Mechanistic Evidence Routing in Transformers: Margin-Gated Intervention for Retrieval-Conflict Hallucination

Anonymous ACL submission

Abstract

When a language model ignores retrieved evidence and answers from memory instead, this is not a knowledge failure—the correct information is present. It is a *routing failure*: the model commits to its parametric answer despite the evidence signal. We show this distinction is mechanistically meaningful, localizable , and correctable at inference time without training data or weight modification.

A single prefill-margin computation detects routing failures (AUC $\geq$ 0.93 across five model–dataset pairs) and gates a format-appropriate logit-bias correction. On MedQA multi-option conflicts, this achieves **77.5%** restoration at zero harm, exceeding the strongest prompting baseline by 67.5 pp and matching domain-matched ITI without training data. On open-ended PopQA (3,141 items, three models), balanced alias-promotion achieves 80–88% restoration at near-zero harm, Pareto-dominating prompt-only baselines in five of six cross-model conditions. The central mechanistic finding is that effective late-layer intervention sites are structurally distinct from the causal *commitment locus* (a five-head cluster, joint flip rate 88.6%, $p < 10^{-9}$), suggesting that steering can succeed through residual-stream write access without exactly identifying the full decision circuit.

1 Introduction

Retrieval-Augmented Generation (Lewis et al., 2020) grounds language model outputs in external evidence. Yet models frequently generate from parametric memory even when the retrieved passage contains the correct answer (Longpre et al., 2021; Li et al., 2023a).

The standard framing treats this as a retrieval or prompting problem. We argue it is neither. When a model ignores evidence it just read, the failure is not that it lacks access–it has it. The failure is that the model *chose* not to use it. This choice is made inside the forward pass, at identifiable computation sites, and is correctable without retraining if those sites can be found.

We localize evidence routing to specific attention heads via per-head logit-difference decomposition, show that a single prefill-margin computation gates the correction reliably, and demonstrate that the resulting training-free intervention Pareto-dominates prompting baselines across three datasets and three architectures. The central finding is that the effective intervention site (any late-layer head) is structurally distinct from the causal commitment locus (a five-head cluster) confirmed by matched causal activation patching. Current approaches operate on inputs (Shi et al., 2024) or outputs (Chuang et al., 2024) without engaging the internal routing decision. The contributions are:

1. **Site–locus dissociation**: single-head patching produces 3–6% flip rates (chance); joint top-5 patching achieves 88.6% ($p < 10^{-9}$, §6), establishing with matched causal controls that steering sites need not coincide with causal decision sites.

2. **Per-head logit-difference decomposition** localizes routing heads (CounterFact: L27/H20, DR$= 6.1\times$) and reveals that localization is task-specific–the same head ranks 664th on MedQA (§4, §6).

3. **Margin-gated intervention**: 77.5% MedQA restoration at zero harm, +67.5 pp over prompting; 71.6 pp gap between label-space and content-space confirms failure is in label selection, not content understanding (§5).

4. **Cross-architecture + open-ended generalization**: DR predicts intervention effectiveness across Mistral/Qwen; balanced alias-promotion achieves 80–88% PopQA restoration at near-zero harm, Pareto-dominant in 5/6 cross-model conditions (§7, §8).

2 Related Work

Knowledge Conflicts in RAG. Longpre et al. (2021) show that LLMs frequently ignore retrieved evidence when it contradicts parametric memory. Xie et al. (2024) construct ConflictQA to evaluate this systematically. Wu et al. (2024) introduce ClashEval for domain-specific evaluation. Bi et al. (2025) study fine-grained knowledge reliance control but require model fine-tuning. Our work provides a *mechanistic* account of why models ignore evidence and corrects it training-free.

Mechanistic Interpretability. The residual stream decomposition follows Elhage et al. (2021). Geva et al. (2023) localize factual recall to attention heads. Meng et al. (2022) localize factual associations to MLP layers via causal tracing and propose targeted weight edits (ROME); our approach modifies no weights. Hernandez et al. (2023) show that relation decoding in transformer LMs is approximately linear, supporting our logit-lens attribution. Activation patching originates with Vig et al. (2020) and is refined for circuit discovery by Conmy et al. (2023). Wang et al. (2023) demonstrate circuit-level IOI analysis; we extend this line to the knowledge-conflict setting, asking not where facts are stored but where the decision to ignore retrieved evidence is made.

Inference-Time Steering. Li et al. (2023b) train probes on attention head activations to steer truthfulness (ITI); we match domain-matched ITI without training data. Zou et al. (2023) propose Representation Engineering, adding activation vectors to the residual stream; unlike RepE, our intervention is per-example gated, confining modification to detected conflict items. Turner et al. (2023) and Rimsky et al. (2024) apply activation addition globally, causing harm on already-correct examples a failure mode the margin gate prevents. Chuang et al. (2024) contrast early and late layer logits to improve factuality; DoLa achieves 3.4% CounterFact restoration with stability 0.250 against our 0.172 restoration at stability 0.997. Shi et al. (2024) amplify context-conditional logits globally without a safety gate, producing stability failures in our experiments.

Knowledge Editing. Meng et al. (2022), Mitchell et al. (2022), and Meng et al. (2023) modify model weights to edit factual associations. Our setting is complementary: we correct per-inference routing failures without persistent modification.

3 Methodology

The framework has two stages. *Localization* identifies which attention heads drive evidence routing via per-head logit-difference decomposition. *Intervention* applies a margin-gated correction only when the model's internal state confirms parametric preference.

3.1 Per-Head Logit-Difference Decomposition

The residual stream decomposes additively as $h^L = h^0 + \sum_{l,h} a_{l,h} + \sum_l m_l$, where h^0 is the embedding, $a_{l,h}$ is the attention head output, and m_l is the MLP output. Since the unembedding W_U is linear, the logit difference $\mathrm{LD} = (W_U[e] - W_U[p]) \cdot h^L$ decomposes as

$$\mathrm{LD} = d_{ep} \cdot h^0 + \underbrace{\sum_{l,h} d_{ep} \cdot a_{l,h}}_{\text{per-head attributions}} + \sum_l d_{ep} \cdot m_l, \quad (1)$$

where $d_{ep} = W_U[e] - W_U[p]$. Each term $d_{ep} \cdot a_{l,h}$ measures how much head (l, h) pushes the residual stream toward evidence over parametric tokens. Comparing these contributions across faithful and hallucinated examples identifies the heads where routing breaks down.

3.2 Margin Gate

The margin measures the model's current preference at the final prefill position:

$$M = \log P_E - \log P_P, \quad P_E = \textstyle\sum_{t \in E} \mathrm{softmax}(z)[t]. \quad (2)$$

$M > 0$ indicates evidence preference; $M < 0$ indicates parametric preference. The gate fires if and only if $M < \tau$. When the gate does not fire, the model generates without any modification and the source of zero intervention-induced change on non-fired examples. Harm on already correct predictions defined as the fraction of C2-faithful items that flip to incorrect under intervention is therefore an empirical property of the selected operating point, not a structural guarantee. When the gate fires, intervention strength is proportional to the deficit, capped at $\lambda_{\max}$:

$$\lambda = \min\Big(\alpha \cdot |M|/\mathrm{gain},\ \lambda_{\max}\Big). \quad (3)$$

In fixed-grid experiments, λ_{pos} and λ_{neg} are selected by Youden's J on the pilot set and held fixed on the evaluation set without post-hoc adjustment.

3.3 Intervention Variants

OV-Circuit Steering (single-token recall). For entity completion, the optimal steering direction through head (L, H)'s output projection is:

$$v^* = W_O^\top d_{ep} \tag{4}$$

Adding $\lambda v^*/\|v^*\|$ to the head output steers its residual-stream contribution toward the evidence token without modifying other components.

Label-Space Logit Bias (categorical decisions). For multi-option decisions:

$$\text{logits}'[t] = \text{logits}[t] + \lambda_{\text{pos}}, \quad t \in E_{\text{letter}}, \tag{5}$$
$$\text{logits}'[t] = \text{logits}[t] - \lambda_{\text{neg}}, \quad t \in P_{\text{letter}}. \tag{6}$$

Token-Level Entity Bias (open-ended and cross-architecture). For open-ended recall and cross-architecture models, the same logit-space bias is applied to the first tokens of all correct entity aliases (E_{alias}) and suppressed for the parametric entity's first token (P_{entity}).following the same form as Eqs. 5–6 with alias-specific token sets (Appendix C). The gate is computed from the localized head or layer margin signal (§7).

