Relation Before Entity: Deferred Commitment in Language Model Factual Recall

Divyansh Agarwal [1]

Abstract

We ask whether relation-type information (e.g., *capital-of*) and entity-specific information (e.g., *France→Paris*) become causally active at the final-token position at the same depth during recall. Using four complementary causal diagnostics across four decoder-only models and eight prompt families, we find a robust temporal asymmetry: **relation information becomes generation-controlling before entity information does**. Relation onset precedes entity onset by 10–16 tested layers (31–44% of network depth) at threshold 0.4, with the ordering holding across all 16 model-threshold combinations for thresholds 0.2–0.5. Critically, entity information is *not* absent early: entity-token patching succeeds at 90–100% in early layers. Instead, entity commitment to generation is **deferred**: entity information is available at the entity-token position but becomes generation-controlling at the final token only after being routed there.

1. Introduction

When a model completes "*The capital of France is* _", does it retrieve *Paris* directly, or does it first settle the relation *capital-of* at the generation position before committing to the entity-specific answer? Prior work has studied *where* factual associations are stored (Meng et al., 2022a;b) and how subject-token representations are enriched with attributes (Geva et al., 2023). Geva et al. (2023) identify a three-stage information flow in factual recall and observe that relation and subject information reach the final token at different points, but do not quantify a layer-wise onset gap across architectures, use direct causal transfer measurement, or separate early entity availability from later final-token commitment. We study the complementary *temporal* question with direct causal measurement: when do relation and entity signals each become generation-controlling at the final token?

We use *entity* broadly to mean the input item supplied to the relation: a country in factual prompts, a verb in tense prompts, a noun in plural prompts, an adjective in lexical prompts, or an element in symbolic prompts. We refer to the prompt position containing this item as the entity-token position.

Core claim. Recall in these controlled prompt families is **temporally factorized**. Relation information becomes causally active at the final-token position in middle layers, whereas entity commitment is deferred to late layers. This staging holds across four architectures, eight prompt families spanning factual, morphological, lexical, and symbolic transformations, and thresholds 0.2–0.5.

Why this matters. If relation and entity information become generation-controlling at different depths, then (1) layer-targeted interventions such as model editing or steering may affect relation-type and answer-identity computation differently, (2) monitoring methods should distinguish information *availability* from causal *commitment*, and (3) similar staging may also matter for chain-of-thought reasoning and multi-hop settings.

Contributions.

1. Transfer-curve patching shows relation onset precedes entity onset by 10–16 tested layers at the primary threshold 0.4, with the ordering robust across thresholds 0.2–0.5 (16/16 model-threshold combinations; §3).
2. Both-change competition pits relation and entity signals directly, showing dominance transitions at the predicted layers (§4).
3. Entity-token patching shows entity information is *available* early at the entity-token position but *committed* late at the final-token position, ruling out the alternative that entity information is simply absent (§5).
4. Steering with relation and entity directions supports the asymmetry via a complementary intervention (§6).

[1]School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia. Correspondence to: Divyansh Agarwal <d.agarwal@student.uq.edu.au>.

Mechanistic Interpretability Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea, 2026.

2. Setup

Models. Llama-3.2-3B (28 layers), Llama-3-8B (32L), Qwen2.5-3B (36L), Phi-2 (32L).

Prompt families. Eight fill-in-the-blank families: *capital*, *language*, *past tense*, *present participle*, *plural*, *opposite*, *comparative*, *chemical symbol* (see Appendix E). Prompt items are programmatically defined; we audit greedy generation and first-answer-token statistics as a sanity check.

Terminology. *Relation*: the transformation or fact type being requested (capital-of, past-tense, opposite). *Entity*: the input item supplied to the relation and determining the answer (e.g., France, *walk*, *cat*, or gold).

Activation patching. At tested layer ℓ, we replace the donor prompt's final-token hidden state into the receiver's final-token position and measure whether the output switches to the donor's answer. We test every other layer.

