DMORE: Differentiable Mixture-of-Reasoning-Experts with Uncertainty-Guided Multi-Level Routing

Roman Sultimov[1,2], **Aleksandr Volkov**[3], **Mariia Kovalchuk**[1,2], **Yury Maximov**[3]
[1]Lomonosov Moscow State University, [2]Moscow Center for Advanced Studies, [3]Interdata Astana
r.sultimov@iai.msu.ru, volkov@iccda.io, m.kovalchuk@iai.msu.ru, yury@iccda.io

Abstract

Large language models (LLMs) often face inputs of widely varying reasoning difficulty, yet most systems allocate a fixed amount of computation per example. We introduce DMORE (Differentiable Mixture-of-Reasoning-Experts), a unified architecture that dynamically routes inputs to specialized reasoning experts via uncertainty-guided, multi-level gating. Unlike prior mixture-of-experts methods that switch among separate models, DMORE activates one, two, or all experts within a single model based on Monte Carlo dropout–derived uncertainty estimates. A three-stage training procedure first specializes experts on domain-specific reasoning tasks, then calibrates the uncertainty estimator, and finally optimizes the full system end-to-end. Experiments on four reasoning benchmarks show that DMORE matches or surpasses strong chain-of-thought baselines while reducing computation by 28%.

1 Introduction

The remarkable success of large language models (LLMs) in complex reasoning tasks has been largely attributed to techniques like chain-of-thought (CoT) prompting [28] and their variants [27, 15]. However, these approaches typically apply uniform computational resources across all reasoning tasks, regardless of their inherent complexity or the specific type of reasoning required. This one-size-fits-all strategy leads to inefficient resource allocation, where simple problems receive excessive computation while complex problems may benefit from additional specialized processing.

Recent advances in mixture-of-experts (MoE) architectures [23, 6] have demonstrated the potential for dynamic computational allocation. However, existing MoE approaches primarily focus on general language modeling tasks and route between separate pre-trained models [2, 1], rather than leveraging specialized reasoning capabilities within a unified architecture. Moreover, current routing mechanisms typically employ fixed expert selection strategies that do not adapt to the uncertainty or complexity of individual instances.

We propose DMORE (Differentiable Mixture-of-Reasoning-Experts), a novel architecture that addresses these limitations through uncertainty-guided multi-level routing within a unified reasoning framework. Our key insight is that different reasoning tasks—mathematical, logical, commonsense, and causal—benefit from specialized architectural components, and the computational allocation should adapt dynamically based on the model's confidence in its reasoning process.

Our DMORE approach introduces three main innovations: (a) *reasoning-specialized experts* with domain-specific architectural priors for mathematical, logical, commonsense, and causal reasoning; (b) *Uncertainty-guided multi-level routing* that adaptively selects one, two, or all experts based on Monte Carlo dropout uncertainty estimation; (c) *unified architecture design* that enables end-to-end optimization while maintaining reasoning specialization.

Our experimental evaluation on four reasoning benchmarks (GSM8K, MATH, LogiQA, CommonsenseQA) demonstrates that DMORE achieves competitive accuracy while significantly improving computational efficiency. Using LLaMA-2-7B as the base model, DMORE attains 84.1% accuracy on GSM8K and 42.3% on MATH, representing improvements of 1.5 and 1.7 percentage points over chain-of-thought baselines (82.6% and 40.6% respectively), while reducing computational costs by 28% through adaptive expert selection.

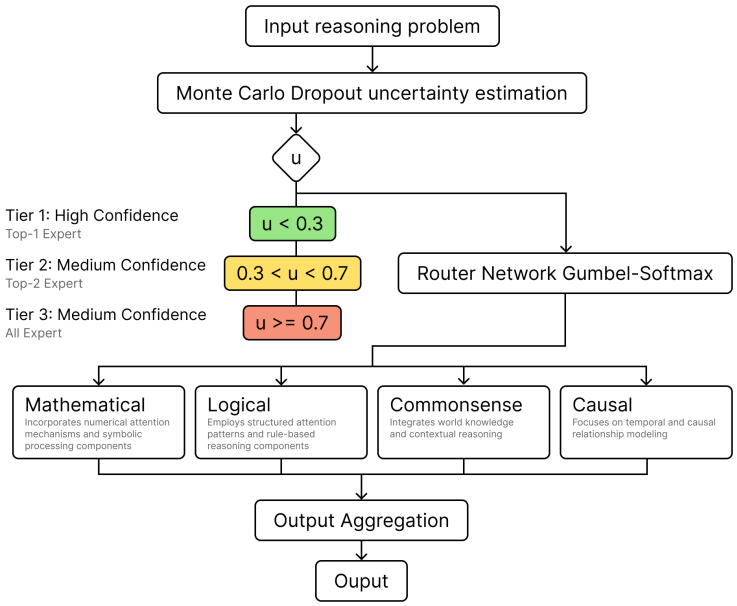

Figure 1: Differentiable Mixture-of-Reasoning-Experts (DMORE) with uncertainty-guided multi-Level routing architecture