4 CounterFact: Single-Token Factual Recall

4.1 Experimental Setup

We use 500 factual conflict items requiring single-token entity completion from CounterFact (Meng et al., 2022). Primary analysis uses Llama-3.1-8B-Instruct (Team, 2024) (32 layers, 32 heads), run on a single NVIDIA H100 (80 GB), seed 1234, bfloat16. We compare against DoLa (Chuang et al., 2024), context-aware decoding (Shi et al., 2024), unconditional head amplification (ERHA), and ITI (Li et al., 2023b). DoLa is evaluated on CounterFact only; preliminary MedQA runs showed stability < 0.5 on the letter-token logit space, consistent with its global contrastive decoding behaviour. CounterFact items are *counterfactual by design*: the retrieved evidence contradicts real-world truth, testing whether the model follows context even against its correct parametric memory. This stresses the mechanism in the opposite direction from PopQA, where evidence is true and parametric memory is wrong.

4.2 Progressive Margin Divergence and Attention Gate Negative

The margin (Eq. 2) is computed at each layer (Figure 1). Faithful and hallucinated trajectories diverge starting at layer 20 and peak at layer 27: AUC $= 0.932$, Cohen's $d = 2.041$. MLP blocks amplify the margin monotonically without sign change, confirming their role as amplifiers, not decision-makers.

We test whether the attention weight from the answer position to the evidence token at L27/H20 could serve as a gate. The gate fire rate is identical for faithful and hallucinated examples at every threshold (AUC ≈ 0.50, Figure 1). The QK circuit attends correctly in both cases; the routing failure lies in what the OV circuit writes from that attended position. This dissociation motivates the margin-based gate.

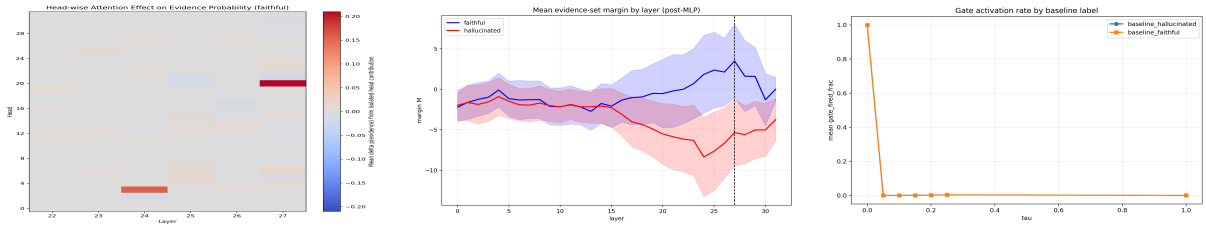
4.3 Intervention Results

Table 1 summarizes the results. The attention mass gate (AG-ERHA) achieves zero restoration because it cannot distinguish faithful from hallucinated examples. OV steering at L27/H20 achieves 17.2% restoration at zero harm and stability 0.997. The remaining 82.8% of unrestored examples have margin deficits too large for output-space correction at safe operating points; deeper residual-stream intervention for these cases is future work. DoLa achieves marginal restoration with severely degraded stability (0.250), consistent with its global application to all items including those that do not require intervention.The full method comparison including stability rates is in Appendix K.

5 MedQA: Multi-Option Categorical Decisions

5.1 Experimental Setup

Dataset. 500 engineered conflict items in a four-option multiple-choice format, constructed from MedQA-USMLE (Jin et al., 2021). Each item contains a clinical guideline supporting the correct letter (A/B/C/D) and a parametric answer the model favors from pre-training. This is not a medical benchmark; it is a controlled testbed for *letter-token commitment* under retrieval conflict. A frozen matched subset of 46 base-wrong items of 500 conflict items, these are the genuine parametric-resistant cases where the model answers incorrectly despite context is defined for main results; a hand-audited clean subset of 40 items is used for kill-test



(a) Head importance. (b) Margin divergence. (c) Attn. gate: AUC$\approx$ 0.50.

Figure 1: CounterFact localization results (Llama-3.1-8B). Left: L27/H20 is the only head with substantial faithful-vs-hallucinated gap (DR= 6.1$\times$). Centre: margin diverges at layer 20, peaks at L27 (AUC= 0.932, $d = 2.041$); MLP blocks amplify only. Right: attention mass gate fails (AUC$\approx$ 0.50) failure lies in OV write, not QK attention.

Method	Rest.	Harm	Stab.
AG-ERHA (attn. gate)	0.000	0.000	1.000
MG-ERHA (single head)	0.057	0.000	0.997
MG-ERHA (multi-head)	0.061	0.000	0.997
DoLa	0.034	—	0.250
OV Steering	**0.172**	**0.000**	**0.997**

Table 1: CounterFact results (Llama-3.1-8B). OV steering at L27/H20 achieves the highest restoration at zero harm. DoLa achieves marginal restoration with severely degraded stability (0.250). Stab. = fraction of initially-correct items remaining correct; a drop in Stab. indicates harm on faithful items.

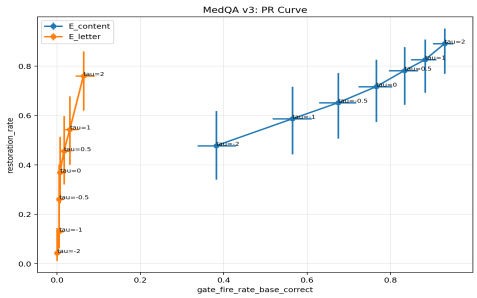


Figure 2: Precision-recall tradeoff over 14 gate configurations ($\tau\times$ bootstrap space). E_{letter} (orange) Pareto-dominates E_{content} (blue) at every operating point. The 71.6 pp gap reveals the routing failure is in label selection, not content understanding. Optimal: $\tau = 2.0$, E_{letter} 76.1% restoration, 6.3% false-alarm rate, zero harm.

comparisons. All results report Wilson 95% CIs; all paired comparisons use McNemar's exact binomial test.

Token Sets. $E_{\text{letter}}/P_{\text{letter}}$: BPE variants of the correct/wrong letter tokens. $E_{\text{content}}/P_{\text{content}}$: content tokens of the correct/wrong option texts.

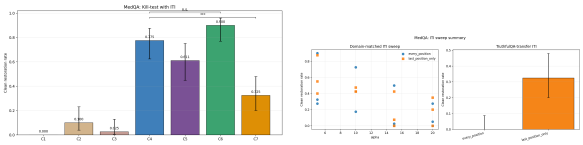
5.2 PR Curve and Answer-Format Dependency

Figure 2 sweeps $\tau \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ and bootstrap space $\in \{E_{\text{content}}, E_{\text{letter}}\}$. E_{letter} Pareto-dominates E_{content} at every operating point. At the optimal configuration ($\tau = 2.0$, E_{letter}), restoration reaches 76.1% [62.1, 86.1] with 6.3% false-alarm rate and zero measured harm.

Finding: Content-space intervention achieves only 5.9% restoration. The 71.6 pp gap is direct evidence that the model has already computed the correct answer internally the failure is exclusively in the final label commitment step, not in content understanding.

5.3 Kill-Test, ITI Comparison, and Dose-Response

Prompt only baselines are C4 (authoritative-quote reframing) and C5 (explicit answer injection + trust instruction, the strongest prompting ceiling). ITI with domain-matched probes (C6) achieves 90.0% at zero harm,statistically indistinguishable from C4 ($p = 0.21$).ITI without domain data (C7) degrades to 32.5% with 75.4% harm,demonstrating that training-free operation is a practical necessity.



(a) Kill-test: all 7 conditions. (b) ITI collapses at α=15.

Figure 3: Left: MedQA kill-test (clean subset, $n = 40$). C4 (mechanism) achieves 77.5% at zero harm; C2/C3 (prompting) cause 5.6–13.8% harm. C5 (mech+trust) drops to 61.1%: the trust instruction miscalibrates the gate, confirming failure is sub-surface. Right: ITI restoration collapses from 90% ($\alpha = 5$) to near-zero ($\alpha = 15$); our method is stable across $\lambda \in [2.5, 10.0]$. Dose-response curve in Appendix J.

Condition	Rest.	Harm	Train	McN.
C1: Vanilla	0.0%	0.0%	None	—
C2: Trust instr.	10.0%	13.8%	None	10^{-27}
C3: Auth. quote	2.5%	5.6%	None	10^{-18}
C4: Mechanism	**77.5%**	**0.0%**	None	—
C5: Mech.+trust	61.1%	0.0%	None	10^{-10}
C6: ITI (matched)	90.0%	0.0%	500	0.21
C7: ITI (transfer)	32.5%	75.4%	TQA	10^{-98}

Table 2: Kill-test on the clean subset (40 items). McNemar p-values vs. C4. **(i)** C2/C3 cause substantial harm (5.6–13.8%), dropping accuracy below baseline. **(ii)** C4 achieves 77.5% at zero harm; McNemar vs. C2: $p = 1.6 \times 10^{-27}$, 90 discordant pairs all in C4's favour. **(iii)** C5 (mechanism + trust instruction) drops to 61.1% ($p = 2.4 \times 10^{-10}$ vs. C4): the trust instruction shifts the margin distribution, miscalibrating the gate and confirming the failure is sub-surface. C6 matches statistically ($p = 0.21$) but requires 500 labelled same-domain training items. C7 shows training-free operation is a practical necessity, not a convenience.

