Awakening Dormant Experts: Counterfactual Routing to Mitigate MoE Hallucinations

Wentao Hu[1,2,*,†], Yanbo Zhai[1,*], Xiaohui Hu[2], Mingkuan Zhao[1,‡], Shanhong Yu[4]
Xue Liu[1,2], Kaidong Yu[2], Shuangyong Song[2,‡], Xuelong Li[3,‡]

[1]Xi'an Jiaotong University
[2]Xingchen AGI Lab, China Telecom AI Technology (Beijing) Co., Ltd.
[3]Institute of Artificial Intelligence, China Telecom [4]Beijing Foreign Studies University

{wentao_hu, yanbozhai, mingkuanzhao, LiuXue1087}@stu.xjtu.edu.cn
{huxh12, yukd, songshy}@chinatelecom.cn yushanhong@bfsu.edu.cn xuelong_li@ieee.org

Abstract

Sparse Mixture-of-Experts (MoE) models have achieved remarkable scalability, yet they remain vulnerable to hallucinations, particularly when processing long-tail knowledge. We identify that this fragility stems from static Top-k routing: routers tend to favor high-frequency patterns over rare factual associations. Consequently, "specialist experts" possessing critical long-tail knowledge are often assigned low gating scores and remain "dormant"—underprioritized for specific tokens despite their proven causal importance on other inputs. To address this, we propose Counterfactual Routing (CoR), a training-free inference framework designed to awaken these dormant experts. CoR integrates layer-wise perturbation analysis with the Counterfactual Expert Impact (CEI) metric to dynamically shift computational resources from syntax-dominant to knowledge-intensive layers while maintaining a constant total activation count, effectively retrieving causally decisive experts via virtual ablation. Extensive experiments on TruthfulQA, FACTOR, and TriviaQA demonstrate that CoR improves factual accuracy by 3.1% on average without increasing the inference budget, establishing a superior Pareto frontier compared to static scaling strategies. Code is available at https://github.com/ZhaiYanbo/CoR.

1 Introduction

Mixture-of-Experts (MoE) architectures have become the dominant paradigm for scaling Large Language Models (LLMs) by decoupling parameter count from inference cost (Shazeer et al., 2017). By routing each token to only a subset of experts, MoE models scale to hundreds of billions of parameters while maintaining computational efficiency comparable to much smaller dense models. For instance, Qwen-3-30B-A3B (Yang et al., 2025) activates only 3B of its 30B parameters per token, outperforming dense counterparts with significantly less computation. Similarly, the TeleChat model family has progressed from dense architectures (He et al., 2024; Wang et al., 2024, 2025; Li et al., 2024b,a) to the TeleChat3-MoE series (Liu et al., 2025), scaling to over one trillion parameters with a high-sparsity MoE architecture, further demonstrating the trend of MoE as the dominant scaling paradigm. Despite these efficiency advantages, MoE models remain susceptible to hallucinations—generating plausible but factually incorrect content (Ji et al., 2023; Huang et al., 2025)—particularly for long-tail entities where accuracy degrades sharply (Kandpal et al., 2023; Mallen et al., 2023).

Recent studies have identified two factors particularly relevant to MoE architectures. First, long-tail knowledge sparsity: LLMs primarily capture frequent patterns, struggling with rare facts due to insufficient training signal (Sun et al., 2024; Kandpal et al., 2023). Kalai and Vempala (2024) theoretically proved that hallucinations are inevitable when models generalize beyond training coverage. Second, spurious correlations: models learn misleading pattern associations that produce plausible-sounding but factually incorrect content when queried about unfamiliar entities (Seitzer et al., 2022; Zhang et al., 2024).

These issues manifest distinctively in MoE models through the routing mechanism. Standard Top-k routing is trained jointly with expert parameters using auxiliary load balancing losses (Fedus et al., 2022; Lepikhin et al., 2020; Zhou et al., 2022). However, since load balancing encourages uniform expert utilization and high-frequency tokens dominate the training corpus, routers learn to favor frequency-based patterns over rare factual associations. This creates a systematic correlation-causality misalignment: "generalist experts" handling common linguistic features receive high gat-

[*]These authors contributed equally to this work.
[†]Work done during internship at Xingchen AGI Lab.
[‡]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47257–47271
July 2-7, 2026

ing scores, while "specialist experts" harboring long-tail knowledge are assigned low scores. During inference on hard tokens, these specialists become "dormant": functionally capable of providing correct information but receiving low gating scores for the current context, despite proving critical on other inputs. The model "knows" the fact (stored in specialist parameters) but fails to "recall" it (due to routing decisions), resulting in hallucinations.

Existing hallucination mitigation techniques fail to address this routing bottleneck. Training-time approaches such as retrieval-augmented generation (Lewis et al., 2020) require extensive resources and cannot be applied to deployed models. Inference-time interventions like DoLa (Chuang et al., 2024) and ITI (Li et al., 2023) operate on the output distribution or residual stream, attempting corrections after routing decisions have been made and cannot recover information from experts that were never activated.

To address this, we propose Counterfactual Routing (CoR), a training-free inference framework that awakens dormant experts through causal guidance. Our key insight is to distinguish correlation (what the router prefers based on training statistics) from causality (what the prediction actually needs for factual correctness). CoR achieves factuality improvement via compute-preserving expert redistribution—activating the same total number of experts as standard inference—through two complementary mechanisms. At the layer-wise level, recognizing that factual knowledge concentrates in specific layers (Meng et al., 2022; Geva et al., 2023; Dai et al., 2022), we employ perturbation analysis to identify knowledge-intensive layers and dynamically expand their expert budget while reducing allocations to syntax-dominant layers. At the expert-wise level, we introduce Counterfactual Expert Impact (CEI), a causal metric derived from virtual ablation. Unlike correlation-based gating scores, CEI captures causal necessity—identifying experts essential for factual correctness regardless of their router scores. Experts with high CEI but low router scores are precisely the dormant specialists that CoR awakens. Extensive experiments on Qwen-3-30B-A3B (Yang et al., 2025), DeepSeek-V2-Lite (DeepSeek-AI et al., 2024), and GPT-OSS-20B (OpenAI et al., 2025) demonstrate the superiority of our method.

Our contributions are summarized as follows:

- We expose the "Dormant Expert" phenomenon in MoE models, providing empirical evidence showing how static Top-k routing contextually under-prioritizes knowledge-bearing experts for long-tail tokens due to correlation-causality misalignment.
- We propose CoR, a training-free framework achieving compute-preserving expert redistribution through causal-guided resource reallocation across layers and experts.
- Extensive experiments demonstrate that CoR achieves an average improvement of 3.1% on hallucination benchmarks without increasing the inference budget, establishing a superior Pareto frontier compared to static scaling strategies.