2 Related Work

High computational demand of modern LLMs have motivated extensive research into efficient inference techniques including model compression through pruning [10], quantization [13], and knowledge distillation [12]. Dynamic inference methods such as early exit mechanisms [25] and adaptive computation time [9] adjust computational allocation based on input complexity. Recent work has explored test-time scaling approaches [18] that increase computation during inference for improved performance on complex reasoning tasks.

Mixture of Experts (MoE) Architectures. MoE [23] enable conditional computation by routing inputs to specialized expert networks. The Switch Transformer [6] demonstrated the effectiveness of sparse expert activation in large-scale language modeling. Recent extensions include GLaM [5] for efficient scaling and specialized routing mechanisms for different modalities and tasks.

Contemporary work has explored MoE approaches specifically for reasoning tasks. Symbolic-MoE [2] routes between heterogeneous pre-trained models based on symbolic skill requirements; CARGO [1], a confidence-aware framework for routing between different LLMs based on embedding-based confidence estimation. Finally, Route-to-Reason [21] jointly selects both language models and reasoning strategies. Unfortunately, these approaches primarily route between separate pre-trained models rather than learning specialized components within a unified architecture.

Reasoning in LLMs. Chain-of-thought prompting [28] has emerged as a fundamental technique for eliciting reasoning capabilities in LLMs. Extensions include zero-shot CoT [15], self-consistency decoding [27], and tree-of-thoughts [29]. Recent work has explored more structured reasoning approaches, including tool-augmented reasoning [22] and program-aided language models [8].

The diversity of reasoning types (mathematical, logical, commonsense, and causal) suggests that specialized approaches may be beneficial. Mathematical reasoning often requires precise symbolic manipulation [3], while commonsense reasoning relies on world knowledge and contextual understanding [24]. This motivates our approach of developing reasoning-specialized experts within a unified architecture.

Uncertainty Estimation in Neural Networks. Uncertainty estimation in deep learning has been extensively studied, with approaches including Bayesian neural networks [20], deep ensembles [16], and Monte Carlo Dropout [7]. For language models, uncertainty estimation has been applied to calibration [4] and selective prediction [14]. Recent work has explored uncertainty-aware routing in neural networks, though primarily for computer vision tasks.

Our work builds upon Monte Carlo Dropout for uncertainty estimation, extending it to reasoning task routing within a mixture-of-experts framework. Unlike previous approaches that use uncertainty for calibration or selective prediction, we leverage uncertainty to guide dynamic expert allocation.

3 Methodology and Evaluation

Problem Setup. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset of reasoning tasks, where x_i represents the input problem and y_i the target solution. We assume that reasoning tasks can be categorized into K types: $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$, where each type t_k corresponds to a specific reasoning domain (e.g., mathematical, logical, commonsense, causal).

Our goal is to learn a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ that can efficiently solve reasoning tasks by dynamically allocating computational resources based on task complexity and type. We formulate this as an optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(f_\theta(x), y) + \lambda \cdot \mathcal{C}(x, \theta)\right] \tag{1}$$

where $\mathcal{L}$ is the loss functio, $\mathcal{C}(x, \theta)$ represents the computational cost for input x, and $\lambda > 0$ controls the efficiency-accuracy trade-off. We modeled computational cost as an expected time or number of tokens proxy learned over training data for each of the experts. We assume linearity of the cost, if multiple experts are involved.

3.1 DMORE Architecture

Reasoning-Specialized Experts. DMORE consists of four specialized experts, each designed for a specific reasoning type.

Mathematical Expert, E_{math}, incorporates numerical attention mechanisms and symbolic processing components with feed forward network (FFN):

$$E_{\text{math}}(x) = \text{FFN}_{\text{math}}(\text{NumAttn}(x) + \text{SymProc}(x)) \tag{2}$$

where $\text{NumAttn}(x) = \text{Attention}(x, W_{\text{num}})$ applies attention with numerical-focused weight matrices, and $\text{SymProc}(x)$ processes symbolic mathematical expressions through specialized embedding layers.