6 Causal Localization and Distributed Commitment

6.1 MedQA-Specific Head Importance

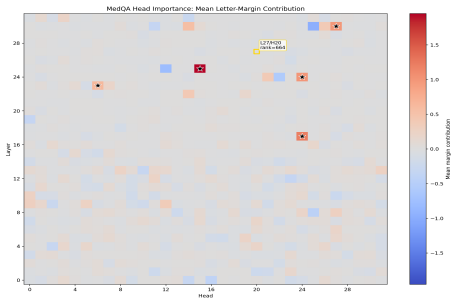


Figure 4: Head importance heatmap on MedQA (32×32 heads, scored by letter-margin contribution). L27/H20-the CounterFact dominant head ranks 664th of 1,024 on MedQA. Top MedQA head is L25/H15 (layer 25, head 15). CounterFact localization does not transfer to multi-option letter decisions, motivating the dual-localization finding (§6).

Scoring all 1,024 heads by letter-margin contribution (Figure 4) reveals that L27/H20 ranks 664th. The top-5 MedQA heads-L25/H15, L17/H24, L24/H24, L30/H27, L23/H6-span layers 17–30. Crucially, intervention at L25/H15 produces *identical* restoration to L27/H20 (76.1% in both cases, overlapping Wilson CIs). The logit-bias intervention is **head-agnostic**: it propagates through the residual stream to the output distribution at any late-layer site, decoupling the choice of intervention site from the choice of localization head.

6.2 Single-Head Patching: Null Result

From the matched subset of 46 base-wrong items, 35 yield valid source–destination pairs: base-correct source items with the same correct letter exist in the dataset for these 35, enabling letter-matched causal patching. Neither L27/H20 (5.7% flip rate, $p = 1.0$ vs. controls) nor L25/H15 (2.9% flip rate, $p \geq 0.5$) produces flips above chance, confirming that no single head causally carries MedQA letter commitment.

6.3 Joint Multi-Head Patching: Positive Causal Evidence

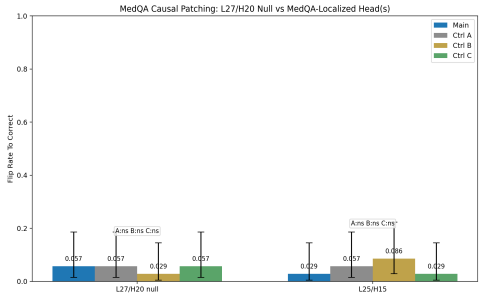


Figure 5: Causal patching results. Single-head patching at L27/H20 and L25/H15 produces flip rates indistinguishable from controls (5.7% and 2.9%, $p \geq 0.5$). Joint patching of the top-5 heads achieves 88.6% [74.0, 95.5] ($p < 10^{-9}$ vs. both controls). Error bars are Wilson 95% CIs.

Simultaneous patching of all five top-ranked MedQA heads achieves a flip rate of 88.6% [74.0, 95.5] (31/35 pairs, Figure 5). Controls (single-head at L27/H20: 5.7%; single-head at L25/H15: 2.9%) show McNemar $p = 3.7 \times 10^{-9}$ and $p = 1.9 \times 10^{-9}$, respectively.

A disruption control patches from mismatched-letter sources (source correct letter $\neq$ destination correct letter): 62.9% of outcomes flip specifically to the *source* letter [46.3, 76.8], 0% to any third letter, 37.1% unchanged. The bimodal outcome (flip to source letter or unchanged, with zero noise) rules out activation-volume dragging and supports the interpretation that the five heads jointly encode the letter commitment, not merely the probability mass (Appendix E).

Finding: No single head causally carries the commitment: single-head patching produces 3–6% flip rates (controls). Joint patching of five late-layer heads achieves 88.6%. Yet logit bias at *any* single late-layer head achieves 76% restoration–the intervention propagates through the residual stream regardless of which site is targeted. Effective steering requires only residual-stream write access, not identification of the exact causal site.

Full statistical definitions are given in Appendix D.

7 Cross-Architecture Routing Regimes

7.1 Three Qualitative Regimes

We extend analysis to Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) (32 layers, 32 heads, standard MHA) and Qwen2-7B-Instruct (Yang et al., 2024) (28 layers, 28Q/4KV heads, GQA 7:1), evaluating on MedQA, CounterFact, and PopQA. Table 3 summarizes the three routing regimes

Regime 1 – Concentrated (Llama). A single head dominates evidence routing (DR $= 6.1$). OV-circuit steering at L27/H20 is effective.

Regime 2 – Distributed (Mistral). No single head dominates (DR $= 1.26$). The margin AUC is strong (0.999 on MedQA, 1.000 on CounterFact) but distributed. True single-head OV hooking fails: 0/57 restored on MedQA, because modifying one head contributing ≈3.9% of the residual stream cannot overcome the collective parametric signal. Token-level logit-bias gated by the localized head margin achieves strong restoration on CounterFact and PopQA, while Mistral MedQA is a partial-transfer case where strong explicit prompting competes with the mechanism (discussed in §7.3).

Regime 3 – GQA-Distributed (Qwen). Grouped-query attention (7:1 ratio) means Q-heads share KV projections in groups of seven, structurally preventing per-head OV routing. Despite this, a layer-level margin signal (L24, residual stream) is highly discriminative (AUC ≥ 0.985). Architecture-appropriate intervention achieves the highest per-dataset restoration of any model.

7.2 Quantitative Cross-Architecture Results

Table 4 shows three headline patterns.

Broad detection transfer. Margin AUC ≥ 0.93 in all five measured cross-model margin settings indicates the conflict-detection signal is not specific to Llama, even when the intervention approach must change.

Near-zero measured harm. The gated mechanism achieves zero measured harm in all Mistral/Qwen cells and 0.07% on Llama PopQA ($\tau = 0.0$, balanced). Prompt-only baselines incur 14–35% harm on CounterFact (§7.4) and 1.7% on Qwen CounterFact C5.

Pareto dominance on CounterFact and PopQA. On PopQA the balanced mechanism Pareto-dominates all prompt-only baselines for all three models. On CounterFact, Mistral and Qwen also Pareto-dominate. Significance: Mistral PopQA C3 vs. C4 $p = 8.1 \times 10^{-28}$; Qwen PopQA C3 vs. C5 $p = 2.5 \times 10^{-7}$. Full McNemar table is in Appendix G.Additional cross-architecture plots are in Appendices N and O. PopQA stratified analyses are in Appendix M. Supra-threshold decomposition controls are in Appendix F.

Note: In cross-architecture evaluation, C4 denotes the authoritative-quote prompt baseline and C5 the explicit-answer injection baseline, following the PopQA and CounterFact convention; these differ from the MedQA kill-test condition labels in Table 2.

Property	Llama	Mistral	Qwen
Regime	Concentrated	Distributed	GQA-Dist.
DR	6.1	1.26	—
Localization	Head L27/H20	Head L29/H2	Layer L24
AUC (MedQA)	0.932	0.999	n/r
AUC (CF)	—	1.000	1.000
AUC (PopQA)	0.959	0.994	0.985
OV hook?	✓	✗	✗
Best impl.	OV steering	Tok. logit	Layer/tok.

Table 3: Three routing regimes. DR = dominance ratio (top head vs. mean). OV hook fails for Mistral (distributed, DR=1.26) and for Qwen (GQA shared KV projections).n/r = not reported; Qwen gate uses layer-level L24 residual-stream signal rather than a head-level margin (see Appendix P). AUC ≥ 0.93 in all five measured settings supports broad conflict-detection transfer.

7.3 Mistral MedQA: Negative Transfer Confirms DR Prediction

Mistral MedQA is the one Pareto failure: C3=35.1% vs. C4=75.4%, C5=91.2%, all at zero harm. This is the exact result the dominance ratio predicts: with DR$= 1.26$, each of Mistral's 32 heads contributes ≈3.9% of the residual stream. In closed MCQ, an explicit-answer prompt supplies sufficient concentrated signal to override the distributed parametric attractor; a single-head logit correction cannot.The mechanism provides unique value in open-ended settings where prompting plateaus–Mistral PopQA (80.5%) and CounterFact (91.7%) are both Pareto-dominant, and the DR boundary is predictable before running any experiment.

