Reverse Engineering a Stateful Reasoning Circuit

Akshit Kumar
IIIT Hyderabad
akshit.kumar@research.iiit.ac.in

Dipti Sharma
IIIT Hyderabad
dipti@iiit.ac.in

Parameswari Krishnamurthy
IIIT Hyderabad
param.krishna@iiit.ac.in

Abstract

We study Gemma-2-2B on a controlled role-gated retrieval task where a prepositional gate (`to` or `from`) selects which of two entities is correct. On 60 single-token name pairs the model attains 100% accuracy with a mean flip magnitude ≈ 3.5 (sum of per-condition correctness margins). Using causal tracing, we identify a *Query-Gated Courier* circuit with three stages: (1) a gate token (from/to) writes a role feature at the answer; (2) this feature perturbs late-layer courier queries, shifting their $q{\cdot}k$ preference; (3) couriers attend to the correct name and inject it via OV, raising its logit. Gate-residual swaps flip predictions, and a compact nine-head keep set reproduces the behavior with high fidelity. The circuit gives a potential algorithm for role tracking and aligns with the Paninian Kāraka analysis, mapping `to` to *sampradāna* and `from` to *apādāna*.

1 Introduction

Large language models such as Gemma-2-2B exhibit structured in-context behavior (Elhage et al., 2021; Olsson et al., 2022; Nanda et al., 2023). We probe this with a *role-gated retrieval* task: the model reads a short context containing two names and a preposition and must produce the correct name at the final position. Behaviorally, it achieves 100% accuracy on 60 single-token pairs with a large flip magnitude (≈ 3.5); see Fig. 1. These margins suggest stable control, motivating a mechanistic study.

Our goal is to explain how the model implements this behavior. We apply causal tracing (Meng et al., 2022) and activation patching (Wang et al., 2022) in a 26-layer decoder-only transformer, localizing control to the prepositional gate and identifying a small set of late attention heads that set the final logit. We then provide causal tests that isolate the medium of control and reconstruct the behavior with a compact subcircuit. The remainder of this paper details this discovery. We first present the behavioural setup and results, then introduce our full circuit hypothesis and the causal evidence that verifies it. We conclude by connecting our findings to the Paninian Kāraka framework (Begum et al., 2007; Bharati et al., 1995) and discussing the limitations of our work.

2 Reverse-Engineering the Role-Gated Retrieval Circuit

2.1 Behavioral Finding

Stimuli and metric. Two prompts:

P_{to}: "A moved the opal to B. Later, Owen took the opal from ___."

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

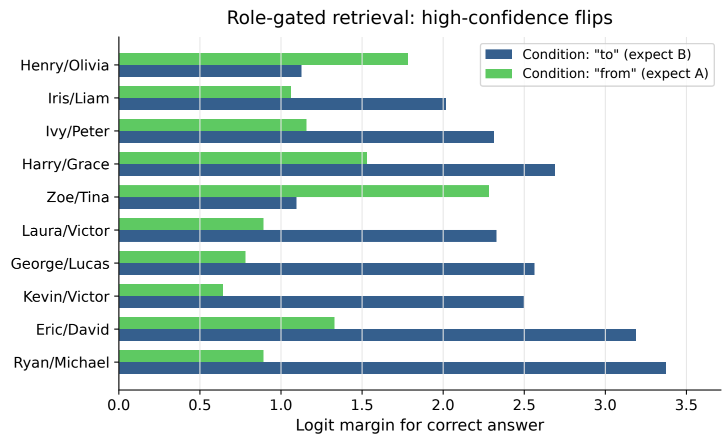

Figure 1: Behavioral result on random pairs: horizontal bars show correctness margins for `to` (expect B) and `from` (expect A).

P_{from}: "A moved the opal from B. Later, Owen took the opal from ___."

Let ℓ be logits at the answer position. Define *correctness margins*

$m_{\text{to}} := \ell_B - \ell_A$ (positive iff the expected B wins), $m_{\text{from}} := \ell_A - \ell_B$ (positive iff the expected A wins).

We summarize gate sensitivity by the *flip magnitude*

$$\Phi := m_{\text{to}} + m_{\text{from}}.$$

Result. Accuracy is 100% in both conditions on 60 pairs. Mean per-condition margins are $\mathbb{E}[m_{\text{to}}] \approx 2.190$ and $\mathbb{E}[m_{\text{from}}] \approx 1.314$. The mean flip magnitude is $\mathbb{E}[\Phi] \approx 3.504$ (Fig. 1).

