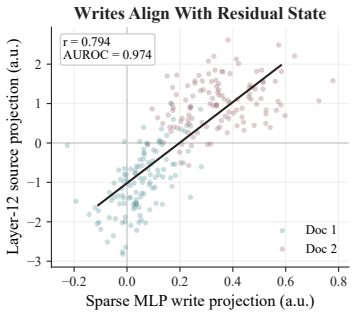
In Qwen, CETT repeatedly selects about 156 neurons across five balanced splits, mostly in layers 11–14 and roughly balanced in sign. This sparse set predicts held-out chosen side with mean AUROC 0.900 and F1 0.808, and its signed CETT score tracks the dense layer-13 residual projection (AUROC 0.995; $r = 0.879$). The stronger test is the actual write projection: after summing selected-neuron writes and averaging over decision tokens, the upstream MLP write score separates doc2-choice from doc1-choice rows at AUROC 0.934 and correlates with the dense residual projection at $r = 0.755$.

Gemma gives the same bridge with a different sparse set. Its localized neurons sit earlier than Qwen's, but their actual MLP writes still point into Gemma's independently learned source-choice direction — the writer projection separates answered doc2 from answered doc1 at AUROC 0.974, while layer-matched random writer sets average 0.546. Matched controls in Figures 7 and 8 further show that the signed magnitude is stronger under contradiction than under exact-copy agreement, paraphrased agreement, or single-document controls, including on closed-book unknown rows. The localized MLP signal therefore looks like a conflict-conditioned source-commitment signal rather than a response to document count, length, or duplicate-document structure.

These middle-layer MLP writes provide the component-level bridge between sparse localized neurons and the downstream source-choice residual state.



(a) Qwen 0.8B



(b) Gemma 3 1B

Figure 5. Sparse MLP writes align with the residual source-selection state. CETT-localized writes project onto the learned source-choice direction in both models. Layer-distribution histograms are in Figure 14.

5. Conflict Surfacing Is a Separate Transparency State

After the source-choice readout, steering, and sparse-write analyses, the remaining question is whether the same mechanism also controls user-facing disclosure. The source-selection axis tells us which document-conditioned continuation the model is moving toward. Disclosure asks whether the response will say that this choice was made under contra-

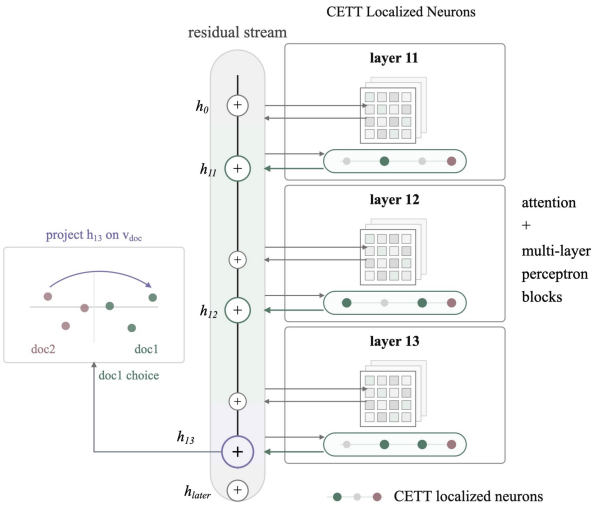


Figure 6. Sparse MLP writers feed the source-choice residual state. CETT localizes MLP neurons whose activations distinguish source choice, and the write-projection test asks whether their residual updates align with the learned source-choice direction.

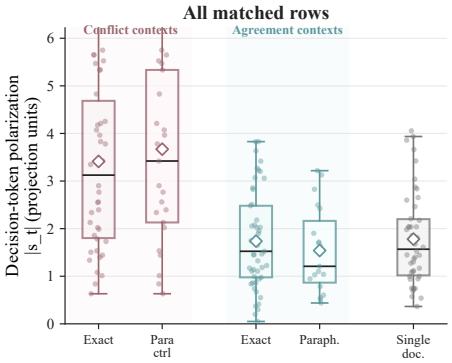


Figure 7. Matched controls for conflict specificity. Agreement and single-document controls reduce sparse-write magnitude relative to contradiction.

diction. A surfacing-friendly prompt helps but leaves many hidden-commitment failures, yielding only 35.3% broad conflict surfacing and 22.4% strictly unresolved surfacing on the 1500-row regenerated corpus. The steering experiment starts from the harder prompt-matched failure subset, where the same prompt produced a plain one-sided answer. After filtering to closed-book `parametric_unknown` rows, this leaves 109 reproducible failures with 0% prompt-only surfacing by construction.

The best Qwen surfacing readout appears before the response begins, at `pre_response_last` in layer 17, while source choice is strongest around answer tokens at layer 13. Gemma shows the same separation in a data-limited readout, with source choice peaking at layer 12 around the answer span, whereas surfacing peaks at layer 17 before generation. Broad surfacing reaches held-out AUROC 0.820 on 259 rows, and strict unresolved surfacing reaches AUROC 0.907 on 240 rows. The signals therefore look like different inter-

Figure 8. Closed-book unknown subset. The contradiction gap remains when closed-book recall is unlikely to dominate.

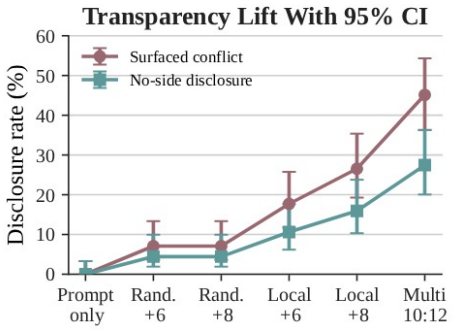
nal states rather than two labels for the same computation.

Broad surfacing also mixes two behaviors: strict unresolved disclosure, where the model names the contradiction and refuses to choose, and leaning disclosure, where it names the contradiction but still gives a one-document answer. The layer-17 direction separates strict unresolved surfacing from plain answers much more cleanly than leaning surfacing from plain answers (AUROC 0.854 versus 0.615), so the intervention uses one direction for disclosure and another to resist collapse back to a one-sided answer.

