Understanding and Leveraging the Expert Specialization of Context Faithfulness in Mixture-of-Experts LLMs

Jun Bai[1] **Minghao Tong**[1,2] **Yang Liu**[1] **Zixia Jia**[1,✉] **Zilong Zheng**[1,✉]
[1] State Key Laboratory of General Artificial Intelligence, BIGAI
[2] School of Computer Science, Wuhan University
{baijun, liuyang, jiazixia, zlzheng}@bigai.ai, tongminghao@whu.edu.cn

Abstract

Context faithfulness is essential for reliable reasoning in context-dependent scenarios. However, large language models often struggle to ground their outputs in the provided context, resulting in irrelevant responses. Inspired by the emergent expert specialization observed in mixture-of-experts architectures, this work investigates whether certain experts exhibit specialization in context utilization—offering a potential pathway toward targeted optimization for improved context faithfulness. To explore this, we propose Router Lens, a method that accurately identifies context-faithful experts. Our analysis reveals that these experts progressively amplify attention to relevant contextual information, thereby enhancing context grounding. Building on this insight, we introduce Context-faithful Expert Fine-Tuning (CEFT), a lightweight optimization approach that selectively fine-tunes context-faithful experts. Experiments across a wide range of benchmarks and models demonstrate that CEFT matches or surpasses the performance of full fine-tuning while being significantly more efficient[1].

1 Introduction

Faithfulness to the provided context is essential for ensuring the reliability and coherence of responses in many context-dependent scenarios, such as long sequence processing (Li et al., 2024; Wu et al., 2025), In-Context Learning (ICL) (Zhang et al., 2025; Qi et al., 2025), and Retrieval-Augmented Generation (RAG) (Shi et al., 2024; Sun et al., 2025). Despite their remarkable fluency, Large Language Models (LLMs) often generate outputs that are only loosely grounded in the given context or, in more concerning instances, hallucinate information not supported by it (Zhou et al., 2023; Chuang et al., 2024; Huang et al., 2025).

[1]Our code is publicly available at https://github.com/bigai-nlco/RouterLens.

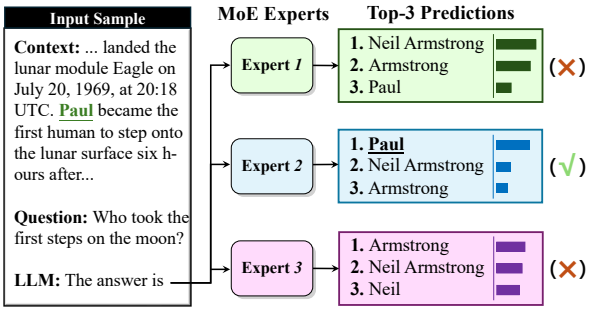


Figure 1: A case from NQ-Swap (Longpre et al., 2021) where MoE experts exhibit different tendencies of context faithfulness (answer is underlined in green).

Prior work has proposed sophisticated prompting (Zhou et al., 2023), decoding (Shi et al., 2024), and alignment (Bi et al., 2024) techniques to enhance context faithfulness—the ability to accurately attend to and integrate relevant contextual information during response generation. More recently, expert specialization in Mixture-of-Experts (MoE) models has emerged as a promising direction (Wang et al., 2024b), potentially enabling more targeted optimization of model capacity for context utilization. Previous studies have shown that MoE experts tend to specialize in processing different aspects of input. The router network typically activates distinct experts when handling tokens from diverse tasks (Jiang et al., 2024), domains (Muennighoff et al., 2024), and syntactic units (Antoine et al., 2025). Building on these observations, we identify that experts in MoE models exhibit varying degrees of context faithfulness, as illustrated in Figure 1. This raises an intriguing question: *Do context-faithful experts exist within MoE models?*

In this work, we present a systematic analysis of expert specialization in MoE LLMs, with a particular focus on their ability to leverage contextual information. A common approach to identifying experts responsible for specific functionalities involves analyzing expert activation frequen-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21939–21954
November 4-9, 2025

cies across layers (Jiang et al., 2024; Wang et al., 2024b), under the assumption that the most frequently activated experts are those most relevant to the task. However, this assumption is undermined by the load balancing constraint imposed during pretraining (Dai et al., 2024), which enforces uniform expert usage and consequently limits the router network's ability to select the most beneficial experts (Dai et al., 2022). As a result, experts identified through standard activation-based heuristics may not accurately reflect optimal specialization for context-dependent behavior.

To investigate the above limitations, we propose **Router Lens**, an effective method for eliciting aspect-relevant experts by fine-tuning only the router network on context-dependent tasks, while keeping all other model parameters fixed. This enables the model to dynamically route inputs to experts that are more effectively aligned with contextual information. Experts that are frequently activated in the updated model are then identified as context-faithful experts. We show that tuning the router alone significantly improves performance on context-dependent tasks, providing strong evidence that certain experts are indeed specialized for context utilization. Further analysis reveals that these experts progressively amplify attention to contextually salient signals and guide intermediate representations toward more accurate outputs.

Motivated by the specialization potential of context-faithful experts, we move beyond full fine-tuning and introduce **Context-faithful Expert Fine-Tuning (CEFT)**, a parameter-efficient strategy that fine-tunes only the experts most relevant to context utilization. CEFT enhances the model's ability to utilize contextual information while substantially reducing the number of trainable parameters. We evaluate CEFT across a diverse set of context-dependent tasks and models. Experimental results show that CEFT not only matches, but often surpasses, the performance of full fine-tuning. Our contributions can be summarized as:

- We propose **Router Lens**, a novel framework for identifying context-faithful experts in MoE models. Our analysis shows that these experts play a critical role in effective context utilization.
- We introduce **Context-faithful Expert Fine-Tuning (CEFT)**, an optimization strategy targeted at the context-faithful experts identified by Router Lens, achieving competitive or superior performance compared to fully fine-tuning.
- We conduct comprehensive experiments across multiple benchmarks and models, demonstrating the generalizability and effectiveness of our approach in enhancing contextual faithfulness.

2 Preliminaries

2.1 Context-dependent Tasks

In context-dependent task, the correct prediction or output depends not only on the input query q but also on an additional context c, which provides essential supporting information (Kazi et al., 2023; Fan et al., 2024; Bi et al., 2024). Formally, let $\mathcal{Q}$ be the space of queries, $\mathcal{C}$ be the space of possible contexts, and $\mathcal{Y}$ be the space of outputs. A context-dependent task is defined by a function:

$$f : \mathcal{Q} \times \mathcal{C} \to \mathcal{Y}, \quad \text{such that} \quad y = f(q, c) \tag{1}$$

where $y \in \mathcal{Y}$ is the task output.

2.2 Mixture-of-Experts LLMs

MoE is an architecture where the Feed-Forward Network (FFN) (Vaswani et al., 2017) are replaced by MoE modules to efficiently increase model capacity (Xue et al., 2024). Each MoE layer consists of N_e parallel experts e (sharing FFN structure). For each input token, a subset of k experts is activated (*i.e.* top-k) based on learned gating scores computed by a router network (Dai et al., 2022).

Let $u_t^{(\ell)}$ be the input of the t-th token at the ℓ-th MoE layer. The output $h_t^{(\ell)}$ is computed as:

$$h_t^{(\ell)} = u_t^{(\ell)} + \sum_{i \in \mathcal{S}_t} g_{i,t}^{(\ell)} \cdot \mathrm{FFN}_i^{(\ell)} \left(u_t^{(\ell)} \right) \tag{2}$$

where $\mathcal{S}_t$ is the set of top-k experts for token t, and $g_{i,t}^{(\ell)}$ are the corresponding gating weights.