Logical Expert , E_{logic}, employs structured attention patterns and rule-based reasoning components, implementing hierarchical attention patterns optimized for logical structure parsing, and incorporating learned representations of logical rules.

Commonsense Expert , E_{common}, integrates world knowledge and contextual reasoning attending to knowledge-relevant tokens and capturing contextual relationships.

Causal Expert , E_{causal}, focuses on temporal and causal relationship modeling temporal dependencies and causal relationships.

Uncertainty-Guided Multi-Level Routing. The core innovation of DMORE is its uncertainty-guided routing mechanism that adaptively determines the number of experts to activate. We employ

Monte Carlo Dropout [7] to estimate epistemic uncertainty $u(x) = \frac{1}{T}\sum_{t=1}^{T} \text{Var}[p_t(y|x)]$, where $p_t(y|x)$ represents the predictive distribution from the t-th Monte Carlo sample with dropout enabled.

Based on the uncertainty estimate, we define three routing tiers:

Tier	Condition	Uncertainty Estimate
Tier 1 (High Confidence)	$u(x) < \tau_1$	$f_{\text{tier}_1}(x) = \max_k \text{Router}(x)_k \cdot E_k(x)$
Tier 2 (Medium Confidence)	$\tau_1 \leq u(x) < \tau_2$	$f_{\text{tier}_2}(x) = \sum_{k \in \text{top-2}} \text{Router}(x)_k \cdot E_k(x)$
Tier 3 (Low Confidence)	$u(x) \geq \tau_2$	$f_{\text{tier3}}(x) = \sum_{k=1}^{4} \text{Router}(x)_k \cdot E_k(x)$

The router network $\text{Router}(x)$ is implemented as:

$$\text{Router}(x) = \text{Softmax}(\text{MLP}([\text{CLS}(x); u(x)])) \tag{3}$$

where $\text{CLS}(x)$ is the classification token representation and $u(x)$ is the uncertainty estimate, and MLP stands for multi-layer perceptron.

Adaptive Gumbel-Softmax for Differentiable Routing. To enable end-to-end training, we employ an adaptive Gumbel-Softmax mechanism for differentiable expert selection:

$$\text{GumbelSoftmax}(\text{Router}(x), \tau_{\text{temp}}) = \frac{\exp((\log(\text{Router}(x)_k) + g_k)/\tau_{\text{temp}})}{\sum_{j=1}^{n_{exp}} \exp((\log(\text{Router}(x)_j) + g_j)/\tau_{\text{temp}})} \tag{4}$$

where $g_k \sim \text{Gumbel}(0, 1)$ and τ_{temp} is the temperature parameter that adapts based on training.

3.2 Training

DMORE employs a three-stage training procedure designed to optimize both specialization and routing efficiency:

Stage 1 - Expert Specialization: each expert is independently fine-tuned on domain-specific datasets:

$$\mathcal{L}_{\text{spec}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[\text{CrossEntropy}(E_k(x), y)]$$

where $\mathcal{D}_k$ contains examples of reasoning type k.

Stage 2 - Uncertainty Calibration: the router and uncertainty estimation components are trained to predict task difficulty:

$$\mathcal{L}_{\text{calib}} = \mathbb{E}_{(x,y)}[\text{MSE}(u(x), \text{difficulty}(x, y))]$$

where $\text{difficulty}(x, y)$ is computed based on expert disagreement and solution complexity.

Stage 3 - End-to-End Optimization: the entire system is jointly optimized with a multi-objective loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{efficiency}} + \lambda_2 \mathcal{L}_{\text{balance}}, \qquad \text{with } \mathcal{L}_{\text{task}} = \mathbb{E}_{(x,y)}[\text{CrossEntropy}(f_\theta(x), y)]$$

and $\mathcal{L}_{\text{efficiency}} = \mathbb{E}_x[\text{NumExperts}(x)]$, $\mathcal{L}_{\text{balance}} = \text{Var}[\mathbb{E}_x[\text{Router}(x)_k]]_{k=1}^4$.

3.3 Experimental Setup

Table 1 contains the datasets which we used for numerical evaluation of the methods.

Dataset	Description
GSM8K [3]	Grade school math word problems requiring multi-step arithmetic reasoning. Contains 7,473 training and 1,319 test problems
MATH [11]	Competition-level mathematics problems spanning algebra, geometry, number theory, and other areas. Includes 7,500 training and 5,000 test problems
LogiQA [19]	Logical reasoning questions requiring deductive and inductive reasoning. Contains 8,678 training and 651 test examples
CommonsenseQA [24]	Multiple-choice questions requiring commonsense reasoning about everyday situations. Includes 9,741 training and 1,221 test examples.