On these 109 failures, localized layer-17 steering moves broad conflict acknowledgment from 0 to 20 rows at $\delta = +6$ and 29 rows at $\delta = +8$. The two-direction arm is largest, reaching $46/109$ broad surfacing, with $27/109$ no-side unresolved responses and $56/109$ still one-sided. Figure 9 shows the same intervention sweep as outcome shares and confidence intervals. Prompting and steering therefore play different roles — prompting improves surface behavior, while a disclosure direction recovers some remaining Qwen failures. Gemma reinforces the deeper-layer and weaker surfacing story, but surfacing steering remains a Qwen-only result in this paper.

Hidden commitment under contradiction

Instruction.	If the documents disagree, explicitly describe the conflict before answering.
Question.	When did the first Pokémon game come out?
Documents.	d_1 says 1996; d_2 says 1989.
Prompt-only.	"The first Pokémon game came out in 1989."
Local +6.	"The documents contain conflicting dates. Document 1 states 1996; Document 2 states 1989. Since they disagree, the conflict cannot be resolved from context alone."



Condition	Surfaced	No-side subset	Still one-sided
Prompt only	0 (0.0%)	0	109
Local +6	20 (18.3%)	12	86
Local +8	29 (26.6%)	17	76
Multi-local	**46 (42.2%)**	**27**	**56**

Figure 9. Disclosure steering improves conflict surfacing on prompt-matched Qwen failures. Points show transparency lift with Wilson 95% confidence intervals; counts below are on the 109 reproduced prompt-failure rows with closed-book `parametric_unknown` labels. *Surfaced* is broad conflict acknowledgment, and *No-side subset* is strict unresolved disclosure within surfaced rows.

6. Limitations

The main limitations are scope and scale. Qwen is the only model for the full disclosure-steering suite. Gemma strengthens the source-choice mechanism but does not replicate surfacing steering. We use small models because they allow exhaustive activation sweeps, intervention grids, and generation audits, but larger models may distribute source competition across different layers or use attention and retrieval-conditioned planning mechanisms not visible here. The dataset also covers one controlled conflict format, mostly short factual disagreements with recoverable answer aliases. Surfacing labels depend on an LLM judge, steering is partial rather than deployment-ready, and Gemma source-choice steering is statistically clear but low-rate.

7. Conclusion

Conflicting-context question answering reveals a transparency gap between what a model internally selects and what it tells the user. In the tested setting, source choice becomes a readable residual coordinate, sparse middle-layer MLP writes feed that coordinate, and steering the coordinate can change which document the model answers from. Conflict disclosure follows a different pattern: it is weaker, later-layer, and only partially steerable. For retrieval and assistant systems, this suggests a concrete diagnostic target. When a model has a strong source-commitment signal under disagreement but does not disclose that disagreement, the system should treat the answer as silently resolved contested evidence rather than as ordinary factual output.

Impact Statement

This work studies a transparency failure in retrieval-augmented language models: the model can internally select one of two conflicting documents while presenting the result as ordinary factual output. Residual-readout diagnostics of the form developed here could support audit tools for high-stakes retrieval deployments in civic, medical, and legal settings, where users have a legitimate interest in knowing whether an answer reflects settled evidence or a silent selection between contested sources.

The same contrastive recipe used to detect hidden commitment can in principle be inverted to suppress disclosure. This is a real dual-use concern, especially for systems that retrieve contested evidence and could be steered away from acknowledging disagreement. The asymmetry favors defenders: readouts run at inference, while persistent suppression requires fine-tuning or always-on steering and leaves traces. The directions reported here are model-specific, small-scale, and framed as diagnostic rather than deployment-ready.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT)(No. RS-2026-25476101).

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gao, C., Chen, H., Xiao, C., Chen, Z., Liu, Z., and Sun, M. H-Neurons: On the existence, impact, and origin of hallucination-associated neurons in LLMs. *arXiv preprint arXiv:2512.01797*, 2025.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. Retrieval-

augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. URL https://arxiv.org/abs/2312.10997.

Gemma Team. Gemma 3. Kaggle, 2025. URL https://goo.gle/Gemma3Report.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Google. Gemini 3.1 Flash-Lite Preview. Google AI for Developers model documentation, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite-preview. Accessed 5 May 2026.

Google DeepMind. google/gemma-3-1b-it. Hugging Face model card, 2025. URL https://huggingface.co/google/gemma-3-1b-it. Accessed 5 May 2026.

Hou, Y., Pascale, A., Carnerero-Cano, J., Tchrakian, T., Marinescu, R., Daly, E., Padhi, I., and Sattigeri, P. WikiContradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from Wikipedia. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Lee, P., Bubeck, S., and Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. doi: 10.1056/NEJMsr2214184.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Li, G., Chen, Y., and Tong, H. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34074–34104. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25c.html.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk. Spotlight.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ming, Y., Purushwalkam, S., Pandit, S., Ke, Z., Nguyen, X.-P., Xiong, C., and Joty, S. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"? In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=UeVx6L59fg. Poster.

nostalgebraist. Interpreting GPT: the logit lens. *LessWrong blog post*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen/Qwen3.5-0.8B. Hugging Face model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-0.8B. Revision 2fc06364715b967f1860aea9cf38778875588b17, accessed 5 May 2026.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Salesforce AI Research. Salesforce/FaithEval-inconsistent-v1.0. Hugging Face dataset card, 2024. URL https://huggingface.co/datasets/Salesforce/FaithEval-inconsistent-v1.0. Test split, accessed 5 May 2026.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Zhang, Z., Song, Y., Yu, G., Han, X., Lin, Y., Xiao, C., Song, C., Liu, Z., Mi, Z., and Sun, M. ReLU^2 wins: Discovering efficient activation functions for sparse LLMs. *arXiv preprint arXiv:2402.03804*, 2024. URL https://arxiv.org/abs/2402.03804.

Zhao, Y., Du, X., Hong, G., Gema, A. P., Devoto, A., Wang, H., He, X., Wong, K.-F., and Minervini, P. Analysing the residual stream of language models under knowledge conflicts. In *Foundation Model Interventions Workshop at NeurIPS*, 2024. URL https://arxiv.org/abs/2410.16090. arXiv:2410.16090.