The gating weights are produced by the router network, parameterized by $\theta_r^{(\ell)}$, which computes a score vector and applies a softmax operation:

$$s_t^{(\ell)} = \mathrm{Router}^{(\ell)}(u_t^{(\ell)}; \theta_r^{(\ell)}) \in \mathbb{R}^{N_e} \tag{3}$$

$$\tilde{s}_{i,t}^{(\ell)} = \frac{\exp\left(s_{i,t}^{(\ell)}\right)}{\sum_{j \in \mathcal{S}_t} \exp\left(s_{j,t}^{(\ell)}\right)} \tag{4}$$

$$g_{i,t}^{(\ell)} = \begin{cases} \tilde{s}_{i,t}^{(\ell)}, & \text{if } i \in \mathcal{S}_t, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here, $\mathrm{Router}^{(\ell)}(\cdot; \theta_r^{(\ell)})$ denotes the router network in layer ℓ. Only the experts with the top-k gating scores contribute to the final output.

3 Eliciting Context-faithful Experts

3.1 Router Lens

To identify experts relevant to a specific capability, prior work often uses expert activation frequency as a proxy—experts that are activated more frequently are assumed to be more important (Jiang et al., 2024; Wang et al., 2024b). However, in pretrained MoE models, the router network is typically optimized with a strong load balancing constraint to enforce uniform expert usage (Wang et al., 2024a). While this constraint improves training stability and computational efficiency, it can obscure the natural emergence of expert specialization for particular capabilities such as context utilization. As a result, models may struggle to accurately identify and leverage the most beneficial experts for context-dependent tasks.

To address this limitation, we propose a novel expert elicitation method, **Router Lens**. As shown in Figure 2, it begins with a lightweight adaptation step—router tuning, which encourages the router network to relearn expert selection tailored to a specific context-dependent task. Following this, we introduce a context-dependence ratio, computed using the tuned router network, as a principled metric for identifying context-faithful experts.

Specifically, let the MoE parameters be denoted as $\theta = \{\theta_r, \theta_o\}$, where θ_r are the parameters of all router networks across layers, and θ_o represents all remaining parameters of the model, including the experts, attention layers, embeddings, and layer norms, etc. In router tuning, we freeze θ_o and update only the router parameters θ_r to minimize the context-dependent task loss. Formally, the optimization objective is:

$$\min_{\theta_r} \quad \mathcal{L}_{\text{task}}\left(f(x; \theta_r, \theta_o)\right) \quad \text{s.t. } \theta_o \text{ fixed}, \tag{6}$$

where x is the model input, $f(\cdot)$ is the MoE model forward function, and $\mathcal{L}_{\text{task}}$ is the supervised loss specific to the context-dependent task. In this way, gradient descent optimization is applied only with respect to θ_r:

$$\theta_r \leftarrow \theta_r - \eta \cdot \nabla_{\theta_r} \mathcal{L}_{\text{task}}(f(x; \theta_r, \theta_o)) \tag{7}$$

where η denotes the learning rate.

To identify the context-faithful experts, we define the **Context-dependence Ratio** $r_i^{(\ell)}$ for expert e_i in the ℓ-th layer, measuring the frequency for which expert i is selected by the tuned router network across the training dataset. Formally:

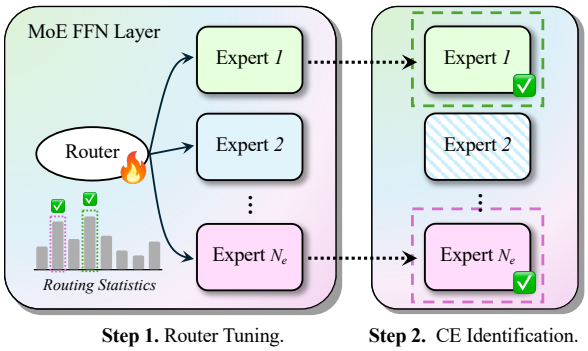


Figure 2: Illustration of Router Lens, where the Context-faithful Experts (CE) in each layer are identified by the tuned router network.

$$r_i^{(\ell)} = \frac{1}{N_s} \sum_{j=1}^{N_s} \frac{1}{L_j} \sum_{t=1}^{L_j} \frac{\mathbb{1}\left(g_{i,t}^{(\ell,j)} > 0\right)}{k} \tag{8}$$

where N_s is the total number of input samples, L_j is the number of tokens in the j-th sample. The value of indicator function $\mathbb{1}(g_{i,t}^{(\ell,j)} > 0)$ equals 1 if expert i is selected (i.e., receives a non-zero gating weight), and 0 otherwise. A higher $r_i^{(\ell)}$ indicates that expert i is selected more frequently across samples of context-dependent tasks, reflecting its importance to context utilization. Then, we identify the experts with top-k high context-dependent ratio as context-faithful experts.

3.2 Experimental Setting

Below, we list the datasets, metrics, and models used in the empirical study. For implementation details, please refer to the Appendix A.

Dataset and Metrics We evaluate our approach on several widely used datasets for context-dependent tasks, including SQuAD (Rajpurkar et al., 2016), NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), NQ-Swap (Longpre et al., 2021), and ConfiQA (Multi-Conflicts subset) (Bi et al., 2024). Dataset statistics are summarized in Table 6. Among these, SQuAD, NQ, and HotpotQA are classic question answering benchmarks that span a spectrum of reasoning challenges upon context, from local comprehension to multi-hop reasoning. Moreover, we also include two counterfactual QA benchmarks: NQ-Swap and ConfiQA, offering a more challenging and complementary evaluation of a model's ability to identify and rely on the correct evidence. For all datasets, we report

Models	SQuAD		NQ		HotpotQA		NQ-Swap		ConfiQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
OLMoE-1B-7B	26.6	49.4	18.3	39.9	27.0	46.1	28.1	40.5	38.7	49.9
w/ Router Tuning	**80.5**	**88.1**	**62.4**	**75.3**	**60.4**	**76.1**	**76.4**	**77.8**	**76.9**	**79.9**
DeepSeek-V2-Lite	25.2	48.6	25.9	48.1	30.8	50.4	27.4	38.0	19.1	35.3
w/ Router Tuning	**83.6**	**91.1**	**65.1**	**77.5**	**61.7**	**77.7**	**82.2**	**84.3**	**75.3**	**78.2**
MiniCPM-MoE-8x2B	45.8	65.1	35.1	55.8	38.9	57.3	42.8	50.5	38.6	48.2
w/ Router Tuning	**80.5**	**89.0**	**61.1**	**74.3**	**60.7**	**76.6**	**71.7**	**74.0**	**65.7**	**70.3**
Mixtral-8x7B	21.1	43.5	19.6	41.3	25.5	43.4	16.3	29.9	12.3	20.3
w/ Router Tuning	**49.6**	**63.2**	**44.0**	**62.4**	**57.0**	**73.2**	**64.1**	**67.1**	**67.6**	**74.9**

Table 1: The performance comparison between untuned and router tuned MoE models on context-dependent tasks.

Models	Methods	EM	F1
OLMoE-1B-7B	Base	24.1	34.6
	RT	**100**	**100**
MiniCPM-MoE-8x2B	Base	37.0	45.2
	RT	**99.8**	**99.8**

Table 2: The performance of Router Tuning (RT) on CounterFact dataset that needs no complex reasoning.

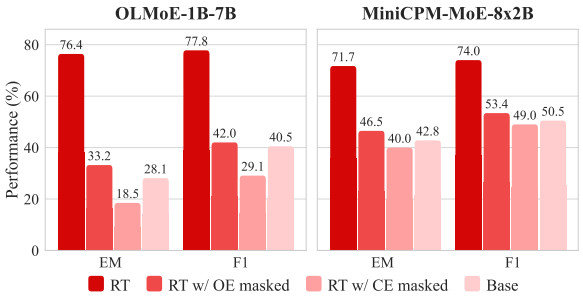


Figure 3: Comparison of the performance impact on NQ-Swap when masking k original experts (OE) vs. the top-k context-faithful experts (CE), evaluated on router-tuned (RT) OLMoE-1B-7B and MiniCPM-MoE-8x2B models, relative to their respective base models (Base).

Exact Match (EM) and token-level F1 scores as evaluation metrics. For additional dataset details, please refer to Appendix B.

