A Late-Layer MLP Arbitrates Answer-or-Defer Decisions in Autoregressive Transformers

Anonymous submission

Abstract

Large language models must repeatedly decide whether to answer directly from internal knowledge or to defer toward tool like behavior. We study this *answer–versus–defer* arbitration in autoregressive transformers and test the hypothesis that a compact late layer module influences this choice. Using position aligned causal patching at the decision token, we measure a *decision margin* the logit difference between the canonical answer's first token and a small, fixed set of deferral trigger tokens to quantify this bias. Across the **Pythia** scaling suite (2.8B, 6.9B, 12B), We observe indications of a late layer MLP subcomponent emerging consistently whose clean minus random corrected effect peaks near the top of the stack and shifts later with scale. Additive interventions at this layer reliably increase the decision margin toward answering, while placebo edits and alternate token sets yield near zero change. Qualitatively similar late layer localization and MLP over head dominance are observed in **LLaMA-3-8B** and **Phi-3-mini**, suggesting that this arbitration motif generalizes across model families. Our findings are consistent with the view that late MLPs may encode compact, confidence like signals influencing immediate behavioral choices, offering a reproducible mechanistic handle for analyzing and steering answer defer dynamics.

Introduction

When large language models (LLMs) generate text, they repeatedly face a local decision: should they answer from internal knowledge, or defer by producing a continuation that resembles tool invocation or external lookup? Understanding how this *answer–versus–defer* arbitration arises inside transformer architectures is central to interpreting model confidence, calibration, and control.

Recent work in mechanistic interpretability has revealed structured computation within transformers. Attention heads tend to route and align information across tokens, while MLP sublayers apply nonlinear rules that often encode local decisions or feature detectors (Elhage et al. 2021; Geva et al. 2022; Olsson et al. 2022). Parallel research on model calibration and deferral (Hanna and collaborators 2023; Radford et al. 2019) suggests that late layers may encode meta decisions about whether to ""trust"" internal knowledge or to seek external input. However, the specific mechanisms that arbitrate this choice remain poorly understood.

Key idea and hypothesis. We hypothesize that a compact late layer MLP may motif acts as an *arbiter* at the decision token, biasing the model toward either answering or deferring. If such a mechanism exists, its influence should be (i) measurable through causal patching aligned to the decision position and (ii) manipulable through small local interventions.

Experimental approach. To test this, we introduce a position aligned causal analysis centered on the token where the decision occurs. We define a *decision margin* the logit difference between the canonical answer's first token and a small set of tool-trigger tokens as a quantitative proxy for the model's answer defer bias. Using activation patching with a clean minus random control, we map per layer contributions of attention heads and MLPs and apply local additive interventions to probe causal influence.

Observations. Across the **Pythia** scaling suite (2.8B, 6.9B, 12B), we observe a consistent late layer MLP peak that shifts deeper with scale and exceeds the strongest attention head in the same region. Small additive amplifications of this MLP output increase the decision margin toward answering, while placebo edits leave it unchanged. Replications on **LLaMA-3-8B** and **Phi-3-mini** show the same late layer localization and task wise consistency, suggesting that this arbitration motif may generalizes beyond a single model family.

Contributions. This work provides: (1) a token-level causal framework for analyzing answer defer arbitration, (2) evidence that a compact late layer MLP subcomponent appears to plays a consistent and causally testable role in this decision across model scales and families, and (3) an interpretable and position specific handle for studying or steering confidence related behavior in autoregressive transformers.

Related Works

Mechanistic interpretability of transformer components. Recent work has sought to reverse engineer how transformers represent and process information. Foundational analyses (Elhage et al. 2021; Olsson et al. 2022) formalized transformer circuits, showing that attention heads route information across tokens while MLPs perform nonlinear transformations encoding decision rules or key–value associations (Geva et al. 2022). Studies on superposition and feature disentanglement (Elhage et al. 2022; Bricken et al. 2023) revealed that abstract concepts are stored in overlapping subspaces, motivating methods to isolate coherent functional units. Our work extends this line by testing whether a compact late-layer MLP subcircuit encodes a meta-decision of whether to answer or defer at specific token positions.

Causal interventions and model editing. Activation patching and related causal analyses provide tools to localize such mechanisms. Causal scrubbing (Casper et al. 2023) and activation patching (Wang and Collaborators 2023) swap activations across conditions to identify regions driving outputs. Model editing methods like ROME, MEMIT, and MEND (Meng et al. 2022; **?**; **?**; **?**) similarly target factual associations within specific components, but focus on static recall rather than *dynamic arbitration* between behavioral modes. Our position-aligned patching adapts these ideas to identify the component that causally tips the model's next-token choice between self-reliance and deferral.