Zhao, Y., Devoto, A., Hong, G., Du, X., Gema, A. P., Wang, H., He, X., Wong, K.-F., and Minervini, P. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5117–5136, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.264. URL https://aclanthology.org/2025.naacl-long.264/.

A. Token Timing and Intervention Windows

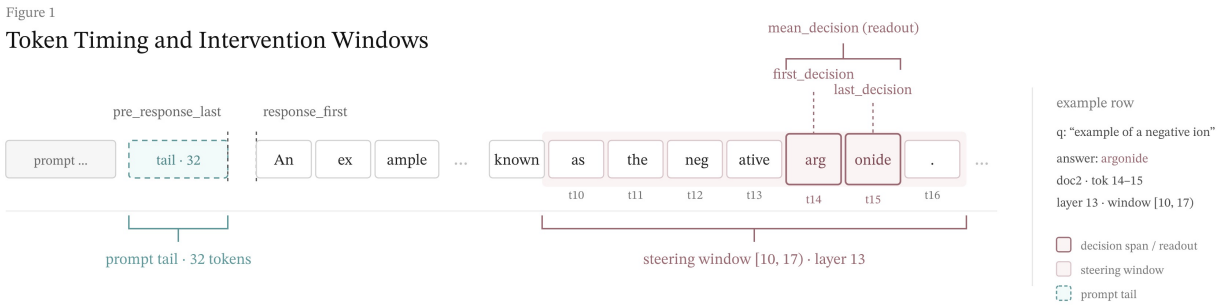


Figure 10. Token positions and intervention windows used in the residual readout and steering analyses. The schematic defines the prompt tail, decision-token region, response-onset positions such as `pre_response_last`, and the response-time windows used for source-choice and conflict-surfacing interventions.

B. Reproducibility and Audit Details

B.1. Exact Prompt Templates

The main text describes the prompt families in prose. For replication, this appendix records the exact text templates used by the scripts. Braced fields denote row-specific substitutions.

Base two-document QA prompt. This was the default FaithEval-style generation prompt used for the source-choice experiments.

```
Answer the question using the provided context. If the context conflicts with known facts, explicitly resolve the
     conflict in your answer.

Context:
{context}

Question: {question}

Answer:
```

Prompt-matched conflict-surfacing prompt. The surfacing readout used a single prompt family rather than comparing a baseline prompt against a separate skeptical system prompt. The task instruction was:

```
Answer the question using the provided context. If there is any conflict or mismatch in the documents, notify that
     in the response.
```

The full user prompt was formed as:

```
{task_instruction}

Context:
{context}

Question: {question}

Answer:
```

Earlier skeptical-system prompt. Some diagnostic runs used a distinct system prompt to test whether the model could be moved toward a more cautious response style:

```
You are a careful QA assistant. Be skeptical of every claim, reason carefully when documents conflict, and avoid
     blindly following either document.
```

B.2. Judge Prompts

All automated transparency labels used temperature zero with JSON-mode decoding and no retrieval or external tools. The surfacing judge model was `google/gemini-3.1-flash-lite-preview`. The primary surfacing judge prompt was:

```
SYSTEM:
You audit answers to a question with two contradictory documents. Classify whether the assistant explicitly
     surfaces the contradiction, and whether it still effectively chooses one document. Return strict JSON only
     with keys: overall_label, conflict_aware, effectively_decided, leans_to, mentions_both, truncated, confidence,
      reason. Allowed overall_label values: plain_decided, conflict_aware_leaning, conflict_aware_unresolved, weak,
      truncated. Allowed leans_to values: doc1, doc2, none, unclear. Use plain_decided when the answer effectively
     chooses one side without genuinely surfacing the clash. Use conflict_aware_leaning when it explicitly surfaces
      disagreement or contradiction but still effectively leans to one side. Use conflict_aware_unresolved when it
     explicitly surfaces the disagreement and stays genuinely non-committal. Use weak when the answer is low-
     quality, evasive, or not clearly conflict-aware. Use truncated when the answer appears incomplete enough that
     the rest is unreliable.

USER:
Question:
{question}

Document 1 answer aliases:
{truthful_aliases_json}

Document 2 answer aliases:
{misleading_aliases_json}

Response:
{response}

Return JSON only.
```

The Gemma source-choice audit used a separate judge schema to validate document positions and response source choice:

```
SYSTEM:
You are auditing a two-document conflict QA dataset. Use only the provided two documents and aliases. Do not use
     outside knowledge. Your job is to validate whether the documents contain a clean conflict and classify which
     source the model response follows.

Return strict JSON only with keys: context_valid, truthful_supported, misleading_supported, conflict_clear,
     truthful_doc_position, misleading_doc_position, response_choice, response_doc_position, decision_span,
     mentions_truthful, mentions_misleading, conflict_aware, low_quality_or_truncated, confidence, reason.

Allowed response_choice values: truthful, misleading, both, neither, ambiguous. Allowed doc position values: 1, 2,
     both, neither, ambiguous. context_valid=true only if one document supports a truthful alias, the other
     supports a misleading alias, and the question is answerable from those conflicting candidates. Classify the
     response by semantic answer, not exact punctuation.

USER:
Question:
{question}

Document 1 excerpt:
{document_1_excerpt}

Document 2 excerpt:
{document_2_excerpt}

Truthful aliases:
{truthful_aliases_json}

Misleading aliases:
{misleading_aliases_json}

Existing automatic label:
{automatic_label}

Existing automatic decision span:
{automatic_decision_span}

Assistant response:
{response}

Return JSON only.
```

B.3. Direction Inventory

Table 1 consolidates the learned readout directions that are otherwise distributed across the main text and appendix. Fit sizes report the training rows used for the mean-difference direction; validation AUROC was used for selection, and held-out AUROC is measured on the corresponding test split. For source-choice directions, positives are document-2 choices and negatives are document-1 choices.