MoE Models We select four widely used open-source MoE-based LLMs for evaluation: OLMoE-1B-7B (Muennighoff et al., 2024), DeepSeek-V2-Lite (DeepSeek-AI et al., 2024), MiniCPM-MoE-8x2B (Hu et al., 2024), and Mixtral-8x7B (Jiang et al., 2024). These models cover a diverse range of configurations in terms of the number of experts (from 8 to 64) and overall model sizes (from 7B to 47B), providing a comprehensive testbed for analyzing expert behavior in context-dependent scenarios. Detailed configurations for each model are listed in Table 7. To ensure consistent and deterministic outputs across models, we adopt greedy decoding (Germann, 2003) for all experiments. For additional model details, please refer to Appendix C.

3.3 Results of Expert Elicitation

In this section, we present the results of router tuning and analyze the role of context-faithful experts in enhancing context utilization.

Table 1 summarizes the model performance on a variety of context-dependent benchmarks, both before and after router tuning. Across all evaluated models and tasks, router tuning consistently yields substantial improvements over the base models. This demonstrates that modifying only the expert selection mechanism significantly boosts the performance of context-dependent tasks, indicating the presence of **context-faithful experts**.

To assess whether the performance gains from the newly selected experts arise solely from enhanced compositional reasoning rather than effective context utilization, we also conduct experiments on the CounterFact dataset (Meng et al., 2022), where the model must rely on the provided counterfactual context to answer a simple question correctly. The task does not require complex reasoning, making it well-suited for isolating and evaluating the role of context-faithful experts. Table 2 reports the EM and F1 scores of OLMoE-1B-7B and MiniCPM-MoE-8x2B on CounterFact. We observe substantial performance improvements after router tuning, further providing strong evidence for the existence of context-faithful experts.

To examine the importance of these experts, we conduct a causal intervention experiment by masking the identified context-faithful experts on the router-tuned models. Specifically, by setting their gating weights to zero and measuring the resulting performance degradation. We conduct this experiment on the NQ-Swap dataset using the

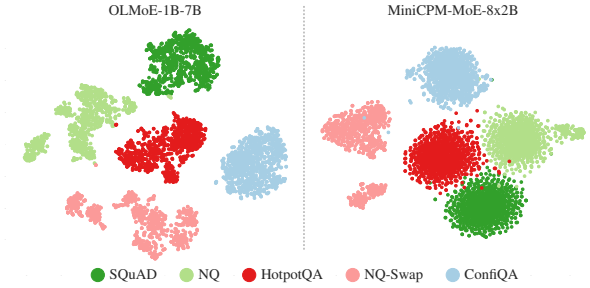


Figure 4: t-SNE visualization of context-faithful expert activation patterns in OLMoE-1B-7B and MiniCPM-MoE-8×2B. For each model, 1,000 examples per dataset are randomly selected for projection.

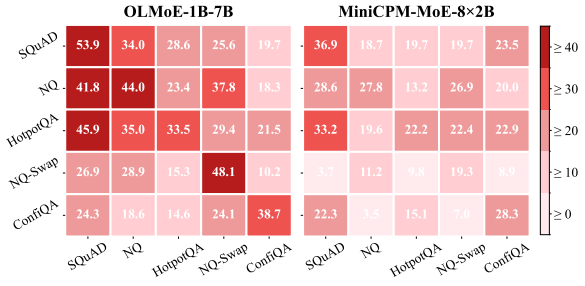


Figure 5: Cross-task transfer performance of router-tuned models. Each cell shows the absolute EM score improvement over the base model, where the model is trained on the dataset in row i and evaluated on the dataset in column j.

OLMoE-1B-7B and MiniCPM-MoE-8x2B models. As illustrated in Figure 3, masking the context-faithful experts leads to substantial drops in performance, with EM score decreasing by 73.2% for OLMoE-1B-7B and 44.2% for MiniCPM-MoE-8x2B. In contrast, masking an equal number of originally selected experts results in smaller performance declines. These findings highlight that *context-faithful experts play a critical role in solving context-dependent tasks*.

3.4 Consistency of Context-Faithful Experts

We further investigate whether the same context-faithful experts are consistently activated across different context-dependent tasks. To this end, we analyze expert activation patterns in OLMoE-1B-7B and MiniCPM-MoE-8×2B. For each input sample, we calculate the activation frequency of context-faithful experts across all layers and concatenate these values into a feature vector. We then apply t-SNE (Van der Maaten and Hinton, 2008) to project these vectors into a 2D space for visualization. As shown in Figure 4, samples from different datasets form clearly separable clusters in both OLMoE-1B-7B and MiniCPM-MoE-8×2B. This indicates that the router learns task-specific activation patterns for context-faithful experts, demonstrating that the model adapts its routing behavior based on the contextual requirements of each task, highlighting both the interpretability and task-awareness enabled by the router-tuning mechanism.

While the results above suggest that different tasks benefit from distinct expert configurations, an important question remains: *Can a tuned router network generalize its ability to activate context-faithful experts to unseen tasks?* To explore this, we assess the transferability of the tuned router by applying model router-tuned on one dataset to other datasets, without any additional adaptation. As shown in Figure 5, router-tuned models consistently outperform their base counterparts on unseen tasks. This suggests that the learned routing strategies capture generalizable and context-aware behaviors, enabling effective expert selection even outside the training domain.

3.5 Inner Working Mechanism of Context-faithful Experts

Moreover, we investigate how context-faithful experts contribute to improving context faithfulness. In Transformer-based models, the self-attention mechanism plays a central role in perceiving and integrating contextual information (Sun et al., 2025). By assigning higher attention scores to relevant context tokens, the model can more effectively utilize the provided context. To assess whether context-faithful experts enhance this mechanism, we examine whether their activation leads to increased attention over contextual tokens compared to the untuned model. To this end, we introduce the **Context Attention Gain (CAG)** metric.

Let $\mathbf{A}_i^{(\ell)} \in \mathbb{R}^{N_h \times L_s \times L_s}$ denote the attention matrices at layer ℓ for the i-th sample, where N_h is the number of heads and L_s is the sequence length of input. Let $\mathcal{C}_i \subseteq \{1, \ldots, L_c\}$ represent the set of context token indices in sample i. We average attention across all heads for the last token:

$$\bar{\mathbf{a}}_i^{(\ell)} = \frac{1}{N_h} \sum_{h=1}^{N_h} \mathbf{A}_i^{(\ell,h)}[L_s, :] \tag{9}$$

Then, compute the attention mass assigned to all context tokens:

$$\alpha_i^{(\ell)} = \sum_{k \in \mathcal{C}_i} \bar{\mathbf{a}}_i^{(\ell)}[k] \tag{10}$$

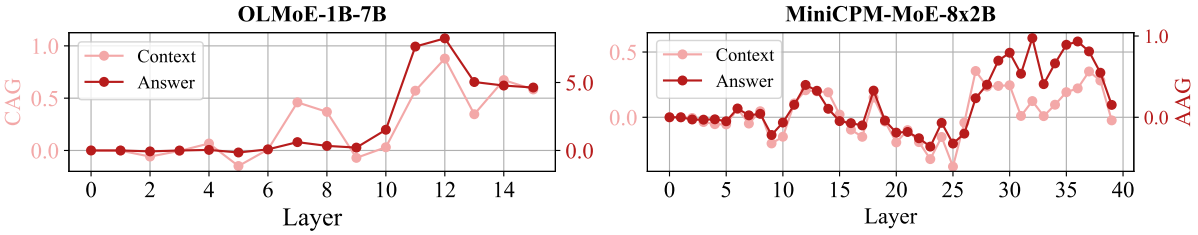
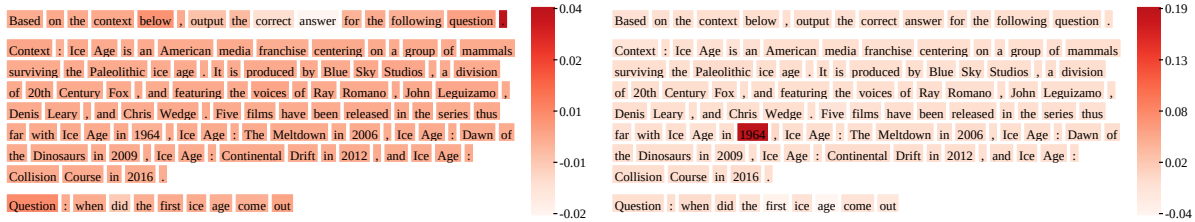


Figure 6: Visualization of layer-wise attention gain on context and answer (CAG and AAG) for the router-tuned model over the untuned model on the NQ-Swap test set.