2.2 Circuit Hypothesis: Query-Gated Couriers

We claim the model uses a three-stage algorithm. Gating: a prepositional gate token $t_{\text{gate}} \in \{\texttt{to}, \texttt{from}\}$ writes a directional role feature r_{gate} into the residual stream at the answer position ans. Steering: r_{gate} additively perturbs the query vectors of specific late-layer courier heads at ans. Retrieval: the perturbed queries steer those heads to attend to the correct name and inject its value via OV, which raises the target logit via the unembedding. Fig. 2 presents the full diagram. The remainder of this section provides causal evidence using activation patching and ablations.

2.3 Causal Evidence for the Circuit

We present the evidence in the computational order suggested by the circuit in Fig. 2 and the attention readouts in Fig. 3: *couriers are the final actors*, the *gate is the switch that controls them*, and the *query is the mechanism that connects the switch to the actors*. Metrics are those defined in Sec. 2.1. Behavioral context appears in Fig. 1.

Couriers are the final actors. Late heads L22H4 and L18H6 concentrate causal effect at the answer position. Two lines of evidence support this. (i) **Attention readout.** For a representative pair, each courier's attention at the answer flips to the correct entity under the gate condition: to selects B, from selects A (Fig. 3). This pinpoints where information is retrieved. (ii) **Local necessity.** Head–OV patching at the answer highlights late layers with peaks on these couriers, and zeroing their output at the answer reduces both per–condition correctness margins and the flip magnitude Φ. These findings place the final name injection on a small set of late heads, consistent with standard head-level decompositions.

The gate is the switch that controls them. Let L_c be the first layer where courier effects appear and let T be the answer index. We patch only the residual stream at the gate token, at (L_c, T), from

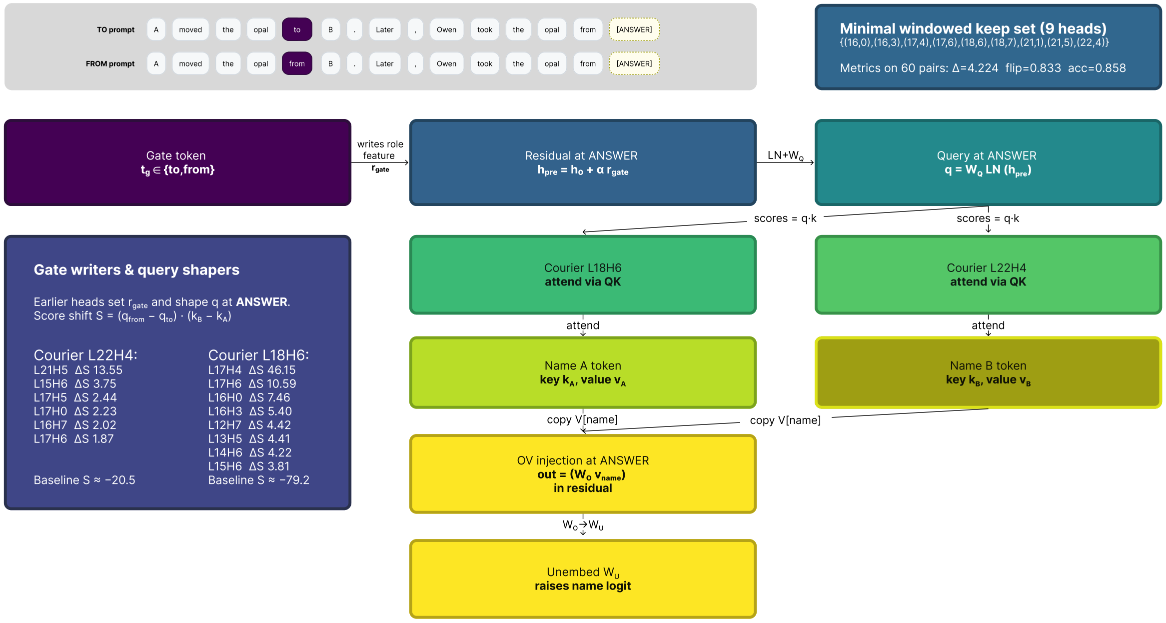

Figure 2: Query-Gated Courier circuit. Gate writes r_{gate} at ans, which steers courier heads (L22H4, L18H6). Couriers attend to the correct name and inject via OV; unembedding raises the correct logit.