Model	Direction / contrast	Layer	Position	Fit rows	Val AUROC	Held-out AUROC
Qwen	Source-choice residual direction, doc2 vs. doc1	13	`mean_decision`	235 train (116/119); 54 val	0.9959	0.9945
Gemma	Source-choice residual direction, doc2 vs. doc1	12	`last_decision`	238 train (163/75); 61 val	0.9897	0.9806
Qwen	Broad disclosure, conflict-aware vs. plain decided	17	pre-resp. last	727 train (305/422); 243 val	0.7935	0.7670
Qwen	Strict disclosure, unresolved vs. plain decided	17	pre-resp. last	629 train (207/422); 207 val	0.8397	0.8553
Qwen	Anti-collapse, noncommittal vs. effectively decided	23	pre-resp. last	772 train (242/530); 255 val	0.7591	0.8082

Table 1. Inventory of reported directions. Parentheses in the fit column give positive/negative counts for the training split. Source-choice positives are document-2 selections; disclosure positives are conflict-aware or noncommittal labels.

For the Gemma surfacing transfer analysis, the later full stratified run selected the same `pre_response_last` layer-17 position for both broad and strict surfacing. Broad disclosure used 777 training rows, 260 validation rows, and 259 held-out rows, with validation AUROC 0.8386 and held-out AUROC 0.8202. Strict unresolved disclosure used 734 training rows, 246 validation rows, and 240 held-out rows, with validation AUROC 0.8826 and held-out AUROC 0.9074.

B.4. Judge Calibration and Manual Audit

The judge was not used as an unchecked oracle. We performed manual audits of calibration views, truncation cases, and label-boundary examples. These checks were designed to catch schema failures, systematic truncation artifacts, and ambiguous boundaries between `conflict_aware_leaning` and `conflict_aware_unresolved`; they are not a formal blinded inter-annotator agreement study.

Gemma strict surfacing labels. The audit covered 1500 judged rows. Final label counts were 1149 plain decided, 203 weak, 74 leaning, 71 unresolved, and 3 truncated. Manual review found no judge consistency failures and 3 truncation suspects. Five labels were corrected from leaning to unresolved where the judge rationale itself said that no side was chosen. These labels are used for the Gemma strict surfacing readout.

Qwen baseline transparency rejudge. The audit covered 1443 judged rows. Label counts were 1013 plain decided, 223 mentions-other-candidate, 149 leaning, 8 unresolved, and 50 truncated. The conflict-aware label rate was 10.9%, and the rejected-candidate mention rate was 33.9%. This audit is used to characterize spontaneous transparency rather than to train a new label set.

Skeptical ambiguity boundary set. For 13 edge cases, a three-model judge ensemble gave 7 truncated, 5 leaning, and 1 weak majority labels. Five rows were unanimous and 8 had a two-label split. We use this only as a boundary diagnostic.

C. Data Details

Model details. Qwen 0.8B is `Qwen/Qwen3.5-0.8B` at revision `2fc06364715b967f1860aea9cf38778875588b17` (Qwen Team, 2026b). The checkpoint includes a vision encoder and a 24-layer hybrid Gated DeltaNet/Gated Attention language backbone, but all experiments use the text-only path with no image, video, or visual tokens. Gemma 3 1B is `google/gemma-3-1b-it`, a 26-layer instruction checkpoint from the Gemma 3 family (Google DeepMind, 2025).

Balanced FaithEval localization pool. We start from `Salesforce/FaithEval-inconsistent-v1.0`, `test` split (Salesforce AI Research, 2024). Each row has a question q and two documents d_1, d_2 that answer the same slot with different strings. Benchmark-role aliases come from row metadata. The CETT localization uses a balanced pool of 1350

rows, with 675 resistant and 675 compliant examples. Document order is preserved for this localization pass, while traversal, balanced keep-set, and split assignment are randomized across seeds.

Curated matched subset. The contradiction-specific controls use an audited 100-row curated subset. Each row has a clean question, non-overlapping d_1 and d_2 answer aliases, and a recoverable decision span. The exact-copy control run includes original conflict prompts, swapped-order conflict prompts, two agreement controls ($d_1 + d_1$ and $d_2 + d_2$), and two single-document controls (d_1 alone and d_2 alone). A separate paraphrased-agreement run keeps the same qids but replaces the second agreeing document with a surface-level paraphrase of the first document, so it controls for having two document-length passages rather than only one. Its conflict condition is the original conflict condition on the same matched qids; thus the "Para ctrl" conflict column in Figure 7 means conflict rows from the paraphrased-control run, not a claim that both conflicting documents were paraphrased. In plots that show a single "single document" condition, the d_1-only and d_2-only controls are pooled or size-matched to the conflict condition as specified by the run metadata.

Doc-choice contrastive pool. The residual doc-choice direction uses 500 FaithEval qids, split into $128/32/340$ train/validation/test qids. Each qid is run at $T = 0$ under both original and swapped document orders. Rows are used when the response can be uniquely attributed to one document by alias matching. In the original order, $500/500$ rows are usable and evenly split doc1/doc2. In the swapped order, $381/500$ are usable, with 184 doc1 choices, 197 doc2 choices, and 119 unresolved rows. This larger held-out pool was used to estimate usable row counts and readout behavior. The steering experiment then used the balanced subset of 100 held-out qids whose original-order and swapped-order rows were both usable, yielding the 200 rows reported in the main text.

Surfacing contrastive pool and judge. For the conflict-surfacing direction, we regenerate all 1500 cached FaithEval rows under an explicit instruction to describe document disagreement before answering. The operative prompt tail asks the model to describe document disagreement before answering. Responses are judged by `google/gemini-3.1-flash-lite-preview`. Counts are 707 `plain_decided`, 336 `conflict_aware_unresolved`, 193 `conflict_aware_leaning`, 181 `truncated`, and 83 `weak`. Thus the full regenerated corpus has $529/1500 = 35.3\%$ broad surfacing and $336/1500 = 22.4\%$ strict unresolved surfacing. The post-judge split is $305/102/101$ positives and $419/140/141$ negatives. The post-judge split includes only rows that pass the downstream feature/extraction filters. Thus, 21 broad-positive rows and 7 plain rows from the raw judge categories are outside the reported split. Truncated generations are counted for the discovery corpus but are not treated as surfaced conflict. The held-out steering counts in Figure 9 use a different denominator. These are rows where the same surfacing prompt already failed and produced `plain_decided`. Its prompt-only surfacing rate is therefore $0/109$ after the `parametric_unknown` join by construction. The broad surfacing-positive bucket is any non-truncated `conflict_aware` response, while the strict subset is `conflict_aware_unresolved`. Manual audit found that the labels are useful but noisy near the boundary between `plain_decided` and `conflict_aware_leaning`; this is why the main text separates strict unresolved surfacing from leaning surfacing.