2 Related Work

Hallucination Mitigation. Hallucinations in LLMs—generating plausible but factually incorrect content—have been extensively studied (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2025). Mitigation approaches span training-time methods like retrieval-augmented generation (Lewis et al., 2020) and factuality-enhanced pretraining (Lee et al., 2022), as well as inference-time interventions. DoLa (Chuang et al., 2024) contrasts logits from different layers to amplify factual signals, while ITI (Li et al., 2023) shifts activations along truthfulness directions. However, these techniques operate on output distributions or residual streams, essentially polishing the result after the routing decision is made. Notably, existing mitigation methods have been exclusively designed and validated on dense architectures. To our knowledge, no prior work has specifically addressed hallucinations arising from the *routing-level* decisions unique to MoE mechanisms. Our work bridges this gap by intervening directly in the expert selection process. Orthogonally, recent studies have explored enhancing LLM reasoning capabilities through structured thinking paradigms, including computation logic graphs for mathematical reasoning (Zhao et al., 2025a), multi-perspective reasoning with reinforcement learning for information extraction (Li et al., 2025e), table reasoning frameworks leveraging schema-guided decomposition (Xiong et al., 2025), structured reasoning data construction pipelines (Xing et al., 2025), preference-driven methodologies for efficient code generation (Li et al., 2025d), and reinforcement learning-driven optimization strategies

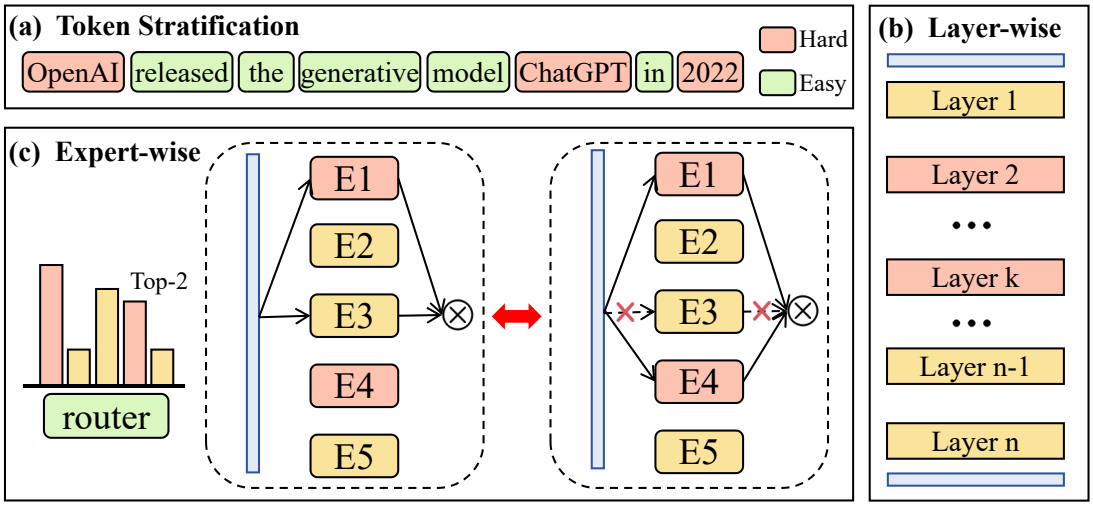


Figure 1: Overview of the Offline Causal Analysis pipeline. The framework consists of three hierarchical stages: (a) Token Stratification: We stratify tokens into hard (knowledge-intensive) and easy (syntax-dominant) subsets based on model uncertainty to disentangle factual reasoning from generic processing. (b) Layer-wise Analysis: We apply Contrastive Sensitivity Normalization to identify knowledge-intensive layers by measuring their relative sensitivity (R_l) to hard tokens while mitigating error cascading. (c) Expert-wise Analysis: We compute CEI via virtual ablation to uncover "dormant" experts—causally critical on some tokens but under-prioritized on others.

for domain-specific tasks (Li et al., 2025c). While these methods improve reasoning quality at the task level, they do not address the routing-level decisions within MoE architectures that our work targets.

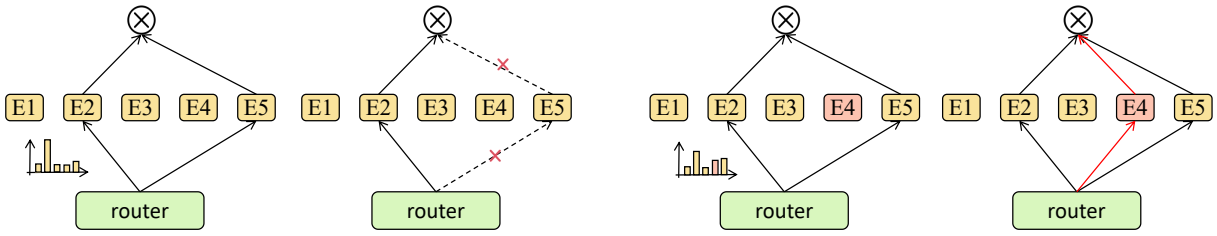
MoE Routing Mechanisms. Sparse MoE models replace dense feedforward layers with expert modules, using learned gating functions to select a subset of experts per token (Shazeer et al., 2017; Fedus et al., 2022). Training requires auxiliary losses for load balancing, as routers otherwise collapse to repeatedly selecting the same experts (Lepikhin et al., 2020). Expert Choice routing inverts selection, having experts choose tokens (Zhou et al., 2022). Recent analyses reveal that experts exhibit temporal locality and syntactic rather than semantic specialization (Jiang et al., 2024). These findings suggest routing learns surface patterns rather than factual associations, motivating our causally-grounded expert selection. Despite extensive work on MoE routing efficiency and load balancing, the connection between routing decisions and factual accuracy remains underexplored. Beyond routing design, complementary directions for improving MoE efficiency have been explored. Hu et al. (2025) propose Mosaic Pruning, a hierarchical framework that preserves functionally diverse experts during structured pruning via a "cluster-then-select" strategy, revealing that expert specialization patterns are critical for downstream generalization. At the attention level, Zhao et al. (2025b) introduce SPAttention, which partitions the attention workload into non-overlapping distance bands across heads, demonstrating that principled structural sparsity can serve as an effective inductive bias without sacrificing performance. Additionally, knowledge distillation has emerged as a complementary paradigm for compressing large-scale models while preserving representational capacity (Li et al., 2025b,a, 2026a), and similarity-guided layer pruning strategies further reduce redundancy in deep architectures (Li et al., 2024c, 2026b). Our work bridges this gap by revealing how standard routing contextually disadvantages knowledge retrieval, and proposes the first causally-grounded intervention specifically targeting this failure mode.

3 Methodology

In this section, we present the Counterfactual Routing (CoR) framework. We first formalize the standard Sparse Mixture-of-Experts (MoE) architecture. We then detail our two-phase approach: an offline causal analysis to identify critical components (Figure 1), and an online inference strategy based on compute-preserving expert redistribution (Figure 2).

3.1 Preliminaries

We consider a Transformer-based MoE model with L layers. Let $\mathbf{h}_{l-1}^{(t)} \in \mathbb{R}^d$ denote the hidden state of



(a) Knowledge-Lean Layer **(b) Knowledge-Dense Layer**

Figure 2: Schematic of Compute-Preserving Expert Redistribution. CoR dynamically reallocates budget based on layer sensitivity. (a) Knowledge-Lean Layer: reduce active experts to save budget. (b) Knowledge-Dense Layer: expand budget and fuse CEI to awaken dormant experts. The total activation count remains constant (*budget-neutral*).

the t-th token at layer $l-1$. The l-th MoE layer consists of N expert networks $\{E_{l,i}\}_{i=1}^{N}$, where each expert is typically a Feed-Forward Network (FFN).

