GRIN: *GR*adient-*IN*formed MoE

Liyuan Liu* Young Jin Kim Shuohang Wang Chen Liang
Yelong Shen Hao Cheng Xiaodong Liu Masahiro Tanaka
Xiaoxia Wu Wenxiang Hu Vishrav Chaudhary
Zeqi Lin Chenruidong Zhang Jilong Xue
Hany Awadalla Jianfeng Gao* Weizhu Chen*

Microsoft

Abstract

Mixture-of-Experts (MoE) models scale more effectively than dense models due to sparse computation through expert routing, selectively activating only a small subset of expert modules. However, sparse computation challenges traditional training practices, as discrete expert routing hinders standard backpropagation and thus gradient-based optimization, which are the cornerstone of deep learning. To better pursue the scaling power of MoE, we introduce GRIN (*GR*adient-*IN*formed MoE training), which incorporates sparse gradient estimation for expert routing and configures model parallelism to avoid token dropping. Applying GRIN to autoregressive language modeling, we develop a top-2 16×3.8B MoE model. Our model, with only 6.6B activated parameters, outperforms a 7B dense model and matches the performance of a 14B dense model trained on the same data. Extensive evaluations across diverse tasks demonstrate the potential of GRIN to significantly enhance MoE efficacy, achieving 79.4 on MMLU, 83.7 on HellaSwag, 74.4 on HumanEval, and 58.9 on MATH.

1 Introduction

The success of large-scale pre-training highlights the importance of model scalability (OpenAI, 2023; Touvron et al., 2023). Mixture-of-Experts (MoE) models have emerged as a promising approach, selectively activating only a small subset of modules on specific inputs through an *expert routing* process, thus improving the model scalability by orders of magnitude (Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022).

However, the sparse activation mechanism of MoE presents several challenges to model training. For example, while the discrete routing function produces non-differentiable outputs, backpropagation, the cornerstone of deep learning, is exclusively compatible with differentiable functions (Rosenblatt, 1957; Bengio et al., 2013). Consequently, backpropagation cannot be directly applied for gradient computation of expert routing.

To fully leverage the scaling potential of MoE, we study gradient estimation for expert routing and configure model parallelism to avoid token dropping in this work. Extending Liu et al. (2023a,b), we propose SparseMixer-v2 to estimate gradient for expert routing, which differs from conventional practices that use the gating gradient as a proxy for the routing gradient. Additionally, we propose a scalable MoE training recipe that uses pipeline

*Correspondence to {lucliu, jfgao, wzchen}@microsoft.com

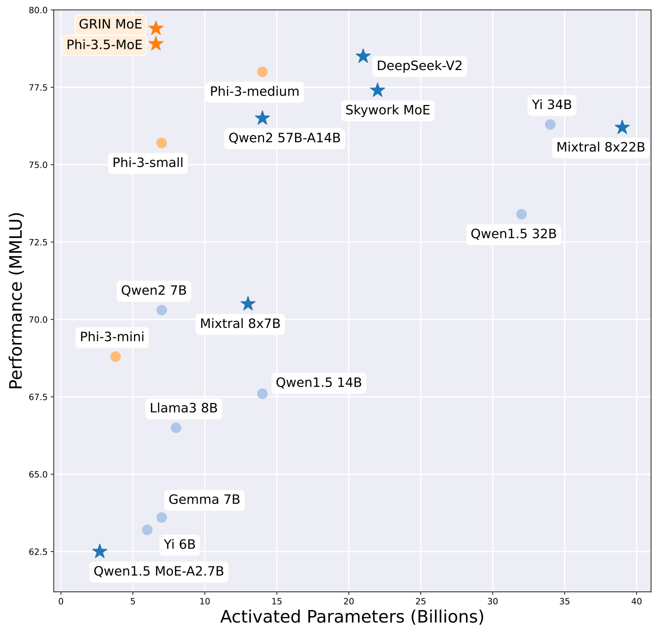

Figure 1: MMLU accuracy and activated parameters.

Parameter	Value
General	
`vocab_size`	32064
`n_position`	4096
`n_layers`	32
`n_embed`	4096
`normalization`	LayerNorm
Attention specific	
`window_size`	2048
`n_head`	32
`n_kv_head`	8
`head_dim`	128
`rotary_dim`	128
Expert specific	
`activation`	SwiGLU
`inner_dim`	6400
`n_expert`	16
`top_k`	2
`moe_mod`	every layer

Table 1: GRIN MoE Specifics.

parallelism and tensor parallelism instead of expert parallelism, thus eliminating the needs of using a capacity factor and dropping tokens during training.

We have applied our method, *GR*adient-*IN*formed MoE, to autoregressive language modeling and developed a 16x3.8B MoE model, referred to as GRIN MoE[2]. The model incorporates MoE layers as its feedforward network, conducting top-2 routing among 16 experts at each layer. Specifically, each expert is implemented as a GLU network and the routing module selectively activates 2 out of 16 expert networks for each token at each layer. As a result, the 16×3.8B model has 42B parameters in total and 6.6B activated parameters for inference.

Our evaluations across a diverse set of tasks reveal that GRIN MoE achieves exceptionally good performance, particularly in coding and mathematics tasks. For example, our model scores 79.4 on MMLU, 74.4 on HumanEval, 80.3 on MBPP, and 58.9 on MATH. GRIN MoE also compares favorably with dense models trained on the same data. For instance, GRIN MoE achieves an average score of 79.58 in Table 2, outperforming 7B dense model (average score of 75.74) and matching the 14B dense model (average score of 78.46) trained on the same data.

