Dissociating Circuit-Level and Distribution-Level Effects of Knowledge Conflicts in LLMs

Anonymous ACL submission

Abstract

We investigate whether large language models resolve conflicts between parametric knowledge and contextual evidence by altering the parametric circuit's logit contribution or through competition at the output distribution. We apply a three-level analysis framework, decomposing conflict resolution into circuit-level (parametric-site logit contribution, Φ_P), pathway-level (MLP vs. attention), and output-level (token rank) effects across four transformer architectures spanning three attention types and two scale points. We find that Φ_P is near-invariant under conflict ($\leq$1.8% variation), while output token ranks shift substantially, establishing that conflicts resolve through additive logit-space competition rather than parametric suppression. Activation patching confirms parametric sites are causally sufficient and necessary, with contextual heads exerting 30–40$\times$ weaker influence. Layer-by-layer analysis reveals three architecturally determined resolution mechanisms, and parametric–contextual coupling decreases substantially with scale, exceeding random-subspace predictions. Domain generalization shows parametric sites are stratified by knowledge complexity with no overlap, yet the resolution mechanism generalizes across domains.

1 Introduction

Large language models store vast amounts of factual and entity-level knowledge in their parameters during pretraining (Petroni et al., 2019). At inference time, this *parametric knowledge* can conflict with *contextual evidence* provided in the input, for instance, when retrieved documents contain outdated information (Lewis et al., 2020), when dialogue context attributes behavior to the wrong entity, or when a user asserts a counterfactual (Longpre et al., 2021; Xie et al., 2024). Understanding how models resolve such conflicts is critical to the reliability of retrieval-augmented generation, multi-turn dialogue, and any application in which parametric and contextual signals may disagree.

Behavioral studies have established that models sometimes follow parametric knowledge and sometimes defer to context, depending on factors such as entity popularity, context length, and prompt phrasing (Longpre et al., 2021; Xie et al., 2024; Pham et al., 2024). The recent survey by Xu et al. (2024) extensively catalogs these behavioral patterns. However, behavioral evaluations cannot distinguish between two fundamentally different mechanistic accounts of where and how conflicts are resolved internally. Under *circuit-level suppression*, conflicting context modulates the residual stream representations flowing into parametric memory sites, reducing their activation-level contribution to the output. Under *distribution-level competition*, the parametric circuit produces the same contribution regardless of context, and the conflict is resolved through additive competition in the logit space before the softmax.

Recent mechanistic work has begun to address this gap. Ortu et al. (2024) traced the layer-by-layer dynamics of factual recall competing with in-context copying, identifying specialized attention heads that mediate the competition. Zhao et al. (2024) showed that the residual stream registers a detectable conflict signal in intermediate layers. Dotsinski et al. (2025) extended the competition-of-mechanisms analysis to LLaMA 3.1, finding reduced attention head specialization in grouped-query attention architectures. Yet what is missing is whether the *magnitude* of the parametric circuit's contribution changes under conflict.

In this work, we organize our analysis at three levels that directly measure parametric circuit contributions under controlled knowledge conflict:

- **Circuit-level:** We define Φ_P, the logit contribution of causally identified parametric sites toward the parametric prediction, and measure its variation across congruent, incongruent, anonymous,

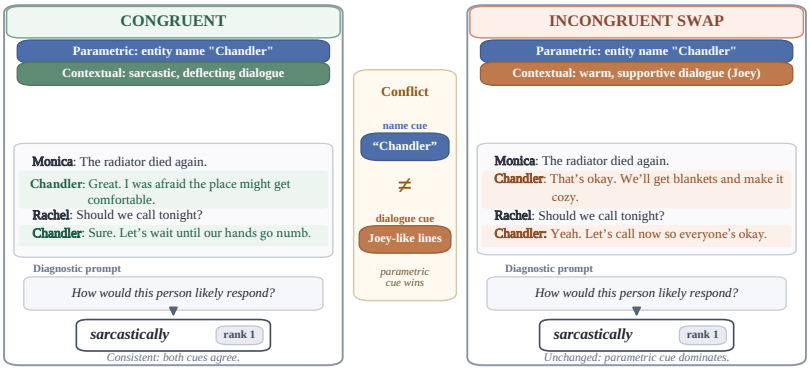

Figure 1: Simulated dialogue representative of the entity-swap stimuli from FriendsQA. In the congruent condition, Chandler's dialogue matches his parametric profile (sarcastic). In the incongruent swap, Joey's warm dialogue is attributed to Chandler, creating a conflict between the entity name (parametric cue) and dialogue content (contextual cue). The model predicts "sarcastically" (rank 1) in both cases, following parametric memory regardless of context.

and name-only conditions.

- **Pathway-level:** We decompose the total contribution into MLP and attention components (ΣMLP vs. ΣAttn), revealing which computational pathway carries parametric knowledge under each architecture.
- **Output-level:** We track token ranks and probabilities of the parametric and contextual predictions across conditions, capturing where the conflict becomes visible in the model's behavior.

We apply this framework to four models: Pythia-1.4B, Pythia-6.9B (Biderman et al., 2023), Mistral-7B (Jiang et al., 2023), and LLaMA-3-8B (Grattafiori et al., 2024), spanning full multi-head attention, grouped-query attention (GQA), and GQA with sliding window attention (SWA), using entity-swap stimuli derived from multi-party dialogue and counterfactual assertions from CounterFact (Meng et al., 2022).

We find that Φ_P is near-invariant under conflict ($\leq$1.8% variation), while output token ranks shift substantially, establishing that conflicts resolve through additive logit-space competition rather than circuit-level modulation (§4.2). Activation patching confirms the causal primacy of parametric sites (§4.3), and logit lens analysis reveals three architecturally determined resolution mechanisms (§4.4). Within the Pythia family, parametric–contextual coupling decreases 12-fold with scale (§4.2), and domain generalization experiments show that parametric sites are stratified by knowledge complexity, with zero overlap between atomic facts and complex entity associations (§4.5).