3. Experiment 1: Transfer Curves and Onset

Design. **Relation transfer**: donor and receiver share entity, differ in relation. Receiver *capital(France)*, donor *language(France)*; success = *French*. Donor relation applied to receiver entity tests whether the final-token state is controlled by relation-type information. **Entity transfer**: same relation, different entity. Receiver *capital(France)*, donor *capital(Japan)*; success = *Tokyo*. Tests whether the final-token state is controlled by the specific answer information.

We report *pair-balanced* relation curves (equal weighting across six relation pairs) and *family-balanced* entity curves (equal weighting across eight families), preventing record-count weighting artifacts.

Onset is the first tested layer exceeding 0.4 for two consecutive tested layers, capturing stable causal commitment while avoiding single-layer spikes. Threshold sensitivity (0.2–0.5) is in Appendix C.

Results. Figure 1 shows relation transfer rising before entity transfer in every model: entity transfer remains near 0% for multiple layers while relation transfer already exceeds 0.4–0.8. Table 1 reports pair-balanced onset layers at threshold 0.4.

Ruling out answer copying. A *wrong-entity* control patches a different entity of the same donor family (e.g., donor *language(Japan)* into receiver *capital(France)*). This separates relation-only transfer from donor-answer copying: output *French* indicates the donor relation applied to the receiver entity, whereas output *Japanese* would indicate copying the donor's specific answer. At the peak relation-only transfer layer for each model, relation-only transfer reaches 0.79–0.86 while donor-answer copying remains $\leq$ 0.11; donor-answer copying rises only in later layers once entity commitment takes over. This indicates that mid-layer patches carry relational structure rather than simply copying specific donor answers. The full wrong-entity control table is provided in Appendix D.

Table 1. Pair-balanced onset at threshold 0.4. Relation onset precedes entity onset in all models. Ordering holds at all thresholds 0.2–0.5 (16/16).

Model	Rel. onset	Ent. onset	Gap	Depth
Llama-3.2-3B	$L6$	$L18$	12	43%
Llama-3-8B	$L10$	$L20$	10	31%
Qwen2.5-3B	$L16$	$L32$	16	44%
Phi-2	$L14$	$L24$	10	31%

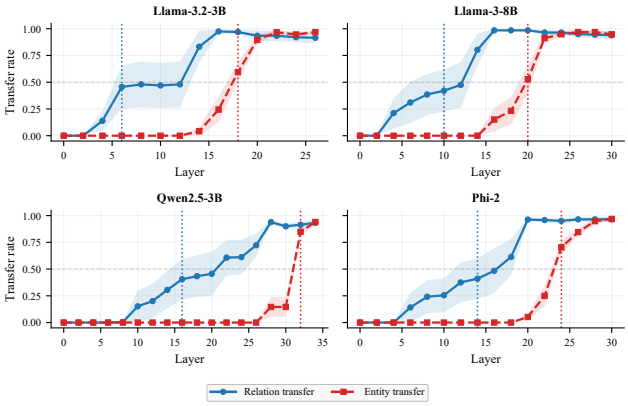

Figure 1. **Transfer curves.** Relation transfer (blue) rises before entity transfer (red) in every model. Entity transfer remains at or near 0% through multiple consecutive layers while relation transfer already exceeds 0.4–0.8. Shading = SEM across relation pairs or entity families. Dotted lines = onset layers.

4. Experiment 2: Both-Change Competition

Design. Donor and receiver differ in *both* relation and entity, forcing direct competition. Receiver *capital(France)*, donor *language(Japan)*. Outputs are classified as: *original retained* (Paris), *relation wins* (French—donor relation applied to receiver entity), *entity wins* (Japanese—the donor's answer), or *mixed*. To our knowledge, this provides a direct way to measure which signal dominates when relation and entity information are placed in causal conflict across depth.

Results. Relation wins dominate middle layers and entity wins dominate late layers across all four models (Figure 3; Appendix A). Table 2 shows that the crossover layer closely tracks the entity onset from Experiment 1: the layer where entity transfer becomes active is also where entity begins winning under direct competition.