the `from` run into the `to` run. This single-position swap flips the prediction in `to` and moves m_{to} toward the `from` value. We quantify control with a recovery fraction

$$\rho_{\text{gate}}(m) \;:=\; \frac{m(\text{to with gate swap}) - m(\text{to})}{m(\text{from}) - m(\text{to})}.$$

Empirically $\rho_{\text{gate}}(m_{\text{to}}) \approx 1$ and $\rho_{\text{gate}}(\Phi) \approx 1$, which localizes the switch to the gate token's residual stream. In the circuit diagram (Fig. 2), this is the edge that writes the role feature r_{gate} at ans.

The query is the mechanism that connects switch to actors. We test whether the gate-written signal influences queries, keys, or values of the couriers at the answer. We patch only the *courier queries* at T from the `from` run into the `to` run and observe large recovery of m_{to} and Φ. Patching keys at the name positions produces near-zero change. Patching values sourced at the gate is negligible. These results isolate a query-centric steering mechanism: the gate writes r_{gate} at the answer, r_{gate} perturbs the courier queries, and the perturbed q shifts the $q \cdot k$ score row toward the correct name. A first-order linearization that connects the gate residual to a query perturbation and then to the name–name score difference appears in Appendix B.3.

2.4 Paninian Interpretation

Minimal glossary, further discussed in Appendix A *Kāraka* = semantic role linked to the verb's action; *vibhakti* = overt marking that encodes a kāraka; *sampradāna* = recipient/beneficiary (≈ dative/'to'); *apādāna* = source/separation (≈ ablative/'from').

The preposition `to` selects *sampradāna* (recipient) and `from` selects *apādāna* (source). Mechanistically, the prepositional token functions as a vibhakti-like marker: it writes a role feature at the answer position that steers courier-head queries toward the argument consistent with the selected kāraka; OV then injects that name so the unembedding raises its logit (Figs. 1–2). This matches the Paninian view that surface vibhakti cues the semantic role realized by an NP. Whether this alignment generalizes across roles, languages and models remains an open question for future work.

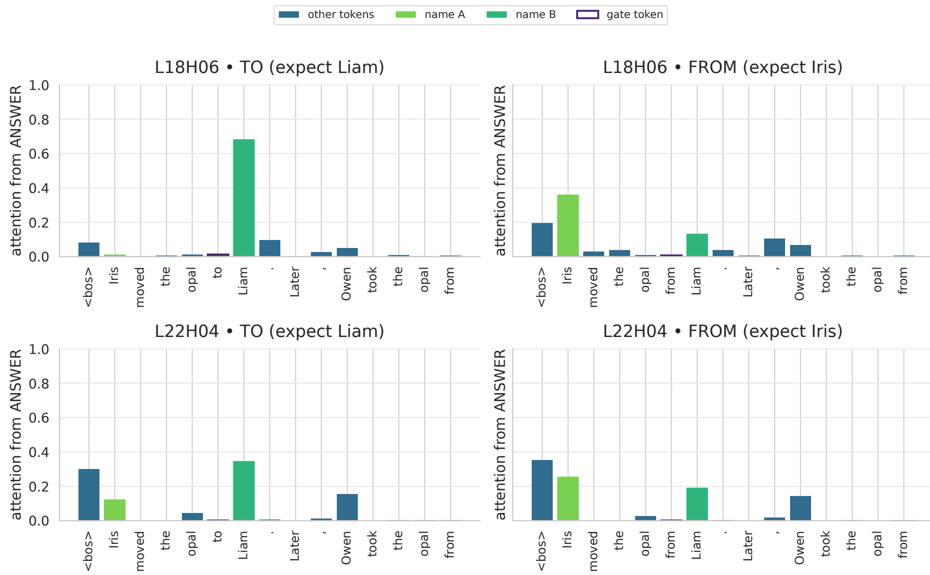

Figure 3: Attention at the answer for courier heads L18H6 and L22H4. Bars show attention from the final token to each source; A and B are colored, the gate token is outlined. Attention shifts to B in "to" and to A in "from," showing gate-controlled retrieval.