In addition, we have conducted an in-depth analysis to shed light on why GRIN MoE works well. We show that GRIN MoE can answer sophisticated questions correctly and quickly pick up on hints, and that GRIN MoE exhibits different expert assignment patterns for different tasks and domains, indicating that its experts have developed specialized expertise and that the router can effectively compose these experts.

2 Model Architecture

Similar to existing state-of-the-art MoE models, GRIN MoE is based on a transformer architecture (Vaswani et al., 2017) with a stack of transformer blocks.

Transformer. The Transformer network is constructed by stacking Transformer blocks, each consisting of an attention layer and a feedforward layer. Residual connections and layer normalization are applied to all sub-layers in a Pre-LN manner.

[2]GRIN MoE weights: `https://huggingface.co/microsoft/GRIN-MoE`. Note a different version of mid-training and post-training, emphasizing long context and multilingual ability, has been conducted and has been released at `https://huggingface.co/microsoft/Phi-3.5-MoE-instruct`.

Table 2: Model Performance on Popular Benchmarks

	MoE				Dense								
	GRIN MoE 16×3.8B	Phi-3.5 16×3.8B	Mixtral 8×7B	Mixtral 8×22B	Phi-3 14B	Phi-3 7B	Mistral 7B	Gemma 7B	Llama3 8B	Llama3 70B	GPT-3.5 v1106	GPT-4o 2024	Gemini 1.5-Flash
MMLU (5-Shot)	79.4	78.9	70.5	76.2	78.0	75.7	61.7	63.6	66.5	80.2	71.4	86.9	79.4
HellaSwag (5-Shot)	83.7	83.8	70.4	79.0	82.4	77.0	58.5	49.8	71.1	82.6	78.8	91.7	-
ANLI (7-Shot)	60.6	59.8	55.2	65.2	55.8	58.1	47.1	48.7	57.3	68.3	58.1	75.7	65.6
GSM-8K (8-Shot; CoT)	90.4	88.7	64.7	83.8	91.0	89.6	46.4	59.8	77.4	93.5	78.1	93.8	82.4
MedQA (2-Shot)	70.4	70.5	62.2	67.9	69.9	65.4	50.0	49.6	60.5	78.5	63.4	88.9	-
AGIEval (0-Shot)	48.2	50.3	45.2	54.0	50.2	45.1	35.1	42.1	42.0	56.9	48.4	37.6	45.2
TriviaQA (5-Shot)	73.9	71.6	78.5	82.2	73.9	58.1	75.2	72.3	67.7	84.5	85.8	66.0	-
Arc-C (10-Shot)	92.0	91.0	87.3	91.3	91.6	90.7	78.6	78.3	82.8	93.0	87.4	97.0	88.3
Arc-E (10-Shot)	98.0	97.1	95.6	96.9	97.7	97.0	90.6	91.4	93.4	98.2	96.3	99.0	97.1
PIQA (5-Shot)	89.0	88.6	86.0	85.0	87.9	86.9	77.7	78.1	75.7	85.3	86.6	92.9	87.5
SociQA (5-Shot)	79.5	78.0	75.9	78.2	80.2	79.2	74.6	65.5	73.9	81.1	68.3	81.4	77.8
BigBench-Hard (3-Shot; CoT)	81.4	79.1	69.7	81.8	81.4	79.1	57.3	59.6	51.5	80.2	68.3	81.2*	-
WinoGrande (5-Shot)	81.4	81.3	62.0	75.3	81.5	81.5	54.2	55.6	65.0	83.3	68.8	89.3	74.7
OpenBookQA (10-Shot)	89.8	89.6	85.8	88.6	87.4	88.0	79.8	78.6	82.6	91.8	86.0	95.2	89.0
BoolQ (2-Shot)	83.4	84.6	77.6	82.7	86.5	84.8	72.2	66.0	80.9	89.1	79.1	90.6	85.8
CommonSenseQA (10-Shot)	81.8	83.5	78.1	82.0	82.8	80.0	72.6	76.2	79.0	84.4	79.6	88.5	84.0
TruthfulQA (10-Shot; MC2)	74.5	77.5	60.1	67.4	75.1	70.2	53.0	52.1	63.2	81.9	85.8	85.6	76.6
HumanEval (0-Shot)	74.4	70.7	37.8	39.6	62.2	61.0	28.0	34.1	60.4	78.7	62.2	92.1	64.4
MBPP (3-Shot)	80.3	80.8	60.2	70.7	75.2	71.7	50.8	51.5	67.7	81.3	77.8	90.4	77.5
Average	79.58	79.23	69.62	76.20	78.46	75.74	61.23	61.73	69.40	82.78	75.27	85.70	

Attention. Following Mistral (Jiang et al., 2023b), we implement the attention layer with grouped-query attention (Ainslie et al., 2023) and sliding window attention (Child et al., 2019). Both techniques are computationally efficient and allow GRIN MoE to attend information beyond the window size. RoPE is adopted for the position encoding to enable long context encoding after pretraining(Su et al., 2024). Our implementation is mostly based on FlashAttention 2 (Dao, 2023).

Mixture of Experts. Different from conventional Transformer models, we construct the feedforward layer as a Mixture-of-Experts layer, employing a router network to sparsely activate selected networks for each input.