2 Related Work

Knowledge Conflicts in LLMs. Large language models encode factual and entity-level knowledge in their parameters during pretraining (Petroni et al., 2019), but this parametric knowledge can conflict with information provided in the input context. Longpre et al. (2021) first demonstrated that models exhibit mixed behavior under such conflicts, sometimes following parametric memory and sometimes deferring to context. Subsequent work established that the resolution depends on entity popularity, context length, and retrieval quality (Xie et al., 2024), and that even simple name-based ambiguities significantly degrade performance in retrieval-augmented settings (Pham et al., 2024). Xu et al. (2024) provide a comprehensive survey organizing these findings into a taxonomy of conflict types, behavioral outcomes, and mitigation strategies. These studies characterize the *behavioral* manifestation of knowledge conflicts but do not identify the internal mechanisms responsible for resolution.

Mechanistic Interpretability of Factual Knowledge. Geva et al. (2021) showed that feed-forward layers operate as key-value memories, and Meng et al. (2022) introduced causal tracing to localize factual associations in mid-layer MLPs. Geva et al. (2023) further dissected the recall process, identifying distinct roles for attention (subject enrichment) and MLPs (attribute extraction). The logit lens (nostalgebraist, 2020) and its calibrated variant (Belrose et al., 2023) enable tracking of intermediate predictions across layers. These methods provide the technical foundation for our

analysis but have been applied primarily to single-mechanism settings rather than to scenarios where two knowledge sources compete.

Mechanistic Studies of Knowledge Conflicts. The closest prior work to ours is Ortu et al. (2024), who studied the competition between factual recall and in-context copying using logit inspection and attention head scaling in GPT-2 and Pythia-6.9B. They traced which mechanism dominates at each layer but did not measure whether the factual recall circuit itself changes magnitude under conflict. Dotsinski et al. (2025) reproduced and extended this analysis to LLaMA 3.1 8B, finding that attention head specialization is substantially reduced in grouped-query attention architectures, though without explaining the architectural mechanism behind this reduction. Zhao et al. (2024) probed the residual stream under knowledge conflict and showed that LLMs can internally detect the presence of conflict, with a logistic regression probe achieving 90% detection accuracy from intermediate activations. Jin et al. (2024) proposed a method for mitigating conflicts by intervening on specific attention heads. Our work fills the gap left by these studies: we directly measure the parametric circuit's *activation-level contribution* (Φ_P) across conflict conditions, establish its near-invariance through both correlational and causal evidence, and explain the cross-architectural differences that prior work observed but could not account for.

3 Experimental Setup

3.1 Models

We study four autoregressive transformer language models spanning three attention architectures and two scale points (Table 1). Pythia-1.4B and Pythia-6.9B (Biderman et al., 2023) share the same architecture (full multi-head attention) and training data (The Pile; Gao et al. 2020), differing only in parameter count, enabling controlled analysis of scale effects. Mistral-7B (Jiang et al., 2023) uses grouped-query attention (GQA) with sliding window attention (SWA), restricting each layer's attention to a local context window. LLaMA-3-8B (Grattafiori et al., 2024) uses GQA with rotary position embeddings (Su et al., 2024) and was trained on approximately 15T tokens with extensive quality filtering. The cross-family comparison (Pythia-6.9B vs. Mistral-7B vs. LLaMA-3-8B) isolates architectural effects at comparable parameter counts. All models are loaded via TransformerLens (Nanda and Bloom, 2022) in half-precision.

	Pythia 1.4B	**Pythia 6.9B**	**Mistral 7B**	**LLaMA-3 8B**
Attention	Full MHA	Full MHA	GQA +SWA	GQA
d_{model}	2048	4096	4096	4096
Layers	24	32	32	32
Training	300B	300B	∼2T	∼15T

Table 1: Models studied. The Pythia pair enables controlled scaling analysis with identical architecture and training data. The cross-family comparison isolates attention architecture at matched scale.

3.2 Knowledge Conflict Stimuli

We construct conflict stimuli from two domains to test both complex entity associations and atomic relational facts.

Entity knowledge. We use the Friends TV series dataset, combining FriendsQA (Yang et al., 2019) for dialogue scenes and MELD (Poria et al., 2019) for emotion annotations. From 150 scenes containing at least two of six main characters, we construct four experimental conditions (Table 2). In the CONGRUENT condition, dialogue is attributed to its original speaker, so parametric and contextual signals agree. In INCONGRUENT_SWAP, character names are exchanged between two speakers within the same scene, creating a direct conflict between the entity name (parametric cue) and the dialogue style (contextual cue). The ANONYMOUS condition replaces all character names with generic labels ("Speaker A", "Speaker B"), removing the parametric trigger while preserving contextual information. The NAME_ONLY condition presents only the character name with no supporting dialogue, isolating pure parametric recall. Each stimulus includes a diagnostic prompt probing entity-associated attributes such as speech style.

Atomic factual knowledge. To test domain generalization, we use the CounterFact dataset (Meng et al., 2022), which provides subject-relation-object triples (e.g., "The Eiffel Tower is located in [Paris]") paired with counterfactual objects (e.g., "Rome"). We prepend a context asserting the counterfactual association and measure whether the model predicts the parametric (true) or contextual (counterfactual) completion. We process 50–100 items per model. As a negative control, we generate 25–50 synthetic dialogues with fictional entity

Condition	n	Description
CONGRUENT	450	Original attribution; parametric and contextual signals agree
INCONG. SWAP	2151	Names swapped; parametric and contextual signals conflict
ANONYMOUS	450	Names removed; contextual signal only
NAME_ONLY	18	Name only; parametric signal only

Table 2: Stimulus conditions for entity knowledge conflicts. INCONGRUENT_SWAP is the critical collision condition. Each of the 150 scenes appears across conditions, enabling paired comparisons.

names (e.g., "Zarvex Thornweave") for which no parametric knowledge exists, verifying that the observed effects require genuine parametric associations.

3.3 Analysis Framework

We analyze conflict resolution at three levels, each providing a distinct view of the model's internal computation.