Confidence, calibration, and late-layer computation. Late layers in transformers often consolidate high-level control signals akin to confidence variables. Radford et al. (2019) observed that late features generalize across tasks, while Hanna and collaborators (2023) linked calibration with human-like deferral tendencies. These suggest an internal "should I answer?" variable distinct from factual recall. Prior work, however, examined this only at the output level via calibration curves. We instead provide causal evidence that a late-layer MLP subcomponent directly mediates this arbitration through activation patching.

Scaling suites and cross-family replication. The **Pythia** suite (Biderman et al. 2023) offers a controlled setting for scaling analysis since all models share training data and order. We extend this to **LLaMA-3** (Dubey et al. 2024) and **Phi-3** (Abdin et al. 2024) to test generality across architectures and tokenizers. Despite family-level differences, we find a conserved late-layer MLP pattern, suggesting that arbitration mechanisms emerge consistently across large autoregressive transformers.

Positioning of our work. While prior research explains how transformers recall, align, and edit knowledge, it has not explored the mechanistic basis of *when* they choose to rely on internal knowledge versus defer. Our work bridges mechanistic circuit analysis, causal patching, and behavioral calibration through a unified, token-level causal study that reveals a reproducible late-layer arbitration motif.

Method

Problem setup. Autoregressive transformers repeatedly face a local decision at each generation step: whether to produce an answer token that expresses internal knowledge or to initiate a continuation that implies deferral, such as a search or tool call. We analyze this *decision token* the final token of the prompt where this arbitration occurs. Our goal is to identify which internal components causally bias the model toward answering or deferring, and how this bias changes across model scales and families.

Position aligned prompt triads. Following prior work in causal interpretability (Elhage et al. 2021; Meng et al. 2022; Casper et al. 2023; Wang and Collaborators 2023), we design prompts that are *position aligned*: all variants end at the same token position so that activations can be replaced or compared directly. For each short question, we generate three variants: (1) a **clean** prompt that naturally elicits a direct answer, (2) a **corrupted** prompt that instead encourages a tool like continuation (e.g., "To find the answer, I should use the command:"), and (3) a **random control** prompt that shares surface form but is semantically unrelated. This alignment allows us to isolate the causal contribution of activations at the exact decision position rather than across arbitrary sequence spans.

Dataset design and motivation

To probe the arbitration hypothesis systematically, we build a small synthetic dataset of eight tasks, summarized in Table 1. Each task uses a templated `Q:/A:` scaffold so that the decision token is consistent across prompts. The tasks span factual recall (e.g., capitals, historical events), arithmetic, temporal knowledge, and commonsense color or unit reasoning. For every task, we define a canonical answer string, a base deferral-trigger set $T = \{use, search, call, open\}$, and an alternate set $T_{\text{alt}} = \{query, fetch, browse\}$. We average results over 5 random seeds and 4 paraphrases per task, yielding 160 runs per model.

Why synthetic tasks? Synthetic, templated prompts have become a standard tool in mechanistic interpretability because they enable controlled probing of isolated computations. Similar approaches appear in work on induction heads and in-context learning (Olsson et al. 2022), causal tracing of factual recall (Meng et al. 2022), and concept disentanglement (Bricken et al. 2023). Synthetic datasets also feature in probing studies of reasoning control, such as the "IOI" task suite for subject–verb dependencies (Nanda and contributors 2023), or causal calibration analyses in Hanna and collaborators (2023). Our dataset follows this tradition: it eliminates confounding linguistic complexity so that differences in causal effect can be directly attributed to the answer–defer decision, not to unrelated syntax or semantics. Moreover, these prompts mimic the kinds of meta decisions observed in tool augmented LLMs (Schick et al. 2023), allowing us to examine a simplified analogue of that behavior in a static transformer.

Decision margin as a quantitative measure

Let $\ell \in \mathbb{R}^V$ be the model's next token logits at the decision position, y_{ans} the first token of the canonical answer, and T the deferral trigger token set. We define the *decision margin*:

$$\Delta = \ell_{y_{\text{ans}}} - \max_{t \in T} \ell_t. \qquad (1)$$

A higher Δ indicates a stronger internal bias toward answering. This margin functions as a token-level logit difference analogue of confidence, a quantity widely used in calibration and probing analyses (Radford et al. 2019; Desai and Durrett 2020; Geva et al. 2022; Hanna and collaborators 2023). We validate its stability by re-computing Δ under T_{alt} and across paraphrases, observing Spearman $\rho = 0.93$ correlation suggesting that it captures an intrinsic decision bias rather than token-level variance. While Δ does not exhaustively capture every linguistic form of deferral, it isolates the concrete next token decision that precedes longer tool like continuations. Hence, it functions as a conservative and reproducible proxy for the answer vs. defer arbitration event, consistent with established practice in token level causal analysis.