2.5 Sufficiency via a Compact Keep Set

We test sufficiency by restricting attention scores at the answer to a compact set of heads $\mathcal{K}$ from layer L_c onward using a scores–keep mask. With

$$\mathcal{K} = \{(16,0),(16,3),(17,4),(17,6),(18,6),(18,7),(21,1),(21,5),(22,4)\}$$

we retain high accuracy ($\approx 86\%$) and large Φ relative to the baseline run. A small set of late heads is therefore sufficient when upstream query formation is intact. This keep-only reconstruction corresponds to the highlighted courier path in Fig. 2.

Within this reconstruction, zero-ablating the primary courier heads (L22H4 or L18H6) individually was sufficient to reduce the flip magnitude (Φ), confirming their necessity to the reconstructed circuit function.

3 Limitations and Future Work

We present this as a case study in a simplified laboratory setting, and claims should be interpreted within that scope. The primary limitation is the small ($N = 60$) set of controlled prompts, which enabled clean causal analysis but may bias toward one circuit instantiation. We ask: is the "Query-Gated Courier" a tractable, general motif, or an artifact of this setup?

Further limitations include:

- **Scope and generalization:** Results are for a single model (Gemma-2-2B). Head identities may not transfer across scales or trainings. Behavior is sensitive to prompt structure; minor lexical changes can alter contributing heads.
- **Tokenization and morphology:** We use single-token English names; multi-token entities and inflected forms are untested.
- **Intervention bias:** Activation patching and scores-keep are interventions that can shift distributions; recovery fractions are under intervention, not true counterfactuals.
- **Unquantified components:** We focus on attention heads; MLP contributions and nonlinear effects are not fully quantified.

- **Metrics and reproducibility:** We report flip magnitude $\Phi = m_{\mathrm{to}} + m_{\mathrm{from}}$ (see Sec. 2.1), not average per-condition margins. Exact reproduction may require our seeds, pinned library versions, and cached activations due to floating-point variance.

Future work: (1) cross-model and cross-scale replication (layer timing, OV alignment, minimal keep-set size); (2) stress tests with multi-token entities, inflection, paraphrases, distractors, and long-context ledgers; (3) stronger identification with two-copy interventions and norm-matched control patches; (4) extend to additional Kārakas and multilingual settings.

4 Conclusion

We reverse-engineer a *Query-Gated Courier* circuit for role-gated retrieval: a prepositional gate writes a role feature at the answer, this feature steers late courier queries, and couriers attend to and inject the correct name via OV so the unembedding raises its logit. Behaviorally the model reaches 100% accuracy on 60 pairs with mean $\Phi \approx 3.6$; mechanistically, gate-residual swaps flip predictions, query-only patches recover the effect while key and gate-value patches are negligible, and a nine-head scores-keep subcircuit suffices (Figs. 1, 2, 3). In this microdomain the circuit aligns with the Paninian Kāraka analysis (as described in 2.4); whether this alignment and the motif generalize remains open. The repository with code, seeds and pinned packages has been made available, more details in Appendix C.1.

Author Contributions

All authors were involved in initial formulation of the project. A.K. designed and conducted all experiments, performed the causal analysis and wrote the manuscript. D.S. and P.K. provided guidance on the linguistic framing and the final interpretation of the results.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful and constructive feedback, which will significantly help in future extension of this work. The mechanistic analysis was made possible by the open source TransformerLens (Nanda and Bloom, 2022) library. We also acknowledge Google for releasing the Gemma-2-2B (Google, 2025c) model under open terms.

References

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. Dependency annotation scheme for indian languages. In *Proceedings of the Linguistic Annotation Workshop (LAW '07)*, pages 149–156, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Akshar Bharati and Rajeev Sangal. Parsing free word order languages in the paninian framework. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 105–111, 1993. doi: 10.3115/981574.981589. URL `https://aclanthology.org/P93-1015`.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi, 1995.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Chris Olah, et al. A mathematical framework for transformer circuits. Transformer Circuits blog, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Google. Gemma terms of use. `https://ai.google.dev/gemma/terms`, 2025a. Last updated 2025-08-07.

Google. google/gemma-2-2b. Hugging Face, 2025b. URL `https://huggingface.co/google/gemma-2-2b`.

Google. Gemma 2 model card. `https://ai.google.dev/gemma/docs/core/model_card_2`, 2025c.

Subhash C. Kak. The paninian approach to natural language processing. *International Journal of Approximate Reasoning*, 1:117–130, 1987. ISSN 0888-613X. doi: 10.1016/0888-613X(87)90007-7. URL `https://www.sciencedirect.com/science/article/pii/0888613X87900077`.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf`.

