CausalMoE: Enhancing Expert Specialization and Interpretability through Cognitively-Inspired Causal Routing

Qian Huang[1], Jing Zhang[1], Peng Zhang (pzhang@tju.edu.cn)[1], Hui Gao[1], Xindian Ma[1], Changhao Song[1] & Liuxian Ge[1]

[1]Tianjin University, Tianjin, China

Abstract

Mixture-of-Experts (MoE) architectures enable scalable language modeling via sparse, input-dependent computation, but suffer from weak expert specialization and limited routing interpretability. From a cognitive perspective, this contrasts with theories of modular cognition that emphasize selective recruitment of functionally specialized systems. We propose CausalMoE, a causality-aware MoE framework that aligns expert activation with causal structure in language. CausalMoE introduces a specialized Causal Expert and a Causal Routing mechanism that identifies tokens along causally salient dependency structures inferred from attention patterns. The Causal Expert is specialized in processing sequences with strong causal and logical dependencies. To further encourage functional differentiation, we introduce a Causal Auxiliary Loss that reduces representational redundancy between experts. Integrated into canonical MoE architectures, including Switch Transformer and GLaM, CausalMoE achieves an average relative perplexity reduction of 18.42% on logic-intensive code datasets and a mean improvement of 6.83% on causal reasoning benchmarks, demonstrating structure-sensitive and interpretable computation allocation.

Keywords: Mixture-of-Experts; Causal Reasoning; Interpretability; Expert Specialization; Modular Cognition.

Introduction

The remarkable capabilities of large language models (LLMs) are largely driven by massive parameterization and large-scale training data. However, continued scaling of dense architectures raises fundamental limitations: all parameters are activated uniformly, incurring high computational cost and forcing heterogeneous knowledge into an undifferentiated parameter space, reducing flexibility and robustness in multi-domain reasoning.

From a cognitive science perspective, this dense and monolithic processing paradigm contrasts sharply with longstanding theories of human cognition. Classical accounts of modular cognition argue that the human mind is composed of partially specialized subsystems, each adapted to process particular types of information efficiently (Barkow et al., 1992; Fodor, 1983). Rather than engaging all cognitive resources uniformly, humans selectively recruit functionally specialized mechanisms depending on task demands.

Mixture-of-Experts (MoE) architectures provide a computational analogue to this modular view of cognition. By dynamically routing inputs to a sparse subset of expert subnetworks, MoE models substantially increase representational capacity without proportional increases in computation. This selective activation resembles cognitive resource allocation mechanisms in which attention and executive control guide processing toward task-relevant subsystems (Corbetta & Shulman, 2002). As such, MoE models constitute a promising architectural framework for exploring cognitively inspired large-scale models.

Despite these advantages, current MoE implementations fall short of their cognitive inspiration in two critical respects. First, expert specialization often fails to emerge reliably. In practice, commonly used routing mechanisms—such as fixed Top-K routing (Zoph et al., 2022) in Switch Transformer—frequently lead to expert collapse, where a small subset of experts dominates processing while others remain underutilized. Although auxiliary load-balancing losses are typically introduced to mitigate this issue, they often enforce overly uniform routing, paradoxically discouraging meaningful functional differentiation among experts (H. Guo et al., 2025).

Second, routing decisions remain largely opaque (Xue et al., 2024). Most existing routers operate as black-box statistical functions, offering limited insight into why specific inputs are assigned to particular experts or what functional role each expert ultimately learns. This lack of interpretability stands in contrast to human cognition, where selective processing is guided by structured reasoning mechanisms, including causal inference and goal-directed attention.

In human language comprehension and reasoning, causal structure plays a central organizing role. Extensive empirical and theoretical work demonstrates that readers actively construct mental or situation models that encode causal dependencies among events when processing narratives (Fletcher & Bloom, 1988; Johnson-Laird, 1983). Rather than treating all tokens or propositions as equally informative, humans preferentially attend to elements that participate in causal chains, as these are critical for coherence, prediction, and explanation. Dual-process theories further suggest that explicit causal tracking engages more deliberate analytical processes (Evans, 2008; Kahneman, 2011). Motivated by these cognitive principles, we propose CausalMoE, a causality-aware Mixture-of-Experts framework that explicitly integrates causal reasoning into the routing process. CausalMoE is designed to address both insufficient expert specialization and routing opacity by aligning expert activation with cognitively meaningful structures in the input.