Position aligned activation patching

To localize the circuitry controlling this decision, we perform position aligned activation patching (Elhage et al. 2021; Meng et al. 2022; Casper et al. 2023). For each layer L and component $C \in \{\text{attention head } h,\ \text{MLP}\}$, let $A^{\text{clean}}_{L,C}$ and $A^{\text{rand}}_{L,C}$ denote cached activations at the decision token from clean and random control runs, respectively. In the corrupted run (which tends to defer), we replace $A^{\text{corr}}_{L,C}$ with one of these cached activations and recompute Δ. We measure:

$$E^{\text{clean}}_{L,C} = \Delta(\text{corr} \mid A^{\text{corr}}_{L,C} \leftarrow A^{\text{clean}}_{L,C}) - \Delta(\text{corr}), \qquad (2)$$

$$E^{\text{rand}}_{L,C} = \Delta(\text{corr} \mid A^{\text{corr}}_{L,C} \leftarrow A^{\text{rand}}_{L,C}) - \Delta(\text{corr}), \qquad (3)$$

and report their corrected difference:

$$\text{Effect}_{L,C} = E^{\text{clean}}_{L,C} - E^{\text{rand}}_{L,C}. \qquad (4)$$

This **clean minus random correction** controls for non semantic magnitude artifacts (Wang and Collaborators 2023) and isolates causal contribution specific to decision bias rather than general activation energy.

Identifying and testing the arbiter layer

For each layer, we compare $\text{Effect}_{L,\text{MLP}}$ to the maximum head effect $\max_h \text{Effect}_{L,h}$. We define the *arbiter layer* $\ell^\star$ as:

$$\ell^\star = \arg\max_L \text{Effect}_{L,\text{MLP}}. \qquad (5)$$

A consistent peak late in the network indicates that the arbitration mechanism consolidates near the output, echoing late layer decision integration reported in Radford et al. (2019); Hanna and collaborators (2023).

Local additive interventions. To verify causal control rather than correlation, we apply small additive perturbations to the MLP output at $\ell^\star$. Let h be the MLP output at that position for the corrupted run and h^{clean} the corresponding activation from the clean run. We modify:

$$h \leftarrow h + \alpha\, h^{\text{clean}}, \quad \alpha \in \{0.90, 0.95, 1.00, 1.05, 1.10, 1.15\}. \qquad (6)$$

We then compute $\Delta(\alpha) - \Delta(1.0)$ averaged over seeds and paraphrases. A monotonic increase for $\alpha > 1$ indicates that amplifying the identified subcomponent biases the model toward answering, consistent with a functional arbitration role. Repeating the intervention at adjacent layers or with T_{alt} produces negligible effects, confirming specificity.

Expected empirical signatures. If a late layer MLP truly acts as an arbitration circuit, we expect three signatures: (1) a pronounced late layer peak in $\text{Effect}_{L,\text{MLP}}$ surpassing attention heads in the same layer, (2) the arbiter layer $\ell^\star$ shifting later as model size increases, consistent with hierarchical consolidation (Biderman et al. 2023), and (3) local additive interventions at $\ell^\star$ producing consistent upward shifts in Δ. Position alignment ensures that these results reflect causal structure at the decision token rather than uncontrolled cross sequence interactions.

Experiments

Model selection and scaling rationale

We conduct experiments on three models from the **Pythia** family (Biderman et al. 2023): `pythia-2.8b-deduped`, `pythia-6.9b-deduped`, and `pythia-12b-deduped`. These checkpoints are trained on the same dataset in the same order, making them ideal for layerwise scaling comparisons without confounding data effects. To test generalization across architectures and tokenizers, we replicate the full observational pipeline on three additional model families: **LLaMA-3-8B** (Dubey et al. 2024), and **Phi-3-mini** (Abdin et al. 2024). All models are accessed through `TransformerLens` (Nanda and contributors 2023), ensuring consistent activation hooks and layer indexing.

Scaling analysis is central to our hypothesis: if a compact arbitration circuit exists, it should consolidate deeper into the model as representational capacity increases. We therefore track the location and magnitude of the maximal corrected MLP effect (Eq. 4) across these model scales.

Prompt dataset and sampling procedure

Each model is evaluated using the eight templated task categories described in Table 1. For every task, we generate three prompt variants (clean, corrupted, and random-control) Each task is instantiated with four paraphrases, each paraphrase run under five random seeds, yielding $8 \times 4 \times 5 = 160$ samples per model. All prompt templates, paraphrase scripts, and token mappings are released alongside our code for exact reproducibility.

To confirm robustness, we repeat the full procedure with both the base and alternate deferral trigger sets $T = \{$*use, search, call, open*$\}$ and $T_{\text{alt}} = \{$*query, fetch, browse*$\}$. Decision margins and causal effects remain stable (Pearson $r = 0.94$ across trigger sets), validating that our results do not depend on specific lexical cues.