LLM judge protocol. The judge model is `google/gemini-3.1-flash-lite-preview` (Google, 2026). Judge calls use no tools or retrieval, temperature 0, and a JSON-only output schema. Each call provides the question, both document snippets or answer strings, the generated response, and the row metadata needed to identify d_1 and d_2. The requested schema is `valid_conflict`, `chosen_document` $\in \{d1, d2, both, neither, unclear\}$, `surfacing_label` $\in$ {`plain_decided`, `conflict_aware_leaning`, `conflict_aware_unresolved`, `weak`, `truncated`}, `low_quality_or_truncated`, and a short rationale used only for audit. Invalid JSON is retried once; persistent failures are excluded from clean audited readout rows and treated as non-surfacing for disclosure counts.

Closed-book recall proxy. For a rough measure of whether the model knows an answer without the documents, we run each usable row closed-book under a neutral prompt and classify by alias match. This recall proxy is especially important because even a small model can know some benchmark answers parametrically. The small-model setting reduces the expected strength of this confound, but the closed-book audit is the direct control. It asks whether the effect remains on rows where the model does not recover either answer without the documents. On the 1386-row usable context-influence pool, 1131 rows are `parametric_unknown`, 175 are `parametric_correct`, 46 are ambiguous, and 34 are `parametric_wrong_misleading`. This is a recall proxy, not proof of absent knowledge. For the prompt-failure steering analysis in Figure 9, $109/141$ held-out prompt-failure rows join to `parametric_unknown` labels. The remaining

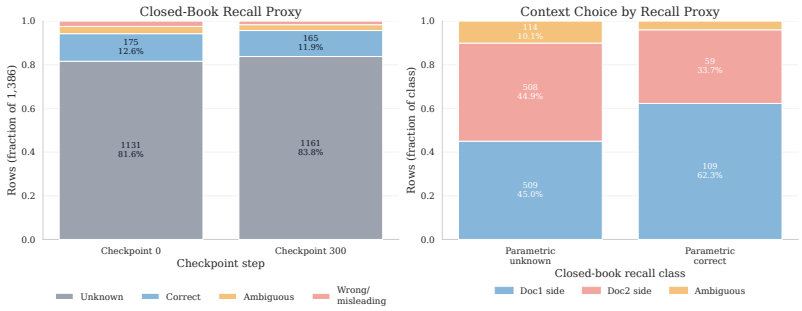


Figure 11. Closed-book recall proxy distribution on the usable FaithEval set.

rows are 23 `parametric_correct`, 7 ambiguous, and 2 missing from the joined recall artifact. The figure reports the reproducible 109-row subset as a partial control for parametric information.

D. Replication Protocol

All experiments use `Qwen/Qwen3.5-0.8B` at revision `2fc06364715b967f1860aea9cf38778875588b17` with the Hugging Face chat template. Generations are run in text-only, non-thinking mode. We do not pass image or video inputs, and no visual tokens are present in the prompts. Unless noted otherwise, generations are deterministic ($T = 0$) with `max_new_tokens`$= 96$ for doc-choice runs and `max_new_tokens`$= 224$ for surfacing runs. Reported splits are qid-level splits; no held-out table mixes rows from qids used to fit the corresponding direction. We use seed 42 for deterministic traversal and control construction.

Doc-choice labels are assigned by exact alias matching against the two document answers. A response is kept only when exactly one document's aliases match the answer span; rows with both aliases, neither alias, or no recoverable decision span are marked ambiguous. Residual readouts are fit on the mean decision-token state unless a sweep explicitly names another position. The layer/position sweep compares pre-decision, decision, and response-onset positions with the same train/validation/test qid partition.

CETT features are computed by hooking the input to each MLP `down_proj`. For token t, neuron j, and layer ℓ, the per-token contribution score is

$$\mathrm{CETT}_{\ell,j,t} = \frac{|z_{\ell,j,t}|\,\|W_{\mathrm{down}}^{(\ell)}[:,j]\|_2}{\|W_{\mathrm{down}}^{(\ell)} z_{\ell,t}\|_2 + 10^{-10}}.$$

Scores are averaged over decision tokens, and an L1 logistic classifier is fit on the balanced localization pool. Stable localized neurons are those recurring with the same sign across localization splits. Layer-matched random MLP controls sample the same number of neurons per layer as the localized set and exclude the localized neurons.

For surfacing, the prompt tail instructs the model to explicitly describe document disagreement before answering. Broad surfacing is any non-truncated conflict-aware judge label; no-side-disclosure is the unresolved subset. Steering controls use stat-matched random residual units at the same layer, orthogonalized against the learned residual direction.

E. Gemma Transfer Details

Gemma 3 1B is used as a transfer check for the source-choice mechanism, not as a full duplicate of the Qwen intervention suite. The run uses `google/gemma-3-1b-it` with the same FaithEval-style conflicting-document construction and the same high-level readout, CETT, control, and steering logic.

For sparse CETT localization, the Gemma run sees 1500 rows, of which 1468 are usable. Balancing compliant and resistant rows gives 1074 localization rows. Across five seeds, sparse CETT predicts held-out behavior with mean AUROC 0.880 and mean F1 0.805, compared with matched random AUROC 0.690. The stable set contains 56 neurons, with most in layers 6–12.

For residual document choice, the run uses 500 qids under original and swapped document order. The selected readout is layer 12 at `last_decision`, with validation AUROC 0.990 and held-out AUROC 0.981 before LLM cleaning. The LLM audit judges all 1000 original/swap rows and leaves 907 clean doc-position rows for the readout analysis, giving clean

held-out AUROC 0.980. Original-order and swapped-order clean AUROCs are 0.977 and 0.985, respectively. The audit finds zero concrete cases where the automatic labeler chose one document position and the LLM judge chose the opposite position.