7.4 Harm Analysis for Prompt-Only Baselines

On CounterFact, authoritative prompt (C4) causes 17–35% harm. This is not a semantic failure but a

Model	Impl	MedQA Rest.	MedQA Harm	CounterFact Rest.	CounterFact Harm	PopQA Rest.	PopQA Harm	Note
Llama-3.1-8B	format-aware bias	76.1 [62–86]	0.0	17.2 [9–29]	0.0	88.1 [83–92]	0.1	primary
Mistral-7B	token logit bias	35.1 [24–48]	0.0	91.7 [65–99]	0.0	80.5 [72–87]	0.0	partial†
Qwen2-7B	layer/token bias	85.2 [68–94]	0.0	96.4 [82–99]	0.0	86.4 [76–93]	0.0	
Best prompt-only baseline per model–dataset pair (shown for reference)								
Mistral (C4)	auth. quote	75.4	0.0	25.0	35.1	41.6	0.0	
Mistral (C5)	expl. inject.	91.2	0.0	83.3	14.0	40.7	0.0	
Qwen (C4)	auth. quote	59.3	0.0	14.3	17.1	45.5	0.0	
Qwen (C5)	expl. inject.	66.7	0.0	57.1	1.7	43.9	0.0	

Table 4: Cross-architecture results. Mechanism harm is 0.0% in all cells except Llama PopQA (0.07%, reported as 0.1%; see §8). †Mistral MedQA is a partial-transfer case (§7.3). C4/C5 harm on CounterFact reflects format drift from authoritative prompting (§7.4). Pareto dominance (mechanism rest. > all baselines *and* harm ≤ all baselines) holds in five of six cross-model cells; Mistral MedQA is the exception. CIs from Wilson 95% method; cross-architecture matched subsets differ from Llama ($n = 57$ for Mistral MedQA).

format disruption: the authoritative framing shifts generation from bare entity completion ("Paris") to numbered-list format ("1. Paris"), changing the first generated token. Under the strict first-token evaluation protocol, this counts as harm. The gated mechanism avoids this entirely because it fires only on the 12–28 detected conflict items, leaving all faithful items unmodified.

8 PopQA: Open-Ended Entity Recall

8.1 Setup

PopQA (Mallen et al., 2023) tests open-ended entity recall (e.g., "Who directed The Doctor?"). Unlike MedQA, the answer space is unconstrained with multiple valid surface forms and aliases. We construct a C2-filtered conflict dataset per model using a two-stage pipeline: (1) filter candidates with parametric confidence ≥ 0.65; (2) retain only items where the model *still* generates the parametric answer after context is provided (genuine C2-parametric conflicts). This yields 243 genuine conflicts for Llama (7.7% of 3,141 evaluated), 113 for Mistral, and 66 for Qwen. Evidence token sets include first tokens of all correct aliases rather than only the canonical answer form.

8.2 Balanced vs. Suppression-Only Ablation

Table 5 shows that promotion is necessary, not optional, for open-ended settings. Suppression-alone displaces the model to a random alternative rather than redirecting to the evidence token. The 44–51 pp gain from adding evidence promotion directly answers whether restoration reflects the mechanism or mere attractor removal: it requires the mechanism to actively redirect, not simply suppress.

Model	Mode	C3	C4	Pareto
Llama	Suppress-only	43.6%	80.7%	✗
Llama	**Balanced**	**88.1%**	80.7%	✓
Mistral	Suppress-only	29.2%	41.6%	✗
Mistral	**Balanced**	**80.5%**	41.6%	✓
Qwen	**Balanced**	**86.4%**	45.5%	✓

Table 5: Suppression-only vs. balanced intervention on PopQA. Suppression-only fails Pareto dominance for Llama and Mistral. Balanced intervention (promotion + suppression) recovers 44.5 pp on Llama and 51.3 pp on Mistral, confirming that restoration requires active evidence redirection, not mere attractor removal.

8.3 Llama PopQA Harm Analysis

The Llama balanced run achieves 88.1% restoration on 243 genuine conflict items with 2/2,792 faithful-item harm (0.07%, CI: 0.01–0.26%). Manual inspection: one alias-format drift and one false-positive gate degradation neither reverts to the parametric answer. Forcing zero harm requires $\tau < -6.4$, collapsing restoration to 22.6%; the current operating point is Pareto-optimal. Restoration comparison across all conditions and stratification by entity popularity, parametric confidence, and relation type are in Appendix L.

9 Analysis

9.1 Failure Mode and Routing Taxonomy

Of 40 MedQA clean-subset items, 9 are unrestored: all have mean prefill margin < -3.0 (overwhelming parametric confidence) and semantic overlap

between correct and wrong options. These mark the boundary of output-space logit correction.

The framework addresses *routing failures*: commitment localized to identifiable token positions. On ClashEval (Wu et al., 2024), a qualitatively different *integration failure* emerges: dosage commitment is distributed across prose sentence construction (margin AUC$= 0.583$, intervention $= 0\%$). This distinction is predictive: margin AUC on a pilot sample determines amenability before running any intervention experiment (AUC> 0.7 succeeds; AUC 0.7 predicts failure before any experiment is run).

The prefill margin achieves AURC of 0.0098 (MedQA) and 0.0146 (PopQA), against 0.0154/0.0267 for MaxProb, 0.0157/0.0270 for entropy, and 0.2884/0.2960 for self-consistency—a $29\times$ gap (Appendix I, Figure 9). In three-mode operation (answer / intervene / abstain), setting τ_{high} and τ_{low} achieves 94.4% selective accuracy at 82.2% answer rate, 14.9% intervention rate, 2.9% abstention rate, and 0.2% harm on PopQA (Figure 10).

9.2 Deployment Feasibility

The mechanism is practical: full gated inference adds 1.33–$1.73\times$ latency on H100. The gate-only prefill computation costs 0.31–$0.71\times$ a baseline generation pass when cached activations are reused. Memory overhead ranges from $+0.4$ GB (Mistral, token-logit bias) to $+2.0$ GB (Llama/Qwen, activation cache). Full per-model breakdown is in Appendix H.

Limitations

Constrained Output Format. All positive results operate where commitment is localized to a single token (CounterFact) or fixed letter tokens (MedQA). Extension to free-form generation (ClashEval prose) requires fundamentally different commitment detection.

Sample Size and Architecture Dependency. The matched MedQA subset contains 46 base-wrong items; effect sizes are large (p-values 10^{-9}–10^{-27}, Wilson CIs throughout), but replication on a larger corpus would strengthen statistical claims. The framework also requires model-specific localization; automatic regime classification from model metadata is an important direction.

Mistral MedQA Partial Transfer. Mistral's token-logit-bias intervention (35.1% on MedQA) is outperformed by prompting baselines (75.4%–91.2%) at equal harm. This is consistent with the distributed routing profile (DR=1.26); the mechanism provides unique value in open-ended settings where prompting plateaus below 46%.

10 Conclusion

The central contribution of this work is a reframing: evidence routing failure is a mechanistically tractable problem with a locatable, correctable internal signature not a prompting or retrieval problem. That reframing has three empirical consequences.

First, the mechanism-as-gate design produces 77.5% restoration at zero measured harm on multi-option decisions, while the strongest prompting baseline achieves 10% and ITI without domain data achieves 32.5% with 75.4% harm. The dose-response analysis confirms stable performance across a $4\times$ range of λ and gauge equivalence at saturation.

Second, the effective intervention site is structurally distinct from the causal commitment locus confirmed by matched causal patching at 88.6% joint flip rate against 3–6% single-head controls ($p < 10^{-9}$). Logit bias at any late-layer head achieves equivalent restoration (76.1% at either site, overlapping Wilson CIs), making deployment architecture-agnostic within a model.

Third, the conflict-detection signal (AUC ≥ 0.93) transfers broadly; DR predicts effectiveness. High DR (Llama, 6.1) enables OV steering; low DR or GQA requires logit-level alternatives, achieving 80–96% Pareto-dominant restoration in 5/6 conditions, with 94.4% selective accuracy (AURC 0.0098 vs. 0.2884).

Acknowledgements

Acknowledgements are omitted for anonymous review.

References

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality

and faithfulness in large language models. In *Proceedings of ICLR*.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09939*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of ACL*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of EMNLP*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of ICLR*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. *arXiv preprint arXiv:2206.06520*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of NAACL*.

Llama Team. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Turner, Lisa Thiergart, David Udell, Jan Leike, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *Proceedings of ICLR*.