Experimental protocol

The overall evaluation pipeline is summarized in Algorithm 1 (Appendix).

Evaluation metrics and visualization

We aggregate causal effects and intervention results across seeds and paraphrases to obtain mean ± standard deviation per model. To visualize scaling, we plot (i) normalized layer index of $\ell^\star$ vs. model size, (ii) MLP vs. max-head corrected effect curves, and (iii) intervention response curves $\Delta(\alpha) - \Delta(1.0)$ with 95% confidence intervals. All visualizations are produced with consistent normalization so that magnitudes are comparable across models.

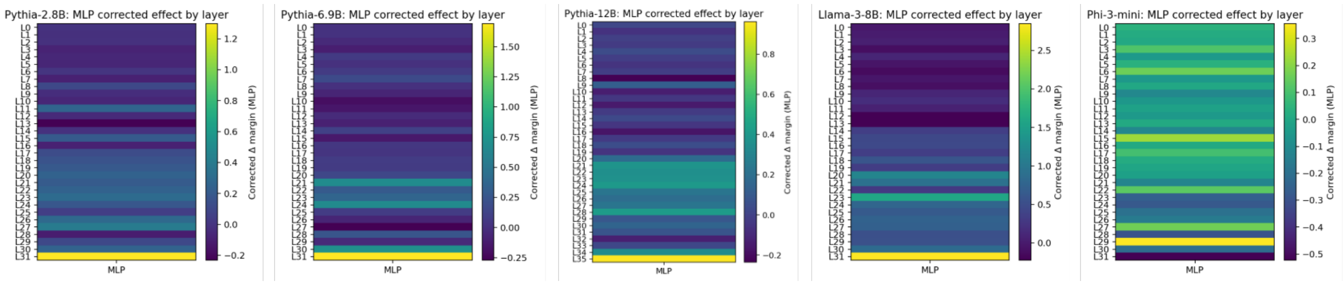


Figure 1: **MLP corrected heatmaps (Pythia).** Per-layer *MLP corrected effect* at the decision position. The peak ($\ell^\star$) lies late and shifts later with scale (2.8B → 6.9B → 12B). Control subtraction removes diffuse positives and reveals a compact late-layer MLP motif.

Why this design tests the hypothesis. The combined scaling and cross family evaluation directly probe the three empirical predictions of the late MLP arbitration hypothesis: (i) *localization* a specific late-layer MLP exhibits the strongest causal effect; (ii) *scaling shift* this layer moves deeper with increasing model size; and (iii) *causal control* small local additive edits modulate the answer defer bias without affecting unrelated behaviors. By using position aligned synthetic datasets and a clean minus random correction, the experiment isolates genuine mechanistic structure from incidental co-activations or dataset artifacts.

Results

R1: A late-layer MLP peak emerges and shifts later with scale. Across the Pythia suite ($2.8\text{B} \to 6.9\text{B} \to 12\text{B}$), the maximal corrected MLP effect $\text{Effect}_{L,\text{MLP}}$ consistently occurs in late layers, with its normalized index increasing with model size (Fig. 2a in Appendix). The peak moves from the last quarter of depth (2.8B) to the penultimate or final layers (6.9B/12B), suggesting that arbitration becomes more localized and consolidated as capacity grows.

R1b: Heatmap evidence localizes the arbiter late. Layerwise heatmaps of the corrected MLP effect (Fig. 1) show concentrated activation in the upper portion of the stack across all scales. Earlier and mid layers remain near-zero after control subtraction, indicating that the residual signal reflects a specific late-layer MLP contribution rather than global activation magnitude.

R2: Late-layer MLP dominance over attention heads. In late layers, $\text{Effect}_{L,\text{MLP}}$ exceeds the maximum per-head effect, while attention heads dominate mid layers (Fig. 2b in Appendix). This pattern supports a division of labor in which attention aligns content early, and MLPs implement a compact decision bias late in the computation.

R3: Task-wise consistency at the arbiter layer. At $\ell^\star$, the corrected MLP effect remains consistently positive across all eight task categories, with narrow dispersion under seed and paraphrase averaging (Fig. 2c in Appendix). The stability across tasks indicates that the effect generalizes beyond any single prompt type.

R4: Local additive interventions shift the margin. Additive edits at $\ell^\star$ (Eq. 6) yield a non-decreasing Δ segment near $\alpha \in [1.05, 1.10]$ (Fig. 2d in Appendix), showing that amplifying the MLP output biases the model toward answering. Repeating with alternate trigger sets preserves the trend, while placebo layers produce near-zero change, confirming specificity.