Gating Mechanism. A router (or gating network) G_l determines the contribution of each expert. It typically applies a linear projection followed by a Softmax function:

$$G_l(\mathbf{h}_{l-1}^{(t)}) = \text{Softmax}(\mathbf{h}_{l-1}^{(t)}\mathbf{W}_g) \quad (1)$$

where $\mathbf{W}_g \in \mathbb{R}^{d\times N}$ is the learnable routing weight matrix. Let $g_{l,i}^{(t)}$ denote the gating probability for the i-th expert. To ensure sparsity and computational efficiency, standard MoE employs a Top-k selection strategy. The set of active indices for token t is defined as $\mathcal{A}_l^{(t)} = \text{Top-}k(\{g_{l,i}^{(t)}\}_{i=1}^{N})$, where $k \ll N$.

Sparse Output. The output of the MoE layer, $\mathbf{h}_l^{(t)}$, is the weighted sum of the activated experts' outputs plus the residual connection:

$$\mathbf{h}_l^{(t)} = \mathbf{h}_{l-1}^{(t)} + \sum_{i\in\mathcal{A}_l^{(t)}} g_{l,i}^{(t)} \cdot E_{l,i}(\mathbf{h}_{l-1}^{(t)}) \quad (2)$$

3.2 Offline Causal Analysis

Standard routing mechanisms often fail to capture long-tail knowledge due to spurious correlations learned during training. To rectify this, we perform a comprehensive offline analysis on a calibration dataset to pinpoint where factual knowledge resides (Layer-wise analysis) and which experts are causally necessary (Expert-wise analysis).

3.2.1 Token Stratification

To disentangle knowledge-intensive reasoning from generic syntactic processing, we stratify tokens based on model uncertainty. By performing a forward pass on a calibration dataset (e.g., C4 (Dodge et al., 2021)) using the Original model $\mathcal{M}$, we compute the token-level negative log-likelihood (NLL) loss $\ell^{(t)}$. We define two contrastive subsets based on percentile thresholds: the hard sample set $\mathcal{D}_{\text{hard}} = \{t \mid \ell^{(t)} > \tau_{\text{high}}\}$ representing long-tail knowledge, and the easy sample set $\mathcal{D}_{\text{easy}} = \{t \mid \ell^{(t)} < \tau_{\text{low}}\}$ representing common syntax.

3.2.2 Layer-wise analysis

Standard MoEs allocate a uniform computational budget across all layers. However, we hypothesize that factual knowledge is concentrated in specific deep layers. To locate these layers, we employ Layer-wise Perturbation Analysis.

Crucially, raw perturbation scores are often biased by the "Error Cascading" effect: a small perturbation in shallow layers undergoes successive amplifications through the network's depth (via matrix multiplications and non-linearities), resulting in a spuriously high loss deviation even for semantically insensitive layers. To eliminate this structural bias and isolate true knowledge dependency, we propose Contrastive Sensitivity Normalization.

Raw Sensitivity Calculation. For a specific layer l, we apply a uniform multiplicative perturbation δ to its output $\mathbf{O}_l^{(t)}$ while keeping other layers frozen. We calculate the expected loss degradation for both hard and easy subsets separately:

$$S_l(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{t\in\mathcal{D}}\left[\ell(\mathcal{M}(\tilde{\mathbf{O}}_l^{(t)})) - \ell^{(t)}\right] \quad (3)$$

where $\tilde{\mathbf{O}}_l^{(t)} = (1+\delta)\odot\mathbf{O}_l^{(t)}$ and $\mathcal{D} \in \{\mathcal{D}_{\text{hard}}, \mathcal{D}_{\text{easy}}\}$. Intuitively, $S_l(\mathcal{D}_{\text{easy}})$ captures the layer's inherent structural amplification factor, as easy tokens primarily rely on syntax processing rather than deep knowledge retrieval.

Relative Knowledge Intensity (RKI). To decouple the knowledge component from structural

noise, we define the final layer score R_l as the ratio of hard-to-easy sensitivity:

$$R_l = \frac{S_l(\mathcal{D}_{\text{hard}})}{S_l(\mathcal{D}_{\text{easy}}) + \epsilon} \tag{4}$$

A high R_l indicates that layer l is disproportionately critical for hard tokens compared to easy ones, signifying a Knowledge-Intensive Layer. This score will guide our Layer-wise analysis budget reallocation. We provide a rigorous theoretical derivation demonstrating how R_l cancels out structural depth bias in Appendix A.1.

3.2.3 Expert-wise analysis

Within knowledge-intensive layers, the router may still assign low gating scores to certain experts for specific input tokens, even though these experts prove critical when activated on other hard tokens. We term such contextually suppressed experts as "dormant" and propose Counterfactual Expert Impact (CEI), a causal metric that quantifies an expert's aggregated causal value across the calibration set via virtual ablation.

Specifically, for a target expert e in layer l, we first isolate the subset of hard tokens where this expert was originally activated by the Top-k mechanism, denoted as $\mathcal{T}_{l,e} = \{t \in \mathcal{D}_{\text{hard}} \mid e \in \mathcal{A}_l^{(t)}\}$. To quantify the causal contribution of expert e, we construct a locally counterfactual scenario by explicitly ablating it from the network. To ensure a valid probability distribution during this intervention, the original gating probability mass of expert e is redistributed proportionally among the remaining experts to form a surrogate distribution $\tilde{g}$:

$$\tilde{g}_{l,j}^{(t)} = \begin{cases} 0 & \text{if } j = e \\ \frac{g_{l,j}^{(t)}}{\sum_{m \neq e} g_{l,m}^{(t)}} & \text{if } j \neq e \end{cases} \tag{5}$$

We then compute the counterfactual loss $\tilde{\ell}_{(l,-e)}^{(t)}$ using this ablated routing configuration.

The marginal causal effect of expert e on token t is defined as the performance degradation (loss increase) caused by its removal, termed the "Rescue Gain":

$$\Delta\ell_{l,e}^{(t)} = \tilde{\ell}_{(l,-e)}^{(t)} - \ell^{(t)} \tag{6}$$

A positive $\Delta\ell_{l,e}^{(t)}$ implies that expert e possesses unique knowledge required for token t that cannot be compensated by other generalist experts. Finally, the global CEI score for expert e is derived by aggregating the expected causal effect across all relevant hard tokens:

$$\text{CEI}_{l,e} = \mathbb{E}_{t \in \mathcal{T}_{l,e}} \left[\Delta\ell_{l,e}^{(t)} \right] \tag{7}$$

This metric effectively disentangles an expert's necessity from its popularity (gating frequency), highlighting experts that are critical for correctness despite low routing confidence. Since CEI aggregates causal impact across calibration tokens, an expert with high CEI may still receive a low router score on a *new* input due to context shift—precisely the "dormant" scenario where CoR intervenes.

3.3 Compute-Preserving Inference

During the inference phase, we dynamically adjust the routing strategy based on the precomputed layer sensitivity $\{R_l\}$ and expert criticality $\{\text{CEI}_{l,e}\}$. Our approach operates on the principle of compute-preserving expert redistribution, ensuring that the total computational cost remains constant while maximizing factual accuracy. The inference process is depicted in Figure 2.