(a) Layer 6 (b) Layer 12

Figure 7: Case study of attention gain from context-faithful experts in OLMoE-1B-7B on an NQ-Swap example. At (a) Layer 6 and (b) Layer 12, i.e., mid-level layer and deeper layer, the router-tuned model progressively increases attention to the context and answer tokens (i.e., "1964"), illustrating a "think twice" mechanism. Notably, the base model fails on this example, while the router-tuned model provides the correct answer.

Let $\alpha^{(\ell)}_{i,\text{base}}$ and $\alpha^{(\ell)}_{i,\text{rt}}$ denote the attention to context tokens for the base and router-tuned models, respectively. Finally, compute the average difference in context attention over all N_s samples:

$$\text{CAG}^{(\ell)} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{\alpha^{(\ell)}_{i,\text{rt}} - \alpha^{(\ell)}_{i,\text{base}}}{\alpha^{(\ell)}_{i,\text{base}}} \tag{11}$$

Similarly, we define **Answer Attention Gain (AAG)** indicator for analyzing the effect of context-faithful experts on the answer within context.

As shown in Figure 6, we visualize the CAG and AAG scores across all layers for OLMoE-1B-7B-Instruct and MiniCPM-MoE-8×2B, evaluated on the NQ-Swap test set. We observe that in both mid and deep layers, the router-tuned models allocate more attention to the overall context and, in particular, to the key answer tokens within the context, compared to their untuned counterparts. This amplification suggests that context-faithful experts effectively recalibrate the model's focus toward the most relevant parts of the context.

We hypothesize that this layer-wise attention amplification reflects a two-stage reasoning process. As illustrated in Figure 7, context-faithful experts in the mid-level layers help the model initially broaden its attention across the entire context—effectively "scanning" and identifying potentially relevant information. In the deeper layers, the model then narrows its focus, concentrating attention on the most critical spans within that context. In other words, the model appears to "think twice": first by attending broadly to gather context, and then by selectively reinforcing attention on crucial details to inform its final decision (For more cases, please refer to Figure 14).

In addition, we define **Answer Probability Gain (APG)** metric to examine how context-faithful experts influence the model's intermediate decision-making process. Let $\mathbf{h}^{(\ell)}_{i,last} \in \mathbb{R}^d$ be the hidden state output by the MoE module at layer ℓ for the last token of input i where d is hidden size, and let y_i be its true answer tokens. We employ Logit Lens (Dar et al., 2023), using the projection matrix $\mathbf{W} \in \mathbb{R}^{V\times d}$ (where V is the vocabulary size) of language model head, to compute the probability $p^{(\ell)}_m(y_i)$ assigned to y_i at layer ℓ under model $m \in \{\text{base}, \text{rt}\}$ where rt refers to "router-tuned". We then compute the average probability gain of the correct answer over all N_s samples:

$$\text{APG}^{(\ell)} = \frac{1}{N_s}\sum_{i=1}^{N_s}\frac{p^{(\ell)}_{\text{rt}}(y_i) - p^{(\ell)}_{\text{base}}(y_i)}{p^{(\ell)}_{\text{base}}(y_i)} \tag{12}$$

This metric quantifies how much context-faithful experts increase the model's implicit confidence in the correct answer at each layer.

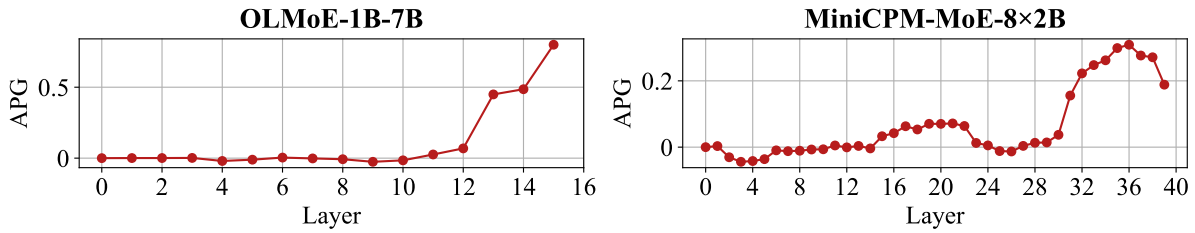


Figure 8: Layer-wise visualization of the Answer Probability Gain (APG) of the router-tuned models (OLMoE-1B-7B and MiniCPM-MoE-8x2B) over their untuned counterparts on the NQ-Swap test set, illustrating how context-faithful experts progressively enhance the model's confidence in the correct answer across layers.

Figure 8 visualizes the layer-wise APG scores for OLMoE-1B-7B and MiniCPM-MoE-8×2B on the NQ-Swap test set. We observe that, due to the context-faithful experts' effective amplification of attention to both the broader context and the key answer tokens, the models are able to more accurately integrate relevant contextual information, leading to a substantial increase in the predicted probability of the correct answer in the deeper layers.

4 Context Faithfulness Optimization

4.1 Context-faithful Expert Fine-Tuning

Our previous analysis reveals a clear distinction between context-faithful experts and others: the former selectively amplify attention to relevant contextual information and key answer spans. By focusing on fine-tuning these high-impact experts, it is expected to achieve: (1) **Efficiency:** Significantly reduce the number of trainable parameters. (2) **Robustness:** Mitigate overfitting by preserving the pretrained weights of non-critical experts. (3) **Effectiveness:** Exploit the specialized routing behavior to further improve context utilization.

Motivated by this consideration, we propose **Context-faithful Expert Fine-Tuning (CEFT)** — a two-stage training strategy (Algorithm 1). CEFT first identifies context-faithful experts using Router Lens analysis, then fine-tunes only these selected experts while freezing the rest of the model.

Algorithm 1 CEFT

Require: Training dataset $\mathcal{D}_{\text{train}}$, MoE model $\mathcal{M}$, number of selected experts k
Ensure: Fine-tuned MoE model $\mathcal{M}^*$

// Phase 1: Expert Identification
Freeze all parameters of $\mathcal{M}$ except the router
Optimizing router parameters on $\mathcal{D}_{\text{train}}$
for each layer ℓ in the MoE model **do**
  for each expert e_i in layer ℓ **do**
    Compute context-dependence ratio $r_i^{(\ell)}$
  end for
  Select top-k experts with the highest $r_i^{(\ell)}$ as context-faithful experts $\mathcal{E}_{\text{context}}^{(\ell)}$
end for
// Phase 2: Identified Expert Fine-Tuning
Freeze all parameters of $\mathcal{M}$ except the experts in $\mathcal{E}_{\text{context}}^{(\ell)}$ for each layer ℓ
Train $\mathcal{M}$ on $\mathcal{D}_{\text{train}}$ to obtain final model $\mathcal{M}^*$

4.2 Empirical Results

Main Performance To evaluate the effectiveness of the proposed CEFT approach, we compare it against two baseline strategies: standard **Fully Fine-Tuning (FFT)** and **Expert-Specialized Fine-Tuning (ESFT)** (Wang et al., 2024b). While ESFT also targets expert-level adaptation by only tuning a subset of experts deemed relevant to the task, it relies on the original, untuned router network to identify these experts, which may result in sub-optimal expert selection due to the influence of load balancing constraint and the limited task-awareness of the untuned router. In contrast, CEFT first adapts the router to reveal truly context-faithful experts—those that significantly enhance context-dependent reasoning—and then fine-tunes only these high-impact experts for more precise and efficient adaptation. For implementation details, please refer to the Appendix A. As shown in Table 3, CEFT consistently matches or surpasses FFT and ESFT across MoE models and benchmarks, indicating its parameter-effectiveness in leveraging contextual information for improving generalization in context-dependent tasks.