Table 2. Both-change crossover and peak layers. Crossover aligns with Experiment 1 entity onset within ≤ 2 tested layers in every model.

Model	Rel. peak	Cross.	Ent. peak	Ent. onset
Llama-3.2-3B	$L14$	$L18$	$L26$	$L18$
Llama-3-8B	$L14$	$L18$	$L24$	$L20$
Qwen2.5-3B	$L26$	$L32$	$L34$	$L32$
Phi-2	$L20$	$L24$	$L28$	$L24$

Controls. Noise-patch controls (random Gaussian vectors) yield near-zero structured wins (≤ 0.008), ruling out generic perturbation as explanation. Self-patch controls confirm hook stability ($\geq 99.2\%$ original retained). Alternate-donor controls show the same qualitative pattern, supporting robustness to donor choice. Unrelated-donor diagnostics reveal high late entity-like overwrite but relation-wins remain near zero (≤ 0.030), ruling out generic patching as explanation for mid-layer relation dominance. Full control diagnostics are provided in Appendix D.

5. Experiment 3: Entity-Token vs. Final-Token Patching

Motivation. A key alternative explanation is that entity information is simply absent until late layers, making the delay trivial. Experiment 3 directly tests this.

Design. For same-relation/different-entity pairs, we compare: (a) patching the donor's *final-token* state into the receiver's final-token position, and (b) patching the donor's *entity-token* state into the receiver's entity-token position (83 records, eight families). Here, the entity-token position is the prompt position containing the input entity, corresponding to the subject-token position in factual prompts. If entity information is present early at this position, entity-token patching should succeed early even when final-token patching fails.

Results. Figure 2 shows a clean double dissociation. Entity-token patching succeeds at 90–100% in early and middle layers while final-token patching remains $\approx 2.4\%$. In late layers the pattern reverses: final-token patching rises to 88–97% and entity-token patching collapses to 2–4%. Table 3 shows the crossover matches entity onset from Experiments 1–2 in every model.

The alternative explanation is ruled out: entity information is not absent early. It is available at the entity-token position from the earliest layers, but becomes generation-controlling at the final token only after being routed there. *Deferred entity commitment is delayed routing, not delayed knowledge.*

Table 3. Entity-token/final-token crossover versus Experiment 1 entity onset. Three complementary diagnostics converge on the same transition layers in every model.

Model	Switch layer	Ent. onset
Llama-3.2-3B	$L18$–$L20$	$L18$
Llama-3-8B	$L20$–$L22$	$L20$
Qwen2.5-3B	$L32$	$L32$
Phi-2	$L24$–$L26$	$L24$

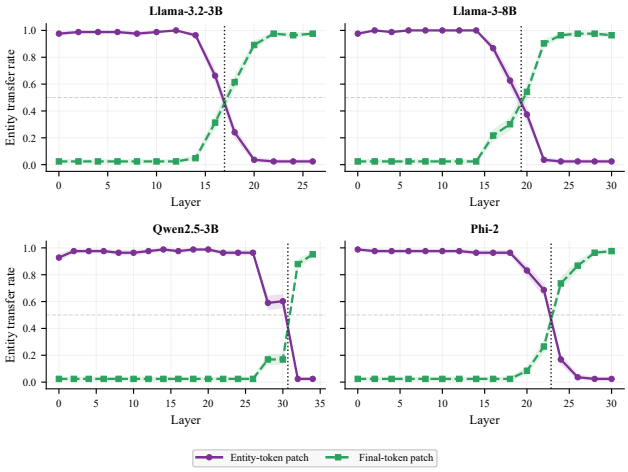

Figure 2. **Entity-token vs. final-token patching.** Entity-token patch (purple) transfers entity identity at 90–100% in early layers. Final-token patch (green) is near zero early. The pattern reverses late. Entity information exists early but reaches the final token only later. Dotted vertical line marks the switch layer.