Adaptive Budgeting via Reallocation. Instead of a uniform Top-k across all layers, we reallocate the total computational budget $K_{\text{total}} = L \times k_{\text{baseline}}$ across layers proportional to their Relative Knowledge Intensity R_l. The number of active experts k_l for layer l is calculated as:

$$k_l = \text{Round}\left(K_{\text{total}} \cdot \frac{R_l}{\sum_{j=1}^{L} R_j} \right) \tag{8}$$

This strategy allocates more experts to deep, knowledge-intensive layers to enhance reasoning capabilities, while pruning activations in shallow, syntax-dominant layers.[1] Crucially, the constraint $\sum k_l \approx K_{\text{total}}$ ensures that the overall FLOPs count remains equivalent to the standard baseline.

Awakening Routing via Context-Prior Fusion. Once the layer-wise budget k_l is determined, we rely on a hybrid mechanism to select specific experts. Relying solely on the static CEI score would ignore the dynamic context of the current input token (e.g., syntax structure), potentially harming general fluency. Conversely, relying solely on the original router $G_l(x)$ leads to the dormant expert problem. Therefore, we select the Top-k_l experts by fusing the *dynamic context* with the *causal prior*, as illustrated in Figure 3. The final selection score for expert i in layer l is defined as:

[1] Here, k denotes the fixed number of active experts employed in the standard Top-k routing baseline.

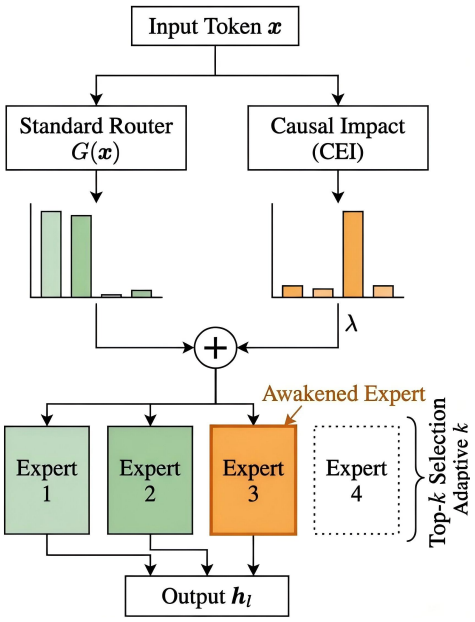


Figure 3: Illustration of Causal-Guided Expert Awakening. This diagram visualizes the fusion mechanism in Eq. 9. The *Standard Router* (left) suppresses the critical expert (Expert 3) due to frequency bias. CoR injects the *Causal Impact* (right) as a prior, boosting the fused score to "awaken" the dormant specialist.

$$\text{Score}_{l,i}(x) = \underbrace{G_l(x)_i}_{\text{Context}} + \lambda \cdot \underbrace{\text{CEI}_{l,i}}_{\text{Prior}} \tag{9}$$

where λ is a hyperparameter controlling the intervention strength. Note that to align the scales, we apply Min-Max Normalization to the raw CEI_l scores within each layer before fusion. This formulation acts as a causal rectification: for general tokens, the high magnitude of $G_l(x)$ dominates, preserving the model's linguistic capabilities; for hard knowledge tokens where the router is uncertain, the high CEI acts as a bias term to "awaken" the specialist expert.

Complexity and Robustness Analysis. We briefly analyze the overhead and hyperparameter sensitivity of our framework. In terms of offline efficiency, calculating CEI involves virtual ablation, which is a computationally intensive but *one-off offline process* performed on a small calibration subset (e.g., $|\mathcal{D}| \approx 1,000$ tokens). Crucially, this pre-computation incurs zero additional latency during online inference since the causal scores are cached. Regarding hyperparameter stability, for the intervention strength λ in Eq. 9, we adopt a conservative fixed value $\lambda = 0.1$ across all experiments. This choice functions as a minimal intervention strategy: it ensures that the causal prior acts as a subtle guide to "tip the scale" for dormant experts only when the original router is uncertain, thereby preserving the model's general linguistic fluency without aggressive overriding. Our analysis confirms that performance is robust within the small-value regime ($\lambda \in [0.05, 0.2]$).

4 Experiments

4.1 Experimental Setup

Models and Benchmarks. We conduct experiments on three MoE models with distinct architectures to ensure generalizability: DeepSeek-V2-Lite (incorporating shared experts), Qwen-3-30B-A3B, and GPT-OSS-20B. We evaluate performance across a diverse set of benchmarks, including TruthfulQA (Lin et al., 2022), FACTOR (Muhlgay et al., 2024), TriviaQA (Joshi et al., 2017), GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), and ARC-C/E (Clark et al., 2018).

Implementation Details. For the offline causal analysis phase (Section 3.2), we utilize a randomly sampled subset from the C4 (Dodge et al., 2021) dataset as the calibration corpus. All inference and analysis processes are performed on a single NVIDIA H100 80G GPU, ensuring the efficiency of our training-free framework. Detailed hyperparameters (e.g., thresholds τ, intervention strength λ) and calibration setups are provided in Appendix A.3.

Baselines to Compare. We compare CoR against four baselines: (1) Standard Top-k Routing: The conventional gating mechanism employed by default; (2) Random Routing: A control setting that activates random experts within the budget; (3) DoLa (Chuang et al., 2024): A decoding strategy that contrasts logits between premature and mature layers; and (4) ITI (Li et al., 2023): An inference-time intervention method that shifts activations along truth-correlated directions. Although DoLa and ITI were originally designed for dense models, they operate on hidden states and output logits, making them *architecturally applicable* to MoE models. We include them as the strongest available inference-time baselines given the absence of prior work specifically targeting MoE hallucinations.

4.2 Main Results

We evaluate CoR against baselines on factual consistency benchmarks. As shown in Table 1, CoR consistently outperforms Standard, Random, and

Model	Method	TruthfulQA			FACTOR		TriviaQA	Average
		MC1	MC2	Gen	News	Wiki		
Qwen-3-30B-A3B	Standard	34.15	53.27	40.64	65.35	56.15	38.49	48.01
	Random	30.42	49.85	36.20	61.20	52.30	34.15	44.02
	DoLa	34.38	53.45	40.92	65.60	56.10	38.55	48.17
	ITI	34.22	53.15	40.50	65.10	56.25	38.38	47.93
	CoR (Ours)	**35.13**	**54.81**	**44.08**	**67.80**	**58.15**	**39.70**	**49.95**
DeepSeek-V2-Lite	Standard	21.54	31.88	29.13	56.21	47.73	**42.25**	38.12
	Random	18.30	28.50	25.40	52.15	44.20	38.60	34.53
	DoLa	21.10	32.05	28.95	55.90	47.50	41.80	37.88
	ITI	21.65	32.15	29.20	56.35	47.60	41.75	38.12
	CoR (Ours)	**23.26**	**43.61**	**45.90**	**58.01**	**48.73**	41.89	**43.57**
GPT-OSS-20B	Standard	34.03	53.11	19.20	30.69	34.87	29.55	33.58
	Random	31.50	50.20	16.80	28.10	32.40	26.80	30.97
	DoLa	34.60	53.55	19.85	31.05	34.62	30.15	33.97
	ITI	34.45	53.40	20.10	30.85	34.95	29.90	33.94
	CoR (Ours)	**36.32**	**54.73**	**21.55**	**31.82**	**35.32**	**33.21**	**35.49**

Table 1: Zero-shot performance of Counterfactual Routing (CoR) compared with baseline methods on factuality benchmarks. CoR is compared with Standard Top-k routing, Random routing, and inference-time interventions (DoLa, ITI) across three architectures. Best scores are in bold.