CausalMoE introduces two key components. First, we pro-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

pose a Causal Expert, a specialized expert dedicated to processing token sequences that exhibit strong causal structure. This design mirrors the idea of functionally specialized cognitive subsystems that are preferentially recruited for causal and logical reasoning. Second, we introduce a Causal Routing mechanism that constructs a token-level causal graph using attention-based inter-block dependencies and identifies critical tokens along causal paths. This process is operationalized by integrating the Causal Explanations from Attention in Neural Networks (CLEANN) algorithm (Rohekar et al., 2023), which enables principled identification of causally salient elements—consistent with findings that causal chains are primary determinants of text comprehension (Fletcher & Bloom, 1988). Tokens identified as causally central are routed to the Causal Expert, while remaining tokens are processed by general-purpose experts.

By grounding routing decisions in causal structure rather than purely statistical correlations, CausalMoE shifts MoE routing toward a causality-aware and interpretable selection mechanism. Outputs from the Causal Expert are subsequently integrated with those of general experts to form the final MoE representation.

To further promote functional specialization, we introduce a Causal Auxiliary Loss that penalizes parameter similarity between the Causal Expert and general experts. This objective encourages representational decoupling and reduces redundant knowledge acquisition, facilitating the emergence of distinct expert roles—analogous to how cognitive specialization is maintained through functional differentiation.

Integrated into classic MoE architectures such as Switch Transformer and GLaM, CausalMoE achieves an 18.42% reduction in perplexity on code generation tasks, alongside an average accuracy improvement of 6.83% on pure causal reasoning tasks. These results demonstrate its ability to perform dynamic and interpretable computation allocation based on input structure.

Related Work

Mixture-of-Experts

The concept of Mixture-of-Experts (MoE) was first introduced by Jacobs et al. (1991), as a dynamic mechanism for combining multiple specialized sub-models through a routing mechanism. Modern advancements in MoE have significantly scaled its architecture to support large-scale models, with the router serving as the core component in MoE systems for large language models, directly influencing both efficiency and performance. Top-K routing is one of the earliest and most widely adopted mechanisms; however, it suffers from a lack of adaptability. Specifically, it fails to account for variations in input difficulty, which may require different numbers of experts. To address the rigidity of Top-K routing, dynamic routing mechanisms have emerged (Zeng et al., 2024). For instance, Huang et al. (2024) proposed a dynamic gating mechanism (similar to the Top-P strategy) that dynamically increases the number of experts based on the confidence of current expert selections, stopping the selection when the cumulative probability exceeds a threshold. Although dynamic routing methods have made progress in improving efficiency and adaptability, their decision-making process is usually black-boxed, lacking interpretability. In addition, the unbalanced routing mechanism can lead to redundant information received by experts, which in turn causes the problem of insufficient expert specialization (Dai et al., 2024).

Applications of Causal Reasoning

Causal reasoning, in contrast to mere correlation, aims to uncover the underlying causal relationships between variables. Causal reasoning frameworks are also increasingly being applied to transformer models themselves to improve their internal mechanisms and outputs. For example, the team from the Massachusetts Institute of Technology proposed a synthesis framework based on interleaved Markov chains, which forces Transformers to learn context-sensitive causal relationships by dynamically adjusting causal structures (D'Angelo et al., 2025). In addition, Rohekar et al. (2023) demonstrated that self-attention patterns in pre-trained Transformers can be leveraged to infer causal salience among tokens, yielding graph structures that reflect which elements are most explanatory for model predictions. Importantly, these inferred structures should be interpreted as proxies for causally informative dependencies, rather than faithful reconstructions of the true data-generating causal process. Their method supports both single-sample causal structure learning and zero-shot causal discovery, showing strong interpretability and effectiveness in tasks such as sentiment classification and recommendation. These findings highlight the great potential of causal reasoning for enhancing the interpretability and controllability of large language models.