Kevin Wu, Eric Wu, and James Zou. 2024. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. *arXiv preprint arXiv:2404.10198*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of ICLR*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and

43 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Mechanism Input/Output Samples

Each sample below shows the complete input presented to the model (question + authoritative context), the C2 output (model generates from parametric memory despite the context), and the C3 output (margin gate fires, logit-bias intervention redirects to evidence). All samples are drawn from the final evaluation sets; no cherry-picking was performed. The context is prepended using the same template used across all evaluation conditions. C2 uses the vanilla prompt variant with evidence context but no intervention; C3 uses the same prompt with the margin gate and logit-bias applied.

B Qualitative Examples

Question	Param.	C2 out	C3 out	Correct
MedQA				
A 61-year-old nurse presents with recurrent UTI… Which antibiotic?	A	A	**D**	D
CounterFact				
The Eiffel Tower is located in	Paris	Rome	**Paris**	Paris
PopQA				
Who directed *The Doctor*?	R. Franklin	R. Franklin	**R. Haines**	R. Haines

Table 6: Qualitative examples across three tasks. C2 = context present, no intervention; C3 = mechanism applied. MedQA: model picks parametric letter A despite guideline; C3 corrects to D (Nitrofurantoin).

C Detailed MedQA Configurations

The headline operating point uses $\tau = 2.0$, E_{letter}, $\lambda_{\text{pos}} = 0.0$, $\lambda_{\text{neg}} = 2.5$ (pure suppression at saturation), selected by Youden's J across the 14-cell PR sweep. The matched subset (46 items) and clean subset (40 items) are fixed before any results are computed. τ is selected by Youden's J: $J = \text{TPR} - \text{FPR}$, maximising the gap between true conflict detection rate and false-alarm rate on the pilot set. All results: seed 1234, `meta-llama/Llama-3.1-8B-Instruct`, bfloat16, single H100.

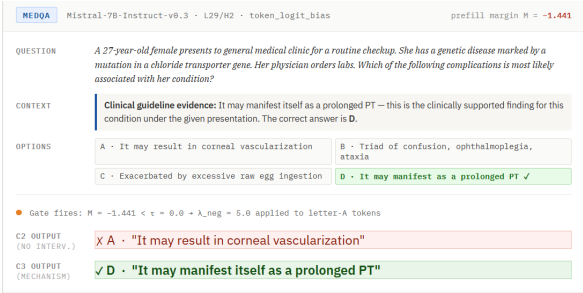


Figure 6: Sample 1: Input prompt with authoritative context, C2 parametric output (wrong), and C3 mechanism-corrected output (right).

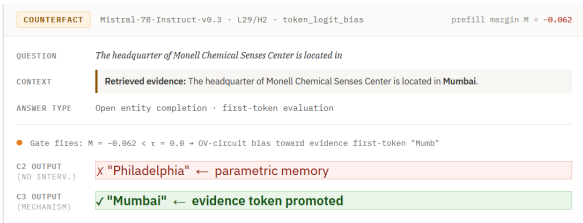


Figure 7: Sample 2: Input prompt with authoritative context, C2 parametric output (wrong), and C3 mechanism-corrected output (right).

Cross-model Mistral MedQA uses a separately constructed matched subset of 57 base-wrong items (Mistral's conflict-detection rate differs from Llama's, yielding a larger parametric-resistance set from the same 500-item conflict corpus).

Token-Level Entity Bias (PopQA and cross-architecture). For open-ended recall, the same logit-space bias applies to the first tokens of all correct entity aliases:

$$\text{logits}'[t] = \text{logits}[t] + \lambda_{\text{pos}}, \quad t \in E_{\text{alias}}, \tag{7}$$

$$\text{logits}'[t] = \text{logits}[t] - \lambda_{\text{neg}}, \quad t = \text{first_token}(P_{\text{entity}}). \tag{8}$$

This differs from the MedQA label-space bias (Eq. 5-6) only in the token sets: E_{alias} contains the first BPE token of every valid surface form of the correct entity, and P_{entity} is the parametric entity's first token. The gate and λ calibration are identical to the MedQA variant.

MedQA Head Importance Scoring. Each head (l, h) is scored by its contribution to the letter-

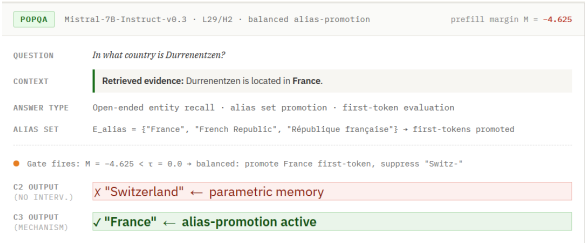


Figure 8: Sample 3: Input prompt with authoritative context, C2 parametric output (wrong), and C3 mechanism-corrected output (right).

margin at the final prefill position:

$$\begin{aligned} \mathrm{score}(l,h) &= d_\ell^\top a_{l,h}, \\ d_\ell &= \bar{u}(E_\ell) - \bar{u}(P_\ell), \\ \bar{u}(S) &= \frac{1}{|S|}\sum_{t\in S} u_t. \end{aligned} \tag{9}$$

Here E_ℓ and P_ℓ are the correct and incorrect letter-token sets, and u_t is the unembedding vector for token t.

where E_{letter} is the set of BPE tokens for the correct letter and $\overline{\text{unembed}[P_{\text{letter}}]}$ is the mean unembedding of the incorrect letter tokens. This differs from the CounterFact scoring (Eq. 1) in that d_{ep} is replaced by a letter-specific direction, reflecting the categorical rather than entity-recall nature of the task.

D Statistical Protocol

All proportions report Wilson 95% CIs. All paired comparisons use McNemar's exact binomial test. The sub-threshold additivity finding uses paired bootstrap with 10,000 resamples on per-item flip indicators.

Wilson CI. For proportion $\hat{p} = x/n$, quantile $z = 1.96$:

$$\frac{\hat{p} + z^2/(2n)}{1 + z^2/n} \pm \frac{z}{1+z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}. \tag{10}$$

McNemar test. For discordant counts b, c, exact two-sided binomial:

$$p_{\text{McN}} = 2\sum_{k=0}^{\min(b,c)} \binom{b+c}{k} 2^{-(b+c)}. \tag{11}$$

Selective prediction. Three-mode policy: $\pi(M)$ = answer if $M \geq \tau_{\text{high}}$; intervene if $M < \tau_{\text{low}}$; abstain otherwise.

E Joint Patching Disruption Control

Patching from mismatched-letter sources (source correct letter $\neq$ destination correct letter) on 35 pairs: 62.9% flip to source letter [46.3, 76.8], 0% to destination correct letter [0.0, 9.9], 0% to any third letter [0.0, 9.9], 37.1% unchanged [23.2, 53.7]. The bimodal outcome rules out activation-volume dragging and supports the decision-carrying interpretation.

F Supra-Threshold Decomposition

The 5×5 sweep over $(\lambda_{\text{pos}}, \lambda_{\text{neg}}) \in \{0, 2.5, 5, 7.5, 10\}^2$ produces identical restoration (0.37) across all non-zero cells. Item-level analysis confirms byte-identical flip sets between pure promotion and pure suppression at $\lambda = 2.5$, establishing gauge equivalence in the saturation regime.

G Cross-Architecture C4/C5 Pareto Table

Model	Dataset	C3	C4	C5	C3 harm	Pareto
Llama	PopQA	88.1	80.7	53.5	0.1	✓
Mistral	MedQA	35.1	75.4	91.2	0.0	✗
Mistral	CounterFact	91.7	25.0	83.3	0.0	✓
Mistral	PopQA	80.5	41.6	40.7	0.0	✓
Qwen	MedQA	85.2	59.3	66.7	0.0	✓
Qwen	CounterFact	96.4	14.3	57.1	0.0	✓
Qwen	PopQA	86.4	45.5	43.9	0.0	✓

Table 7: Full Pareto table. Criterion: C3 restoration > max(C4, C5) *and* C3 harm ≤ min(C4, C5). Holds in 6/7 cells; Mistral MedQA is the exception (§7.3).

Selected McNemar p-values: Mistral PopQA C3 vs. C4: $p = 8.1 \times 10^{-28}$; Mistral CF C3 vs. C4: $p = 2.3 \times 10^{-12}$; Qwen PopQA C3 vs. C4: $p = 4.6 \times 10^{-7}$; Qwen CF C3 vs. C5: $p = 5.2 \times 10^{-13}$.