Mitigation of Catastrophic Forgetting We further evaluate OLMoE-1B-7B and MiniCPM-MoE-8x2B on MMLU (Hendrycks et al., 2021) to assess the impact of RT, FFT, and CEFT on the models' original capabilities. As shown in Table 4, the decline in MMLU performance is approximately

Methods	SQuAD		NQ		HotpotQA		NQ-Swap		ConfiQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
OLMoE-1B-7B										
FFT	81.6	88.8	62.1	75.3	63.3	78.8	88.3	88.8	85.5	88.9
ESFT	82.6	89.7	62.6	75.5	63.0	78.7	**91.4**	**91.7**	86.9	89.2
CEFT	**83.1**	**90.3**	**64.1**	**76.9**	**63.8**	**79.1**	90.5	90.8	**87.1**	**89.4**
DeepSeek-V2-Lite										
FFT	85.4	92.0	67.7	79.6	62.2	78.3	91.3	92.5	**87.3**	**90.6**
ESFT	86.2	92.9	68.5	80.3	**65.9**	**81.6**	91.7	92.5	86.3	88.8
CEFT	**87.0**	**93.5**	**69.0**	**81.2**	63.7	80.1	**92.7**	**93.6**	**87.3**	89.4
MiniCPM-MoE-8x2B										
FFT	81.5	89.6	62.8	75.9	62.8	78.6	90.4	91.0	**86.3**	**88.8**
ESFT	83.3	90.7	65.9	78.6	64.3	80.4	88.1	89.1	84.4	87.2
CEFT	**83.8**	**91.6**	**66.6**	**79.3**	**64.9**	**80.9**	**90.7**	**92.0**	84.8	87.5
Mixtral-8x7B										
FFT	65.3	76.5	45.3	59.7	56.3	73.6	68.0	69.6	**83.1**	87.8
ESFT	66.1	77.2	45.9	60.3	**56.8**	73.7	67.2	68.4	79.3	85.9
CEFT	**68.8**	**78.9**	**47.1**	**61.1**	56.7	**74.1**	**68.7**	**71.4**	81.7	**87.9**

Table 3: Performance comparison of Fully Fine-Tuning (FFT), Expert-Specialized Fine-Tuning (ESFT) (Wang et al., 2024b), and Context-faithful Expert Fine-Tuning (CEFT). **Bold** numbers indicate the best performance, while underlined numbers denote the second-best.

Training Dataset	Base	FFT	CEFT	RT
OLMoE-1B-7B				
NQ-Swap	50.5	32.1	43.6	48.0
ConfiQA	50.5	45.1	48.3	49.6
MiniCPM-MoE-8x2B				
NQ-Swap	55.7	46.0	54.1	55.5
ConfiQA	55.7	53.3	55.4	55.2

Table 4: Performance of OLMoE-1B-7B and MiniCPM-MoE-8x2B models on MMLU benchmark after employing different training methods.

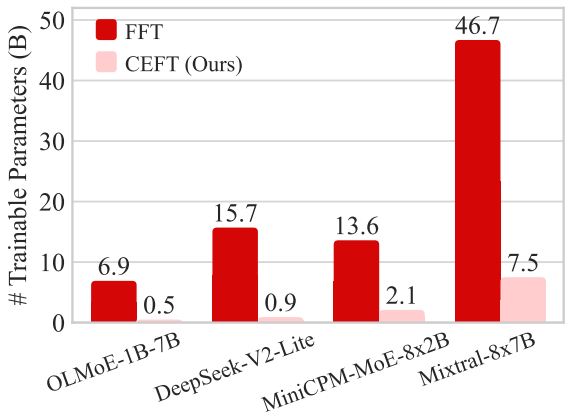


Figure 9: Comparison of trainable parameters (fewer is better) between Fully Fine-Tuning (FFT) and Context-faithful Expert Fine-Tuning (CEFT).

proportional to the number of trainable parameters, with CEFT being substantially less susceptible to catastrophic forgetting compared to FFT.

Training Efficiency Figure 9 illustrates the substantial reduction in trainable parameters achieved by CEFT. For instance, OLMoE-1B-7B with FFT requires 6.9B parameters, whereas CEFT needs only 0.5B—a 13.8× reduction. This highlights CEFT as an efficient approach for adapting MoE models to context-dependent tasks, delivering comparable performance with far fewer trainable parameters than FFT.

Effect of Trainable Expert Count We conduct an ablation study on the number of training experts used in CEFT for OLMoE-1B-7B, using the NQ-Swap and ConfiQA datasets. Model performance is evaluated with varying numbers of experts: 1, 4, 8, 16, and 32. As shown in the Table 5, increasing the number of experts generally improves performance up to a point, after which the gains plateau or slightly decline. This behavior can be attributed to the following: (1) Training too few experts may fail to capture enough context-faithful experts, limiting the model's ability to leverage context information. (2) Training too many experts may involve irrelevant or noisy experts, leading to overfitting and degraded generalization. These findings suggest that using a moderate number of experts provides a favorable trade-off between performance and effi-

#Experts	NQ-Swap		ConfiQA	
	EM	F1	EM	F1
1	87.5	88.5	84.4	88.0
4	90.2	91.0	85.4	88.3
8	90.5	90.8	87.1	89.4
16	88.8	89.4	88.0	91.1
32	88.1	88.8	85.6	88.8

Table 5: Performance of OLMoE-1B–7B model on NQ-Swap and ConfiQA with varying numbers of experts trained by CEFT.

ciency in CEFT. Notably, we observe that setting the number of training experts to match the actual number of activated experts in the model often yields strong results (8 for OLMoE-1B-7B).

5 Related works

Context Faithfulness Context faithfulness refers to the extent to which a model's output remains accurate, consistent, and grounded in the provided context. Unlike factual consistency, it emphasizes alignment specifically with the input context rather than with external world knowledge (Zhou et al., 2023; Xu et al., 2024). In tasks such as RAG, maintaining context faithfulness is critical to ensure the reliability and trustworthiness of generated content (Zhou et al., 2024). Recent studies have shown that LLMs often generate responses that are fluent yet contextually unfaithful (Du et al., 2024; Xie et al., 2024). Various techniques, including prompt engineering (Zhou et al., 2023), decoding constraints (Shi et al., 2024), attention amplification (Sun et al., 2025), mechanistic approaches (Yu et al., 2023; Ortu et al., 2024; Minder et al., 2025), and learning-based methods such as RAG-fashioned fine-tuning (Zhang et al., 2024) and reinforcement learning with context-aware rewards (Bi et al., 2024), have been proposed to improve grounding in the given context. Differently, our work focuses on an under-explored direction: leveraging the expert specialization in terms of the context utilization to improve the MoE performance on context-dependent tasks.

Expert Specialization The MoE architecture replaces the standard FFN layer (Vaswani et al., 2017) with a modular MoE layer comprising multiple parallel experts and a router network to sparsely activate top-k experts (Xue et al., 2024). During the pretraining stage of MoE, a load balancing loss is typically incorporated to force balanced expert utilization (Dai et al., 2024), which encourages the formation of expert specialization. Recently, a lot of work has focused on the understanding of expert specialization, finding that different experts are responsible for different tokens (Muennighoff et al., 2024), domains (Jiang et al., 2024), tasks (Wang et al., 2024b), semantic units (Li and Zhou, 2025), or syntactic units (Antoine et al., 2025). These findings facilitate the targeted optimization of the specialized experts (Wang et al., 2024b). Inspired by these, our work investigates the expert specialization of context faithfulness.

6 Conclusion

In this work, we investigate expert specialization of context faithfulness in MoE models. We propose Router Lens to uncover context-faithful experts. Building upon this, our CEFT approach selectively updates the model parameters, yielding strong empirical gains while maintaining efficiency.