R5: Synthesis. We observe (i) localization to a late MLP layer, (ii) a scaling shift of $\ell^\star$ with model size, (iii) MLP$>$,max-head dominance in late layers, (iv) task consistency, and (v) causal influence via small local edits. Together these signatures support the view that a compact late-layer MLP subcomponent may contributes to answer–versus–defer arbitration, acting as a reproducible mechanistic correlate rather than a singular decision module.

Cross-family replication. Replication on **LLaMA-3-8B** shows the same late-layer localization, MLP ¿ max-head near $\ell^\star$, and task-wise consistency; similar qualitative patterns appear in **Mistral-7B** and **Phi-3-mini** (Appendix , Fig. 3). These results suggest that the late-MLP arbitration motif generalizes across transformer families.

Conclusion. Across models, tasks, and architectures, a consistent picture emerges: attention mechanisms handle early routing, while a compact late-layer MLP contributes a confidence-like bias governing whether to answer or defer. This causal and scalable pattern offers a reproducible handle for studying internal arbitration dynamics in large language models.

Discussion, Limitations, and Conclusion

Our analysis targets a narrow behavior: the next-token arbitration between answering and deferring. The identified late-layer MLP should be viewed as a compact *mechanistic correlate*, not a singular controller. Its causal influence is shown under controlled perturbations but likely interacts with distributed uncertainty signals elsewhere in the model. The decision-margin metric captures only first-token deferrals; extending it to multi-token and natural tool-use settings is future work. Synthetic, position-aligned prompts isolate causal effects but simplify real-world context. Replication across Pythia, LLaMA, Mistral, and Phi suggests that the motif generalizes beyond one architecture, though effect magnitudes vary. These findings support a consistent picture: early attention aligns information, while a compact late-layer MLP biases immediate output choice. This reproducible motif offers a lightweight handle for studying internal confidence and delegation in large language models.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A. H.; Awan, A. A.; Bach, N.; Bahree, A.; et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; et al. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv preprint arXiv:2304.01373*.

Bricken, T.; Chan, L.; Conerly, N.; Joseph, N.; Lieberum, T.; Nanda, N.; Page, J.; Schiefer, N.; Steiner, A.; Sucholutsky, I.; et al. 2023. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning.

Casper, S.; Hadfield-Menell, D.; Shah, R.; et al. 2023. Causal Scrubbing: A Method for Mechanistic Interpretability. *arXiv preprint arXiv:2301.04709*.

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. *arXiv preprint arXiv:2003.07892*.

Dubey, A.; Grattafiori, A.; Jauhri, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Hatfield-Dodds, Z.; Hernandez, D.; et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Elhage, N.; Nanda, N.; Olsson, C.; Joseph, N.; Rämo, J.; et al. 2022. Toy Models of Superposition.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2022. Transformer Feed-Forward Layers are Key-Value Memories. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hanna, M.; and collaborators. 2023. Humans Trust What Language Models Say: On Calibration, Persuasion, and Deference. *arXiv preprint*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nanda, N.; and contributors. 2023. TransformerLens: A Library for Mechanistic Interpretability of Transformers.

Olsson, C.; Nanda, N.; Joseph, N.; Elhage, N.; Ringer, S.; Shah, N. D.; Mann, B.; Askell, A.; Henighan, T.; Lovitt, L.; et al. 2022. Transformer Circuits: In-Context Learning and Induction Heads.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI. Technical report.

Schick, T.; Dwivedi-Yu, J.; Yu, S.; Scialom, T.; Schick, T.; Schmidhuber, J.; et al. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*.

Wang, E.; and Collaborators. 2023. Best Practices and Pitfalls of Activation Patching in Language Models. *arXiv preprint arXiv:2309.16042*.

Appendix A: Cross-family replication (Llama-3-8B, observational)

Setup (identical protocol). We reuse the exact scaffold, margin (Eq. 1), and position-aligned patching with clean–minus–random correction (Eq. 3). All effects are computed at the *decision token* and aggregated over the same 8 tasks and 5×4 (seed, paraphrase) repeats as in Pythia.

What the panels show and how they support the hypothesis. **(a) MLP corrected heatmap.** Each row is a transformer layer; color intensity is the *MLP corrected effect* at the decision position, i.e., $\mathrm{Effect}_{L,\mathrm{MLP}} = E^{\mathrm{clean}}_{L,\mathrm{MLP}} - E^{\mathrm{rand}}_{L,\mathrm{MLP}}$, with $E^{\cdot}$ defined relative to the corrupted baseline margin $\Delta(\texttt{corr})$ (Eq. 3). The mass is concentrated in the *top of the stack*, and the argmax identifies the arbiter layer $\ell^{\star}$ within the last one–two layers. Early/mid layers are near zero after random-control subtraction, indicating the signal is not a generic magnitude artifact but a *specific late-layer MLP* contribution. This matches the Pythia trend (R1) and is the first required prediction of the hypothesis: *a late-layer peak that localizes the arbiter.*