The idea of MoE is originally discussed in Jacobs et al. (1991) and Jordan & Jacobs (1994), which integrates separate networks together and uses each to handle a separate subset of training cases. Recently, many attempts have been made to leverage MoE for scaling large language models (Shazeer et al., 2017; Lepikhin et al., 2021; Lewis et al., 2021; Kim et al., 2021; Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022).

For each MoE layer, the model picks from a set of distinct feedforward networks for every input tokens, which is determined by a router network. Particularly, given n expert parameters $\{\boldsymbol{w}_0, \cdots, \boldsymbol{w}_{n-1}\}$, the output of one MoE module for inference is

$$\textstyle\sum_{i=0}^{n-1} Gating(\boldsymbol{z})_i \cdot TopK(\boldsymbol{z})_i \cdot Expert(\boldsymbol{x}, \boldsymbol{w}_i), \tag{1}$$

where $\boldsymbol{z} = Router(\boldsymbol{x}, \boldsymbol{r})$, $\boldsymbol{r}$ is the router parameters, $Gating(\cdot)$ is a gating function (usually softmax), and $Expert(\cdot)$ is a FNN. In our study, we define use a linear network as the router, i.e., $Router(\boldsymbol{x}, \boldsymbol{r}) = \boldsymbol{x} \cdot \boldsymbol{r}^T$ As to $TopK(\boldsymbol{z})$, it is the TopK function, i.e., $TopK(\boldsymbol{z})_i := 1$ if $\boldsymbol{z}_i$ is among the TopK coordinates of $\boldsymbol{z}$ and $TopK(\boldsymbol{z})_i := 0$ otherwise.

During model training, different MoE algorithms may produce different outputs, as we will discuss in detail in Section 3.

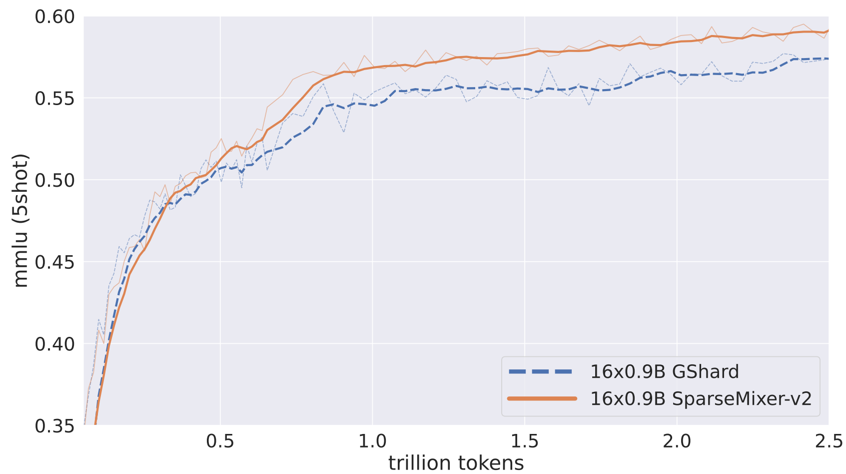

Figure 2: Controlled Comparisons of SparseMixer-v2 and GShard on 16×0.9B MoE.

3 GRIN MoE

In this section, we discuss in detail the two key techniques used in GRIN MoE:

- We propose SparseMixer-v2 to estimate the gradient related to expert routing, while the conventional MoE training treats expert gating as a proxy for the gradient estimation.
- We scale MoE training with neither expert parallelism nor token dropping, while the conventional MoE training employs expert parallelism and deploys token dropping.

3.1 Gradient Estimation for Expert Routing

Expert routing produces discrete expert assignment, allowing the network to be sparsely activated and thus bringing the great scaling potential. At the same time, such a routing process is not differentiable, thus making it infeasible to directly apply the vanilla backpropagation and obtain reliable gradient estimation.

Conventional MoE Training. Conventional MoE training uses the router outputs to compute gating results, treating the gating gradients as a proxy of the router gradient. Particularly, as in Equation 1, the output of the MoE module is: $\sum_{i=0}^{n-1} Gating_i(\boldsymbol{z}) \cdot TopK(\boldsymbol{z})_i \cdot Expert(\boldsymbol{x}, \boldsymbol{w}_i)$, where $\boldsymbol{z} = Router(\boldsymbol{x}, \boldsymbol{r})$. Conventional MoE training views $TopK(\boldsymbol{z})_i$ as constants and only back-propagates through $Gating(\boldsymbol{z})$ to compute router weight gradients. In this way, it treats the gating gradient as a proxy of the router gradient, i.e., $\nabla_{\text{conventional}} \boldsymbol{r} := \nabla Gating(\boldsymbol{z}) \cdot \frac{\partial Gating(\boldsymbol{z})}{\partial \boldsymbol{r}}$.

SparseMixer-v2. Inspired by recent advances in Straight-Through gradient estimators (Bengio et al., 2013; Liu et al., 2023a), we proposed the SparseMixer-v2 method, an extension of SparseMixer (Liu et al., 2023b), to obtain scalable and reliable gradient estimations in this study.

We briefly introduce the SparseMixer-v2 method below, and leave a detailed description to Appendix A. We first replace the $TopK(\cdot)$ function as random sampling of discrete variables in model training. Then, following Liu et al. (2023a) and Liu et al. (2023b), we apply Heun's third order method to approximate the expert routing gradient and construct a modified back-propagation to give a mathematically sound gradient estimation for expert routing.