Looking forward, we see several exciting directions: combining mechanistic interpretability techniques such as SAE (Kang et al., 2025) to further unravel MoE experts; extending our methods to other forms of specialization such as reflection (Li et al., 2025) and reasoning (Liu et al., 2025). We hope this work inspires further research into aspect-specific expert discovery and optimization.

Limitations

The proposed Router Lens relies on fine-tuning the router network to discover context-faithful experts. Thus, it cannot identify such experts in a zero-shot or training-free setting. Moreover, improving the context utilization ability of these experts also depends on training procedures; we currently lack a method to directly enhance their behavior without additional supervision or optimization.

Ethics Statement

All experiments in this work are conducted on publicly available datasets commonly used in natural language processing research. No human subjects, personal data, or sensitive content are involved. We acknowledge that large-scale language models may carry risks of misuse, bias, or hallucination, and our work aims to improve their transparency and controllability, which may help mitigate such issues in the long term.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62376031).

References

Elie Antoine, Frédéric Béchet, and Philippe Langlais. 2025. Part-of-speech sensitivity of routers in mixture of experts models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6467–6474.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2024. Context-dpo: Aligning language models for context-faithfulness. *CoRR*, abs/2412.15280.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Stablemoe: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7085–7095.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, and 81 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *CoRR*, abs/2405.04434.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer C. White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Ulrich Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. 2025. Are the values of llms structurally aligned with humans? A causal perspective. In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 23147–23161.

Samreen Kazi, Shakeel Ahmed Khoja, and Ali Daud. 2023. A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, 56(S2):2509–2569.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti,

Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2025. Reflectevo: Improving meta introspection of small llms by learning self-reflection. In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 16948–16966.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Ziyue Li and Tianyi Zhou. 2025. Your mixture-of-experts LLM is secretly an embedding model for free. In *The Thirteenth International Conference on Learning Representations*.

Yang Liu, Jiaqi Li, and Zilong Zheng. 2025. Rulereasoner: Reinforced rule-based reasoning via domain-aware dynamic sampling. *CoRR*, abs/2506.08672.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*.

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

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. Controllable context sensitivity and the knob behind it. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2024. Olmoe: Open mixture-of-experts language models. *CoRR*, abs/2409.02060.

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

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. 2025. In-context editing: Learning knowledge from self-induced distributions. In *The Thirteenth International Conference on Learning Representations*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, pages 783–791.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. 2024a. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *CoRR*, abs/2408.15664.

Zihan Wang, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, and Yu Wu. 2024b. Let the expert stick to his last: Expert-specialized fine-tuning for sparse architectural large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 784–801.

Tong Wu, Junzhe Shen, Zixia Jia, Yuxuan Wang, and Zilong Zheng. 2025. From hours to minutes: Lossless acceleration of ultra long sequence generation up to 100k tokens. *CoRR*, abs/2502.18890.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. Openmoe: An early effort on open mixture-of-experts language models. In *Forty-first International Conference on Machine Learning*.

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. *CoRR*, abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.

Jianfei Zhang, Bei Li, Jun Bai, Rumei Li, Yanmeng Wang, Chenghua Lin, and Wenge Rong. 2025. Selecting demonstrations for many-shot in-context learning via gradient matching. In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 11686–11704.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: adapting language model to domain specific RAG. *CoRR*, abs/2403.10131.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 14544–14556.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *CoRR*, abs/2409.10102.

Prompt template used in experiments

Based on the context below, output the correct answer for the following question.

Context: ...Mission commander Paul and pilot Buzz Aldrin, landed the lunar module Eagle on July 20, 1969, at 20:18 UTC. Paul became the first human to step onto the lunar surface six hours after...

Question: Who took the first steps on the moon in 1969?

Figure 10: Prompt template used during training and evaluation for context-dependent tasks.

A Implementation Details

Prompt Template To evaluate MoE models on context-dependent tasks, we adopt a unified prompt template as illustrated in Figure 10. This template explicitly instructs the model to generate an answer based on the given context and question.

Router Tuning We fine-tune the model for 1 epoch on SQuAD, NQ, and HotpotQA datasets, and 3 epochs on NQ-Swap and ConfiQA datasets. All experiments use the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 5e-4, a warmup ratio of 0.1, and cosine learning rate decay. The batch size is set to 8, and the maximum sequence length is 300. All the experiments are run on one NVIDIA A100 80GB GPU.

Context-faithful Expert Fine-Tuning The hyperparameters are identical to those used in Router Tuning, except that CEFT fine-tunes the parameters of the context-faithful experts instead of the router.

B Datasets

SQuAD It is a widely used benchmark for extractive question answering (Rajpurkar et al., 2016). Each example consists of a question and a context paragraph from Wikipedia, with the answer being a span within the paragraph (CC BY-SA 4.0 license).

NQ It is a large-scale dataset collected from real user queries issued to the Google search engine (Kwiatkowski et al., 2019). Each question is paired with a Wikipedia page. Compared to SQuAD, NQ is more challenging due to its longer contexts and more diverse question types (Apache-2.0 license).

HotpotQA It is a multi-hop question answering dataset where answering each question requires reasoning over multiple paragraphs including both supporting facts and distractor documents (CC BY-SA 4.0 license) (Yang et al., 2018).

Dataset	#Train	#Val	#Test
SQuAD	82,258	4,330	10,507
NQ	98,867	5,204	12,836
HotpotQA	69,281	3,647	5,904
NQ-Swap	3,000	746	1,000
ConfiQA	4,000	500	1,500

Table 6: The statistics of datasets used in this work.

Models	Params (A/F)	#Experts (A/F)
OLMoE-1B-7B	1.3B / 6.9B	8 / 64
DeepSeek-V2-Lite	2.4B / 15.7B	6 / 64
MiniCPM-MoE-8x2B	4B / 13.6B	2 / 8
Mixtral-8x7B	12.9B / 46.7B	2 / 8

Table 7: Overview of the MoE models used in this work. A and F refer to "Activated" and "Full".

NQ-Swap It is a variant of the NQ (Longpre et al., 2021), where part of the relevant document context is swapped with similar-looking but semantically misleading content, forcing the model to rely more heavily on context utilization (MIT License).

ConfiQA It is designed to evaluate the context-faithfulness when knowledge conflicts (Bi et al., 2024). The dataset is constructed by sampling real-world facts, generating multi-hop paths, and creating counterfactual contexts by replacing entities. We use its MC (Multi-Conflicts) subset that includes multiple counterfactuals (MIT License).

C Models

OLMoE-1B-7B OLMoE-1B-7B is a fully open-source MoE developed by researchers from the Allen Institute for AI (Muennighoff et al., 2024). This model consists of 16 layers where 8 out of 64 experts are activated in each, and has 7 billion parameters but activates only 1 billion parameters per input token. We use its Instruct version in this work: OLMoE-1B-7B-0924-Instruct.

DeepSeek-V2-Lite This is a small version of the DeepSeek-V2 model (DeepSeek-AI et al., 2024). It has 27 layers where each MoE layer consists of 2 shared experts, 64 routed experts, and 6 activated experts. This configuration results in a total of 15.7 billion parameters, with 2.4 billion activated for each token. We use the parameters of its chat version: DeepSeek-V2-Lite-Chat.

MiniCPM-MoE-8x2B MiniCPM-MoE-8x2B is a MoE variant of the MiniCPM model (Hu et al., 2024). It initializes using sparse upcycling, replacing MLP layers with MoE layers. With two 2 of 8 experts activated, it results in approximately 4B activated parameters. This approach significantly boosts performance across various benchmarks, while maintaining computational efficiency.

Mixtral-8x7B Mixtral-8x7B is built upon the Mistral 7B model but incorporates 8 experts in each layer (Jiang et al., 2024). This design allows each token to access a vast pool of 47B parameters, yet only 13B parameters (2 experts in each layer) are actively used during inference. We use its Instruct version, Mixtral-8x7B-Instruct-v0.1.