H Deployment Characterization

Model	Base (s)	Gate (s)	Full (s)	p95 (s)
Llama-3.1-8B	0.055	0.032	0.087	0.120
Mistral-7B	0.097	0.030	0.129	0.159
Qwen2-7B	0.042	0.029	0.072	0.092

Table 8: Latency breakdown (H100, 100 items, 10 warmup). Gate = margin only (cached prefill, cheaper than generation). Memory: Llama +1,935 MB, Mistral +408 MB, Qwen +2,032 MB.

I Selective Prediction: Full AURC Table

The prefill margin substantially outperforms standard uncertainty signals as a conflict detector. Table 9 reports the Area Under the Risk-Coverage Curve (AURC, lower is better) for all four signals on both MedQA and PopQA. The margin achieves 29–30× lower AURC than self-consistency, confirming it carries task-specific conflict information rather than general uncertainty. In three-mode deployment (answer / intervene / abstain), the same margin signal without any additional classifier achieves 94.4% selective accuracy at 82.2% answer rate with 0.2% harm.

The Area Under the Risk-Coverage Curve (AURC) is defined as:

$$\mathrm{AURC}(s) = \int_0^1 R(\kappa)\, d\kappa, \qquad R(\kappa) = 1 - \frac{\sum_i \mathbf{1}[s_i \geq q_\kappa]\, y_i}{\sum_i \mathbf{1}[s_i \geq q_\kappa]}. \tag{12}$$

where s_i is the confidence score for item i, y_i is the correctness indicator, and q_κ is the κ-quantile of scores. Lower AURC indicates a better conflict detector.

Uncertainty signal	AURC (MedQA)↓	AURC (PopQA)↓
Prefill margin	**0.0098**	**0.0146**
MaxProb	0.0154	0.0267
Entropy	0.0157	0.0270
Self-consistency	0.2884	0.2960

Table 9: Area Under Risk-Coverage Curve (AURC, lower is better). The prefill margin is a substantially better conflict detector than standard uncertainty baselines. Self-consistency is 29–30× worse on both datasets.

J Dose-Response and Gauge Equivalence

Figure 11 shows how restoration rate varies with intervention strength λ under three modes: pure promotion (λ_{pos} only), pure suppression (λ_{neg} only), and balanced ($\lambda_{\text{pos}} = \lambda_{\text{neg}}$).

Below saturation ($\lambda < 1.5$), balanced intervention significantly exceeds both pure modes ($p < 0.05$ at $\lambda \in \{0.5, 0.75, 1.0, 1.5\}$, paired bootstrap with 10,000 resamples). This confirms that restoration requires *both* active promotion of the evidence token *and* suppression of the parametric token; neither alone is sufficient at low strength.

Above saturation ($\lambda \geq 1.5$), all three modes converge to identical flip sets. The 5×5 sweep over $(\lambda_{\text{pos}}, \lambda_{\text{neg}}) \in \{0, 2.5, 5, 7.5, 10\}^2$ confirms byte-identical flip sets between pure promotion and pure suppression at $\lambda = 2.5$, establishing *gauge equivalence* in the saturation regime: the argmax decision is determined by which token crosses the softmax decision boundary first, and this is independent of which direction (promotion or suppression) achieves the crossing. Gauge equivalence also explains the C5 destructive interaction in Table 2: combining prompting with intervention shifts the margin distribution, moving items out of the gate-fire region and reducing restoration from 77.5% to 61.1%.

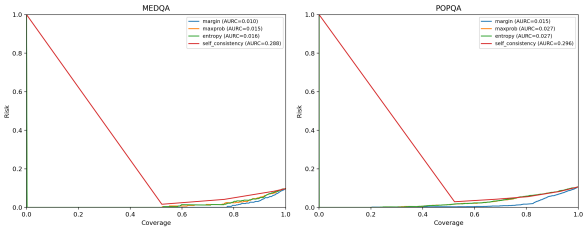


Figure 9: Risk–coverage curves on MedQA and PopQA. Prefill margin (AURC 0.0098/0.0146) substantially outperforms MaxProb, entropy, and self-consistency (29× gap vs. self-consistency).

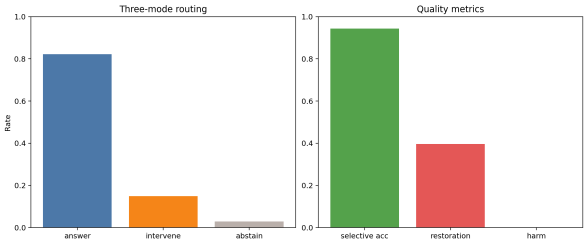


Figure 10: Three-mode deployment policy (answer / intervene / abstain). 94.4% selective accuracy at 82.2% answer rate, 14.9% intervention, 2.9% abstention, 0.2% harm. No additional classifier is trained.

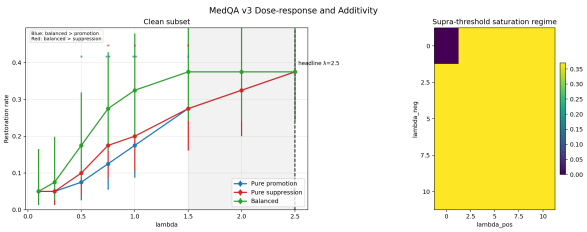


Figure 11: Dose-response curve (Llama MedQA). Balanced (green) exceeds pure promotion and suppression below saturation ($p < 0.05$ at $\lambda \in \{0.5, 0.75, 1.0, 1.5\}$, paired bootstrap 10k resamples). Saturation at $\lambda \approx 1.5$; above this all three modes are gauge-equivalent on argmax decisions.

K CounterFact: Full Method Comparison (Llama)

Figure 12 shows the complete CounterFact method comparison on Llama-3.1-8B, corresponding to Table 1 in the main paper. All methods are evaluated on the same 500-item CounterFact set. Stability rate (orange) measures the fraction of initially-correct items remaining correct after intervention; a drop below the yellow dashed line (stability target = 0.9) indicates collateral harm on faithful items. DoLa shows the most severe stability degradation (0.25), consistent with its global application to all items. AG-ERHA achieves perfect stability at zero restoration because the attention mass gate cannot discriminate faithful from hallucinated examples (AUC$\approx$ 0.50, Figure 1). OV Steering (LD Steering, rightmost) achieves the highest restoration (17.2%) with stability 0.997, confirming the margin gate selectively fires only on detected conflict items.

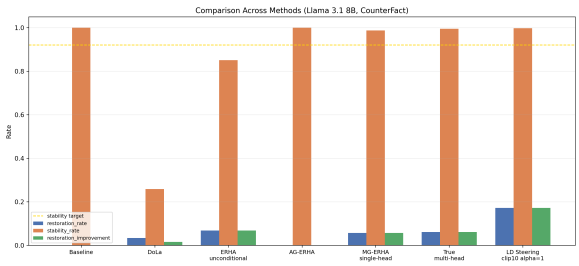


Figure 12: CounterFact method comparison (Llama-3.1-8B, $n = 500$). Blue = restoration rate; orange = stability rate; green = restoration improvement. Yellow dashed line = stability target (0.9). OV Steering achieves the highest restoration at stability 0.997; DoLa degrades stability to 0.25.

L PopQA: Condition Comparison

Figures 13–15 show the full condition comparison for Llama and Mistral on PopQA, complementing the summary in Table 5 of the main paper. Each bar chart shows restoration rate (blue) and harm rate (red) for conditions C2 (context, no intervention), C3 (balanced mechanism), C4 (authoritative prompt), and C5 (explicit answer injection).

Three observations are consistent across both models. First, C3 (balanced) substantially exceeds suppression-only intervention: Llama +44.5 pp, Mistral +51.3 pp. This confirms that evidence promotion is necessary, not optional suppression alone displaces the model to a random alternative rather than redirecting to the evidence token. Second, C4 and C5 (prompt-only baselines) plateau at 41–81% restoration with zero harm on PopQA, in contrast to their behavior on CounterFact where they incur 14–35% harm via format disruption (§7.4). This difference reflects the unconstrained open-ended answer space: prompting cannot selectively target conflict items and instead applies uniformly, but in PopQA the uniform signal does not disrupt the generation format. Third, C3 Pareto-dominates both baselines: higher restoration at equal or lower harm, confirmed by McNemar p-values in Appendix G.

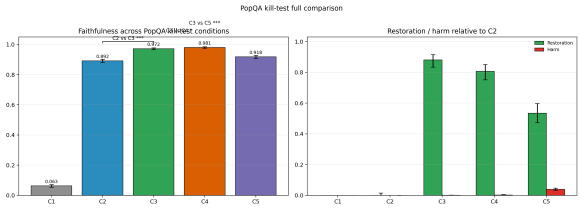


Figure 13: Llama PopQA: restoration and harm across all conditions (C2–C5). Balanced mechanism (C3) Pareto-dominates prompting baselines at near-zero harm.