Effectiveness of SparseMixer-v2. In Liu et al. (2023b), the effectiveness of SparseMixer is demonstrated on the neural machine translation task and the ELECTRA language model training. However, it has not been applied to autoregressive language model training at a large scale. In the development of GRIN MoE, we conducted controlled experiments that showd promising results for SparseMixer-v2. The result motivates us to apply this algorithm to training GRIN MoE.

Table 3: Training Throughput of Dense and MoE on 64 H100 gpus.

	Total Parameters	Active Parameters	Throughput Per GPU	Relative Throughput
Dense	1.6B	1.6B	34222	
MoE	10B	1.6B	27962	81.71%
Dense	6.6B	6.6B	8176	
MoE	42B	6.6B	7077	86.56%

Particularly, we trained two 16×0.9B MoEs with 2.5T tokens. One of them follows the same recipe used in GRIN MoE, and the other replaces SparseMixer-v2 with the conventional GShard method. As shown in Figure 2, the performance boost of SparseMixer-v2 generalizes to the autoregressive language model training at the 16×0.9B scale: although GShard performs better at the first 0.5T tokens, SparseMixer-v2 achieves stronger performance in the later stage of training. It is worth mentioning the similar phenomenon observed in small scale experiments on ELECTRA pretraining that Switch tends to perform better in the beginning while SparseMixer comes from behind in the late stage (Liu et al., 2023b). We suspect this is due to the model architecture difference, e.g., it introduces more randomness to training by replacing $TopK(\cdot)$ function with random sampling of discrete variables, which may slow down the training in the begining. Also, it is worth mentioning that, such extra randomness also makes it harder to compare the training loss to the GShard. Semi-controlled experiment results of a larger scale are discussed in Section 5.

3.2 Implementation and Scaling

Comparing to conventional models that activate all parameters for all inputs, MoE models have more parameters for the same FLOPs due to their structured sparse computation, significantly impacting computational efficiency. Conventional MoE training distributes different expert networks across devices (i.e., expert parallelism) and employs strategies like token dropping to facilitate the training process.

As our first step towards pursuing the scalability brought by MoE, we focus on MoE training with a relative small number of experts (i.e., top2 routing over 16 experts). Leveraging recent engineering advances, we avoid expert parallelism and eliminate the need for capacity factor or token dropping. In the end, we are able to achieve over 80% relative training efficiency improvement, compared to a dense model with the same active parameters, for GRIN MoE training.

MoE Implementation. For MoE computation without expert parallelism, we find the Megablocks (Gale et al., 2023) package to be very helpful. Particularly, we find its `grouped_GEMM` kernels and wrappers outperform its `sparse` version, offering substantial performance improvement. In addition, we rely on data parallelism, pipeline parallelism, and activation checkpointing in the training of GRIN MoE, which lead to the best throughput for our 16×3.8B model.

Training Throughput Comparisons of Dense and MoE Models. To showcase the benefits of MoE training, we compare its training throughput to that of a conventional dense model. Hardware details for these studies are in Appendix B. It is important to note that, the throughput of the dense model is measured under the same parallelism setting as that of the MoE model, and the comparison here is to study the GPU kernel efficiency of densely activated networks (i.e., Dense) and sparsely activated networks (i.e., MoE).

As summarized in Table 3, we compare MoE models of two different sizes to their corresponding dense models with the same number of parameters, measuring their training throughput using the identical hardware. Despite having over six times as many parameters as the dense model, MoE models achieve more than 80% relative throughput in this experiment, confirming the significant computational scaling potential of models with GRIN MoE method.

Additionally, our observations indicate that MoE models do not experience more severe or different throughput degradation compared to dense models when scaling up model size. Both dense and MoE models show similar slowdown patterns in our experiments. For instance,

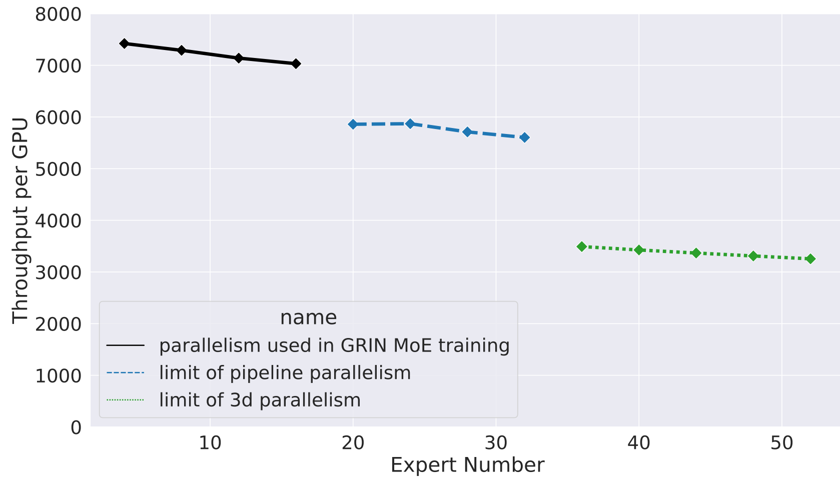

Figure 3: Scaling of Different Parallelism Settings on 64 H100 gpus. The reported throughput for N experts (x-axis) refers to the average training throughput of a 3.8BxN top2 MoE.

the training throughput of the 6.6B dense model is approximately 4.19 times slower than that of the 1.6B dense model, which has 4 times fewer parameters. Similarly, the training throughput of the 42B MoE model is about 3.96 times slower than that of the 10B MoE model, which has 4.2 times fewer parameters.