Circuit-level: parametric-site logit contribution (Φ_P). Using causal tracing (Meng et al., 2022), we first identify the set $\mathcal{P}$ of layer-position pairs with significant causal effect on parametric predictions (§4.1). We then compute the logit contribution of these sites toward the parametric prediction y_{param}:

$$\Phi_P(y_{\text{param}}) = \sum_{(l,t)\in\mathcal{P}} \mathbf{W}_U[y_{\text{param}}] \cdot \mathbf{h}_{l,t} \quad (1)$$

where $\mathbf{W}_U$ is the unembedding matrix and $\mathbf{h}_{l,t}$ is the residual stream contribution from site (l, t). This measures how many logit units the parametric sites collectively contribute toward y_{param}, independently of the final output. Crucially, Φ_P captures *activation-level* variation: while model weights are fixed at inference, the activations flowing through parametric sites depend on the full input, including any conflicting context in the residual stream.

Pathway-level: MLP vs. attention decomposition. We decompose the total logit contribution across all layers into MLP and attention components:

$$\Sigma\text{MLP} = \textstyle\sum_{l=1}^{L} \mathbf{W}_U[y_{\text{param}}] \cdot \text{MLP}_l(\mathbf{x}) \quad (2)$$

$$\Sigma\text{Attn} = \textstyle\sum_{l=1}^{L} \mathbf{W}_U[y_{\text{param}}] \cdot \text{Attn}_l(\mathbf{x}) \quad (3)$$

The ratio ΣAttn / ΣMLP quantifies which computational pathway dominates parametric retrieval under each architecture and condition.

Output-level: token rank and probability. For each stimulus, we record the rank and probability of both the parametric prediction y_{param} and the contextual prediction y_{ctx} in the model's output distribution. A key analytical distinction is that Φ_P can remain constant while token ranks shift, which would indicate that the conflict resolves downstream of the parametric circuit through additive competition in the logit space.

4 Results

4.1 Parametric Circuit Identification

We apply causal tracing (Meng et al., 2022) to identify parametric memory sites $\mathcal{P}$ for entity knowledge (see Appendix A for full details). The procedure recovers the two-site pattern reported for factual triples, but localization varies across architectures: Pythia-6.9B shows a concentrated mid-layer hotspot (L12, IE = 1.45) consistent with MLPs as key-value memories (Geva et al., 2021), while LLaMA-3-8B shows the most diffuse pattern (peak IE = 1.03) with attention contributing more than MLPs. We select the top-10 IE sites per model as $\mathcal{P}$ for all subsequent analyses (Table 6 in Appendix A).

4.2 Parametric Near-Invariance Under Conflict

Having identified the parametric sites $\mathcal{P}$, we now ask the central question: does the parametric circuit's logit contribution change when contextual evidence contradicts parametric memory? We compute Φ_P (Eq. 1) for each stimulus across all four conditions and compare matched scene pairs between CONGRUENT and INCONGRUENT_SWAP (Table 3).

Model	Φ_P^{C}	Φ_P^{I}	$\Delta\%$	p (paired)
Pythia-1.4B	19.98	19.62	1.80	**.0005**
Pythia-6.9B	75.22	74.52	0.93	.708
Mistral-7B	1.61	1.61	0.00	.162
LLaMA-3-8B	11.82	11.84	0.17	**.019**

Table 3: Parametric-site logit contribution Φ_P (mean over 150 matched scene pairs) under congruent (C) and incongruent (I) conditions. $\Delta\%$: absolute percentage change. Paired Wilcoxon signed-rank test; bold indicates $p < .05$. Even where significant, the effect size is $\leq$1.8%.

Across all four architectures, Φ_P varies by at most 1.8% between congruent and incongruent conditions. Two models show significant paired

effects (Pythia-1.4B, $p = .0005$; LLaMA-3-8B, $p = .019$), but magnitudes are small (0.36 and 0.02 logit units on baselines of $\sim$20 and $\sim$12); the two larger models show no significant difference, suggesting residual coupling diminishes with scale. In the NAME_ONLY condition, Φ_P increases by 24–37% across models, driven by 1.2–2.0$\times$ MLP amplification at the same layers identified in §4.1, indicating that contextual input reduces parametric activation to some degree but not in a conflict-specific manner. Despite this circuit-level stability, output-level predictions shift substantially (Figure 2): in LLaMA, y_{param} maintains median rank 1 across all named conditions while y_{ctx} rises from rank 5 to rank 2 in ANONYMOUS; in Mistral, y_{param} drops to rank 2 under conflict. This establishes the central dissociation: the circuit-level contribution is near-constant, but the output distribution shifts through competing contextual signals accumulating in the residual stream.

This near-invariance is *scale-dependent*. Within the Pythia family, residual coupling decreases 12-fold from 1.4B to 6.9B (1.80% to 0.93% variation; $p < .001$ to $p = .708$). If parametric and contextual representations occupied random subspaces, doubling d_{model} (2048 to 4096) would predict only a $1/\sqrt{2} \approx 1.4\times$ reduction (Elhage et al., 2022); the observed $12\times$ reduction suggests that larger models actively learn to orthogonalize these subspaces. Decomposing the total logit contribution into MLP and attention components (Eqs. 2–3) further reveals that attention architecture determines the parametric retrieval pathway: $\Sigma\text{Attn}/\Sigma\text{MLP}$ decreases monotonically from Full MHA (2.01) to GQA (1.23) to GQA+SWA (0.29), as grouped-query and sliding window attention reduce per-head associative capacity and shift retrieval to MLPs (see Appendix C for full decomposition).

4.3 Causal Validation via Activation Patching

The near-invariance of Φ_P is correlational: logit attribution measures contributions but does not establish causal necessity. We validate the circuit-level findings with three activation patching interventions on the 150 matched INCONGRUENT_SWAP scenes.

In **parametric override**, we replace the activations at $\mathcal{P}$ with those from the corresponding CONGRUENT run, asking whether parametric sites are *sufficient* to shift the output toward the parametric prediction. In **parametric suppression**, we replace parametric-site activations with those from the ANONYMOUS condition, where no parametric signal is present, asking whether these sites are *necessary*. In **contextual suppression**, we zero-ablate the entity tracking heads identified in Appendix B, asking how much the contextual signal contributes causally.