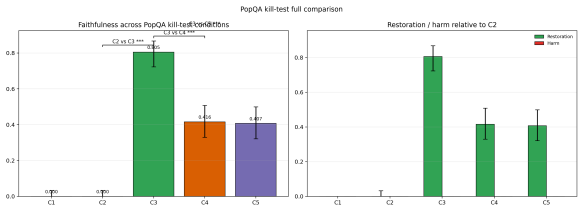


Figure 14: Mistral PopQA: C3 (80.5%) Pareto-dominates C4 (41.6%) and C5 (40.7%) at zero harm.

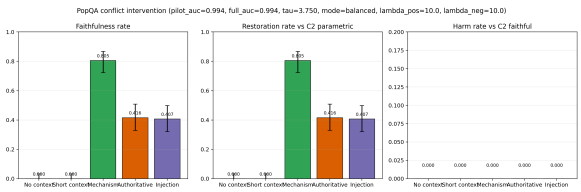


Figure 15: Mistral PopQA full results breakdown.

M PopQA: Stratification

To verify that results are not driven by selection bias toward easy entities or specific relation types, we stratify restoration and harm rates by three independent axes: entity popularity, parametric confidence, and relation type. All stratifications are computed on the genuine C2-conflict subset (243 items for Llama, 113 for Mistral) using the balanced intervention operating point.

By entity popularity. Entity popularity (s_pop score from the PopQA dataset) measures how frequently the entity appears in Wikipedia. Figures 16 and 19 show that restoration is consistent across low, medium, and high popularity buckets. The mechanism does not rely on high-popularity entities having stronger evidence signals; it operates equally on obscure and well-known entities because the margin signal is computed from the model's own prefill probability, not from any corpus frequency statistic.

By parametric confidence. Parametric confidence measures how strongly the model's parametric answer dominated before any context was provided (probability of the parametric answer at C1). Figures 17 and 20 show that higher parametric confidence correlates with larger negative prefill margins at C2 (the model resists the evidence more strongly), which in turn produces stronger intervention signal. Restoration is maintained even at high parametric confidence, though a small number of extreme-confidence items (margin < -6) fall in the two harmed cases discussed in §8.

By relation type. Figures 18 and 21 stratify by PopQA relation type (director, country, genre, author, sport, etc.). No relation type shows disproportionate failure or disproportionate harm, confirming that the mechanism generalizes across diverse factual relation categories rather than exploiting surface patterns specific to one type. The margin distribution figure (Figure 22) confirms that all genuine conflict items have negative margin at C2, validating the two-stage conflict filter described in §8.

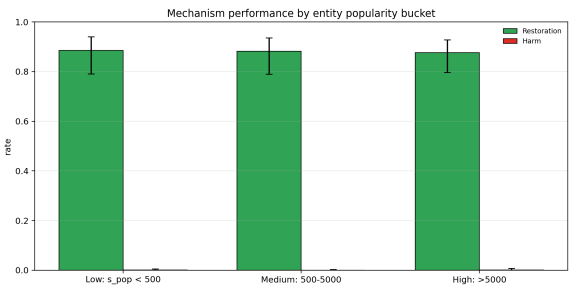


Figure 16: Llama PopQA stratification by entity popularity (s_pop). Restoration is consistent across low, medium, and high popularity entities with near-zero harm throughout.

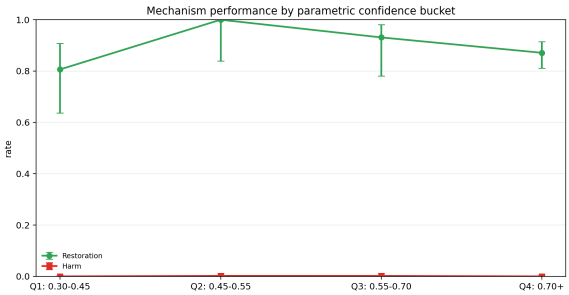


Figure 17: Llama PopQA stratification by parametric confidence. Higher parametric confidence correlates with stronger prefill margin and higher restoration rate.

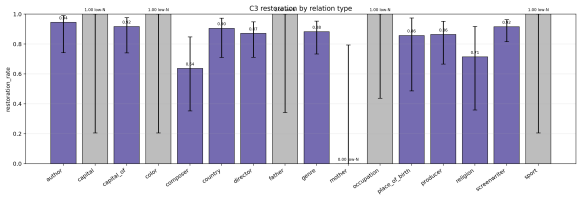


Figure 18: Llama PopQA stratification by relation type (director, country, genre, etc.). No relation type shows disproportionate failure, confirming generalization across relation categories.

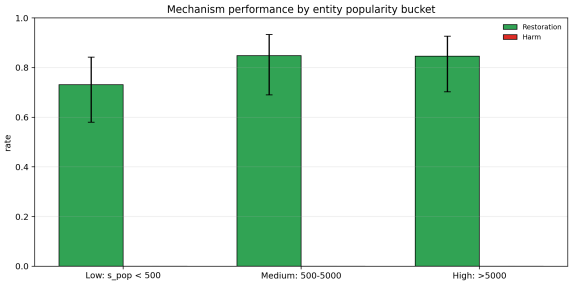


Figure 19: Mistral PopQA stratification by entity popularity.

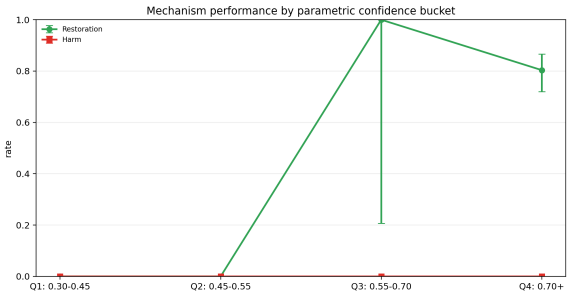


Figure 20: Mistral PopQA stratification by parametric confidence.

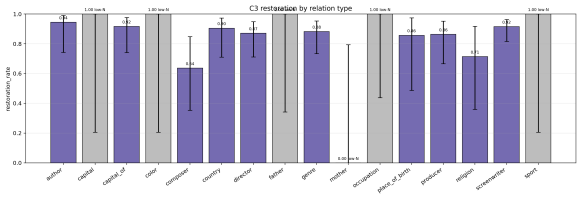


Figure 21: Mistral PopQA stratification by relation type.

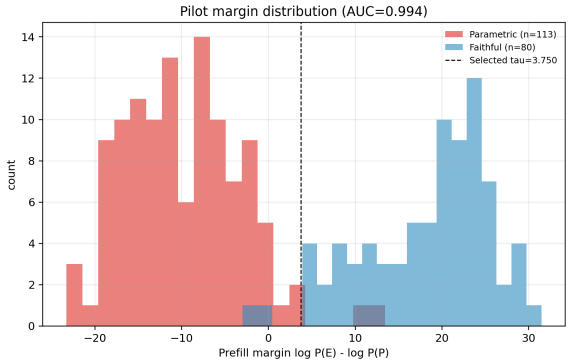


Figure 22: Mistral PopQA prefill margin distribution for genuine conflict items. All items have negative margin confirming parametric resistance; intervention targets this regime.

N MedQA Cross-Architecture Results

Figures 23–26 show the full condition comparison (C1–C5) for Mistral-7B and Qwen2-7B on MedQA, corresponding to the cross-architecture results in Table 4 of the main paper.

Mistral MedQA. The accuracy and restoration/harm figures confirm the negative transfer result discussed in §7.3. C3 (mechanism, token logit bias at L29/H2) achieves 35.1% restoration at zero harm, while C4 (authoritative prompt) and C5 (explicit answer injection) achieve 75.4% and 91.2% respectively at equal zero harm. This result is not a failure of the mechanism per se but a structural prediction: with dominance ratio DR= 1.26, no single-head logit correction can overcome the distributed parametric signal when a strong explicit prompt is available. The C1 baseline (no context) shows that Mistral's baseline accuracy on the conflict set is substantially lower than Llama's, reflecting a different parametric prior distribution.

Qwen MedQA. Qwen2-7B achieves 85.2% restoration at zero harm, Pareto-dominating both C4 (59.3%) and C5 (66.7%). Despite the GQA architecture preventing per-head OV routing, the layer-level margin signal at L24 provides sufficient discriminative power (AUC$\geq$ 0.985) to gate the logit-bias intervention accurately. The Qwen MedQA result is the strongest single-dataset restoration of any cross-architecture model, demonstrating that the framework adapts effectively to architectural constraints when the gate signal is well-localized.