Methods	NQ-Swap			ConfiQA		
	Acc_{LLM}	EM	F1	Acc_{LLM}	EM	F1
FFT	91.9	88.3	88.8	90.5	85.5	88.9
ESFT	93.6	91.4	91.7	90.8	86.9	89.2
CEFT	92.4	90.5	90.8	91.5	87.1	89.4

Table 8: The accuracy (Acc_{LLM}) computed by LLM-as-a-judge, EM, and F1 scores of OLMoE-1B-7B on NQ-Swap and ConfiQA.

D Evaluation by LLM-as-a-judge

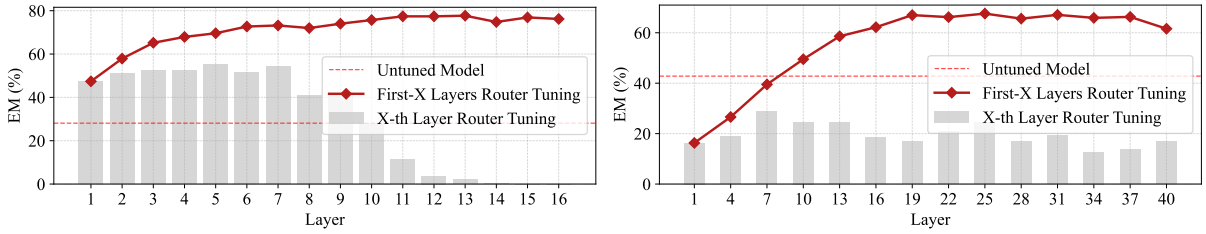
EM and F1 are widely used metrics for evaluating question answering tasks. Both are standard in benchmarks such as SQuAD and NQ, ensuring comparability and providing a comprehensive assessment of model performance. Additionally, we supplement our results with LLM-as-a-judge for OLMoE-1B-7B on the NQ-Swap and ConfiQA datasets. Specifically, we employ Qwen3-32B (Yang et al., 2025) to assess the correctness of the MoE models' predictions.

The Table 8 reports the accuracy computed by the LLM-as-a-judge (Acc_{LLM}), alongside the EM and F1 scores for OLMoE-1B-7B. As expected, Acc_{LLM} tends to be higher than EM, reflecting its more permissive evaluation criteria. However, all three metrics show consistent performance trends across models and datasets.

E Layer-wise Context-faithful Expert Collaboration Analysis

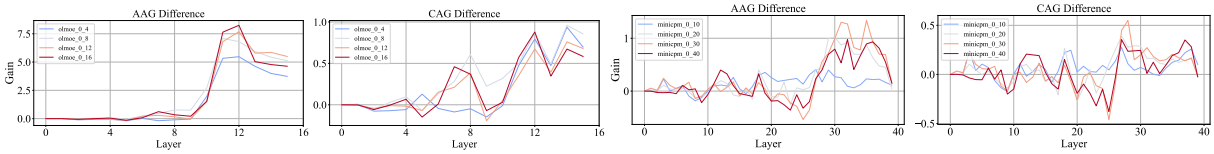
We investigate the collaborative effect of context-faithful experts—specifically, whether activating more such experts leads to greater performance gains on context-dependent tasks.

To this end, we perform two sets of fine-tuning experiments on the NQ-Swap dataset using OLMoE-1B-7B and MiniCPM-MoE-8×2B: (1) X-th Layer Router Tuning, where we fine-tune the router of a single layer at a time, and (2) First-



(a) OLMoE-1B-7B (b) MiniCPM-MoE-8x2B

Figure 11: The performance on NQ-Swap when fine-tuning the router of single layer (X-th Layer Router Tuning) and fine-tuning the routers of first layers (First-X Layers Router Tuning).



(a) OLMoE-1B-7B (b) MiniCPM-MoE-8x2B

Figure 12: Visualization of layer-wise Context Attention Gain (CAG) and Answer Attention Gain (AAG) for router-tuned models with varying numbers of tuned layers, relative to the untuned baseline, on the NQ-Swap test set.

X Layers Router Tuning, where we progressively fine-tune the routers of the first X layers.

As shown in Figure 11, the results reveal clear differences between two models. For X-th Layer Router Tuning, OLMoE-1B-7B shows substantial improvements even when tuning only a single early layer, suggesting that adjusting routing decisions at individual layers can significantly enhance context utilization. In contrast, MiniCPM-MoE-8×2B exhibits negligible gains from tuning single layer, implying that it requires multi-layer router adaptation to benefit from tuning. We hypothesize that this discrepancy arises partly from architectural differences: OLMoE-1B-7B employs more experts per layer (64) and activates more experts per token (8) than MiniCPM-MoE-8×2B, making each routing decision more influential and context-sensitive. For First-X Layers Router Tuning, both models show incremental performance improvements as more layers are tuned, indicating that context-faithful experts across layers collaborate to improve performance. However, the gains gradually saturate with the inclusion of more tunable layers, suggesting that some experts across different layers may perform redundant roles.

We further analyze the attention gain on both the context and the answer across different First-X Layers Router-Tuned models to examine whether different groups of context-faithful experts exhibit similar effects. As shown in Figure 12, the observed "think twice" phenomenon is not an isolated artifact caused by a particular combination of context-faithful experts. Instead, it is a consistent behavior that emerges across models with varying numbers of tuned layers. To better understand the influence of this phenomenon—particularly the substantial increase in attention gain on both the context and the answer—we visualize the relationship via a scatter plot. Additionally, we compute the Pearson correlation coefficient between the attention gain on the answer and the corresponding performance improvements. As illustrated in Figure 13, we observe a strong positive correlation (r = 0.95), indicating that greater attention gain on the answer is closely associated with higher model performance. This result provides further evidence that context-faithful experts enhance the model's ability to focus on the most relevant contextual information in deeper layers, ultimately contributing to more context-faithful predictions.

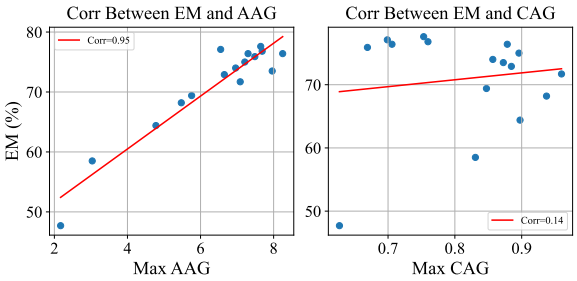


Figure 13: The correlation between the model's EM performance and the answer (context) attention gain.

Methods	BLEU	METEROR	ROUGE-L
OLMoE-1B-7B			
Base	1.9	35.5	21.7
RT	15.9	38.0	39.6
FFT	17.1	38.5	40.8
CEFT	16.2	38.3	40.8
MiniCPM-MoE-8x2B			
Base	1.8	35.6	22.8
RT	11.6	34.8	36.8
FFT	15.6	38.7	40.3
CEFT	15.9	38.3	39.7

Table 9: Performance on Gigaword benchmark.

F Performance on Non-QA Task

We also examine RT and CEFT on other context-dependent task, e.g., summarization. We evaluate OLMoE-1B-7B and MiniCPM-MoE-8x2B on the widely used Gigaword benchmark (Rush et al., 2015), where we sample 2000/500/500 examples as the train/validation/test sets, respectively.

Table 9 reports their performance in terms of BLEU, METEOR, and ROUGE-L scores. Our experiments show that router tuning consistently improves summarization performance over the base model. Furthermore, with CEFT—which selectively tunes the experts identified by the tuned router—we achieve performance comparable to full fine-tuning. These results validate the effectiveness and generalizability of our method beyond contextual QA and reasoning tasks.

G Performance on Task Independent on the Context

We further evaluate CEFT on the MemoTrap dataset (Shi et al., 2024), a benchmark specifically designed to detect whether language models memorize and regurgitate training data. Notably, MemoTrap is independent of any additional context, and therefore does not require context faithfulness.