Methodology

Mixture of Experts

The Mixture-of-Experts (MoE) layer consists of a set of expert groups and a routing mechanism. The output $x \in \mathbb{R}^d$ of the multi-head attention mechanism is fed into the gating network to generate the output y. The output of the MoE layer can be expressed as:

$$y = \sum_{i=1}^{N} g_i(x) e_i(x) \tag{1}$$

Where, N denotes the number of experts, $e_i(x)$ represents the processing result of the i-th expert on the token x, and $g_i(x)$ is the gating threshold of the i-th expert for the token x. For the values of the gating network $g_i(x)$, a Top-K strategy is adopted, which can be expressed as:

$$g(x) = softmax(\mathrm{TopK}(x \cdot W, k)) \tag{2}$$

$$\mathrm{TopK}(p, k)_i = \begin{cases} p_i & \text{if } p_i \text{ is in Top-K elements ,} \\ 0 & \text{otherwise ,} \end{cases} \tag{3}$$

Where, $g(x)$ is the value of the gating network for token x, W is the parameter of the gating network in $\mathbb{R}^{d \times N}$, and

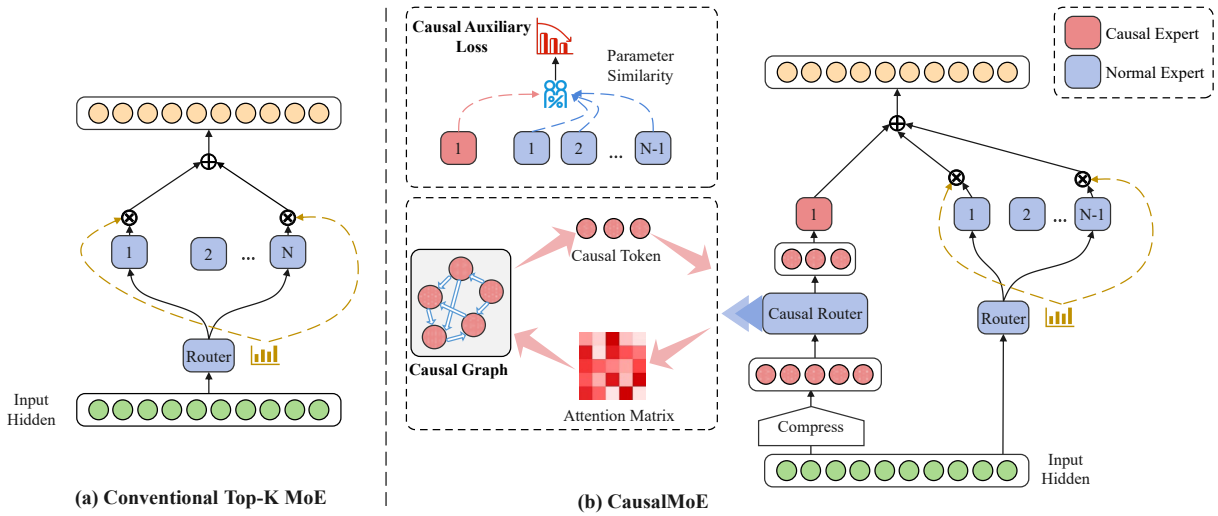

Figure 1: Illustration of the CausalMoE Architecture. Subfigure (a) shows the expert weight layer structure using a traditional Top-K routing strategy, while subfigure (b) presents the structure of CausalMoE, which incorporates Causal Routing and Causal Experts.

TopK indicates selecting only the K experts with the highest similarity.

While sparse routing allows MoE models to scale capacity efficiently, similarity-based routing offers limited interpretability and weak guarantees of functional specialization. These limitations motivate the structured and interpretable routing mechanisms introduced in the following sections.