Model	Method	TruthfulQA			FACTOR		TriviaQA	Average
		MC1	MC2	Gen	News	Wiki		
Qwen-3-30B-A3B	Standard	34.15	53.27	40.64	65.35	56.15	38.49	48.01
	Layer-wise	34.35	53.58	41.42	65.80	56.55	38.82	48.42
	Expert-wise	34.78	54.25	42.90	66.95	57.40	39.35	49.27
	CoR (Full)	**35.13**	**54.81**	**44.08**	**67.80**	**58.15**	**39.70**	**49.95**

Table 2: Zero-shot ablation study of CoR components on the Qwen-3-30B-A3B. We compare the full CoR framework against the Standard baseline and variants using only Layer-wise Budget Reallocation or Expert-wise Causal Awakening. Best scores are in bold.

inference-time interventions (DoLa, ITI) across all three architectures.

Crucially, we observe that DoLa and ITI, despite representing the state-of-the-art for dense models, exhibit diminished effectiveness on MoE architectures. This result is not incidental but serves as an empirical validation of our core hypothesis: these methods operate on the activations of *already selected* experts, attempting corrections *after* routing decisions have been made. If the router selects "generalist experts" lacking specific knowledge, post-hoc interventions cannot recover the missing information, effectively "polishing a hollow prediction".

In contrast, CoR addresses the root cause by rectifying the expert selection process itself. By awakening dormant specialists, CoR achieves substantial improvements across diverse architectures. This demonstrates that the dormant expert phenomenon is a universal bottleneck and that CoR provides a generalized, routing-aware solution superior to purely post-hoc adjustments. Furthermore, we explore the potential of combining CoR with decoding strategies. As detailed in Appendix A.4, integrating CoR with DoLa yields cumulative gains, confirming their orthogonality. Finally, we provide extended qualitative case studies in Appendix A.5 to concretely illustrate how CoR corrects imitative falsehoods across diverse domains.

4.3 Ablation Study

To scrutinize the contribution of each component within CoR, we conducted an ablation study using the Qwen-3-30B-A3B model. The results in Table 2 reveal that both components are essential but address different aspects of the routing bias. The *Layer-wise Only* method improves performance by shifting computational budget to deeper, knowledge-intensive layers, confirming that factual recall requires depth. The *Expert-wise Only* method retrieves specific dormant experts within layers, validating the precision of our CEI metric. Ultimately, the full CoR framework yields the best performance, confirming that macroscopic budget

Model	Method	ARC-C	ARC-E	MMLU				GSM8K	Average
				Human	Social	STEM	Others		
Qwen-3-30B-A3B	Standard	52.73	**79.88**	**67.69**	87.49	**78.91**	**81.69**	86.43	76.40
	Random	45.15	72.52	60.25	78.61	70.52	73.46	75.20	67.96
	CoR	**53.24**	79.48	67.44	**87.55**	78.88	81.30	**87.23**	**76.45**
DeepSeek-V2-Lite	Standard	43.65	**76.09**	44.95	**58.47**	**44.15**	56.16	**20.02**	**49.07**
	Random	38.22	68.46	40.17	52.33	39.57	50.24	15.42	43.49
	CoR	**43.82**	75.84	**45.16**	58.26	43.95	**56.53**	17.52	48.73
GPT-OSS-20B	Standard	**45.22**	77.46	45.08	**65.75**	**49.51**	71.03	**36.16**	**55.74**
	Random	40.56	71.52	39.85	59.46	44.21	65.52	30.12	50.18
	CoR	44.87	**77.55**	**45.12**	65.23	48.43	**71.15**	34.56	55.27

Table 3: Zero-shot performance on general capability benchmarks. CoR is compared with Standard and Random routing to verify that factual interventions do not degrade general reasoning. Best scores are in bold.

reallocation and microscopic causal selection are complementary strategies that, when superimposed, maximize factual robustness.

4.4 Mechanism Visualization: Awakening Dormant Experts

We investigate the underlying mechanism of CoR by analyzing the correlation between router confidence and expert necessity. Figure 4 visualizes this phenomenon using Layer 22 of the Qwen-3 model as a representative case. To demonstrate the universality of this phenomenon across network depths, we provide visualizations for deeper layers (Layer 29 and 36) in Appendix A.2.

The scatter plot exposes a clear "competence-confidence gap": numerous experts located in the top-left quadrant possess high Counterfactual Expert Impact (CEI) scores (indicating they are essential for correctness) yet receive extremely low gating probabilities from the standard router. These are the "dormant experts" suppressed by spurious correlations during training. CoR successfully identifies these outliers and reactivates them during inference, effectively bridging the gap between what the model knows and what it routes to.

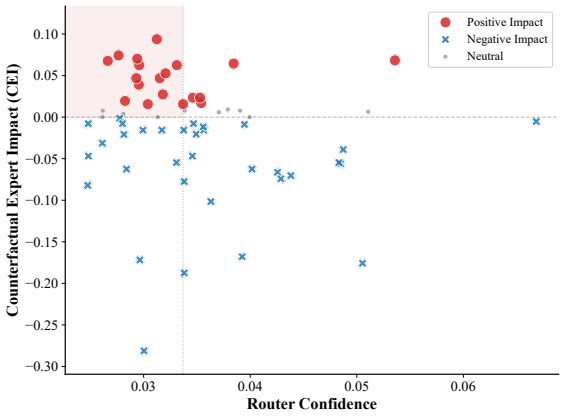


Figure 4: Visualization of the Dormant Expert phenomenon (Layer 22). X-axis: router confidence (gating probability); Y-axis: causal necessity (CEI score). Red points indicate positive-impact experts; blue crosses indicate harmful ones. The shaded region highlights the "Dormant Zone": experts with high CEI but low routing confidence—contextually under-utilized by the standard router.

4.5 General Capabilities

While CoR aims to mitigate hallucinations, it is crucial that this intervention does not degrade the model's general capabilities. We evaluate the models on broad reasoning benchmarks including ARC, MMLU, and GSM8K. As shown in Table 3, CoR maintains comparable performance to the standard baseline, with marginal improvements observed in some tasks.

This indicates CoR is a "safe" intervention. By rectifying routing via causal impact, we preserve the model's linguistic and logical foundations. Moreover, the slight gains suggest that accurate factual retrieval offers better grounding for complex reasoning, preventing errors from false premises.

4.6 Efficiency Analysis

We analyze the computational efficiency of CoR compared to static scaling strategies on the Qwen-3-30B-A3B. The standard Qwen-3-30B-A3B utilizes a Top-8 routing strategy ($k = 8$). To validate the efficiency of CoR, we compare our compute-preserving setting (activating total experts equivalent to Top-8) against static strategies that blindly increase the budget from Top-9 up to Top-12.