Scaling Study and Tensor Parallelism. In this section, we discuss whether it is computationally feasible to train MoE models with a larger number of experts without employing expert parallelism.

First, by relying solely on pipeline parallelism, the maximum number of experts can be extended from 16 to 32 by further partitioning different layers across GPUs. However, increasing the number of experts beyond this would result in too many parameters for a single layer, making it difficult to support without partitioning one layer across multiple GPUs.

To address this challenge, conventional MoE training relies on expert parallelism to further partition the model, which introduces the side effects of capacity factor and token dropping. In our study, we explore the use of tensor parallelism instead of expert parallelism (Narayanan et al., 2021). Similar to expert parallelism, which has two all-to-all communication overheads in both forward and backward computations, tensor parallelism has two all-reduce communication overheads in forward and backward computations. Although all-reduce operations have higher latency than all-to-all operations, we can mitigate these overheads by overlapping communication with computation during backward computation.

As in Figure 3, the maximum number of supported experts is extended to 52 (with 132B total parameters) by combining pipeline parallelism and tensor parallelism (i.e., 3D parallelism). It is worth noting that, since our throughput study hardware setting has only 64 GPUs, we can partition the model into at most 64 stages. With 272 H100 GPUs, the limit of this parallelism can be further extended to 200+ experts.

While this demonstrates the feasibility of scaling MoE training without expert parallelism, we observe that using more complex parallelism typically leads to reduced computational throughput. Correspondingly, one important direction for our future work is to perform MoE training with more experts in a more sparse manner.

3.3 Global Load Balance Loss Adaptations

As previously discussed, our training framework is designed to scale MoE training using tensor parallelism, pipeline parallelism, and data parallelism, but not expert parallelism. Correspondingly, there is no need to employ token dropping or capacity factor, allowing the distribution of activated experts to deviate from a uniform distribution.

Therefore, we adapt the load balance loss to regulate the global expert load balance instead of the local one. Particularly, the popular load balancing loss is defined as :

$$\alpha \cdot n \cdot \sum_{i=1}^{n} f_i \cdot E[\texttt{softmax}(\boldsymbol{z})_i],$$

where α is a hyper parameter, n is the number of experts, and f_i is the fraction of tokens dispatched to expert i (Fedus et al., 2022). Conventionally, f_i is computed at different GPUs locally and thus the load balance loss will regulate the local expert load balance and alleviate token dropping. In our study, we modified the load balance auxiliary loss by computing f_i globally (i.e., all-reduced within the data-parallel process group), regulating the expert load to be balanced globally. Although this adjustment incurs additional communication overheads, similar to tensor parallelism, these communications can be performed in parallel with computations in an asynchronized way, thus largely reducing the additional latency.

4 Experiment

4.1 Training Setting

Pre-training. GRIN MoE is pre-trained on 4T tokens as a Causal Language Model. The same training dataset has been used to train Phi-3 dense models (Abdin et al., 2024).

Post-training. Post-training consists of two stages: Supervised Fine-Tuning (SFT) based on the causal language modeling objective, followed by Direct Preference Optimization (DPO; Rafailov et al., 2024). The model is trained with 24B tokens in SFT, using high-quality data across diverse categories, e.g., math, coding and conversation (Abdin et al., 2024). The DPO dataset contains 1.4B tokens, including safety and identity preference data that is used to align the model output with Microsoft's Responsible AI principles (Haider et al., 2024). We further adopt regularization techniques, such as adding random noise to the input embedding (Jain et al., 2024) and applying dropout in expert layers (Fedus et al., 2022), to improve the model's generalization performance. It is worth mentioning that another version of mid-training and post-training have been conducted with an emphasize on long context and multilingual ability, which has been released as Phi-3.5-MoE (Abdin et al., 2024).

4.2 Evaluation of GRIN MoE

Table 2 summarizes the performance of GRIN MoE on popular benchmarks. Benchmarks and baseline methods are elaborated in Appendix B.

Since both Phi-3 and GRIN MoE models are trained on the same datasets, the effectiveness of our MoE training recipe is easily demonstrated. We can see that GRIN MoE with 6.6B activated parameters performs significantly better than 7B dense model and similar to the 14B dense model. Compared to the 14B dense model, GRIN MoE performs better on math, coding, and MMLU tasks.

Comparing GRIN MoE to Phi-3.5-MoE, which has been developed with a different focus (i.e., multilingual capabilities and long context handling), we find that these two models have distinct strengths. We observed that GRIN MoE excels in math and reasoning tasks, while Phi-3.5-MoE demonstrates superior performance in question-answering (QA). Despite their different strengths, both models yield similar average scores across various benchmarks, which is expected given that both are configured as 16x3.8B MoEs and both are trained with sparse backpropagation. Further comparisons are available in Section4.3.

Our evaluation also shows that GRIN MoE is significantly better than many open-sourced models with a similar number of active parameters, such as Mixtral 8×7B (12.9b activated parameters), Mistral 7B, Gemma 7B, Llama3 8B. And GRIN MoE is better than Mixtral 8×22B on most of the tasks. Nevertheless, GRIN MoE's performance still falls short of Llama3 70B and GPT-4o. This gap is expected, given the substantially larger computational and data resources utilized in training these two latter models.