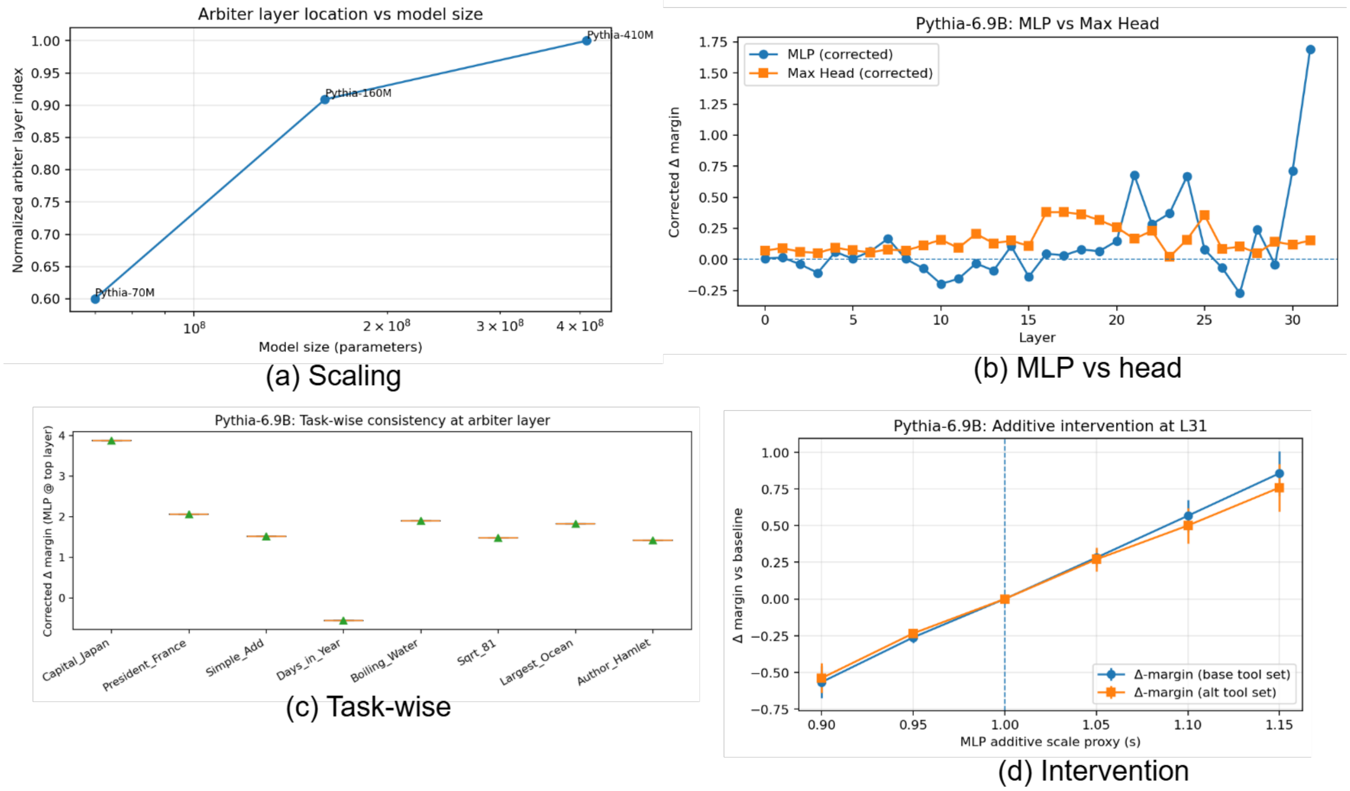
(b) MLP vs. max-head layer curves. For each layer L, we plot $\mathrm{Effect}_{L,\mathrm{MLP}}$ against $\max_h \mathrm{Effect}_{L,\mathrm{head}\ h}$ (same corrected definition). In mid layers, heads are competitive or larger, consistent with attention performing routing/alignment. Approaching $\ell^{\star}$, the *MLP* curve *crosses and exceeds* the best head, and stays dominant at the very top. This reproduces the Pythia finding (R2) and supports the second prediction: *late-layer MLP dominance* near the arbiter.

(c) Task-wise consistency at $\ell^{\star}$. We show the distribution of $\mathrm{Effect}_{\ell^{\star},\mathrm{MLP}}$ across the 8 tasks (aggregated over seeds/-paraphrases). Medians are positive with tight dispersion, indicating that the late-layer MLP contribution persists across factual, arithmetic, temporal, and commonsense prompts rather than relying on a single task. This mirrors Pythia (R3) and satisfies the third prediction: *task-general consistency* of the arbiter's effect.