6. Experiment 4: Steering Temporal Asymmetry

As an independent check, we construct relation steering directions (mean-difference vectors across same-entity/different-relation prompts) and entity steering directions (pair-specific same-relation/different-entity pairs), and apply them at mid and late layer zones (Appendix B). Random-direction baselines are ≤ 0.015, supporting direction-specificity. Relation directions steer effectively in middle layers; entity directions are substantially weaker mid-layer and strongest late. Relation directions also remain effective late but are outcompeted by entity commitment under direct conflict (Experiment 2); relation information persists but does not win.

7. Related Work

Factual recall and model editing. Meng et al. (2022a) use causal tracing to identify MLP computations at subject-token positions that mediate factual recall, enabling single-fact weight editing. Meng et al. (2022b) extend this line of work to mass editing by distributing updates across a

range of critical MLP layers. Our entity-token patching is consistent with this broad picture: entity information can be causally available at the token containing the input entity before it becomes generation-controlling at the final token. These model-editing results are plausibly related to the availability–commitment gap we measure, although we do not directly evaluate weight editing here.

Staged information flow in factual recall. Geva et al. (2023) identify a three-step mechanism for factual attribute extraction: subject enrichment at early MLP layers, relation propagation to the final token, and attribute extraction via attention heads. They find that relation information reaches the final token before subject information, using attention-edge interventions. Our work complements this result with direct causal transfer measurements across four architectures and eight prompt families. We introduce a both-change competition that places relation and entity signals in direct conflict, and an entity-token versus final-token patching analysis showing that entity information can be available early at the entity-token position while becoming generation-controlling at the final token only later.

Relation and task representations. Todd et al. (2024) show that attention heads transport compact, causally effective task representations (function vectors) that generalize beyond the demonstrating context. Hernandez et al. (2024) show that relation-decoding computations are well approximated by linear maps on subject representations. Popovič & Färber (2026) find that per-head attention contributions are comparatively strong features for linear relation classification. These works support the view that relation and task level information can be linearly represented inside transformer models. Our question is complementary: when does relation-level information become generation-controlling relative to entity-specific information at the final prediction position?

Causal intervention methods. Our methodology follows causal mediation analysis, activation patching, and transformer-circuits style analysis. Vig et al. (2020) emphasize interventions on internal model components to distinguish information that is merely decodable from information that causally mediates behavior. Heimersheim & Nanda (2024) offer a systematic treatment of activation patching methodology, including how patching experiments should be applied and interpreted. Wang et al. (2022) provide a detailed circuit-level analysis of indirect object identification in GPT-2 Small. In contrast, we do not claim to identify a complete circuit; instead, we use causal interventions to identify a cross-model, cross-family timing asymmetry in when relation and entity information become generation-controlling.

8. Discussion

A four-stage picture. Together, the results support a four-stage picture: *entity-token availability* $\rightarrow$ *relation becomes final-token active* $\rightarrow$ *entity becomes final-token committed* $\rightarrow$ *late overwrite sensitivity*.

Early layers contain entity information at the entity-token position (90–100% patching success), but this information is not yet generation-controlling at the final token ($\approx 2.4\%$ transfer). Middle layers make relation information final-token active; late layers route and commit entity information to the final token; and the latest layers become broadly overwrite-sensitive, with unrelated donors inducing late entity-like overwrite while unrelated relation-wins remain near zero (Appendix D).

Implications. The availability-commitment gap suggests that monitoring and intervention methods should distinguish *information being represented* from *information controlling generation*. A feature may be decodable at one position or layer without yet causally determining the next token. In practical terms, middle-layer interventions may preferentially affect which relation type is applied (e.g., redirecting from *capital-of* to *language-of*), whereas late-layer interventions may preferentially affect which specific answer is produced (e.g., changing *Paris* to *Tokyo*). This suggests that single depth monitoring tools may conflate relation-level and entity-level commitment, especially if they measure representational availability rather than causal control. For model editing methods such as ROME (Meng et al., 2022a), our results raise the hypothesis that edits at different depths may target different stages of recall: relation selection, entity routing, or final answer commitment. We do not test weight editing directly, but the observed transition zone provides a concrete target for future editing and steering analyses. Whether analogous staging appears in multi-hop reasoning, where models may need to resolve multiple relation-entity stages sequentially, remains important future work.