CausalMoE

Building upon the general MoE architecture, we propose the CausalMoE architecture. As illustrated in Figure 1(b), it comprises two key components: Causal Routing and Causal Experts. These components are designed to enhance the interpretability of the routing mechanism and improve the level of expert specialization.

Construction of the Causal Graph Attention-based Causal-Discovery (ABCD) (Rohekar et al., 2023): If each token is treated as a node in a causal graph, under the assumptions of causal Markov and faithfulness, theoretical derivation shows that the output A of the attention scores from the last layer of the transformer's self-attention can be used to compute the correlation coefficient matrix of the corresponding nodes in the causal graph, which can be expressed as follows:

$$\mathbf{C}_z = \mathbf{A}\mathbf{A}^\top \tag{4}$$

$$\rho_{i,j} = \mathbf{C}_z(i,j) \big/ \sqrt{\mathbf{C}_z(i,i)\mathbf{C}_z(j,j)} \tag{5}$$

Where, $\rho_{i,j}$ represents the correlation coefficient matrix between each pair of nodes in the causal graph. Using this correlation coefficient matrix, the iterative causal discovery(ICD) algorithm (Rohekar et al., 2021) can be employed to learn the causal structure of the input sequence, thereby recovering the equivalence class of the underlying causal graph. The ABCD method leverages pre-trained Transformers for zero-shot causal discovery. We emphasize that the induced causal graph should be interpreted as a proxy for causal salience rather than a faithful recovery of the underlying causal data-generating process.

Sequence Compression To capture the coarse-grained causal logic information between texts, inspired by the work of Yuan et al. (2025), here we compress the query and key into blocks via an MLP before calculating the attention scores, and use the block information to construct the causal graph. Specifically, it can be expressed as follows:

$$\tilde{K}_t^{\text{cmp}} = f_K^{\text{cmp}}(\mathbf{k}_{:t}) = \left\{ \varphi(\mathbf{k}_{il+1:il+l}) \,\middle|\, 0 \leqslant i \leqslant \left\lfloor \frac{t}{l} \right\rfloor \right\} \tag{6}$$

Where, l denotes the block length, t represents the length of the token sequence, and ϕ is a learnable MLP that maps the keys within a block to a single compressed key through intra-block positional encoding. $\tilde{K}_t^{\text{cmp}} \in \mathbb{R}^{d \times \lfloor \frac{t}{l} \rfloor}$ is a tensor composed of these compressed keys, and $\tilde{Q}_t^{\text{cmp}} \in \mathbb{R}^{d \times \lfloor \frac{t}{l} \rfloor}$ can be processed in a similar way. Then, the compressed attention can be expressed as:

$$\tilde{A}_t^{\text{cmp}} = \text{softmax}\left(\frac{\tilde{Q}_t^{\text{cmp}} \cdot (\tilde{K}_t^{\text{cmp}})^T}{\sqrt{d_k}} \right) \tag{7}$$

Where, d_k is the dimension of $\tilde{Q}_t^{\text{cmp}}$ and $\tilde{K}_t^{\text{cmp}}$, and $\tilde{A}_t^{\text{cmp}}$ represents the block-level attention scores obtained after compression.

Causal Routing and Causal Experts CLEANN is a method proposed by Rohekar et al. (2023) that can find the minimum explanation set through a causal graph. CausalMoE constructs a causal graph from the compressed token sequence, then employs the CLEANN algorithm to find the

minimal explanation set for the most critical token sequence blocks, and feeds all tokens in the explanation set to the Causal Expert for processing. This enables MoE to clearly understand the important causal logic in the text through the Causal Expert. It can be specifically expressed as follows:

$$E_{express} = \text{CLEANN}(ABCD(\tilde{A}_t^{\text{cmp}})), \tag{8}$$

$$y_{causal} = \begin{cases} e_{causal}(x) & \text{if } x \text{ is in } E_{express}, \\ 0 & \text{otherwise}, \end{cases} \tag{9}$$

Where, $e_{\text{causal}}(x)$ represents the processing result of the Causal Expert on x, and E_{express} denotes the explanation set.