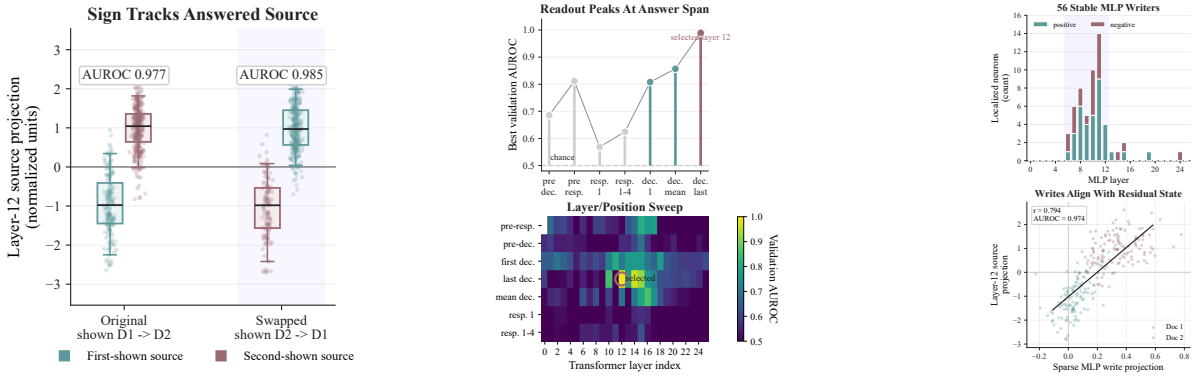
For the sparse-writer bridge, we project the actual upstream MLP writes of the 56 stable Gemma CETT neurons onto the independently learned layer-12 source-choice direction. On 240 balanced held-out rows, the all-sparse upstream write projection separates doc2-choice from doc1-choice rows at AUROC 0.974, with correlation $r = 0.794$ against the dense residual projection. Twenty layer-matched random writer sets have mean AUROC 0.546.

For Gemma source-choice steering, we report only the fixed-hook run. Earlier diagnostic hooks are excluded because even no-op hooks could alter generation. In the valid run, 240 qids are evaluated under original and swapped order with response-only steering at $|\delta| = 0.05$. Localized steering produces $28/960$ source-choice flips, compared with $2/960$ for the stat-matched random unit; $27/28$ localized flips follow the expected sign. The one-sided Fisher test for localized more than random flips gives $p = 3.65 \times 10^{-7}$. This is clean causal evidence for source-choice steering, but it is low-rate, and response text changes more often under localized steering than under the random unit ($83/960$ versus $39/960$).

For surfacing, Gemma provides only a prompt-matched readout/timing check. Source choice still peaks at layer 12 around the answer span, while the later full stratified surfacing run peaks deeper, at layer 17 with `pre_response_last`. Broad surfacing reaches held-out AUROC 0.820 on 259 rows, and strict unresolved surfacing reaches AUROC 0.907 on 240 rows. We use this only to support the claim that surfacing is a weaker and deeper-layer state than source choice, not as a Gemma disclosure-steering replication. The packaged Gemma surfacing panel is an older layer-22 pairwise diagnostic and is therefore not used as visual evidence for the layer-17 stratified result.

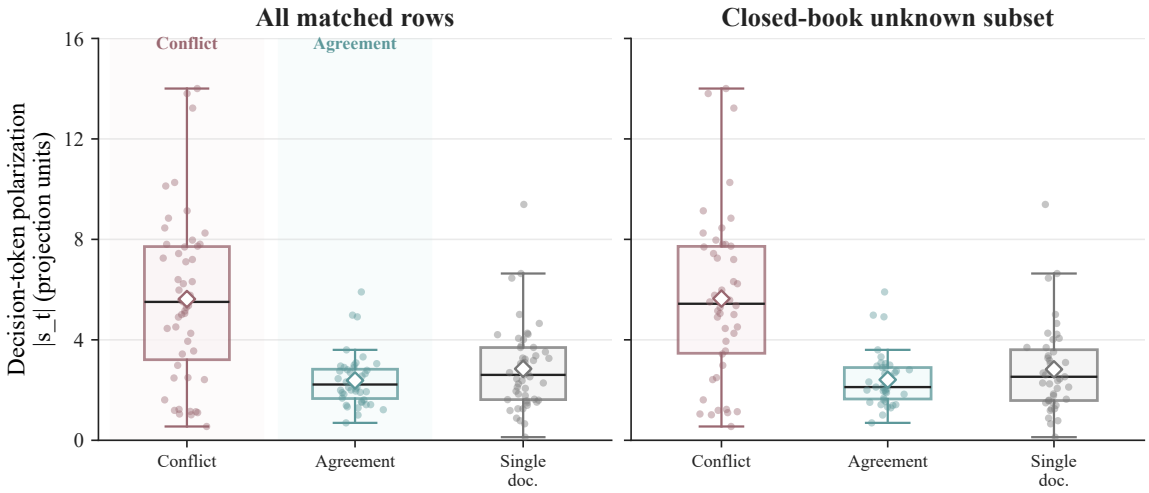
F. Gemma Transfer Figures

The full Gemma panels are included here so that the transfer claim is visually auditable.

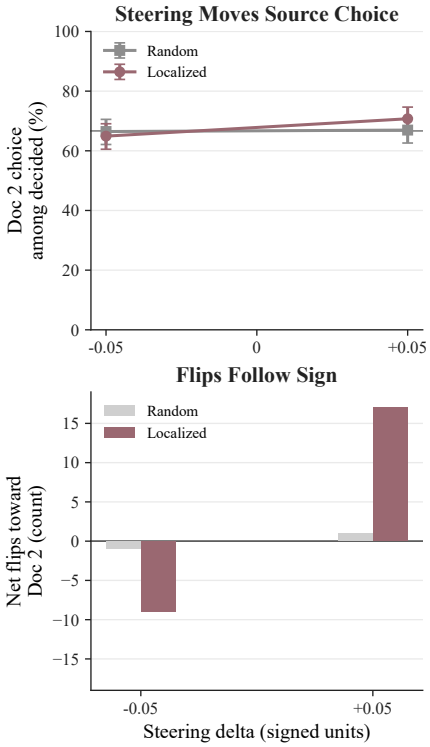


(a) Source-choice sign *(b)* Timing sweep *(c)* Sparse-to-residual bridge

Figure 12. Gemma source-choice transfer. The learned layer-12 source-choice projection tracks answered source under document-order swap, peaks near the answer span, and is recovered from sparse MLP writes.