Limitations. Our experiments use controlled fill-in-the-blank prompt families and greedy first-answer generation; natural-language QA, longer contexts, and free-form reasoning remain untested. The evaluated models are open-weight decoder-only models in the 3B–8B range, so exact layer numbers should not be interpreted as universal. We identify a robust causal timing asymmetry, not a complete circuit: the mechanisms routing entity information from entity-token positions to final-token commitment remain to be localized. The central claim is the ordering (relation before entity commitment), not any single absolute layer index.

9. Conclusion

Recall in controlled prompt families is temporally factorized: relation information becomes generation-controlling at the final-token position before entity information does. At threshold 0.4, relation onset precedes entity onset by 10–16 tested layers (31–44% of network depth), with the ordering holding across all 16 model-threshold combinations for thresholds 0.2–0.5. Entity information is not absent early: it is available at the entity-token position from the earliest layers, but its commitment to generation is deferred until routed to the final token. Four complementary causal diagnostics converge on the same transition layers, providing evidence for **deferred entity commitment** in decoder-only transformers.

Impact Statement

This work aims to improve mechanistic understanding of recall in language models. Better understanding may support safety monitoring, steering, and model editing; we do not identify specific harms requiring further discussion.

Acknowledgements

The author thanks the School of Electrical Engineering and Computer Science at the University of Queensland for computational resources, and the anonymous reviewers for constructive feedback.

References

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Hernandez, E., Sen Sharma, A., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pp. 10504–10526, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Popovič, N. and Färber, M. Tracing relational knowledge recall in large language models. *arXiv preprint arXiv:2604.19934*, 2026.

Todd, E., Li, M., Sen Sharma, A., Mueller, A., Wallace, B., and Bau, D. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33: 12388–12401, 2020.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