The final output of the CausalMoE layer is obtained by dynamically blending the outputs from the standard Top-K experts (y) and the Causal Expert (y_{causal}) using a learnable gating mechanism based on the sigmoid function σ. The blending operation can be expressed as:

$$y_{\text{moe}} = \sigma(y + y_{\text{causal}}) \odot y + \big(1 - \sigma(y + y_{\text{causal}})\big) \odot y_{\text{causal}} \tag{10}$$

where y is the standard MoE output defined in Equation (1), y_{causal} is the output of the Causal Expert as defined in Equation (9), and $\odot$ denotes element-wise multiplication.

Causal Auxiliary Loss (CAL)

To enable the Causal Expert to focus on processing causal logic information, inspired by the research of Y. Guo et al. (2025), we propose a novel auxiliary loss function: Causal Auxiliary Loss. The formula is as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{X}_c \cdot \mathbf{X}_i}{|\mathbf{X}_c||\mathbf{X}i|} \tag{11}$$

Where, $\mathbf{X}_c$ represents the parameter vector of the Causal Expert, and $\mathbf{X}_i$ represents the parameter vector of the i-th general expert. This penalty forces the Causal Expert to learn distinct representations, thereby enhancing its specialization in causal logic rather than simply replicating general knowledge.

Experiments

Our experiments are designed to validate the cognitive plausibility of CausalMoE at two levels: (1) Architectural Efficiency Level: By measuring perplexity on both conventional language modeling tasks and tasks rich in causal structures (code generation), we examine the model's capacity to process inputs with varying logical loads. (2) Core Cognitive Ability Level: We directly evaluate the model's performance on pure causal reasoning tasks.

Data and Metrics

We employ a multi-tiered evaluation strategy to assess both general language modeling and, crucially, the specific causal reasoning capabilities of our architecture.

Datasets. Our evaluation uses three categories of data:

- **Language Modeling**: To evaluate the model's general sequence modeling ability under different structural characteristics, we use:
  - **Text Datasets**: **WikiText-103/2** (Merity et al., 2016) and **Penn Treebank (PTB)** (Marcinkiewicz, 1994), representing natural language with relatively weak explicit causal structures.
  - **Code Datasets**: **HumanEval** (Chen et al., 2021), **MBPP** (Austin et al., 2021), **APPS** (Hendrycks et al., 2021), and **Codeforces** (Penedo et al., 2025), which inherently embody strong logical and causal dependencies due to programmatic structure.
- **Causal Reasoning**: To directly evaluate the model's capacity for explicit causal inference, we use:
  - **Corr2cause** (Jin et al., 2023): Tests formal causal graph reasoning (via d-separation) isolated from world knowledge.
  - **CausalBench** (Wang, 2024): Evaluates causal inference (including counterfactual reasoning) across text, code, and math domains from multiple perspectives.

Table 1: Language Modeling Efficiency on Text and Code Datasets (Perplexity, Lower is Better).

Model	Text (Avg) PPL ↓	Code (Avg) PPL ↓
SMoE-medium	43.786	21.964
CausalSMoE-medium	**42.681**	**20.640**
MomentumSMoE-medium	42.741	22.449
CausalMomentumSMoE-medium	**42.262**	**18.369**
AdamSMoE-medium	42.268	19.161
CausalAdamSMoE-medium	**41.403**	**17.049**
GLaM-medium	52.741	28.674
CausalGLaM-medium	**51.124**	**23.705**
MomentumGLaM-medium	46.792	25.486
CausalMomentumGLaM-medium	**45.720**	**19.590**
AdamGLaM-medium	43.940	22.113
CausalAdamGLaM-medium	**43.908**	**19.673**
MomentumSMoE-large	42.351	15.539
CausalMomentumSMoE-large	**39.738**	**12.808**

Metrics. We report **Perplexity (PPL)** for language modeling. For direct causal evaluation, we use **Accuracy**. This distinguishes general sequence modeling improvements from gains in causal inference.