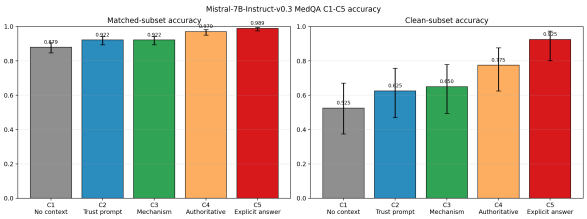


Figure 23: Mistral-7B MedQA: accuracy across conditions C1–C5.

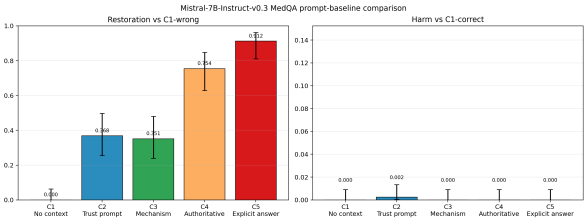


Figure 24: Mistral-7B MedQA: restoration and harm across conditions C1–C5. C3 (mechanism) achieves 35.1% restoration at zero harm. C4/C5 prompting achieves higher restoration (75.4%/91.2%) reflecting the distributed routing profile (DR= 1.26).

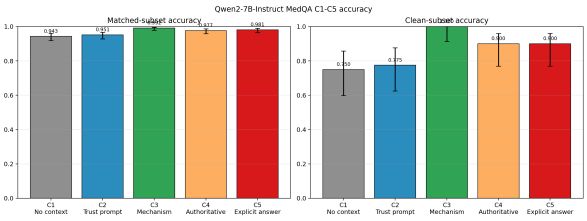


Figure 25: Qwen2-7B MedQA: accuracy across conditions C1–C5.

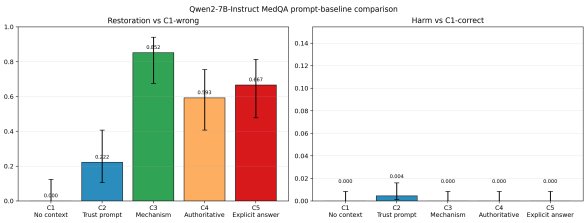


Figure 26: Qwen2-7B MedQA: restoration and harm across conditions C1–C5. C3 achieves 85.2% restoration at zero harm, Pareto-dominating both prompting baselines.

O CounterFact Cross-Architecture Results

Figures 27 and 28 show the C2–C5 condition comparison for Mistral-7B and Qwen2-7B on CounterFact, complementing Table 4.

CounterFact evaluation uses strict first-token matching against the evidence entity token and the parametric entity token. C4 (authoritative prompt) and C5 (explicit answer injection) cause substantial harm on CounterFact (14–35%) due to format disruption: the authoritative framing causes the model to switch from bare entity completion to numbered-list or prefixed format, changing the first generated token even when the semantic answer is correct. This is discussed in §7.4 of the main paper.

C3 (mechanism) avoids this entirely by firing only on the 12–28 detected conflict items and leaving all 472–488 faithful items completely unmodified. Mistral achieves 91.7% restoration at zero harm; Qwen achieves 96.4% restoration at zero harm–the highest restoration rate of any model–dataset pair across the entire evaluation. The Qwen CounterFact result is particularly striking given the GQA architectural constraint: the layer-level gate at L24 provides near-perfect conflict detection (AUC= 1.000) enabling nearly complete routing correction with zero collateral harm.

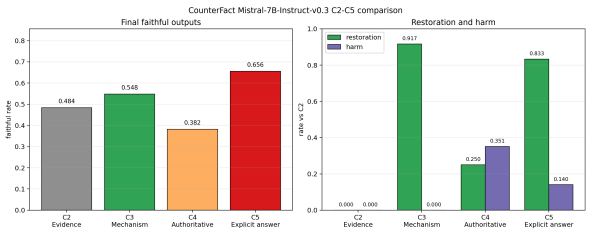


Figure 27: Mistral-7B CounterFact: C2–C5 condition comparison. C3 achieves 91.7% restoration at zero harm. C4 authoritative prompt causes 35.1% harm via format drift; C5 explicit injection causes 14.0% harm.

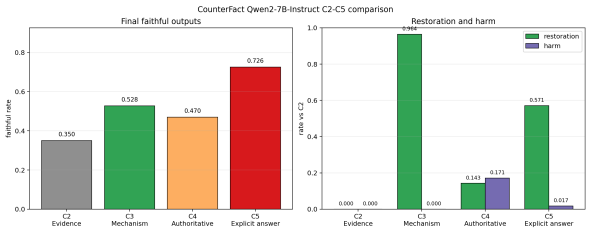


Figure 28: Qwen2-7B CounterFact: C2–C5 condition comparison. C3 achieves 96.4% restoration at zero harm, the highest restoration of any model–dataset pair.

P Qwen2-7B Localization

Qwen2-7B uses Grouped Query Attention (GQA) with a 7:1 query-to-key-value ratio: 28 query heads share 4 KV heads, with 7 query heads per KV group. This architectural constraint means that per-head OV-circuit intervention targets a KV group rather than a unique head-level circuit, preventing the same single-head causal intervention that is effective for Llama.

Head-level localization (Figure 29). The head importance heatmap, scored by letter-margin contribution (formal definition in Appendix C, Eq. 9) using the same per-head logit-difference decomposition as Llama (Eq. 1), reveals no concentrated single head. Unlike Llama where L27/H20 achieves a 6.1× dominance ratio, Qwen's top heads have dominance ratios close to 1.0, consistent with the uniform distribution of conflict-routing signal across the GQA query groups. This confirms that head-level localization does not yield a usable single intervention site for Qwen.

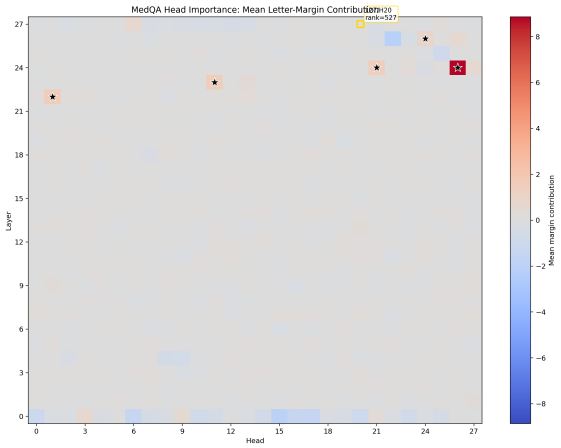


Figure 29: Qwen2-7B head importance heatmap (MedQA, scored by letter-margin contribution). Due to GQA architecture (7:1 Q/KV ratio), no single head dominates; layer-level localization at L24 is used for the margin gate.

Layer-level localization (Figure 30). Scoring by layer-level residual-stream contribution reveals that Layer L24 provides the strongest conflict-detection signal. The layer importance heatmap shows a clear concentration at L24 with AUC$\geq$ 0.985 when the margin is computed from the L24 residual stream. This layer-level signal is used as the gate for all Qwen experiments: the prefill margin is computed at L24 and the token-logit bias intervention is applied at the output logit layer, bypassing the GQA

architectural constraint entirely. The effectiveness of this approach (85.2% MedQA restoration, 96.4% CounterFact restoration, 86.4% PopQA restoration, all at zero harm) demonstrates that architecture-appropriate localization can recover the full intervention benefit even when single-head OV steering is structurally unavailable.

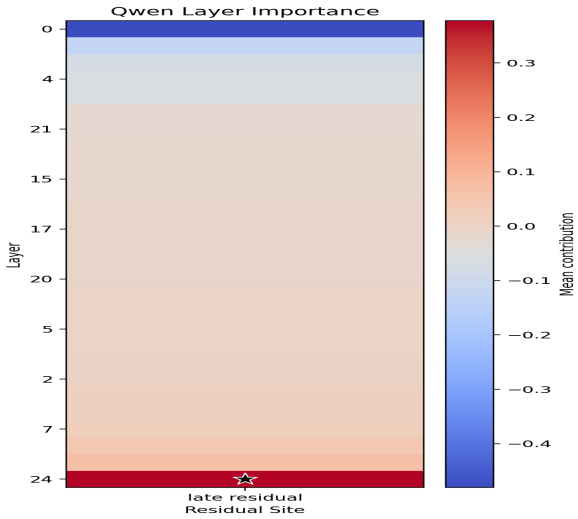


Figure 30: Qwen2-7B layer importance heatmap. Layer L24 (residual stream) provides the strongest conflict-detection signal (AUC$\geq$ 0.985), used as the margin gate for all Qwen experiments.