A. Both-Change Competition Figure

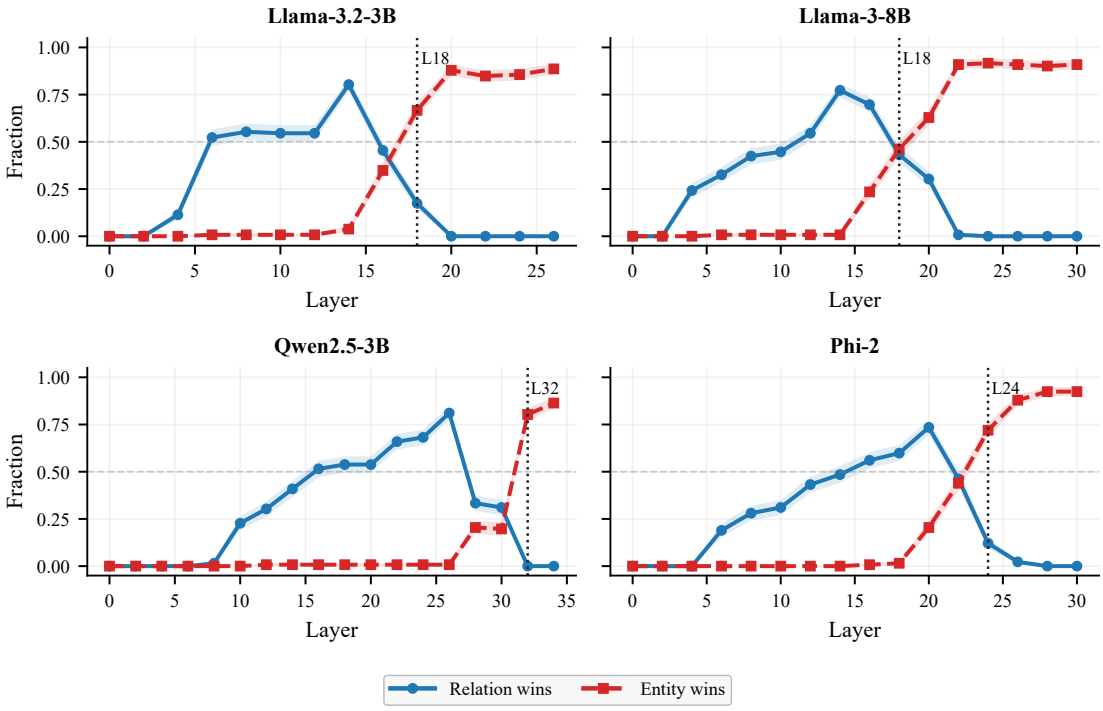

Figure 3. **Both-change competition (all four models).** Relation wins (blue) dominate middle layers; entity wins (red) dominate late layers. Dotted vertical lines = crossover layers, aligning with entity onset from Experiment 1.

B. Steering Figure and Table

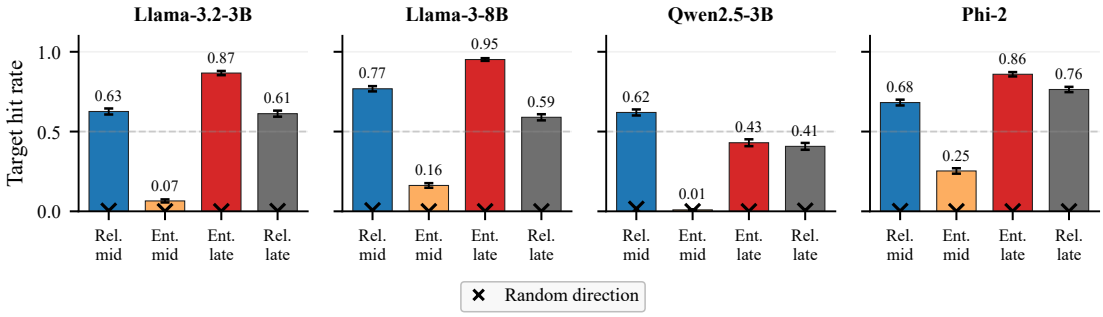

Figure 4. **Steering temporal asymmetry.** Relation directions (blue/orange bars) steer effectively in middle layers; entity directions (red bars) are substantially weaker mid-layer and strongest late. × marks near-zero random-direction baselines (≤ 0.015). Relation directions also remain effective late but are outcompeted by entity under direct conflict (Experiment 2).

Table 4. Steering hit rates at $\alpha = 1.0$. Random baselines ≤ 0.015. R = relation, E = entity. Relation steers mid-layer; entity steers late.

Model	R@mid	E@mid	E@late	R@late
Llama-3.2-3B	0.626	0.065	0.867	0.612
Llama-3-8B	0.768	0.162	0.952	0.589
Qwen2.5-3B	0.620	0.009	0.430	0.407
Phi-2	0.682	0.253	0.859	0.764

C. Threshold Sensitivity

Table 5. Pair-balanced onset across thresholds 0.2–0.5. Relation-before-entity holds in all 16 model-threshold combinations (✓).

Model	Thresh.	Rel. onset	Ent. onset	Gap	Rel.<Ent.
Llama-3.2-3B	0.2	L6	L16	10	√
Llama-3.2-3B	0.3	L6	L18	12	√
Llama-3.2-3B	0.4	L6	L18	12	√
Llama-3.2-3B	0.5	L14	L18	4	√
Llama-3-8B	0.2	L4	L18	14	√
Llama-3-8B	0.3	L6	L20	14	√
Llama-3-8B	0.4	L10	L20	10	√
Llama-3-8B	0.5	L14	L20	6	√
Qwen2.5-3B	0.2	L12	L32	20	√
Qwen2.5-3B	0.3	L14	L32	18	√
Qwen2.5-3B	0.4	L16	L32	16	√
Qwen2.5-3B	0.5	L22	L32	10	√
Phi-2	0.2	L8	L22	14	√
Phi-2	0.3	L12	L24	12	√
Phi-2	0.4	L14	L24	10	√
Phi-2	0.5	L18	L24	6	√