Table 10 reports the accuracy of OLMoE-1B-7B and MiniCPM-MoE-8x2B on MemoTrap. From these results, we observe the following in this context-independent task: (1) Router tuning helps the model activate the appropriate experts for the task and leads to significant performance improvement. (2) CEFT, by selectively training the most relevant experts, achieves performance comparable to full fine-tuning. These findings demonstrate the effectiveness of both Router Lens and CEFT beyond context-dependent tasks.

Models	Base	RT	FFT	CEFT
OLMoE-1B-7B	56.6	89.0	96.5	96.1
MiniCPM-MoE-8x2B	63.5	86.4	96.5	95.1

Table 10: Performance on MemoTrap dataset.

Methods	EM	F1
Base	28.1	40.5
ContextCite	28.9	41.2
CFP	46.1	51.6
CAD	55.9	60.8
Context-DPO	65.7	69.8
CEFT	90.5	90.8
w/ ContextCite	90.7	91.2
w/ CAD	90.8	91.0

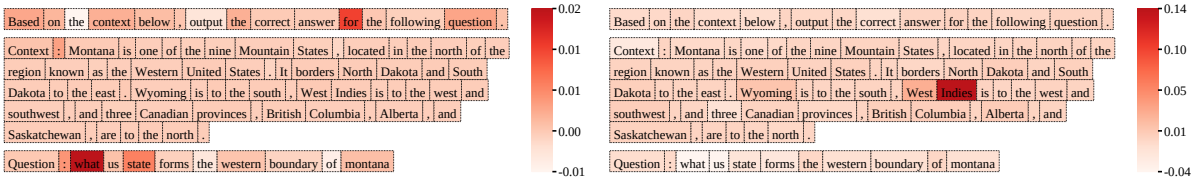
Table 11: Performance comparison of methods designed for improving context faithfulness on NQ-Swap.

H Comparison with Other Methods for Context Utilization

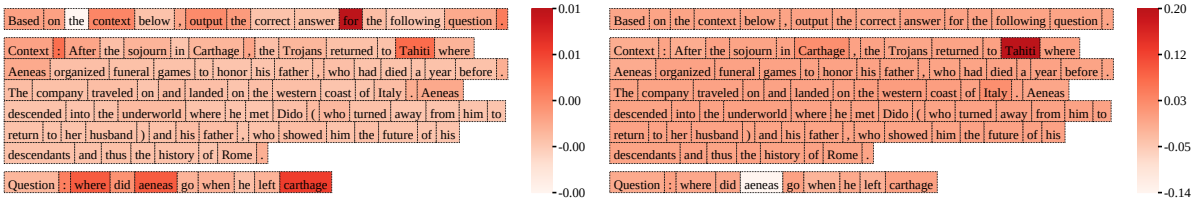
We further compare our CEFT approach with Context-Faithful Prompting (CFP) (Zhou et al., 2023), Context-Aware Decoding (CAD) (Shi et al., 2024), ContextCite (Cohen-Wang et al., 2024), and Context DPO (Bi et al., 2024) on the NQ-Swap and ConfiQA datasets. Among these, CFP, CAD, and ContextCite are unsupervised methods optimizing for prompt engineering, decoding, and context pruning, respectively. In contrast, Context DPO adopts a direct preference optimization to guide LLMs to prefer context-faithful outputs.

Table 11 shows the performance comparison between CEFT and the above context-utilization methods. Across both datasets—NQ-Swap and ConfiQA, CEFT consistently achieves the best results, significantly outperforming all baselines in both EM and F1 scores. These results highlight the effectiveness of CEFT in leveraging context-faithful experts to model context-sensitive behavior more accurately than prior unsupervised prompting or decoding strategies, as well as preference-optimized approaches.

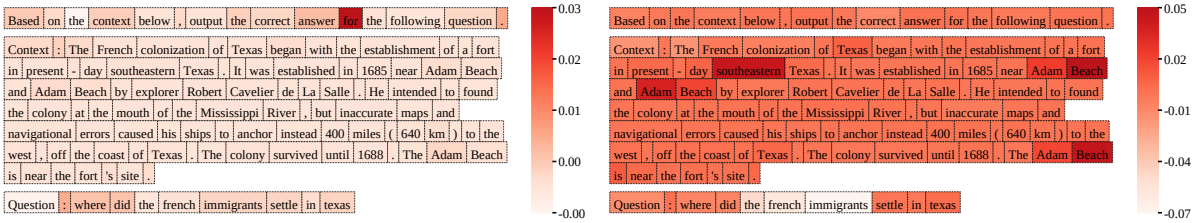
We combine ContextCite and CAD with CEFT to explore potential improvements, respectively. However, neither ContextCite nor CAD yields substantial gains. A possible explanation is that both context-level and decoding-level optimizations have limited capacity to enhance performance in this setting, particularly since they are unsupervised and not directly aligned with the model's training objectives.



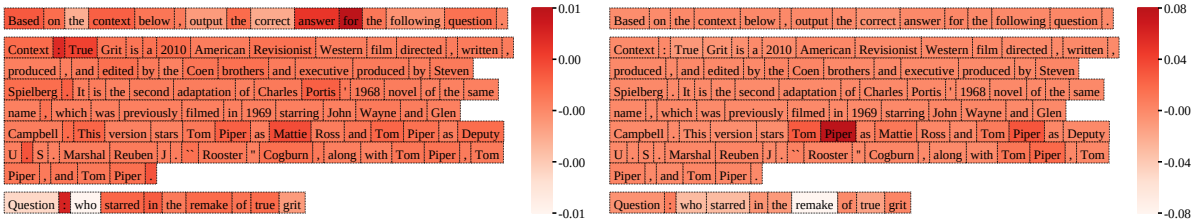
(a) Case 1-Layer 6 (b) Case 1-Layer 12



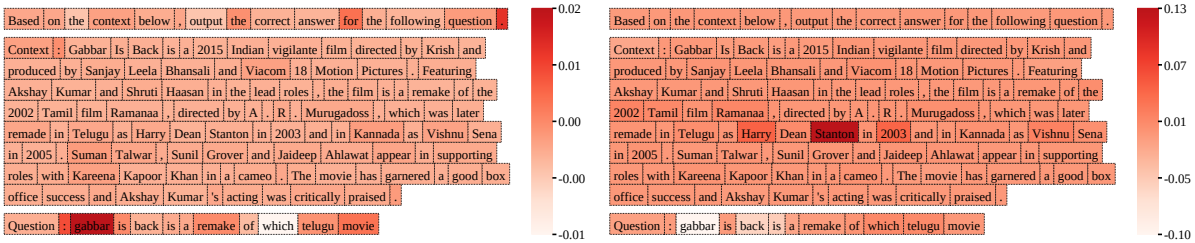
(c) Case 2-Layer 6 (d) Case 2-Layer 12



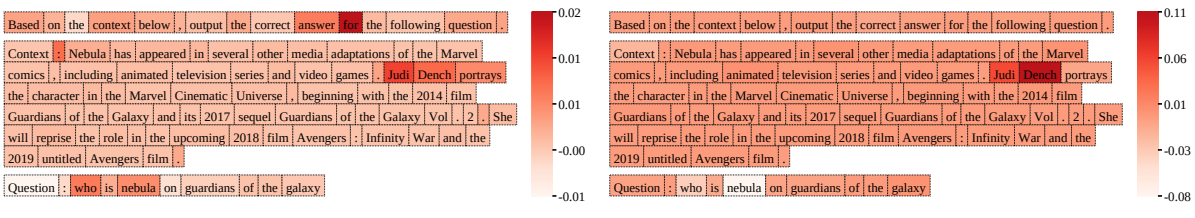
(e) Case 3-Layer 6 (f) Case 3-Layer 12



(g) Case 4-Layer 6 (h) Case 4-Layer 12



(i) Case 5-Layer 6 (j) Case 5-Layer 12



(k) Case 6-Layer 6 (l) Case 6-Layer 12

Figure 14: More cases of attention gain from context-faithful experts in OLMoE-1B-7B on NQ-Swap examples. The correct answers to these examples are “West Indies”, “Tahiti”, “Adam Beach”, “Tom Piper”, “Harry Dean Stanton”, and “Judi Dench”, respectively.