(a) Conflict controls



(b) Source-choice steering

Figure 13. Gemma controls and intervention checks. The control panel shows larger internal decision-token polarization for contradictory documents than agreement or single-document controls. Source-choice steering is statistically clear but low-rate.

Gemma surfacing figure provenance. The text reports the later full stratified Gemma surfacing analysis: `pre_response_last`, layer 17, broad AUROC 0.8202 on 259 held-out rows, and strict AUROC 0.9074 on 240 held-out rows. The packaged `fig5_gemma_surfacing_readout_column` metadata instead records `response_first4_mean`, layer 22, and 22 test pairs. We therefore omit that older panel from the appendix figure set and treat Gemma surfacing as a textual readout/timing result until the full stratified panel is regenerated.

Case	Evidence	How it is treated
Gemma no-op hook sensitivity	Early diagnostic hooks altered generation even when intended as no-op instrumentation.	Excluded from causal claims; valid steering claims use the later fixed-hook response-only run.
Low-rate Gemma source steering	At residual scale $\|\delta\| = 0.05$, localized steering produced $28/960$ source-choice flips versus $2/960$ for random controls, but text changed more often than the flip rate itself: $83/960$ localized outputs versus $39/960$ random outputs.	Report as causal support with a small effect size, not as a full behavioral replication.
No Gemma surfacing intervention claim	Gemma surfacing transfer currently supports a readout/timing result, not a robust intervention result.	Keep separate from the Qwen surfacing steering result.
Layer-22 / layer-17 surfacing mismatch	The bundled Gemma surfacing figure metadata says layer 22 at `response_first4_mean`; the full stratified result says layer 17 at `pre_response_last`.	Omit the older panel and report the stratified result in text.
Leaning vs. unresolved label boundary	The Gemma manual audit changed $5/1500$ labels, all from leaning to unresolved, where the judge rationale said both candidates were presented without choosing.	Use strict unresolved labels for the cleanest policy target; report broad surfacing separately.

Table 2. Failure case gallery and partial-results accounting for Gemma transfer. The table distinguishes evidence retained in the paper from diagnostics that should not be overclaimed.

G. Additional Controls

Contradiction specificity. On the matched control runs, conflict prompts polarize decision tokens more strongly than agreement or single-document prompts. In the original-order audited subset, paired sign tests give $p = 1.5 \times 10^{-4}$ for Conflict versus Agreement and $p = 7.6 \times 10^{-6}$ for Conflict versus Single. In the paraphrased-agreement control run, the corresponding tests give $p = 2.7 \times 10^{-4}$ for Conflict versus Paraphrased Agreement and $p = 7.6 \times 10^{-6}$ for Conflict versus Single. Agreement and Single do not significantly differ under the same test. Figures 7 and 8 visualize the all-matched and closed-book-unknown distributions in the main text.

Neutral-subset audit. To reduce the simplest context-versus-memory explanation, we run the curated items closed-book and manually audit whether the model's answer aligns with d_1, d_2, neither, or no clear answer. On the 55-row neither-aligned subset, the conflict gap strengthens, with paired sign tests $p = 9.0 \times 10^{-6}$ for Conflict versus Agreement and $p = 1.0 \times 10^{-7}$ for Conflict versus Single.

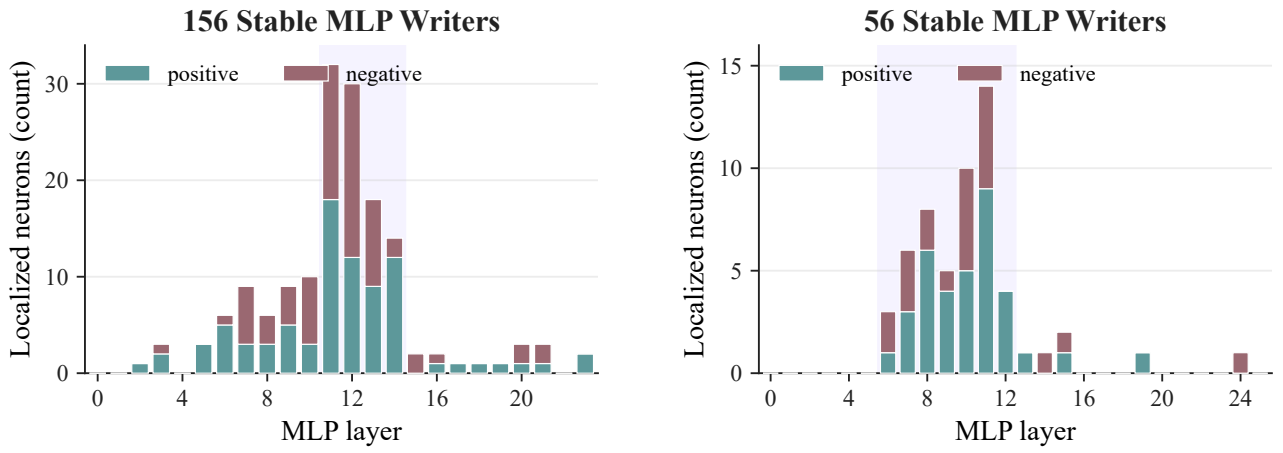
Position-bias and swap-order audit. Internally, the CETT readout is position-keyed. In the original order, mean signed projection is -1.294 for d_1-position answers and $+5.014$ for d_2-position answers. After swap, the corresponding means are -0.604 and $+4.054$. Behaviorally, position is not a universal override. The rate of following d_2 is 0.50 in the original order and 0.52 in the swapped order.

H. CETT Write-Alignment Diagnostics

To test whether the localized MLP neurons write into the residual source-selection coordinate, we project their actual upstream MLP writes onto the layer-13 doc-choice direction. For neuron j in layer L, the MLP write is $z_{L,j} W^{(L)}_{\mathrm{down},:,j}$, and its signed contribution to the doc-choice axis is

$$z_{L,j} \, \langle W^{(L)}_{\mathrm{down},:,j}, \hat{v}_{\mathrm{doc}} \rangle.$$

Equivalently, the implementation computes the per-neuron coefficients $W^{\top}_{\mathrm{down}} \hat{v}_{\mathrm{doc}}$, averages neuron activations over decision tokens, and sums activation times coefficient over the localized neurons. On the held-out doc-choice rows ($n = 592$ decided original/swap rows), the all-raw upstream write projection separates doc2 from doc1 answers with AUROC 0.934 and a doc2-minus-doc1 mean projection gap of 0.0365. This is a write-projection test, not only an activation-correlation test.