The results reveal a stark asymmetry (Figure 3). Parametric override produces output distributions nearly identical to the CONGRUENT baseline (KL ≈ 0; Mistral achieves KL $= 5 \times 10^{-4}$), confirming that the parametric sites are sufficient to determine entity-specific predictions even when contextual evidence disagrees. Parametric suppression produces the largest distributional shift (KL $=$ 5.1–6.1 across models), confirming necessity. Contextual suppression, by contrast, produces 30–40$\times$ smaller effects than parametric suppression. In Mistral, contextual suppression does not reach significance ($p = .174$), consistent with its MLP-dominated architecture ($\Sigma\text{Attn}/\Sigma\text{MLP} = 0.29$): ablating attention heads removes a pathway that contributes minimally to the entity prediction.

The causal data converge with the correlational evidence: parametric sites are the primary drivers of entity-specific predictions, and their influence is robust to conflicting context. The contextual tracking heads maintain entity representations across turns (Appendix B) but exert weak causal influence on the diagnostic token, acting more as representational scaffolding than as direct competitors in the logit space.

4.4 Layer-by-Layer Conflict Resolution Dynamics

To understand *how* the competition between parametric and contextual signals unfolds across depth, we apply logit lens analysis (nostalgebraist, 2020) to the INCONGRUENT_SWAP stimuli. At each layer l, we project the residual stream through the unembedding matrix and record the logit-space gap $\Delta\ell(l) = \text{logit}(y_{\text{ctx}}, l) - \text{logit}(y_{\text{param}}, l)$. We use logit-space rather than probability-space because we find that GQA models exhibit poor logit lens calibration in probability space (see Appendix F), with non-monotonic oscillations in LLaMA between layers 18–29. Logit-space trajectories avoid this artifact and permit direct comparison across architectures.

The trajectories reveal three qualitatively distinct resolution mechanisms (Figure 4):

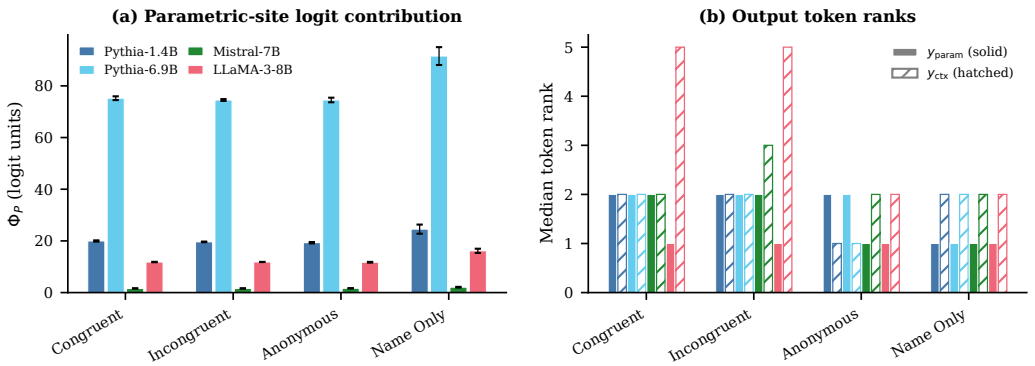

Figure 2: The circuit–distribution dissociation. **(a)** Φ_P by condition for each model; the parametric-site logit contribution is near-identical across CONGRUENT and INCONGRUENT_SWAP (max $\Delta = 1.8\%$), with a notable increase in NAME_ONLY (no contextual input). Error bars: ± 1 SE. **(b)** Median token rank of the parametric prediction (y_{param}) and contextual prediction (y_{ctx}) across conditions. Despite near-constant Φ_P, output ranks shift: y_{ctx} rises in ANONYMOUS (names removed) and drops in NAME_ONLY (names only).

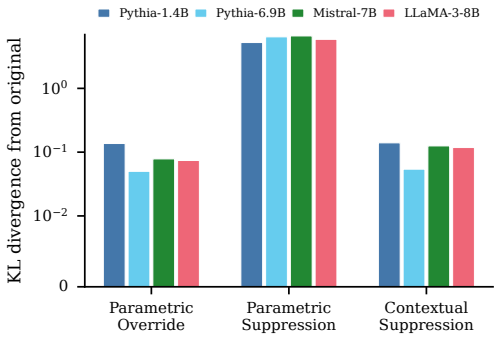

Figure 3: Causal intervention effect sizes. KL divergence from the original output distribution under three interventions across four models. Parametric suppression produces the largest shift; contextual suppression is 30–40× weaker. NS: not significant at $\alpha = .05$.

Gradual accumulation (Pythia, Full MHA). In both Pythia models, $\Delta\ell$ starts near zero and increases gradually across layers, reflecting a race condition in which parametric and contextual signals build up through the same pathway. The normalized switch point l^*/L is 0.40 for Pythia-1.4B and 0.66 for Pythia-6.9B, and context ultimately wins in 73–85% of stimuli. The compression ratio ρ is negative (-0.78 and -0.66 respectively), indicating a context-dominant resolution.

Substrate segregation (Mistral, GQA + SWA). Mistral shows an early and sharp crossover: the contextual signal emerges rapidly in the first third of the network (likely carried by the sliding window attention), then stabilizes as the MLP-stored parametric signal arrives in later layers. The normalized switch point is early ($l^*/L = 0.70$), but 70% of stimuli resolve through this early crossover pattern. The positive compression ratio ($\rho = +0.45$) reflects moderate parametric resistance.

Parametric entrenchment (LLaMA, GQA + large pretraining). LLaMA shows the most extreme pattern: $\Delta\ell$ remains near zero or negative (parametric-dominant) for most of the network, with a late contextual gain emerging only in the final layers. The normalized switch point is $l^*/L = 0.98$, meaning the conflict is not resolved until the penultimate layer. Despite this late gain, context wins in only 2% of entity stimuli (3 of 150). The compression ratio ($\rho = +0.32$) reflects strong parametric entrenchment consolidated through training on 15T tokens.

Pattern	P-1.4B	P-6.9B	Mist.	LLaMA
Ctx. override	85%	73%	29%	2%
Early crossover	5%	10%	70%	16%
Late ctx. gain	3%	7%	0%	77%
Param. dominant	7%	10%	1%	5%

Table 4: Distribution of resolution patterns across models for entity knowledge stimuli.