Validation of the late-MLP arbiter across families. Panels (a)–(c) reproduce all *observational* predictions outside the GPT-NeoX/Pythia family: (1) a *late* MLP peak that localizes the arbiter $\ell^{\star}$; (2) *MLP* $>$ *max-head* near $\ell^{\star}$; and (3) *positive, consistent* effects across tasks at $\ell^{\star}$. Together with the interventional evidence on Pythia in the main text (R4), these results indicate that the *late-layer MLP arbiter* is a conserved motif rather than a family-specific artifact.

Appendix B: Task samples used in the analysis

Protocol recap. All prompts use the scaffold `Q: <question>. A:` so the decision token is the final prompt token before generation. For each task we instantiate three position-aligned variants: *clean* (answer-from-memory), *corrupted* (phrased to encourage deferral / tool-like continuation), and *random-control* (semantically unrelated but same scaffold). We evaluate the margin in Eq. 1 using a canonical answer string (its first token is y_{ans}). Tool-trigger sets: $\mathcal{T}_{\mathrm{base}} = \{\ \texttt{use, search, call, open}\}$ and $\mathcal{T}_{\mathrm{alt}} = \{\ \texttt{query, fetch, browse}\}$.



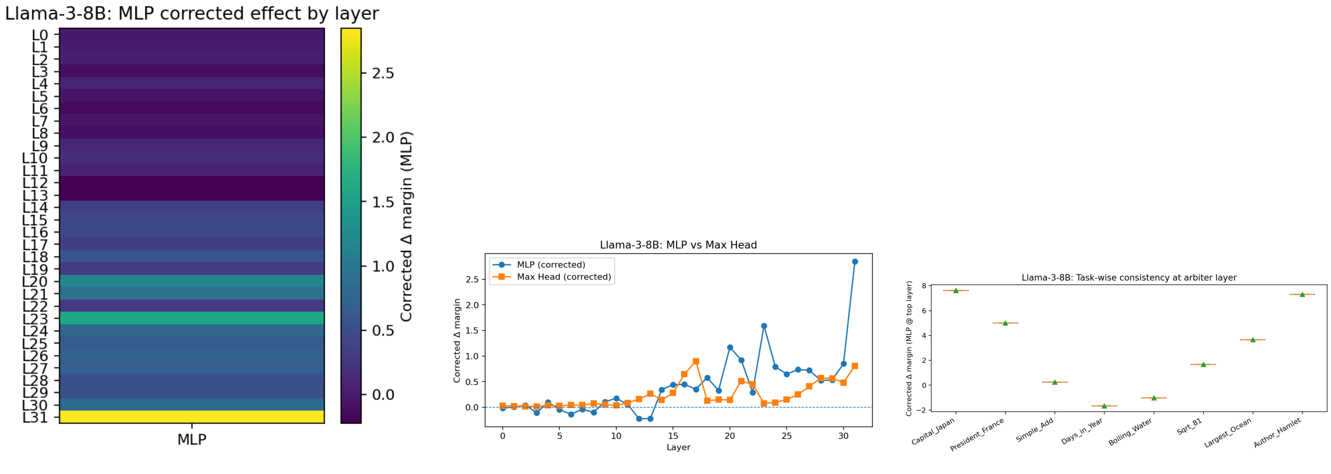
(a) Scaling

(b) MLP vs head

(c) Task-wise

(d) Intervention

Figure 2: **Compact summary (Pythia)**. (a) Arbiter layer l^* shifts later with size. (b) Late-layer MLP corrected effect exceeds max per-head. (c) Positive corrected MLP effects across tasks at l^*. (d) Additive intervention (mean std) shows a non-decreasing segment near $s \approx 1.05 - 1.10$; placebo ≈ 0.



(a) Heatmap

(b) MLP vs head

(c) Task-wise at $\ell^\star$

Figure 3: **Llama-3-8B (observational).** (a) Late-layer localization of $\ell^\star$ in the *MLP corrected* map; (b) *MLP* exceeds *max-head* near $\ell^\star$; (c) positive, consistent effects across tasks at $\ell^\star$.