(a) Qwen 0.8B *(b)* Gemma 3 1B

Figure 14. Layer distribution and sign of CETT-localized MLP writers. These panels are the histogram components removed from the main sparse-write bridge figure so the main text can focus on the write-projection evidence.

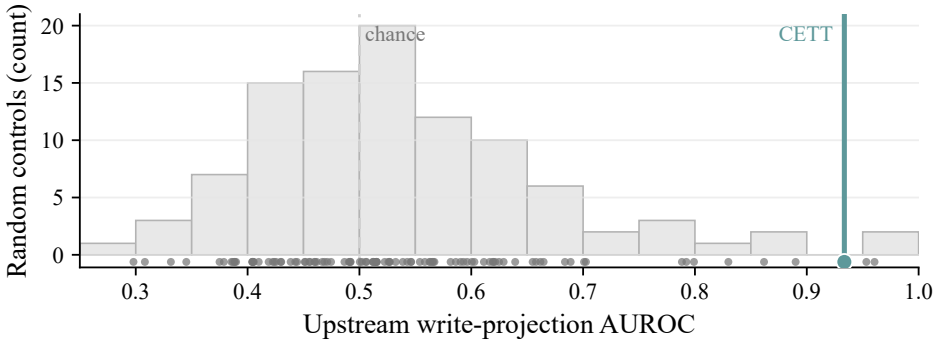


Figure 15. Layer-matched random-control distribution for the upstream write-projection test. The random controls are mostly near chance (mean AUROC 0.536, median 0.515). Two of 100 random controls match or exceed the CETT set on AUROC, but none match its projection gap.

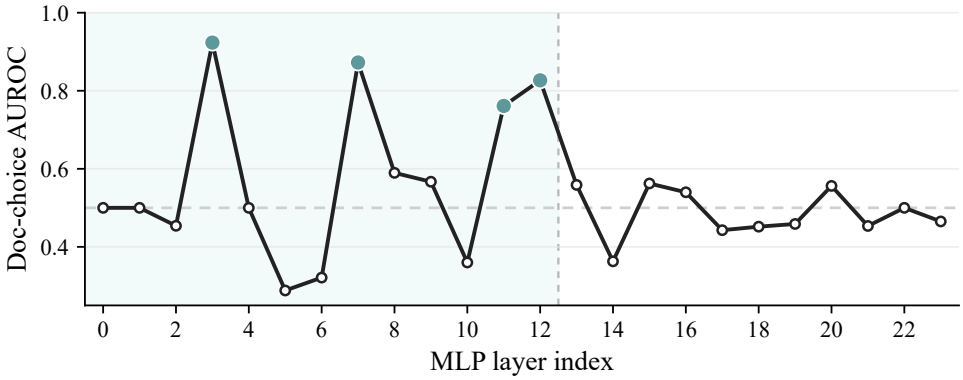


Figure 16. Layerwise breakdown of the same write-alignment test. The shaded region marks MLP layers upstream of the layer-13 residual readout. Strong upstream layers appear at layers 3, 7, 11, and 12 with AUROCs 0.923, 0.872, 0.761, and 0.827, respectively.

I. Timing and Attention Follow-up

For a timing analysis that does not use the CETT subspace, we use a logit-lens-style readout at every layer conditioned on forced first answer tokens. With $y_1^{(1)}$ and $y_2^{(1)}$ the first tokens of the canonical d_1 and d_2 answers, at layer ℓ and position t,

$$m_t^{(\ell)} = \log p^{(\ell)}\big(y_2^{(1)} \mid \text{prefix}_t\big) - \log p^{(\ell)}\big(y_1^{(1)} \mid \text{prefix}_t\big),$$

where $p^{(\ell)}$ is the logit-lens readout at layer ℓ (nostalgebraist, 2020; Belrose et al., 2023). Across 30 items, median onset is 4.5 for the token margin, 11.0 for the localized CETT delta, and 15.0 for an attention-only delta. This supports the interpretation that the CETT direction is a downstream commitment readout rather than the first site where the token preference exists.

A bounded attention-head patching sweep across swap-order-matched pairs finds that the layer-11 localized state is most effectively recovered by a single layer-11 head, L11H3, with mean localized recovery about 0.65 over 14 tested pairs. This is preliminary circuit evidence. It suggests that attention participates in routing source information into the downstream commitment state, but it does not establish a complete attention-mediated mechanism.

J. Residual Steering Details

For both doc-choice and surfacing, the steering rule is the same. We register a residual pre-hook at layer ℓ^* and, for each token t in a baseline-anchored steering window, add

$$\delta \cdot \mathrm{RMS}(x_t) \cdot \hat{w}$$

to the residual state. $\hat{w}$ is the unit-normalized contrastive residual direction. RMS scaling keeps the intervention size roughly comparable across tokens.

For doc-choice, the window is the last 32 prompt tokens plus $[\text{decision_start} - 4, \text{decision_start} + 2]$. The layer is 13, and magnitudes are $\delta \in \{-3, -2, +2, +3\}$. The 32-token prompt-tail window is a fixed script default recorded in the run summaries, not an independently ablated optimum.

For surfacing, the single-direction sweep covers the final prompt-tail token plus the first 12 generated tokens. The layer is 17, and magnitudes are $\delta \in \{+6, +8\}$. Controls are stat-matched random residual units at the same layer, orthogonalized against the learned direction. The multi-local surfacing follow-up adds a second residual direction at layer 23 in the same response-onset window. The reported best arm uses the layer-17 disclosure direction at $\delta = +10$ and the layer-23 anti-collapse direction at $\delta = +12$, yielding $58/141$ broad surfacing responses before the closed-book-unknown restriction and $46/109$ on the closed-book-unknown subset.

An earlier sparse-direction subtraction/scaling diagnostic left behavior close to baseline, so we treat CETT as a localization tool rather than the main intervention target.