4.5 Domain Generalization

To test whether the findings above are specific to entity knowledge from multi-party dialogue, we apply causal tracing and logit lens analysis to atomic relational facts from CounterFact (Meng et al., 2022) (50–100 items per model) and to synthetic stimuli with fictional entity names (25–50 items per model as a negative control).

Parametric storage sites are stratified by knowledge complexity (Figure 5). Complex entity associations occupy late MLP layers (L20–31, cor-

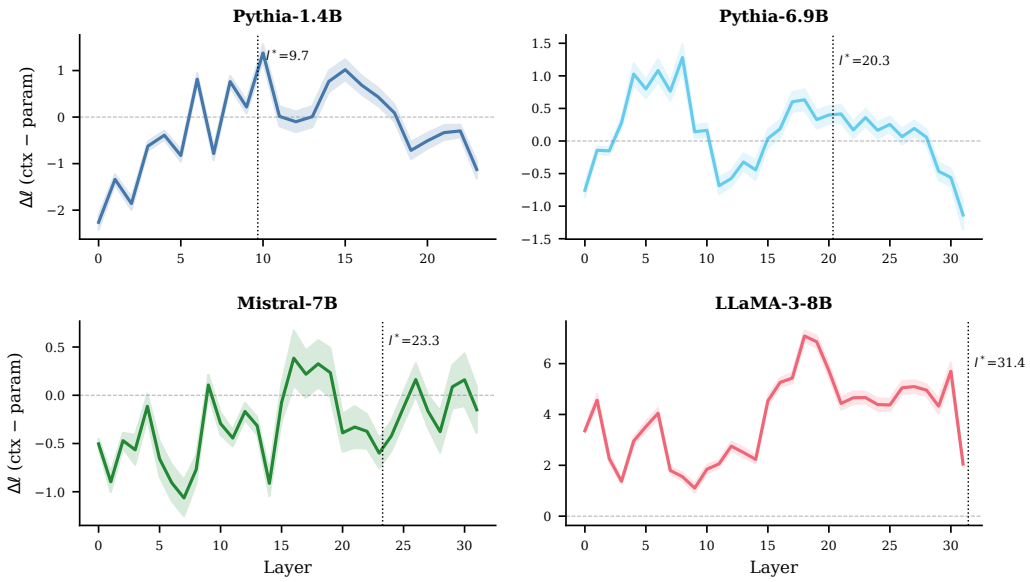

Figure 4: Layer-by-layer logit-space gap $\Delta\ell(l)$ for INCONGRUENT_SWAP stimuli. One panel per model; shaded bands indicate 95% CI across stimuli. Positive values indicate contextual dominance. The three resolution mechanisms are visible: gradual accumulation (Pythia), early crossover with subsequent parametric resistance (Mistral), and persistent parametric dominance with late contextual gain (LLaMA). Dashed lines mark the estimated switch point l^*.

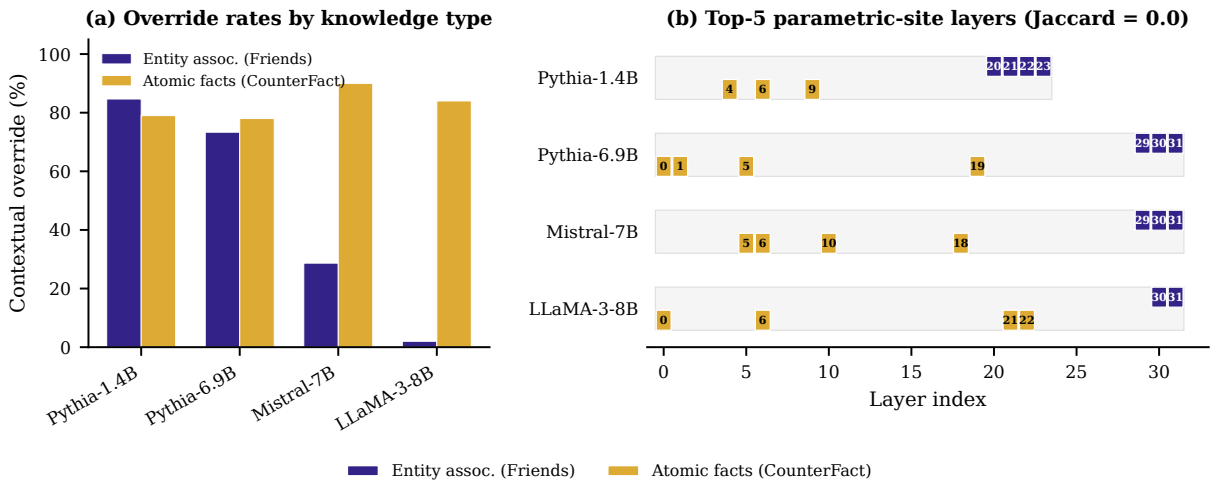

Figure 5: Domain generalization. **(a)** Contextual override rates for entity knowledge vs. atomic factual knowledge across models. Large models show a sharp dissociation: facts are easily overridden while entity associations resist. **(b)** Top-5 parametric-site layers for each domain. Entity and factual knowledge occupy non-overlapping depth bands (Jaccard = 0.0 in all four models), with facts stored early and entities stored late.

responding to 83–97% of network depth), while atomic relational facts occupy early-to-mid layers (L0–19, 0–59% depth). The top-5 parametric layers for entity and factual knowledge show zero overlap (Jaccard = 0.0) across all four architectures, indicating that the model maintains distinct circuits for different knowledge types despite sharing the same residual stream.