Table 1: Reasoning benchmarks covering different reasoning types and used for DMORE evaluation

Implementation Details. We use LLaMA-2-7B [26] as our foundation model, chosen for its strong reasoning capabilities and computational efficiency. In utilized expert architecture: each expert

consists of 4 transformer layers with specialized attention mechanisms and feed-forward networks. The hidden dimension is 4096, and attention heads are 32. We train for 3 epochs in each stage with a learning rate of 2e-5, batch size of 16, and gradient accumulation steps of 4. The uncertainty thresholds are set to $\tau_1 = 0.3$ and $\tau_2 = 0.7$ based on validation performance. We report accuracy on all benchmarks and measure computational efficiency through FLOPs reduction and average number of activated experts.

We compare DMORE against several strong baselines including: (a) *LLaMA-2-7B + CoT* [17]: Chain-of-thought prompting with the base model, achieving 82.6% on GSM8K and 40.6% on MATH; (b) *LLaMA-2-7B (Zero-shot)*: Direct answer generation without reasoning steps; (c) *Static MoE*: A traditional mixture-of-experts with fixed routing, using the same expert architectures as DMORE; (d) *Random Routing*: random selection between experts to isolate the effect of uncertainty-guided routing; (e) *Single Expert*: each specialized expert used independently to assess specialization benefits.

3.4 Results and Analysis

Main Results. Table 2 presents the main experimental results across all benchmarks. DMORE achieves competitive accuracy while significantly improving computational efficiency.

DMORE consistently outperforms the CoT baseline across all benchmarks while achieving substantial efficiency gains. The improvements are particularly notable on MATH and LogiQA, suggesting that specialized experts are especially beneficial for complex mathematical and logical reasoning tasks.

Method	GSM8K	MATH	LogiQA	CommonsenseQA	Efficiency
LLaMA-2-7B (Zero-shot)	18.2	7.8	24.1	45.3	100%
LLaMA-2-7B + CoT	82.6	40.6	67.4	78.9	0%
Static MoE	81.3	39.2	66.8	77.5	15%
Random Routing	79.7	37.9	65.2	76.1	22%
DMORE	**84.1**	**42.3**	**69.1**	**80.2**	**28%**
Improvement over CoT	+1.5	+1.7	+1.7	+1.3	+28%

Table 2: Experimental results on reasoning benchmarks. Efficiency is measured as FLOPs reduction compared to CoT baseline

Ablation Studies. Table 3 shows the performance of individual experts on different reasoning types, confirming the effectiveness of specialization. The results clearly demonstrate that experts perform best on their specialized domains, with the mathematical expert excelling on GSM8K and MATH, the logical expert on LogiQA, and the commonsense expert on CommonsenseQA. Table 3 illustrates the distribution of routing decisions across different uncertainty levels and their impact on performance.

Confidence Level	Avg. Experts	Accuracy	Efficiency
High ($u < 0.3$)	1.0	87.3%	66%
Medium ($0.3 \leq u < 0.7$)	2.0	82.1%	33%
Low ($u \geq 0.7$)	4.0	78.9%	0%
Overall	1.8	84.1%	28%

Expert	GSM8K	MATH	LogiQA	CommonsenseQA
Mathematical	**83.2**	**41.8**	58.3	65.7
Logical	71.4	32.1	**68.9**	72.3
Commonsense	69.8	28.7	61.2	**79.8**
Causal	67.2	30.4	63.5	74.1

Table 3: Left: routing strategy analysis showing expert activation patterns. Right: performance of individual experts on different reasoning types.

The uncertainty-guided routing effectively balances accuracy and efficiency, with high-confidence predictions requiring only a single expert while maintaining high accuracy.

Computational Efficiency Analysis. DMORE achieves substantial computational savings through adaptive expert activation. On average, only 1.8 out of 4 experts are activated per instance, resulting in a 28% reduction in FLOPs compared to the CoT baseline. The efficiency gains are particularly pronounced for simpler problems, where high confidence allows single-expert routing.

Error Analysis and Interpretability. We analyze the types of errors made by DMORE and the interpretability of routing decisions. The uncertainty-guided routing shows strong correlation with problem difficulty, with complex multi-step problems typically routed to multiple experts while straightforward problems are handled by single experts.

Common error patterns include: mathematical errors in complex algebraic manipulations (12% of errors); logical fallacies in multi-premise reasoning (18% of errors); commonsense knowledge gaps (15% of errors); routing errors where uncertainty estimation fails (8% of errors).