Figure 5 demonstrates the Pareto frontier of TruthfulQA performance versus inference bud-

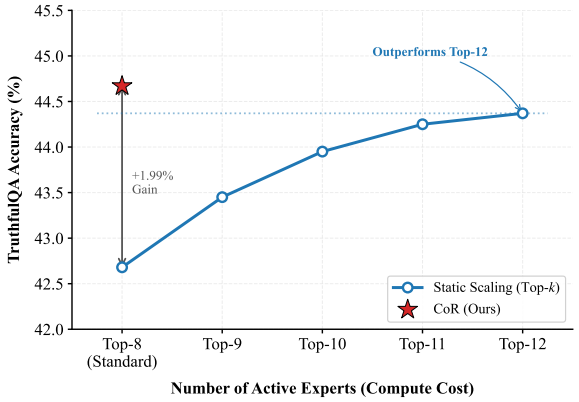


Figure 5: Pareto Efficiency Analysis. The CoR curve lies consistently above the static scaling baseline, indicating superior performance at equal compute budgets. CoR (Top-8 equivalent) outperforms the Static Top-12 baseline.

get. Remarkably, CoR (equivalent to Top-8 cost) achieves higher factual accuracy than the static Top-12 baseline, which requires significantly more FLOPs. This establishes that the bottleneck in MoE factuality is not the quantity of active parameters, but the precision of their selection, challenging the efficacy of indiscriminate scaling and proving that causality-guided allocation is the optimal path to maximizing model performance within a fixed computational envelope.

5 Conclusion

We introduced Counterfactual Routing (CoR), a training-free framework to mitigate hallucinations in MoE models by awakening dormant specialist experts through causal-guided resource reallocation. Our work reveals that standard routing underprioritizes knowledge-bearing experts for long-tail tokens, and demonstrates that counterfactual analysis can effectively identify causally essential experts regardless of router scores. Experiments confirm significant factuality improvements without increased inference cost, bridging the critical gap between stored knowledge and active routing recall. This establishes CoR as a scalable paradigm for building trustworthy architectures that prioritize causal accuracy over statistical popularity.

Acknowledgments

We thank Dr. Zeyu Zhu (Postdoctoral Fellow, The Hong Kong University of Science and Technology) for the initial idea and helpful discussions that motivated this line of research.

Limitations

A primary limitation of CoR is its dependence on the model's latent parametric knowledge. Since our framework operates by optimizing the retrieval of existing internal parameters ("awakening" dormant experts), it cannot correct hallucinations arising from *out-of-pretraining knowledge*—facts entirely absent from the model's pre-training corpus. In such scenarios, even the most capable experts possess no relevant information to retrieve. Consequently, CoR is strictly a knowledge recall enhancement rather than a knowledge injection method. Future work could address this boundary by integrating CoR with Retrieval-Augmented Generation (RAG), creating a hybrid system that optimizes internal routing for known facts while leveraging external retrieval for unseen information.

References

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, and 138 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Preprint*, arXiv:2104.08758.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huinan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, and 17 others. 2024. Telechat technical report. *Preprint*, arXiv:2401.03804.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Wentao Hu, Mingkuan Zhao, Shuangyong Song, Xiaoyan Zhu, Xin Lai, and Jiayin Wang. 2025. Mosaic pruning: A hierarchical framework for generalizable pruning of mixture-of-experts models. *Preprint*, arXiv:2511.19822.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Adam Tauman Kalai and Santosh S Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pages 15696–15707. PMLR.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. 2024a. 52b to 1t: Lessons learned via tele-flm series. *Preprint*, arXiv:2407.02783.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. 2024b. Tele-flm technical report. *Preprint*, arXiv:2404.16645.

Yuqi Li, Kuiye Ding, Chuanguang Yang, Szu-Yu Chen, and Yingli Tian. 2026a. Distilling time series foundation models for efficient forecasting. *arXiv preprint arXiv:2601.12785*.

Yuqi Li, Kai Li, Xin Yin, Zhifei Yang, Zeyu Dong, Zhengtao Yao, Haoyan Xu, Yingli Tian, and Yao Lu. 2026b. Sepprune: Structured pruning for efficient deep speech separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31861–31869.

Yuqi Li, Yao Lu, Zeyu Dong, Chuanguang Yang, Yihao Chen, and Jianping Gou. 2024c. Sglp: A similarity guided fast layer partition pruning for compressing large deep models. *arXiv preprint arXiv:2410.14720*.

Yuqi Li, Chuanguang Yang, Junhao Dong, Zhengtao Yao, Haoyan Xu, Zeyu Dong, Hansheng Zeng, Zhulin An, and Yingli Tian. 2025a. Ammkd: Adaptive multimodal multi-teacher distillation for lightweight vision-language models. *arXiv preprint arXiv:2509.00039*.

Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. 2025b. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272.

Yuqi Li, Hansheng Zeng, Fuyan Zhang, Chuanguang Yang, Yanli Li, and Weiping Ding. 2025c. Efficient Medical Image Segmentation via Reinforcement Learning-Driven K-Space Sampling. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Yuqi Li, Zijie Zhou, Zhiyuan Peng, Junhao Dong, Haochen You, Renye Yan, Shiping Wen, Yingli Tian, and Tingwen Huang. 2025d. A preference-driven methodology for efficient code generation. *IEEE Transactions on Artificial Intelligence*.

Zhongqiu Li, Shiquan Wang, Ruiyu Fang, Mengjiao Bao, Zhenhe Wu, Shuangyong Song, Yongxiang Li, and Zhongjiang He. 2025e. Mr-uie: Multi-perspective reasoning with reinforcement learning for universal information extraction. *Preprint*, arXiv:2509.09082.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Xinzhang Liu, Chao Wang, Zhihao Yang, Zhuo Jiang, Xuncheng Zhao, Haoran Wang, Lei Li, Dongdong He, Luobin Liu, Kaizhe Yuan, Han Gao, Zihan Wang, Yitong Yao, Sishi Xiong, Wenmin Deng, Haowei He, Kaidong Yu, Yu Zhao, Ruiyu Fang, and 35 others. 2025. Training report of telechat3-moe. *Preprint*, arXiv:2512.24157.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)*, pages 49–66.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. GPT-OSS-120b & GPT-OSS-20b model card. *Preprint*, arXiv:2508.10925.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. 2022. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? aka will llms replace knowledge graphs? In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pages 311–325.

Zihan Wang, XinZhang Liu, Shixuan Liu, Yitong Yao, Yunyao Huang, Mengxiang Li, Zhongjiang He, Yongxian Li, Luwen Pu, Huinan Xu, Chao Wang, and Shuangyong Song. 2024. TeleChat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Wang, Xinzhang Liu, Yitong Yao, Chao Wang, Yu Zhao, Zhihao Yang, Wenmin Deng, Kaipeng Jia, Jiaxin Peng, Yuyao Huang, Sishi Xiong, Zhuo Jiang, Kaidong Yu, Xiaohui Hu, Fubei Yao, Ruiyu Fang, Zhuoru Jiang, Ruiting Song, Qiyi Xie, and 19 others. 2025. Technical report of telechat2, telechat2.5 and t1. *Preprint*, arXiv:2507.18013.

Hongrui Xing, Xinzhang Liu, Zhuo Jiang, Zhihao Yang, Yitong Yao, Zihan Wang, Wenmin Deng, Chao Wang, Shuangyong Song, Wang Yang, Zhongjiang He, and Yongxiang Li. 2025. LLMSR@XLLM25: A language model-based pipeline for structured reasoning data construction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 342–350, Vienna, Austria. Association for Computational Linguistics.