This depth stratification predicts an interaction between knowledge type and contextual override. Atomic facts, stored early, leave many subsequent layers through which contextual evidence can accumulate and override the parametric prediction. Complex entity associations, stored late, leave few layers for contextual competition. The data confirm this prediction:

For the Pythia models, override rates are similar across domains (∼78–85%), but for Mistral and LLaMA they diverge sharply: atomic facts are easily overridden (84–90%) while entity associa-

	P-1.4B	P-6.9B	Mist.	LLaMA
Entity override	85%	73%	29%	2%
Fact override	79%	78%	90%	84%
Entity layers	L20–23	L29–31	L29–31	L30–31
Fact layers	L4,6,9	L0,1,5	L5,6,10	L0,6,21

Table 5: Contextual override rates and top parametric-site layers for entity knowledge (Friends) vs. atomic factual knowledge (CounterFact). Override rate: proportion of stimuli where the contextual prediction outranks the parametric prediction.

tions resist (2–29%), with LLaMA showing the most extreme 82 percentage-point gap. Synthetic stimuli with fictional entity names (e.g., "Zarvex Thornweave") serve as a negative control, producing near-chance override rates (56–80%) with no concentrated parametric sites, confirming that the observed effects require genuine parametric associations.

5 Discussion

The additive competition model. Our results support a simple account of conflict resolution: the parametric circuit contributes a near-fixed $\Phi_P(y)$ for a given entity, and the output is determined by $P(y) = \text{softmax}(\Phi_P(y) + \Phi_C(y) + \Phi_{\text{res}}(y))$, where Φ_C varies by condition while Φ_P does not. There is no dedicated conflict-resolution circuit; the softmax selects the winner through arithmetic competition, consistent with the linear representation hypothesis (Park et al., 2024) and additive residual stream updates (Elhage et al., 2021). This explains the near-zero KL in parametric override: patching Φ_P from the congruent run changes nothing because Φ_P was already the same.

Implications for knowledge editing and RAG. Our depth stratification results suggest that ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), which target mid-layer MLPs, are well-calibrated for atomic facts but may miss entity-level knowledge stored in late layers (83–97% depth). The architecture-dependent retrieval pathway (§4.2) further implies that MLP-only edits may suffice for GQA+SWA models but not for full MHA models. For retrieval-augmented generation, the depth-after-injection principle provides a mechanistic explanation for why RAG is more effective at correcting simple facts than at overriding entity-level beliefs (Longpre et al., 2021; Xie et al., 2024): early-stored facts leave many layers for contextual override (78–90%), while late-stored entity knowledge leaves few (2–85%, architecture-dependent).

Implications for safety and trustworthiness. Models that stubbornly maintain entity-level beliefs despite contradicting context may perpetuate outdated or harmful associations (Blodgett et al., 2020). Our results show this risk is architecture-dependent: larger models consolidate entity knowledge more deeply (LLaMA: 2% override), making in-context correction insufficient, while atomic facts remain correctable (78–90% override). Safety-critical applications requiring faithful context adherence should account for the interaction between knowledge complexity and architectural design.

6 Conclusion

We applied a three-level analysis framework to study knowledge conflicts across four transformer architectures. Our central finding is a dissociation: the parametric circuit's logit contribution is near-invariant to conflict ($\leq$1.8%), while the output distribution shifts through additive logit-space competition. Activation patching confirms the causal primacy of parametric sites, logit lens analysis reveals three architecturally determined resolution mechanisms, and domain generalization shows parametric storage is stratified by knowledge complexity with zero layer overlap. The vulnerability of parametric knowledge to contextual override is predictable from storage depth and remaining layers for contextual accumulation, with implications for knowledge editing, RAG, and the safety of deployed systems. Future work should extend this analysis to instruction-tuned models and mixture-of-experts architectures.

7 Limitations

Our entity knowledge experiments are restricted to six entities from a single fictional domain (FriendsQA), and we study only base (pretrained) models; instruction tuning and RLHF may alter conflict resolution dynamics. We do not cover mixture-of-experts or linear attention architectures, which may exhibit qualitatively different storage patterns. The GQA models showed poor logit lens calibration in probability space (Appendix F), which we mitigate by reporting logit-space trajectories.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. In *Advances in Neural Information Processing Systems*, volume 36.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Asen Dotsinski and 1 others. 2025. On the generalizability of "competition of mechanisms: Tracing how language models handle facts and counterfactuals". *arXiv preprint arXiv:2506.22977*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Mor Geva, Avia Caciularu, Kevin Wang, and Yoav Goldberg. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramber, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

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

nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. Association for Computational Linguistics.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151, Miami, Florida, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Rongwu Xu and 1 others. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 396–406. Association for Computational Linguistics.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Analysing the residual stream of language models under knowledge conflicts. In *Foundation Model Interventions Workshop at NeurIPS 2024*. ArXiv:2410.16090.

A Causal Tracing Details

We apply the causal tracing methodology of Meng et al. (2022) to identify parametric memory sites for entity knowledge. For each model, we select up to 40 prompts where the model's expected answer achieves reciprocal rank ≥ 0.25, ensuring that only prompts for which the model has genuine parametric knowledge are included. We corrupt entity name token embeddings with Gaussian noise ($\sigma = 0.1$) and average over 10 noise samples per prompt. For each of six token positions (first, mid, and last subject token, pre-subject, post-subject, and last token) and each layer, we restore the clean activation and measure the indirect effect (IE) on the parametric prediction.

Table 6 summarizes the key characteristics across models. Pythia-6.9B exhibits the cleanest two-site pattern, with a sharp mid-layer hotspot at L12 (IE = 1.45) and early subject-token activity at L1–L4. LLaMA-3-8B shows the most diffuse pattern, with knowledge distributed across all layers and no single restoration recovering more than IE = 1.03. This distribution gradient does not simply scale with model size: Pythia-6.9B (6.9B parameters) is more concentrated than Mistral-7B (7B parameters), suggesting that training data diversity, not parameter count, is the primary determinant of knowledge localization.

For all subsequent experiments, we select the top-10 sites (by IE at the last token position) per model as the parametric site set $\mathcal{P}$. We additionally note that certain entities (Joey, Ross) produce negative IE values in mid-to-late layers across multiple models, indicating non-linear multi-layer processing pathways that single-site restoration cannot fully capture. This motivates the multi-site activation patching approach in §4.3.

	Pythia 1.4B	**Pythia 6.9B**	**Mistral 7B**	**LLaMA-3 8B**
Peak IE	1.00	**1.45**	1.00	1.03
Peak layer	L23	L12	L31	L29
Subject site	L5 (0.42)	L1 (1.15)	L3 (0.57)	L3 (0.54)
Pattern	Diffuse	Concentrated	Late diffuse	Very diffuse
Dominant	Mixed	MLP	MLP	**Attn**

Table 6: Parametric circuit identification via causal tracing. Peak IE: maximum indirect effect at the last token position. Subject site: highest-IE layer at subject token positions, with IE value in parentheses.