Settings

In this study, the total number of experts N for all models is set to 16, with the number of activated experts $K = 2$. For CausalMoE, 15 of the experts are general experts, and 1 is a Causal Expert, whose activation is determined based on the causal graph. Experiments on each dataset are repeated 10 times to calculate the mean. We use the Wikitext-103 dataset for pre-training. A batch size of 48 is adopted and all models undergo 80,000 gradient updates.

Table 2: Causal Reasoning Performance of Different Models (Accuracy, Higher Is Better) with CAL Ablation.

Model	Corr2Cause (Acc. ↑)	CausalBench (Acc. ↑)		
		Code	Math	Text
SMoE-medium	85.018	48.802	84.343	71.401
CausalSMoE-medium(w/o CAL)	91.726	73.354	86.450	72.441
CausalSMoE-medium	**92.294**	**76.269**	**87.408**	**73.295**
MomentumSMoE-medium	83.729	66.306	75.592	70.736
CausalMomentumSMoE-medium(w/o CAL)	92.634	73.532	84.214	71.346
CausalMomentumSMoE-medium	**93.476**	**76.735**	**85.382**	**72.870**
AdamSMoE-medium	80.954	71.586	85.388	72.103
CausalAdamSMoE-medium(w/o CAL)	82.830	74.417	86.219	73.090
CausalAdamSMoE-medium	**88.844**	**78.268**	**87.236**	**73.433**
GLaM-medium	86.622	55.280	84.606	72.634
CausalGLaM-medium(w/o CAL)	89.925	56.981	85.200	73.132
CausalGLaM-medium	**93.305**	**73.161**	**85.926**	**73.554**
MomentumGLaM-medium	86.835	50.528	84.517	72.567
CausalMomentumGLaM-medium(w/o CAL)	88.252	54.699	86.781	73.700
CausalMomentumGLaM-medium	**91.457**	**74.594**	**87.101**	**74.001**
AdamGLaM-medium	83.436	53.061	85.387	72.764
CausalAdamGLaM-medium(w/o CAL)	86.370	54.453	86.073	74.095
CausalAdamGLaM-medium	**89.299**	**56.641**	**87.300**	**74.133**
MomentumSMoE-large	82.678	66.473	84.179	71.600
CausalMomentumSMoE-large(w/o CAL)	91.726	72.203	84.216	71.450
CausalMomentumSMoE-large	**96.108**	**76.269**	**87.870**	**74.218**

Baseline Models

The comparison baselines include SMoE (Fedus et al., 2022), GLaM (Du et al., 2022), and their Momentum/Adam variants (Teo & Nguyen, 2024). During the experiments, two model configurations were considered: medium (6 layers, approximately 220M parameters) and large (12 layers, approximately 390M parameters).

Efficiency and Structure Sensitivity in Language Modeling

CausalMoE consistently outperforms baselines on both natural language and code datasets (Table 1). Its performance gains, however, show strong structural sensitivity: modest on natural language benchmarks (e.g., WikiText, PTB) but substantially larger on code datasets with dense logical and causal dependencies (e.g., HumanEval, MBPP), where average perplexity decreases by 18.42%.

This pattern indicates that CausalMoE's advantage is not a uniform gain in parameter efficiency, but depends on the structural properties of the input. Programming code contains explicit causal dependencies, control flow, and logical constraints, which place higher demands on causal integration and analytical processing. Consistent with task-dependent resource allocation in cognitive theory, CausalMoE preferentially routes such high–causal-load inputs to experts specialized in causal processing.

These results therefore support CausalMoE's core cognitive hypothesis: performance gains arise from selectively allocating resources based on input causal structure. Subsequent expert-usage analyses further support this interpretation.