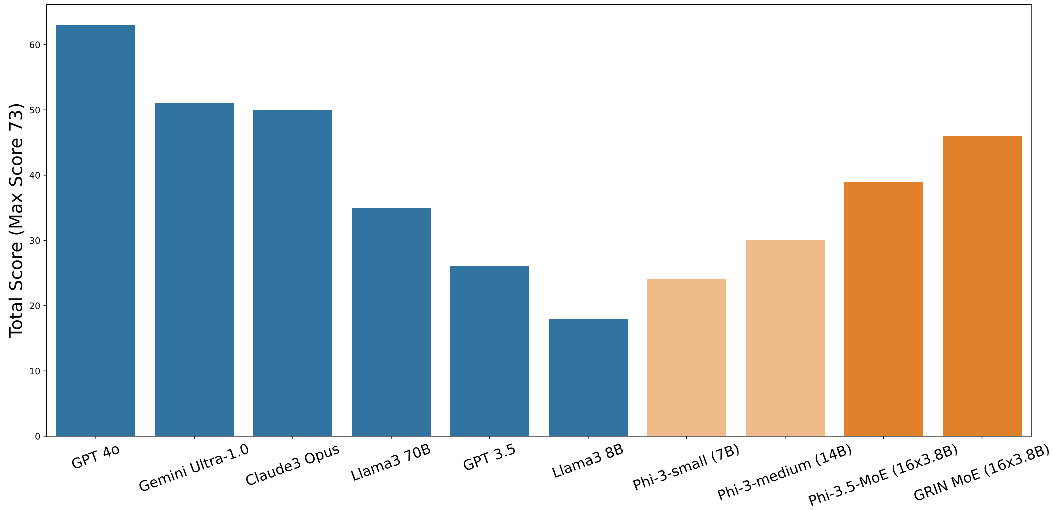

Figure 4: Test Score on Translated 2024 GAOKAO Math-1.

Table 4: GRIN MoE performance on LiveBench-2024-07-25. Models are ranked by their average score (AVG). *Baseline results are referenced from the official benchmark.

	Reasoning	Coding	Mathematics	Data Analysis	Language	IF	AVG
Claude-3-haiku*	29.3	24.5	25.7	41.5	30.1	64.0	35.9
Mixtral-8x22B-instruct-v0.1*	29.3	32.0	28.3	31.7	26.5	63.1	35.2
GPT-3.5-turbo-0125*	26.7	27.7	26.9	41.2	24.2	60.5	34.5
GRIN MoE	35.3	23.7	29.8	32.0	16.9	57.6	32.5
Mistral-small-2402*	26.0	21.2	28.2	31.9	22.1	63.9	32.2
Command-r-plus*	28.7	19.5	24.9	24.6	23.9	71.5	32.2
Gemma-2-9B-it*	17.3	22.5	24.0	35.1	27.6	61.6	31.3

4.3 Math Ability Case Study

Phi-3 data, the training data of GRIN MoE, contains a massive amount of synthetic data, greatly boosting model performance on benchmarks. Despite its effectiveness, it left doubts on the performance of GRIN MoE on real-world tasks. Correspondingly, we conduct case studies on math questions of the newly released GAOKAO exam (i.e., Nationwide Unified Examination for Admissions to General Universities and Colleges), which is the annual national undergraduate admission exam in China. Known for its rigorous security protocols, this exam serves as an ideal ”truly held-out” test bed for assessing AI models’ ability to answer math questions. Note that the training of GRIN MoE concludes on June 3, Pacific Standard Time, and the 2024 GAOKAO starts on June 7, China Standard Time.

Exam Score. To assess the capability of various models in answering math questions, we used translated questions as the input[3], scored their responses manually, and visualized the result in Figure 4 (detailed scoring results and the response of GRIN MoE can be find in Appendix C). GRIN MoE scored 46 out of 73 points on these questions, outperforming Llama3 70B by 11 points, and is only 6 and 5 points away from the Gemini Ultra-1.0 and Claude3 Opus, respectively.

These results demonstrate GRIN MoE ’s strong capacity for mathematical reasoning. As the chance of data leakage in this case study is slim, the result suggests that the capacity of GRIN MoE is likely attributed to the generative distillation approach (Hsieh et al., 2023; Mukherjee et al., 2023), instead of memorization.

GRIN MoE Responses Discussion. Analyzing the response of GRIN MoE to these questions, we have some interesting observations:

[3]We used the translated questions from `https://github.com/zhengaq/GAOKAO-Math24`.

- GRIN MoE is capable of answering challenging math questions. For instance, question 4 requires the model to perform algebraic manipulations involving trigonometric functions with an unknown variable. As in Appendix D.4, GRIN MoE not only produces the correct answer but also presents all the necessary derivations accurately. Notably, GRIN MoE is the only model among those we tested, including Llama3 and the Phi-3 models, to correctly answer this question.
- For single-choice questions and multiple-choice questions, sometimes GRIN MoE produces the right or partially right answers, together with chain-of-thought prompts that have typos (question 5 as in Appendix D.5 and question 9 as in Appendix D.9), skipped steps (question 7 as in Appendix D.7) or errors (question 10 as in Appendix D.10 and question 11 as in Appendix D.11). Additionally, we observe that the output of GRIN MoE may change dramatically for challenging questions. These phenomenons are also observed in the responses of other models.
- GRIN MoE has the ability to quickly catch a hint. For question 13, as in Appendix D.13, GRIN MoE initially makes a mistake by assuming that the point $(0, 1)$ is on the curve $y = \ln(x + 1) + a$, leading to an incorrect answer. However, after incorporating a hint that highlights the relationship between the curve and the point, GRIN MoE adjusts its response significantly and solves the problem correctly (as in Appendix D.15). It is noteworthy that GRIN MoE is the only model to generate the correct answer to this question with the hint, among all Llama3 and Phi-3 dense models.