B Entity Tracking Heads

To identify attention heads that maintain entity identity across dialogue turns, we compute a Character

Tracking Score (CTS) for each head:

$$\mathrm{CTS}(h) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{\alpha_h^{\text{same}}(s)}{\alpha_h^{\text{same}}(s) + \alpha_h^{\text{other}}(s)} \quad (4)$$

where $\alpha_h^{\text{same}}(s)$ and $\alpha_h^{\text{other}}(s)$ are total attention weights from head h at scene s to same-speaker and other-speaker tokens respectively. Across all four models, the top-CTS heads are concentrated in the first 20–35% of network depth. CTS-ranked ablation degrades character identification probe accuracy by 22–37 percentage points, compared to 5–12 for random ablation, confirming that CTS identifies genuinely functional tracking heads. These tracking heads serve as the contextual circuit component in the activation patching experiments (§4.3), though their causal influence on the diagnostic token is 30–40× weaker than parametric sites.

We select the top 10 heads by CTS per model. Pythia-1.4B: 10 heads in layers 2–7; Pythia-6.9B: layers 3–11; Mistral-7B: layers 2–10; LLaMA-3-8B: layers 3–9.

C Architecture and Scale: Full Decomposition

We decompose the total logit contribution toward y_{param} into MLP and attention components using Eqs. 2–3. Figure 6 shows the decomposition across all four conditions and models.

The ratio ΣAttn/ΣMLP decreases monotonically with attention compression: Full MHA (2.01) > GQA (1.23) > GQA+SWA (0.29). In the NAME_ONLY condition, MLP contributions increase by 1.2–2.0× while attention contributions decrease, at the same layers identified by causal tracing (Appendix A), indicating gain modulation of existing circuits rather than recruitment of new ones.

D Parametric Near-Invariance: Paired Analysis

Figure 7 visualizes the near-invariance of Φ_P through paired scatter plots. For each of the 150 matched scenes, we plot Φ_P under the CONGRUENT condition against the mean Φ_P across the corresponding INCONGRUENT_SWAP stimuli.

All four models cluster tightly around the identity line. The two models with significant paired effects (Pythia-1.4B, LLaMA-3-8B) show slightly more scatter below the diagonal, but the maximum per-scene deviation is less than 2 logit units on baselines ranging from 1.6 to 75.

E Domain Generalization: Resolution Patterns

Figure 8 shows the full distribution of resolution patterns across entity associations (Friends), atomic facts (CounterFact), and synthetic controls.

Within the Pythia family, contextual override dominates for both knowledge types (≥73%). Mistral shows 70% early crossover for entities but 90% contextual override for facts, because early-stored facts are overridden before MLP-dominant late layers intervene. LLaMA exhibits 77% late contextual gain for entities but 84% contextual override for facts, confirming that the entity-fact inversion reflects qualitatively different processing dynamics at different depths. The synthetic control shows no late contextual gain pattern, confirming that late-layer dynamics require genuine parametric memory.

F Logit Lens Calibration

The logit lens (nostalgebraist, 2020) projects the residual stream at each layer through the unembedding matrix to obtain intermediate predictions. For full MHA models (Pythia), probability trajectories are well-calibrated, increasing monotonically from near-uniform to confident predictions by layer 3. GQA models exhibit qualitatively different behavior: LLaMA-3-8B shows non-monotonic oscillations in probability space between layers 18–29. These oscillations do not appear in logit space, suggesting an artifact of GQA interaction with the unembedding normalization. We therefore report logit-space trajectories throughout §4.4. This calibration issue does not affect Φ_P (Eq. 1), which is computed directly from residual stream activations, or the activation patching experiments (§4.3), which operate on the full forward pass.

G Full Statistical Tables

Table 7 reports Φ_P across all four conditions and models, including the ANONYMOUS and NAME_ONLY conditions omitted from the main text for space.

H Experimental Details

Compute. All experiments were conducted on a compute cluster using NVIDIA A100-80GB GPUs. Causal tracing and activation patching require `run_with_cache` in TransformerLens (Nanda and Bloom, 2022), consuming approximately 40GB

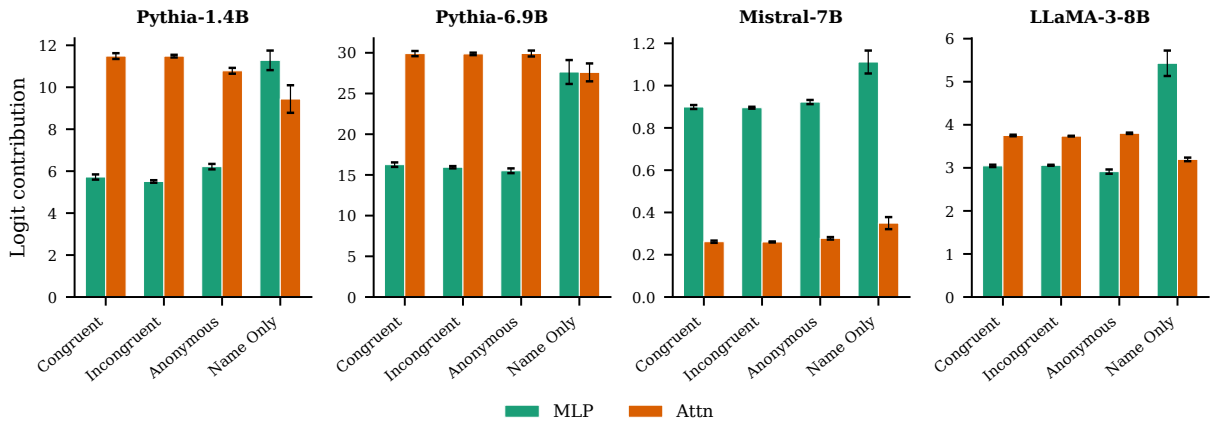

Figure 6: MLP vs. attention logit contribution toward y_{param} across conditions and models. Pythia models (Full MHA) show attention-dominant retrieval, while Mistral (GQA+SWA) is strongly MLP-dominant. In NAME_ONLY, MLP amplifies by 1.2–2.0×, indicating gain modulation of existing circuits.