Neel Nanda and Joseph Bloom. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2301.05217`. ICLR/technical report; unpublished manuscript / preprint.

Catherine Olsson, Nelson Elhage, Neel Nanda, Tom Henighan, Nicholas Joseph, Nova DasSarma, et al. In-context learning and induction heads, 2022. URL `https://arxiv.org/abs/2209.11895`. arXiv preprint arXiv:2209.11895.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

A Paninian roles, vibhakti, and our mapping

Kārakas (semantic roles). Pāṇini introduces six core kārakas that relate participants to the action denoted by the verb: *apādāna* 'that from which departure occurs', *sampradāna* 'recipient/beneficiary', *karaṇa* 'instrument', *adhikaraṇa* 'locus', *karman* 'object/goal', and *kartṛ* 'agent'. These are semantic notions and need not map one-to-one to morphological cases (Kak, 1987).

Vibhakti (case/postpositional marking). In the Paninian tradition, surface markers (*vibhakti*) encode the kāraka borne by a noun phrase; the mapping from kāraka to vibhakti is language-specific. For Indian languages, vibhakti comprises case suffixes and postpositions; in fixed-word-order languages like English, positional information can be treated as part of vibhakti for mapping purposes (Bharati and Sangal, 1993; Begum et al., 2007).

Sampradāna and `to`. *Sampradāna* marks the intended recipient/beneficiary of the action. In many languages it aligns with dative marking and is often expressed by a 'to/for' relation in English. Computational treatments adopt this role as a dependency label for recipients. (Kak, 1987)

Apādāna and `from`. *Apādāna* denotes the source or point of separation, canonically glossed as 'from'. Modern Paninian parsers use the apādāna label for ablative/source dependents.

B Name Inventory and Tokenization Constraints

Scope. This appendix lists the exact single-token names used to form the 60 evaluation pairs. Each name satisfies the tokenizer constraint that both the no-space and leading-space variants tokenize to a single id in Gemma-2-2B, and that the leading-space id is used for next-token prediction at the answer position.

B.1 Full name list (sorted)

- Ada
- Alan
- Alice
- Amy
- Anna
- Ava
- Ben
- Bella
- Carol
- Cindy
- Clara
- David
- Donna
- Ella
- Emily
- Emma
- Eric
- Ethan
- Eva
- Frank
- George
- Grace
- Hank
- Ivy
- Jack
- Jacob
- James
- Jane
- Jason
- John
- Julia
- Kevin
- Lara
- Laura
- Leah
- Leo
- Liam
- Linda
- Lisa
- Mark
- Megan
- Michael
- Nick
- Noah
- Nora
- Oscar
- Owen
- Paula
- Peter
- Rita
- Ryan
- Sara
- Sarah
- Sean
- Susan
- Tina
- Tom
- Vera
- Victor
- Zoe

Pairs used. The 60 evaluation pairs (A, B) were sampled without replacement from the above inventory subject to $A \neq B$ and single-token constraints. The exact list of pairs used in the figures is:

A	B	A	B	A	B
Liam	Rita	James	Susan	Peter	Sara
James	Ivy	Noah	Linda	Michael	Mark
Tom	Oscar	Nick	Tina	George	Paula
Kevin	Megan	Nick	Ava	John	Tina
Emma	Sara	Grace	Ryan	Peter	Ivy
Rita	David	Nick	Michael	George	Linda
Kevin	Sara	Amy	Leo	Anna	Eric
Bella	Frank	Carol	Oscar	Cindy	Jacob
Clara	Sean	Donna	Victor	Ella	Mark
Emily	Nora	Eva	Jack	Alan	Jane
Laura	Ryan	Leah	Peter	Leo	Grace
Lisa	Noah	Lara	James	Megan	David
Michael	Amy	Nora	Kevin	Oscar	Emma
Paula	John	Rita	Tom	Ryan	Susan
Sara	George	Sean	Bella	Susan	Alan
Tina	Eric	Tom	Laura	Vera	Nick
Victor	Anna	Zoe	Kevin	Ava	Peter
Ivy	Frank	John	Sara	Emma	David
Grace	Oscar	Kevin	Linda	Mark	Paula
Noah	Jane	Peter	Megan	Sara	Ryan

Reproducibility. The inventory above is sufficient to regenerate the pairs deterministically using the random seed documented in the code release C.1. Names not conforming to the single-token constraint were excluded during preprocessing.