D. Additional Controls

Wrong-entity relation control. The wrong-entity control patches a different entity of the same donor family, distinguishing two possible outputs: *relation-only transfer* (donor relation applied to the recipient entity, e.g., output = *French*) versus *donor-answer copying* (donor's specific answer for the wrong entity).

Peak relation-only transfer is read from `relation_wrong_entity_summary.csv` by taking the layer with maximum `relation_only_transfer_pct` per model:

Table 6. Wrong-entity control: peak relation-only transfer layer per model. Rel.-only denotes the donor relation applied to the receiver entity; donor-copy denotes copying the donor's specific answer. Donor-copy is near zero at these layers and rises only later once entity commitment takes over.

Model	Peak layer	Rel.-only	Donor-copy	Exp. 2 rel. peak
Llama-3.2-3B	L14	0.818	0.038	L14
Llama-3-8B	L14	0.788	0.000	L14
Qwen2.5-3B	L26	0.811	0.000	L26
Phi-2	L20	0.864	0.114	L20

The peak wrong-entity layer matches the peak relation-wins layer from Experiment 2 in every model. This alignment provides convergent evidence that the wrong-entity control and both-change competition identify the same relation-dominant middle-layer regime. At these layers relation-only transfer reaches 0.79–0.86 while donor-answer copying remains ≤ 0.11; donor-answer copying rises substantially only in later layers once entity commitment takes over at the final token. This rules out donor-answer copying as an explanation for mid-layer relation transfer.

Both-change competition controls. Table 7 summarises the four control conditions for Experiment 2. Noise max is the maximum structured win rate (entity or relation) across all layers under random Gaussian patching. Self orig. is the mean original-retained rate under self-patch. Unrel. ent. is peak entity-like overwrite from unrelated donors, which is high in late layers when the final-token state becomes broadly overwrite-sensitive. Unrel. rel. is peak relation-wins from unrelated donors, which remains near zero throughout.

Table 7. Both-change control diagnostics. Noise patches produce near-zero structured wins; self-patches preserve the original output; unrelated donors show high late entity-like overwrite but near-zero relation-wins, indicating mid-layer relation dominance is not a generic patching artifact.

Model	Noise max	Self orig.	Unrel. ent.	Unrel. rel.
Llama-3.2-3B	0.000	1.000	0.970	0.015
Llama-3-8B	0.008	1.000	0.924	0.030
Qwen2.5-3B	0.008	0.992	0.917	0.008
Phi-2	0.008	0.992	0.932	0.008

Unrelated-donor overwrite diagnostic. Unrelated-donor patches produce high entity-like overwrite only in the late regime, marking broadly overwrite-sensitive final-token states. However, relation-wins from unrelated donors remain ≤ 0.030, ruling out generic patching as an explanation for mid-layer relation dominance.

E. Prompt Family Details

Table 8. Prompt families, templates, and record counts used in the controlled prompt banks.

Family	Template	N
capital	The capital of {X} is	33
language	The official language of {X} is	33
past tense	The past tense of {X} is	21
present part.	The present participle of {X} is	21
plural	The plural of {X} is	27
opposite	The opposite of {X} is	12
comparative	The comparative form of {X} is	12
symbol	The chemical symbol for {X} is	12

Prompt items are programmatically defined from fixed item banks. We audit greedy generation and first-answer-token statistics, including rank and logit margin, to identify ambiguous or unstable items and to document prompt quality.

F. Code and Reproducibility

Code and results available at https://github.com/divyanshddn146/deferred-entity-commitment