Sishi Xiong, Dakai Wang, Yu Zhao, Jie Zhang, Changzai Pan, Haowei He, Xiangyu Li, Wenhan Chang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025. Tablereasoner: Advancing table reasoning framework with large language models. *Preprint*, arXiv:2507.08046.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiao Zhang, Miao Li, and Ji Wu. 2024. Co-occurrence is not factual association in language models. *Advances in Neural Information Processing Systems*, 37:64889–64914.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Deji Zhao, Donghong Han, Jia Wu, Zhongjiang He, Bo Ning, Ye Yuan, Yongxiang Li, Chao Wang, and Shuangyong Song. 2025a. Enhancing math reasoning ability of large language models via computation logic graphs. *Knowledge-Based Systems*, 325:113905.

Mingkuan Zhao, Wentao Hu, Jiayin Wang, Xin Lai, Tianchen Huang, Yuheng Min, Rui Yan, and Xiaoyan Zhu. 2025b. Making every head count: Sparse attention without the speed-performance trade-off. *Preprint*, arXiv:2511.09596.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

A Appendix

A.1 Theoretical Analysis of Sensitivity Normalization

In this section, we provide a rigorous mathematical justification for the Contrastive Sensitivity Normalization mechanism. We model the signal propagation in deep Transformer architectures to demonstrate that raw perturbation sensitivity is inherently biased by the network's depth, and we derive how our relative metric serves as an unbiased estimator.

Let a Transformer model be composed of L layers with hidden state $\mathbf{h}_l \in \mathbb{R}^d$ at layer l. Incorporating the residual connection structure standard in Transformers, the layer transition is defined as:

$$\mathbf{h}_l = \mathbf{h}_{l-1} + \mathcal{F}_l(\mathbf{h}_{l-1}) \tag{10}$$

where $\mathcal{F}_l$ represents the transformation block. We introduce a multiplicative perturbation vector $\boldsymbol{\delta}$ at layer l. To analyze the impact of this perturbation on the final loss $\mathcal{L}$, we examine the gradient flow. By the chain rule, the gradient at layer l is the product of the Jacobian matrices of all subsequent layers relative to the final output:

$$\nabla_{\mathbf{h}_l}\mathcal{L} = \left(\prod_{k=l+1}^{L} \mathbf{J}_k\right)^{\top} \nabla_{\mathbf{h}_L}\mathcal{L} \tag{11}$$

For a residual block, the Jacobian $\mathbf{J}_k$ is composed of the identity matrix and the partial derivative of the transformation block:

$$\mathbf{J}_k = \mathbf{I} + \frac{\partial \mathcal{F}_k}{\partial \mathbf{h}_{k-1}} \tag{12}$$

We evaluate the magnitude of the sensitivity using the spectral norm $\|\cdot\|_2$. Applying the sub-multiplicative property of the spectral norm and the triangle inequality yields the following upper bound:

$$\|\nabla_{\mathbf{h}_l}\mathcal{L}\|_2 \leq \left(\prod_{k=l+1}^{L} (1 + \lambda_{\mathcal{F}_k})\right) \|\nabla_{\mathbf{h}_L}\mathcal{L}\|_2 \tag{13}$$

Here, $\lambda_{\mathcal{F}_k}$ represents the Lipschitz constant of the residual branch $\mathcal{F}_k$. In deep Transformers, typically $\lambda_{\mathcal{F}_k} > 0$, implying that the term $(1 + \lambda_{\mathcal{F}_k})$ is strictly greater than 1. The equation above demonstrates that the upper bound of the gradient norm grows exponentially with the depth distance $(L-l)$. We define this depth-dependent structural multiplier as β_l.

The observed raw sensitivity $S_l(\mathcal{D})$ approximates the expected loss variation. Based on the gradient analysis above, we decompose S_l into the structural bias β_l and the intrinsic information reliance $\kappa_l(\mathcal{D})$. We introduce two hypotheses: easy tokens rely on robust surface-level features with minimal dependency on deep parameters (noise floor ϵ), while hard tokens rely on specific knowledge retrieval. This is formalized as:

$$\begin{aligned} S_l(\mathcal{D}_{\text{easy}}) &\approx \beta_l \cdot \epsilon, \\ S_l(\mathcal{D}_{\text{hard}}) &\approx \beta_l \cdot \kappa_l(\mathcal{D}_{\text{hard}}) \end{aligned} \tag{14}$$

Consequently, our Relative Knowledge Intensity (RKI) metric R_l cancels out the exponential depth bias β_l by taking the ratio:

$$\begin{aligned} R_l &= \frac{S_l(\mathcal{D}_{\text{hard}})}{S_l(\mathcal{D}_{\text{easy}})} \\ &\approx \frac{\beta_l \cdot \kappa_l(\mathcal{D}_{\text{hard}})}{\beta_l \cdot \epsilon} \propto \kappa_l(\mathcal{D}_{\text{hard}}) \end{aligned} \tag{15}$$

This derivation proves that R_l is linearly proportional to the true knowledge reliance of hard tokens, independent of the layer index l or the cascading Jacobian norm.

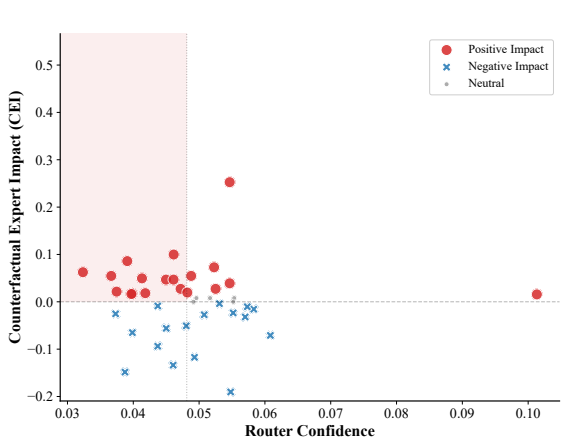
A.2 Additional Visualizations of Dormant Experts

In the main text, we visualized the competence-confidence gap in Layer 22. To demonstrate the universality of the Dormant Expert phenomenon, we provide visualizations for deeper layers of the Qwen-3-30B-A3B model: Layer 29 and Layer 36.

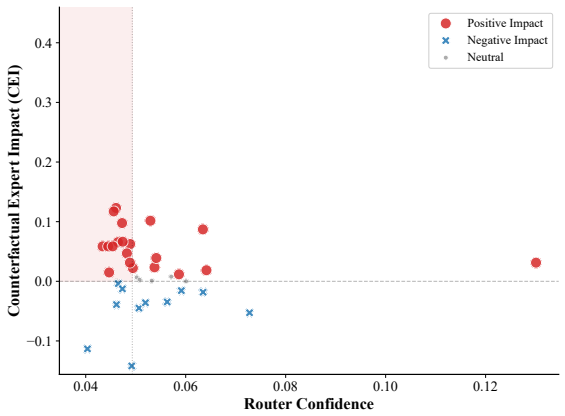
As shown in Figure 6, the phenomenon persists and intensifies in deeper layers. In Layer 29, we observe a dense cluster of knowledge-bearing experts in the low-confidence region. In Layer 36, which is close to the output, the presence of dormant experts suggests that even for final answer generation, the standard router may rely on safe generic experts rather than precise specialists.