B.2 Model and runtime details

All experiments use Gemma-2-2B in float16 inference with a 26-layer decoder-only architecture. We use a single GPU. We disable dropout. We prepend BOS for all prompts. We use the model's native tokenizer.

B.3 Mathematical details for Sect. 2.3

Gate-to-query linearization. Let $r_{\text{gate}} = h^{\text{from}}_{L_c,T} - h^{\text{to}}_{L_c,T}$. A first-order expansion of the courier query at the answer gives

$$\delta q_h \;:=\; q^{\text{from}}_{L_c,T,h} - q^{\text{to}}_{L_c,T,h} \;\approx\; W^Q_{L_c,h}\, J_{\text{LN}}(h_{L_c,T})\, r_{\text{gate}}.$$

The induced change in the score difference between the two name positions (i_B, i_A) is

$$\Delta s_h \;\approx\; \frac{\langle \delta q_h,\; k_{L_c,i_B,h} - k_{L_c,i_A,h}\rangle}{\sqrt{d_h}}.$$

Head-level decomposition and OV alignment. Let $u_X = W_U e_X$. At the answer index T,

$$m_{\text{to}} \;\approx\; \sum_{\ell,h} \langle W^O_{\ell,h} z^{\text{to}}_{\ell,T,h},\, u_B - u_A\rangle \;+\; \text{MLP terms}, \quad m_{\text{from}} \;\approx\; \sum_{\ell,h} \langle W^O_{\ell,h} z^{\text{from}}_{\ell,T,h},\, u_A - u_B\rangle.$$

Define the per–head flip contribution $\phi_{\ell,h}$ by summing these two inner products, so that $\Phi \approx \sum_{\ell,h} \phi_{\ell,h}$. OV alignment at a name position i^* is

$$\alpha_{\ell,h} \;:=\; \langle W^O_{\ell,h} v_{\ell,i^*,h},\, u_{\text{correct}} - u_{\text{other}}\rangle.$$

Scores–keep mask. Let $S_{\ell,h} \in \mathbb{R}^{T\times T}$ be attention scores. The answer-row keep mask is

$$\tilde{S}_{\ell,h}[T,:] = \begin{cases} S_{\ell,h}[T,:], & (\ell,h) \in \mathcal{K}, \\ -10^9\mathbf{1}, & \text{otherwise}, \end{cases} \qquad \ell \ge L_c.$$

Define Φ_{keep} under this intervention and $\eta = \mathbb{E}[\Phi_{\text{keep}}]/\mathbb{E}[\Phi_{\text{base}}]$. Per–head importance within $\mathcal{K}$ is estimated by bootstrapping pairs.

C Compute Resources and Reproducibility

Hardware. All experiments ran on a single GPU: NVIDIA GeForce RTX 2080 Ti (11 GB VRAM). No multi-GPU, no distributed training, no gradient updates. Inference-only.

Software. Python 3.11, PyTorch 2.7.1 with CUDA 12.8, TransformerLens (commit hash provided in the code release), NumPy, Matplotlib, Seaborn. Default inference in fp16. Analysis steps that require exact dot products cast tensors to fp32. uv.lock file provided contains the exact versions of packages which can be used to reproduce the results.

Model and I/O. Gemma-2-2B decoder-only transformer. Batch size 1. Sequence length per prompt ≤ 32 tokens. Name inventory and random seeds are included in the release to reproduce the exact pairs.

Per-experiment resources. Table 1 reports approximate wall-clock times on the 2080 Ti, peak GPU memory, and input sizes. Times vary with driver and library versions. Each entry is the median of three runs after warmup.

Total compute. End-to-end reproduction of all figures and tables requires ≈ 15 to ≈ 25 minutes of GPU time on the 2080 Ti. Peak disk use for activation caches and figures is ≤ 1 GB.

Experiment	Input size	Wall clock
Behavioral evaluation (60 pairs, 2 prompts)	120 prompts	≈ 0.5 min
Attention patterns for L18H6, L22H4	2 heads $\times$ 2 prompts	≈ 0.5 min
Head OV patch heatmap at ANSWER	26 layers $\times$ 8 heads	≈ 4 min
Gate residual swap at gate token	60 pairs	≈ 0.5 min
Query-only patches for couriers	2 heads $\times$ 60 pairs	≈ 2 min
Scores-keep reconstruction (keep set of 9)	60 pairs	≈ 2 min
Greedy keep-set search windowed	up to 60 steps	≈ 3 min
OV alignment and per-head sums	2 heads	≈ 1 min

Table 1: Approximate compute on a single RTX 2080 Ti. No caching across rows except where noted in the release scripts.