When comparing GRIN MoE to Phi-3.5-MoE, different response patterns have been observed. As shown in Appendices E.1 and E.7, Phi-3.5-MoE occasionally generated responses without chain-of-thought prompts, while GRIN MoE consistently provided chain-of-thought responses for all 14 questions. Furthermore, as in Appendices E.12 and E.4, Phi-3.5-MoE sometimes produced responses with repetitive outputs at the end, whereas GRIN MoE maintained proper conclusions for all 14 questions. We suspect this behavior may be related to the different focus of the training in Phi-3.5 models which include handling of a longer context window up to 128K tokens and multilingual capabilities.

Moreover, the answers produced by Phi-3.5-MoE differ significantly from those of GRIN MoE. The models provided identical answers for only 5 out of 14 questions. Even for questions both models answered correctly, they offered different explanations. Additionally, while both models made the same mistake on question 13 (as shown in Appendices D.13 and E.13), GRIN MoE was able to quickly incorporate the provided hint and answer the question correctly (as in Appendix D.15), whereas Phi-3.5-MoE failed (as in Appendix E.15).

4.4 Limitations and Weakness

Since the Phi-3 data, the training corpus of GRIN MoE, is constructed with an emphasize on reasoning and coding abilities, we observe the model to yield a suboptimal performance on natural language tasks. We use the 2024-07-25 release of the LiveBench for model evaluation (White et al., 2024) and summarize the performance of GRIN MoE in Table 4, which also shows the performance of 6 other models that have similar average scores.

Comparing to baselines having similar average score on this benchmark, GRIN MoE achieves better scores on the reasoning, coding, and mathematics. The result is consistent with our case study in Section 4.3. Meanwhile, we observe that GRIN MoE achieves an exceptionally low average score (i.e., 16.9) on natural language tasks. We suspect that this is due to the limitation of the training corpus, since other models trained on the same corpus exhibit similar problems.

5 Analyses

As described in Section 3, we have tailored the training recipe for GRIN MoE, featuring SparseMixer-v2 and load balance loss adaptation. However, due to resource constraints, we were unable to set up a controlled environment to individually study the impact of each variable at the scale of GRIN MoE. Therefore, we conducted a semi-controlled comparison to quantify the effect of the training recipe.

Table 5: Semi-controlled Performance Comparisons. Different from results in Table 2 and Figure 1, the reported performances here are before the posttraining. Detailed settings are elaborated in Section 5.1.

		MMLU	Arc-C	GSM-8K
Dense	7B	75.1	89.8	85.3
	14B	76.4	89.8	85.9
MoE (Control recipe)	16×3.8B	75.1	88.9	79.1
MoE (Main recipe)	16×3.8B	77.4	89.6	89.5

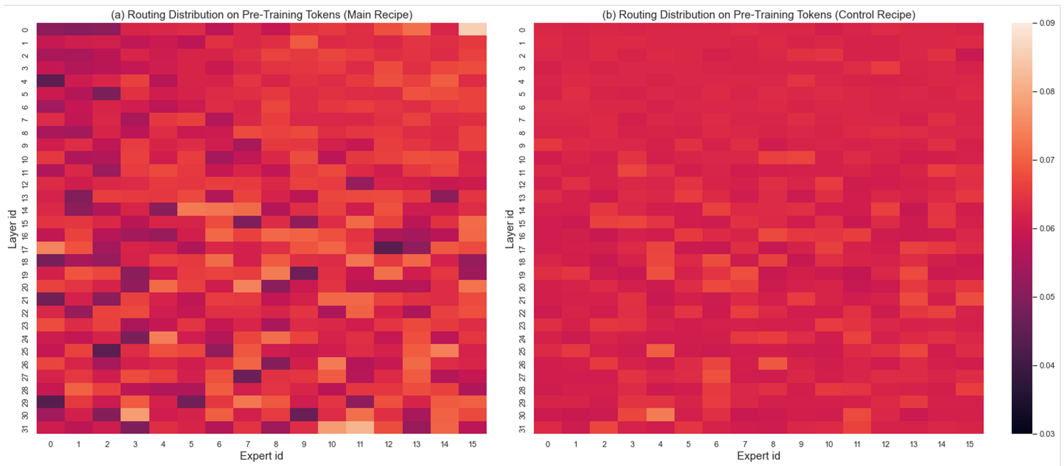

Figure 5: Routing distribution on 2 million pretraining tokens. The model on the left is trained by main recipe and the right is trained by control recipe. The values are normalized per layer. The summation of the values in each row is 1 (perfectly balanced loading would result in a value of 0.0625).

5.1 Semi-controlled Setting

We compare the following two training recipes (the *Main recipe* and the *Control recipe*):

- *Main recipe* is the one used for GRIN MoE training, as described in Section 3.
- *Control recipe* resembles conventional MoE training recipes and is used for comparison. It differs from the main recipe in that the former replaces SparseMixer-v2 used in the main recipe with GShard, exchanges global load balance loss for local load balance loss, and modifies several hyper-parameters.