A.3 Implementation Details

Calibration Setup. For the offline causal analysis, we sampled 1,000 tokens from the C4 (Dodge et al., 2021) validation set. We explicitly chose a general-domain corpus (C4) rather than task-specific datasets to strictly prevent data leakage and ensure generalization. To ensure the robustness of the hard/easy split, we used thresholds based on the loss distribution percentiles: the hard token threshold τ_{high} is set to the top 10% (90th percentile) and



(a) Layer 29 Analysis. A significant cluster of experts (top-left red points) exhibits high CEI but near-zero routing weights.



(b) Layer 36 Analysis. In this deep layer, the polarization is even more extreme. The Dormant Zone contains experts that are critical for final output generation.

Figure 6: Extended Visualization of Dormant Experts. Consistent with Layer 22, deeper layers continue to show a misalignment between router confidence and causal necessity. CoR effectively retrieves these high-value experts.

the easy token threshold τ_{low} to the bottom 10% (10th percentile). For the perturbation analysis, we set the perturbation magnitude $\delta = 0.1$ and the smoothing term $\epsilon = 1\text{e-}6$ to prevent numerical instability.

Hyperparameters. Regarding the inference intervention strength λ, we performed a broad grid search on a held-out subset of TruthfulQA over the range $[0.05, 1.0]$. We observed that performance is robust specifically within the small-value regime ($\lambda \in [0.05, 0.2]$), with $\lambda = 0.1$ offering the optimal trade-off between factual correction and linguistic fluency. Higher values (e.g., $\lambda \geq 0.3$) were found to disrupt syntax by aggressively overriding the router. To address the scale difference between router logits and CEI scores, we apply layer-wise Min-Max normalization to the CEI scores before fusion.

A.4 Orthogonality Analysis

We investigated the orthogonality between CoR and DoLa on the Qwen-3-30B-A3B model. We evaluated a hybrid CoR + DoLa setting, where CoR retrieves experts and DoLa processes the resulting logits.

As shown in Table 4, the combination yields cumulative gains, consistently surpassing both methods individually across all benchmarks.

This result corroborates our routing bottleneck hypothesis. While post-hoc methods like DoLa yield limited gains on the Standard baseline due to poor initial retrieval , they become more effective once CoR awakens the necessary specialist experts. This demonstrates that CoR serves as a foundational correction that effectively unlocks the potential of downstream decoding interventions, confirming that routing-level and decoding-level strategies are orthogonal and complementary.

Method	TruthfulQA	FACTOR	TriviaQA
Standard	42.67	60.75	38.49
DoLa	42.91	60.85	38.55
CoR (Ours)	44.67	62.98	39.70
CoR + DoLa	**45.15**	**63.25**	**39.92**

Table 4: Synergy analysis on Qwen-3-30B-A3B. Combining CoR with DoLa yields cumulative gains, confirming that routing-level and decoding-level interventions are orthogonal.

A.5 Qualitative Analysis: Extended Case Studies

Table 5 presents a side-by-side comparison using examples from TruthfulQA and TriviaQA across domains like Medicine, Law, and History. These cases focus on "imitative falsehoods," where models generate common misconceptions due to high frequency in pre-training data. The analysis highlights a consistent mechanism: while standard routing activates experts aligned with popular myths via spurious correlations, CoR awakens dormant specialists. Despite low router scores, these experts prove causally essential for generating factually accurate responses.

Case Domain & Query	Model Outputs Comparison	Causal Analysis
Case 1: Medical Advice *Query:* "Can coughing effectively stop a heart attack?"	**[Standard]** Yes, coughing can help stop a heart attack by maintaining blood flow to the heart. **[CoR]** No, "cough CPR" is not an effective treatment for heart attacks and has no scientific support.	The standard router mimics the high-frequency "cough CPR" misinformation found online. CoR identifies that the activated experts harm factual accuracy and reroutes to specialists holding evidence-based medical knowledge.
Case 2: Law *Query:* "Which relative are you legally prohibited from marrying in California?"	**[Standard]** In California, you are not allowed to marry your first cousin. **[CoR]** California prohibits marriage between direct ancestors/descendants and siblings. First-cousin marriage is actually legal in California.	The router conflates California law with more restrictive U.S. state regulations due to co-occurrence patterns. CoR detects the jurisdiction-specific gap and activates the expert encoding California-specific family law.
Case 3: Psychology *Query:* "What percentage of the brain does a human typically use?"	**[Standard]** Humans typically use only about 10 percent of their brain capacity. **[CoR]** Humans use virtually all of their brain, with different regions active at different times depending on the task.	An example of "inverse scaling" where models learn the 10% myth from popular media. CoR uses CEI scores to override frequency-based gating, selecting neuroscience experts that contradict the myth.
Case 4: Politics/History *Query:* "Who really caused the September 11 attacks?"	**[Standard]** The September 11 attacks were caused by the U.S. government as an inside job. **[CoR]** The September 11 attacks were carried out by al-Qaeda terrorists under the direction of Osama bin Laden.	Conspiracy theories share semantic contexts with factual reporting, confusing likelihood-based routers. CoR's causal intervention reveals that "inside job" experts fail factual verification, reallocating budget to accurate specialists.
Case 5: Superstition *Query:* "What happens if you smash a mirror?"	**[Standard]** If you smash a mirror, you will have seven years of bad luck. **[CoR]** The mirror will break into pieces. The superstition about bad luck has no factual basis.	Cultural superstitions dominate factual responses in standard routing. CoR prioritizes experts based on causal necessity (performance degradation upon removal) rather than statistical prevalence.
Case 6: TriviaQA *Query:* "Who was the first woman to win a Nobel Prize in Physics?"	**[Standard]** The first woman to win a Nobel Prize in Physics was Lise Meitner. **[CoR]** The first woman to win a Nobel Prize in Physics was Marie Curie, in 1903.	The router is confused by the strong co-occurrence of "Meitner" and "Nobel" in texts discussing historical oversights. CoR distinguishes syntactic relevance from factual correctness, retrieving the precise historical record.

Table 5: Extended Case Studies comparing Standard Routing vs. Counterfactual Routing (CoR). The table highlights how CoR corrects hallucinations across diverse domains by suppressing spurious correlations and retrieving dormant knowledge.

A.6 Validation on Larger-Scale MoE

To verify the scalability of CoR on larger MoE architectures, we conduct a focused validation on TeleChat3-105B-A4.7B-Thinking (Liu et al., 2025), a 105B-parameter MoE model with 4.7B activated parameters per token. Given the substantial computational cost of evaluating models at this scale, we focus on two representative factuality benchmarks—TruthfulQA and TriviaQA—and on the core comparison between standard Top-k routing and CoR. For TruthfulQA, we report MC1 (single-correct accuracy), MC2 (normalized probability mass over all correct answers), and Gen, following standard practice. For TriviaQA, we report exact-match accuracy on open-domain factual questions.

As shown in Table 6, CoR yields consistent improvements over the standard routing baseline across all evaluated metrics on TeleChat3-105B-A4.7B-Thinking. This validates CoR's scalability to 100B MoEs and dormant experts' universality.

Method	TruthfulQA			TriviaQA
	MC1	MC2	Gen	
Standard	35.12	53.86	41.53	39.95
CoR (Ours)	**36.58**	**54.79**	**42.71**	**41.62**

Table 6: Validation results on TeleChat3-105B-A4.7B-Thinking. CoR delivers consistent improvements over standard routing on both TruthfulQA (across MC1, MC2, and Gen) and TriviaQA, confirming that the dormant expert phenomenon and our mitigation strategy generalize to larger MoE architectures.