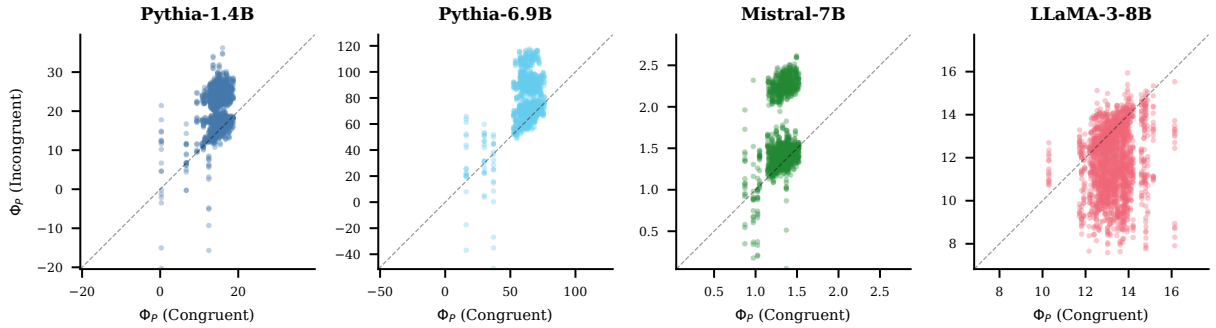

Figure 7: Paired Φ_P values for matched scenes under CONGRUENT (x-axis) and INCONGRUENT_SWAP (y-axis). Points near the identity line (dashed) indicate near-invariance. All four models show tight clustering around the diagonal, with larger models showing less deviation.

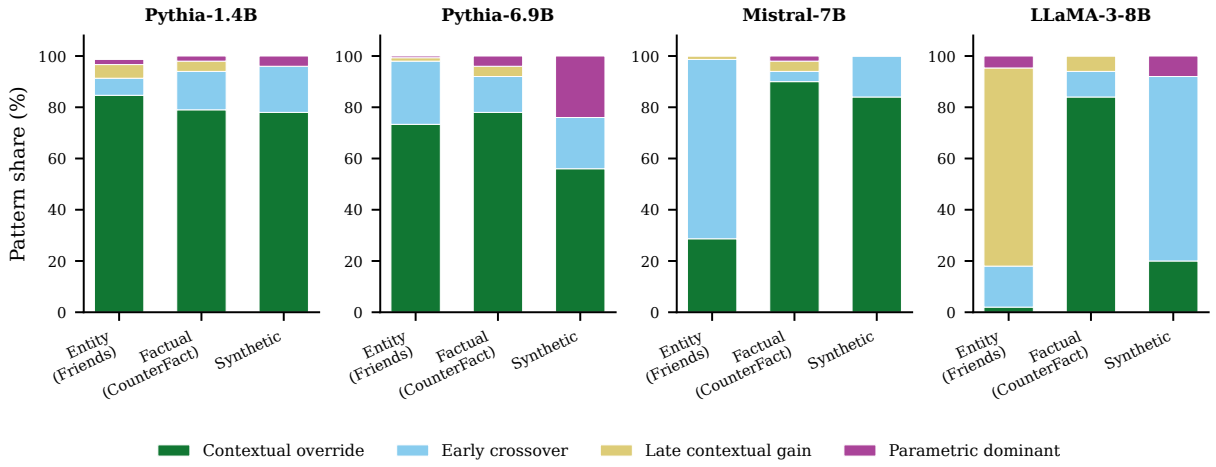

Figure 8: Distribution of conflict resolution patterns across knowledge domains and models. Within the Pythia family, contextual override dominates both domains. Mistral and LLaMA show domain-dependent patterns: entity stimuli resolve through early crossover (Mistral) or late contextual gain (LLaMA), while factual stimuli show contextual override regardless of architecture.

VRAM for the 7–8B parameter models. Total compute for all experiments across four models was approximately 200 GPU-hours.

Diagnostic prompts. Each stimulus includes a diagnostic prompt of the form "How would [entity] likely respond?" probing speech style attributes.

Model	Condition	Φ_P (mean)	n
Pythia-1.4B	Congruent	19.98	450
	Incongruent	19.62	2151
	Anonymous	19.30	450
	Name Only	24.55	18
Pythia-6.9B	Congruent	75.22	450
	Incongruent	74.52	2151
	Anonymous	74.53	450
	Name Only	91.53	18
Mistral-7B	Congruent	1.61	450
	Incongruent	1.61	2151
	Anonymous	1.66	450
	Name Only	2.09	18
LLaMA-3-8B	Congruent	11.82	450
	Incongruent	11.84	2151
	Anonymous	11.73	450
	Name Only	16.16	18

Table 7: Φ_P across all conditions and models. The first three conditions (with dialogue context) show near-identical values, while Name Only (pure parametric recall) shows 24–37% amplification.

We select this diagnostic because speech style reliably differentiates all six characters across all four models (median reciprocal rank ≥ 0.25). Other entity attributes (occupation, relationships) were tested in preliminary experiments but produced weaker differentiation for the smaller Pythia models.

Synthetic controls. Synthetic stimuli use fictional entity names (e.g., "Zarvex Thornweave") embedded in dialogue templates matching the Friends format. These produce near-chance override rates (56–80%) with no concentrated parametric sites, confirming that late-layer dynamics observed for real entities require genuine parametric memory. LLaMA shows a mild parametric bias even for fictional names (probability ratio 1.66), likely reflecting architectural defaults in its late MLP layers rather than entity-specific knowledge.

Stimulus pairing. The 150 base scenes generate 450 CONGRUENT stimuli (3 diagnostic types $\times$ 150 scenes) and 2,151 INCONGRUENT_SWAP stimuli (each scene contributes multiple swap pairs from its $\geq$2 characters). Paired statistical tests in §4.2 match on scene ID, averaging Φ_P across swap variants for each scene.