Disclosure. Preliminary and failed variants, including prompt ablations and alternative ranking heuristics for keep sets, consumed an additional ≈ 3 to ≈ 5 GPU hours on the same hardware.

Determinism. We provide seeds for name sampling and pair ordering, pinned package versions, and optional precomputed caches. Floating-point nondeterminism can cause small changes in margins and head rankings. Reported metrics aggregate over the fixed 60 pairs.

C.1 Reproducibility

The repo can be accessed here: `https://github.com/komikat/prep-gated-circuits`. The *README.md* has instructions on how to reproduce the results mentioned in this paper.

D Gemma License

Gemma-2-2B is released by Google under the Gemma Terms of Use ("Gemma License"). The license permits use, reproduction, modification, and distribution of the weights and model derivatives, provided distributors include the Gemma Terms and the Prohibited Use Policy in downstream terms and attach a notice file; "Model Derivatives" include modified weights and models created using patterns or outputs of Gemma (e.g., distillation or synthetic-data training). Outputs themselves are not claimed by Google. Use is subject to the Prohibited Use Policy, and Google may restrict usage that violates the terms. Access via platforms such as Hugging Face requires explicit acceptance of the Gemma License. Google (2025a,c,b)

E NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our contributions: the discovery and causal analysis of the 'Query-Gated Courier' circuit. The scope is explicitly defined as a case study on a specific model and a controlled dataset, which is reflected throughout the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations and Future Work" (3) section that details the narrow scope of our study (single model, small N) and the methodological limitations of our causal analysis.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: This paper is primarily empirical. All mathematical formalizations of our methodology and metrics are accompanied by their derivations in Appendix B.3.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper fully describes our methodology, prompt structure, and the specific model used (Gemma-2-2B). The full list of 60 name pairs used as data is provided in the appendix (B), making the core experiments fully reproducible.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide an link to a public code repository in the appendix C.1. The repository contains a self-contained Jupyter notebook to reproduce all figures and key results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Yes. Our work involves inference on a pretrained, open-weights model (Gemma-2-2B), so no training was performed. We specify the model used and provide the full dataset of prompts in the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Main results report per-pair values and aggregate means (100% accuracy on 60 pairs; mean flip magnitude) without error bars or confidence intervals. Causal metrics (recovery fractions, ablation drops, keep-set fidelity) are shown without variability quantification. Factors of variability (name-pair sampling, seed effects) are not explicitly modeled; no method for error-bar computation is described.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide a section in the appendix C detailing the compute resources (GPU type, memory) used and the approximate execution time required to run our reproduction notebook.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes, the research conforms to the NeurIPS Code of Ethics. This work is foundational research aimed at increasing the scientific understanding of AI models. It does not involve human subjects or sensitive data, and we have made a best effort to be transparent about our methodology and its limitations.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: Our work is foundational mechanistic interpretability research aimed at increasing the scientific understanding and transparency of existing AI models. It does not introduce a new model or deployable application. The primary positive societal impact is the contribution to a scientific foundation that may enable safer, more reliable AI. Our specific findings on an internal reasoning circuit do not present a direct or foreseeable risk of misuse.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: Our work analyzes an existing, publicly available language model and does not release a new model or dataset that poses a high risk for misuse. The findings are of a scientific nature and do not present a direct avenue for malicious application. Therefore, safeguards for a new asset release are not applicable.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: Yes. The primary asset used is the Gemma-2-2B model, which is cited appropriately (D). The model's license and terms of use are available in appendix (D). Our code is built on open-source TransformersLens library, which has been credited in acknowledgements (4).

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We are releasing the code to reproduce all experiments. The code is provided as a self-contained Jupyter notebook with comments and instructions. The data (the list of 60 name pairs, B) is included directly in the code, and the code is released under an MIT License. C

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [NA]

Justification: This research does not involve crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The core methodology of this paper is the causal analysis of a large language model (Gemma-2-2B). The LLM is the object of study, not a tool used in the research process itself.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.