Table 3.4 presents representative examples of DMORE's routing decisions and reasoning processes.

Problem	Uncertainty	Routing	Reasoning
"What is 15 × 8?"	0.12	Math Expert	Direct calculation
"If all birds can fly, and penguins are birds, can penguins fly?"	0.45	Logic + Common	Logical contradiction with world knowledge
"A train travels 60 mph for 2 hours, then 80 mph for 3 hours. What's the average speed?"	0.73	All Experts	Complex multi-step calculation requiring verification

Table 4: Qualitative examples of DMORE routing decisions.

4 Discussion and Conclusion

Implications for Efficient Reasoning. DMORE demonstrates that uncertainty-guided routing can effectively balance accuracy and computational efficiency in reasoning tasks. The key insight is that not all reasoning problems require the same computational resources, and adaptive allocation based on confidence can yield significant efficiency gains without sacrificing performance.

The success of specialized experts suggests that different reasoning types benefit from distinct architectural components. This finding has implications for future LLM design, suggesting that reasoning-aware architectures may be more effective than general-purpose models for complex reasoning tasks.

Limitations and Future Directions. Our approach has several key limitations, including

Expert Design: The current expert architectures are manually designed based on reasoning type intuitions. Future work could explore automated expert discovery and specialization.

Uncertainty Calibration: While Monte Carlo Dropout provides reasonable uncertainty estimates, more sophisticated uncertainty quantification methods could improve routing decisions.

Scalability: Our experiments focus on a 7B parameter model. Scaling to larger models and more experts presents both opportunities and challenges.

Training Complexity: The three-stage training procedure adds complexity compared to standard fine-tuning. Investigating end-to-end training approaches could simplify the process.

Impact. DMORE efficiency improvements have positive impact on computational requirements for reasoning tasks. The interpretable routing decisions also enhance model transparency, which is valuable for high-stakes applications requiring explainable AI. Although the specialized nature of experts could potentially amplify biases present in domain-specific training data. Careful attention to fairness and bias mitigation will be important in future developments.

Outcome and Future Directions. We introduced DMORE, a novel mixture-of-reasoning-experts architecture that achieves competitive accuracy while significantly improving computational efficiency through uncertainty-guided multi-level routing. Our key contributions include: (1) reasoning-specialized experts with domain-specific architectural priors, (2) uncertainty-guided adaptive expert selection, and (3) a unified architecture enabling end-to-end optimization.

Experimental results demonstrate that DMORE outperforms chain-of-thought baselines across four reasoning benchmarks while reducing computational costs by 28%. The success of uncertainty-guided routing suggests promising directions for future research in efficient reasoning and conditional computation.

Our work opens several avenues for future investigation, including scaling to larger models, automated expert discovery, and applications to other reasoning domains. We believe that DMORE represents a significant step toward more efficient and specialized reasoning architectures for LLMs.

5 Acknowledgments and Disclosure of Funding

This work of Roman Sultimov was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-15-2025-013, dated June 20, 2025, IGK 000000C313925P4B0002).

References

[1] Amine Barrak, Yosr Fourati, Michael Olchawa, Emna Ksontini, and Khalil Zoghlami. Cargo: A framework for confidence-aware routing of large language models. *35th IEEE International Conference on Collaborative Advances in Software and Computing*, 2025.

[2] JCY Chen, S Yun, E Stengel-Eskin, T Chen, et al. Symbolic mixture-of-experts: Adaptive skill-based routing for heterogeneous reasoning. *arXiv preprint arXiv:2503.05641*, 2025.

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

[4] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

[5] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning*, pages 5547–5569, 2021.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 22(120):1–39, 2021.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, pages 1050–1059, 2016.

[8] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *International Conference on Machine Learning*, pages 10764–10799, 2023.

[9] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[10] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

[14] Aditya Kamath, Robin Jia, and Percy Liang. Selective prediction for neural networks. *arXiv preprint arXiv:2010.06924*, 2020.

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[17] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*, 2024.

[18] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Simple test-time scaling for efficient reasoning. *arXiv preprint arXiv:2501.19393*, 2024.

[19] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

[20] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[21] Zhihong Pan, Kai Zhang, Yuze Zhao, and Yupeng Han. Route to reason: Adaptive routing for llm and reasoning strategy selection. *arXiv preprint arXiv:2505.19435*, 2025.

[22] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[23] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[24] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2019.

[25] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. *23rd international conference on pattern recognition (ICPR)*, pages 2464–2469, 2016.

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[27] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Nazneen Sharan, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[29] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.