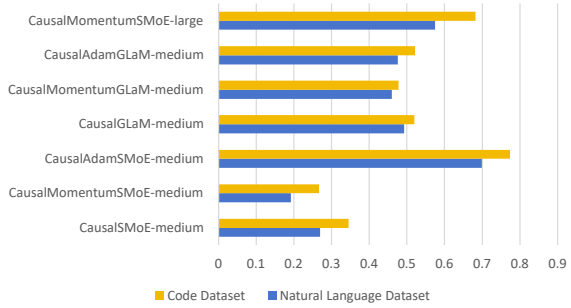

Figure 2: Activation Frequency of the Causal Expert across Natural Language and Code Datasets.

Expert Utilization Analysis

To assess whether CausalMoE allocates computation in a structure-sensitive manner, we analyze the utilization patterns of the Causal Expert across input domains, measuring the proportion of sequences routed to it for natural language text and programming code. As shown in Figure 2, the Causal Expert is activated substantially more frequently on code datasets than on natural language text datasets, a pattern that is consis-

Table 3: Average Code Generation Performance of Baseline, CausalMoE(w/o CAL), and CausalMoE Models on Different Datasets (Perplexity, Lower Is Better).

Dataset	Baseline PPL ↓	CausalMoE (w/o CAL) PPL ↓	CausalMoE PPL ↓
Humaneval	62.417	57.979	**54.733**
MBPP	16.953	15.534	**13.392**
APPS	5.703	5.390	**4.105**
Codeforce	3.718	3.597	**3.103**

tent across all evaluated benchmarks. Notably, this behavior emerges without explicit supervision or task-specific routing rules, suggesting that the routing mechanism has learned to associate latent structural properties of the input with expert selection.

From a cognitive perspective, this utilization pattern provides direct evidence of task-dependent recruitment of specialized processing mechanisms. Unlike natural language, programming code contains explicit causal and logical structures that require more analytical reasoning. CausalMoE reflects this distinction by preferentially allocating causally structured inputs to the Causal Expert, while routing less causally constrained inputs to general-purpose experts. Crucially, this increased activation is not imposed by load-balancing constraints or heuristic rules, but arises from the learned causal routing process itself. Together with the perplexity improvements reported earlier, these findings indicate that CausalMoE's gains stem from selective and interpretable resource allocation aligned with input causal structure, closely mirroring principles of human cognitive organization.

Evaluation on Causal Reasoning Benchmarks

While the preceding experiments evaluate language modeling efficiency and structure sensitivity, they do not directly assess whether CausalMoE improves causal reasoning itself. To address this limitation, we further fine-tune and evaluate CausalMoE on two dedicated causal reasoning benchmarks: Corr2Cause and CausalBench. Importantly, these benchmarks are not designed for next-token prediction or surface-level language modeling; instead, they directly evaluate a model's ability to infer, manipulate, and reason over causal relations.

Corr2Cause focuses on formal causal graph reasoning, requiring models to distinguish correlation from causation and to perform structured inference over explicit causal dependencies. CausalBench extends this evaluation across multiple domains, including code, mathematics, and natural language, testing whether causal reasoning capabilities generalize beyond a single modality or task format.

CausalMoE consistently outperforms strong MoE baselines on both benchmarks (Table 2), with gains persisting across domains and task formats. This indicates improvements are not artifacts of language modeling or domain-specific heuristics.

Unlike perplexity-based evaluations, which can be influenced by statistical regularities, accuracy on Corr2Cause and CausalBench reflects a model's capacity to explicitly reason about causal relations. The observed improvements therefore support the claim that CausalMoE does not merely improve efficiency in language processing, but implements a functionally meaningful mechanism for causal reasoning—consistent with theories that posit specialized cognitive systems for analytical and causal inference.

Ablation Experiments

Due to inherent architectural coupling in the proposed model, our ablation experiments adopt a progressive stacking strategy and only report results for the baseline, baseline+Causal Routing and Causal Experts, and baseline+Causal Routing and Causal Experts+Causal Auxiliary Loss (CAL). Independent ablation of CAL is not feasible with the current architectural design.

As shown in Tables 2 and 3, the removal of the Causal Auxiliary Loss (CAL) results in a consistent performance decline on both causal reasoning benchmarks and code generation tasks. Although the model without CAL still outperforms non-causal MoE baselines, its advantages are significantly diminished. This indicates that architectural causal specialization alone is insufficient to fully exploit the merits of CausalMoE.