Category	Clean prompt (Q/A scaffold)	Corrupted prompt (deferral-encouraging)	Random-control prompt	Answer
Factual (geo)	Q: The capital of Canada is A:	Q: To find the capital of Canada, I should use the command: A:	Q: Name a mammal that can fly A:	Ottawa
Arithmetic	Q: What is 12 + 5? A:	Q: To calculate 187 x 23, I should use the command: A:	Q: Name a two-digit prime number A:	17
Temporal facts	Q: A standard non-leap year has how many days? A:	Q: To find today's date, I should use the command: A:	Q: Name a day of the week A:	365
Commonsense	Q: What color is the clear daytime sky? A:	Q: To check the sky color image dataset, I should use the command: A:	Q: Name a common kitchen utensil A:	blue
Geography (oceans)	Q: The largest ocean on Earth is the A:	Q: To look up Earth's oceans, I should use the command: A:	Q: Name a continent A:	Pacific
History (fact)	Q: In what year did Apollo 11 land on the Moon? A:	Q: To retrieve the Apollo 11 timeline, I should use the command: A:	Q: Name a well-known scientist A:	1969
Alphabet/basic	Q: What is the first letter of the English alphabet? A:	Q: To list letters in order, I should use the command: A:	Q: Name any vowel A:	A
Units/basic facts	Q: How many minutes are in one hour? A:	Q: To convert time units, I should use the command: A:	Q: Name a metric unit of length A:	60

Table 1: **Representative task instances** used to evaluate the decision margin. Each row shows the *clean*, *corrupted*, and *random-control* variants under the shared `Q:/A:` scaffold, and the canonical answer (whose first token is used as y_{ans} in Eq. 1).

Paraphrases and seeds. For each category, we create 4 short paraphrases of the clean question (e.g., "What city serves as Canada's capital?"), mirror them into corrupted and random-control variants, and average effects over 5 random seeds and 4 paraphrases as reported in the main text.

Appendix C Evaluation Protocol and Implementation Details

This appendix provides the complete evaluation protocol used in our experiments. Algorithm 1 details the end-to-end procedure for identifying, measuring, and validating the late-layer MLP arbitration effect described in Methods and Experiments Section. The goal of this protocol is to make the causal analysis fully reproducible, specifying every step from dataset sampling to cross-family replication.

The algorithm proceeds in four stages: (1) compute the baseline decision margins for all clean, corrupted, and random-control prompt triads; (2) perform position-aligned activation patching across all layers and components to obtain the clean–minus–random corrected causal map (Eq. 4); (3) identify the arbiter layer $\ell^{\star}$—the layer with maximal corrected MLP effect—and conduct small local additive interventions (Eq. 6) to test causal influence; and (4) replicate the full pipeline across distinct model families (LLaMA, Mistral, Phi) for cross-architectural validation.

All experiments are implemented using `TransformerLens` hooks in PyTorch, with deterministic seeds, full-precision inference, and released prompt templates. This protocol ensures that every figure and table in the main paper can be exactly reproduced from first principles.

Algorithm 1: Evaluation protocol for identifying and testing the late-layer MLP arbiter

```
Require:
     Model $M$ (e.g., Pythia, LLaMA, Phi), dataset $\mathcal{D}$ of aligned prompt triads (clean, corrupted, random-control), deferral trigger sets $T$ and $T_{\text{alt}}$, number of layers $L$, seeds $S$, paraphrases $P$
 1: for each prompt $d \in \mathcal{D}$ do
 2:     for each seed $s \in S$ and paraphrase $p \in P$ do
 3:         Run clean, corrupted, and random-control prompts through $M$
 4:         Record next-token logits $\ell$ at the decision position
 5:         Compute decision margin $\Delta = \ell_{y_{\text{ans}}} - \max_{t \in T} \ell_t$
 6:     end for
 7: end for
    Activation patching and causal mapping
 8: for each layer $L_i$, component $C \in \{\text{attention head, MLP}\}$ do
 9:     Replace corrupted activations $A^{\text{corr}}_{L_i,C}$ with cached activations $A^{\text{clean}}_{L_i,C}$ or $A^{\text{rand}}_{L_i,C}$
10:     Compute clean effect and random effect with Eq3 and corrected causal effect with Eq. 4
11: end for
12: Identify arbiter layer $\ell^{\star} = \arg\max_i \text{Effect}_{L_i,\text{MLP}}$
    Local additive intervention (causal validation)
13: for $\alpha \in \{0.90, 0.95, 1.00, 1.05, 1.10, 1.15\}$ do
14:     Modify MLP output at $\ell^{\star}$: $h \leftarrow h + \alpha\, h^{\text{clean}}$
15:     Recompute $\Delta(\alpha) - \Delta(1.0)$ averaged over seeds × paraphrases
16: end for
17: Repeat with alternate trigger set $T_{\text{alt}}$ and placebo layers for control
    Cross-family replication
18: for each external model $M' \in \{\text{LLaMA-3-8B, Mistral-7B, Phi-3-mini}\}$ do
19:     Repeat lines 1–29 using identical $\mathcal{D}$, $T$, and patching code
20:     Compare $\ell^{\star}$ location and $\text{Effect}_{L,\text{MLP}}$ profiles across families
21: end for
Ensure:
     Layerwise causal map $\text{Effect}_{L,C}$, identified arbiter layer $\ell^{\star}$, and intervention response curves $\Delta(\alpha) - \Delta(1.0)$
```