We attribute this degradation to weakened functional decoupling: without CAL, the Causal Expert increasingly overlaps with general experts, undermining its specialization. CAL explicitly enforces functional independence by discouraging parameter similarity, paralleling cognitive theories that emphasize interference control in modular systems. These results indicate that CAL is essential for translating causal architecture into effective causal reasoning behavior.

Conclusion

We proposed CausalMoE, a causally grounded Mixture-of-Experts framework that connects sparse neural architectures with core ideas from cognitive science. By integrating Causal Routing, Causal Experts, and a Causal Auxiliary Loss, CausalMoE enables functionally specialized and structure-sensitive computation. Empirically, the model exhibits selective efficiency: performance gains are modest on unstructured natural language but substantially larger on code and logic-intensive inputs with dense causal dependencies. Beyond language modeling, CausalMoE achieves consistent improvements on Corr2Cause and CausalBench, which directly evaluate causal reasoning rather than next-token prediction, providing evidence for improved performance on benchmarks designed to probe causal reasoning. Ablation results further show that causal specialization requires explicit mechanisms for functional decoupling. Overall, this work suggests a path toward cognitively motivated expert architectures where computation is dynamically allocated according to causal structure.

References

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. (2021). Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201–215.

Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al. (2024). Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

D'Angelo, F., Croce, F., & Flammarion, N. (2025). Selective induction heads: How transformers select causal structures in context. *The Thirteenth International Conference on Learning Representations*.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. (2022). Glam: Efficient scaling of language models with mixture-of-experts. *International conference on machine learning*, 5547–5569.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*(1), 255–278.

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

Fletcher, C. R., & Bloom, C. P. (1988). Causal reasoning in the comprehension of simple narrative texts. *Journal of Memory and language*, *27*(3), 235–244.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Guo, H., Lu, H., Nan, G., Chu, B., Zhuang, J., Yang, Y., Che, W., Leng, S., Cui, Q., & Jiang, X. (2025). Advancing expert specialization for better moe. *arXiv preprint arXiv:2505.22323*.

Guo, Y., Cheng, Z., Tang, X., Tu, Z., & Lin, T. (2025). Dynamic mixture of experts: An auto-tuning approach for efficient transformer models. https://arxiv.org/abs/2405.14297

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., & Steinhardt, J. (2021). Measuring coding challenge competence with apps. *NeurIPS*.

Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., Xu, K., Chen, L., Huang, S., & Feng, Y. (2024). Harder tasks need more experts: Dynamic routing in moe models. *arXiv preprint arXiv:2403.07652*.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, *3*(1), 79–87.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., & Schölkopf, B. (2023). Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Marcinkiewicz, M. A. (1994). Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, *273*, 31.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Penedo, G., Lozhkov, A., Kydlíček, H., Allal, L. B., Beeching, E., Lajarín, A. P., Gallouédec, Q., Habib, N., Tunstall, L., & von Werra, L. (2025). Codeforces.

Rohekar, R. Y., Gurwicz, Y., & Nisimov, S. (2023). Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, *36*, 31450–31465.

Rohekar, R. Y., Nisimov, S., Gurwicz, Y., & Novik, G. (2021). Iterative causal discovery in the possible presence of latent confounders and selection bias. *Advances in Neural Information Processing Systems*, *34*, 2454–2465.

Teo, R. S., & Nguyen, T. M. (2024). Momentumsmoe: Integrating momentum into sparse mixture of experts. *Advances in Neural Information Processing Systems*, *37*, 28965–29000.

Wang, Z. (2024). Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 143–151.

Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., & You, Y. (2024). Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., Xie, Z., Wei, Y., Wang, L., Xiao, Z., et al. (2025). Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*.

Zeng, Z., Miao, Y., Gao, H., Zhang, H., & Deng, Z. (2024). Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models. *arXiv preprint arXiv:2406.13233*.

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., & Fedus, W. (2022). St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.