We then trained MoE models using the two recipes on a 4T-token corpus, and compared them with the Phi-3 7B and 14B models trained on a super-set of the 4T-token corpus on downstream tasks without post-training.

Note that comparing to the controlled experiment in Figure 2, the control recipe does not adapts the global load balance loss as in Section 3.3, while the GShard baseline in Figure 2 adapts the global load balance loss adaptation.

5.2 Downstream Performance

The results are presented in Table 5. The model trained using the control recipe matches the performance of the 7B dense model. The main recipe is more effective, resulting in a model whose performance matches that of the 14B dense model. We attribute the effectiveness to the use of SparseMixer-v2 and the adaptive loss modifications.

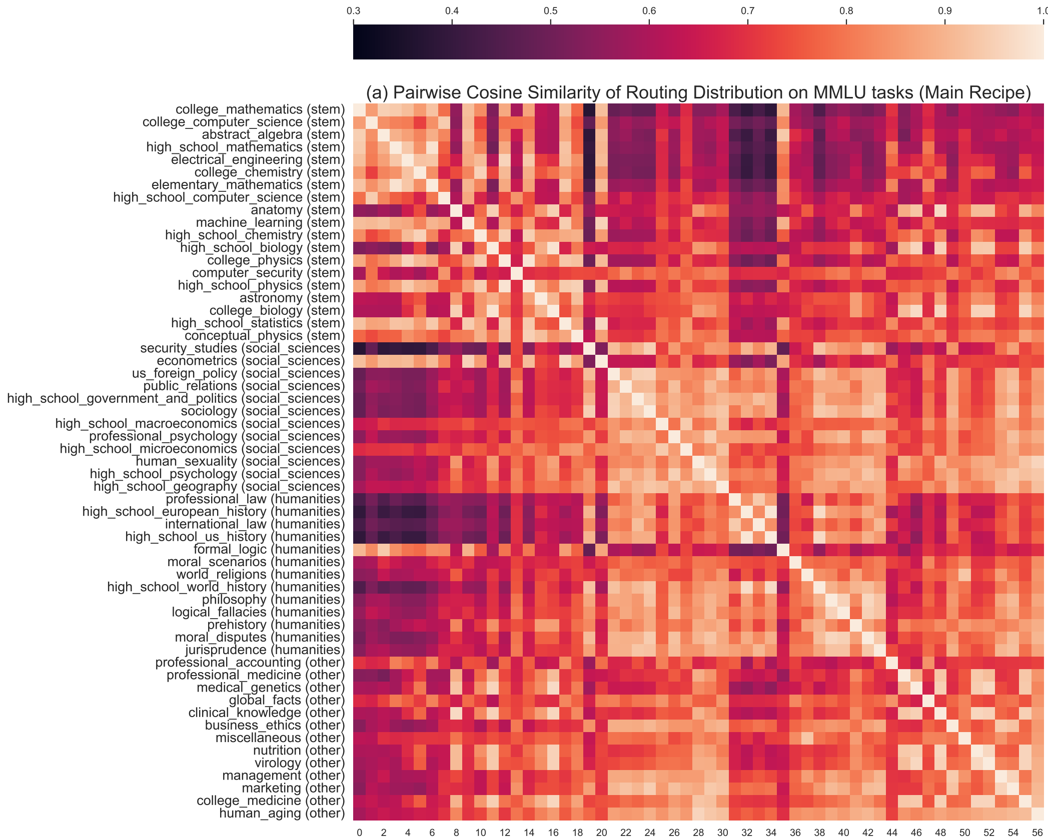

Figure 6 (a): MoE Routing distribution similarity across MMLU 57 tasks for the main recipe. The darker color means lower similarity. X-axel and Y-axel have the same task order.

5.3 Routing Analyses

We analyze the routing distributions of models trained with the main and control recipes. We count how many times each expert (in each layer) is selected by different hidden states as routing distribution.

MoE routing on pretraining data. To analyze the routing distribution on the pretraining dataset, we randomly select 2 million tokens from the pretraining dataset and visualize the expert loading distribution on them in Figure 5. First, we see that all layers for both recipes have reasonably balanced expert loading. Note that the maximum value in Figure 5 is 0.09, which is about 1.44 times of the perfectly balanced expert loading (0.0625). Then, relatively, we observe that the model trained with the main recipe is less balanced than the control recipe.

MoE routing on different MMLU tasks. Next, we empirically verify whether different experts contain domain specific information. We first compare the routing distribution among the MMLU 57 tasks. For each task, we sample 24 prompts with 5-shot. The routing distribution for each task is a vector with 16 (experts per layer) $\times$ 32 (layers) dimensions (total number of experts). We then compute cosine similarity between the routing distribution of different tasks and visualize the similarities as a heatmap in Figure 6 (a) and 6 (b), where we group these 57 tasks into 4 categories based on their meta data. Note that the meta data we used here is provided in Hendrycks et al. (2021),

The MoE trained with our main recipe is shown in Figure 6 (a), we can see the `STEM` category has a clear boundary to `social_science` and `humanities`. Additionally, it is quite reasonable for the two outliers in `social_science` and `humanities`, (i.e., `econometrics` and `formal_logic`), to have a higher similarity to the `STEM` category. This indicates that

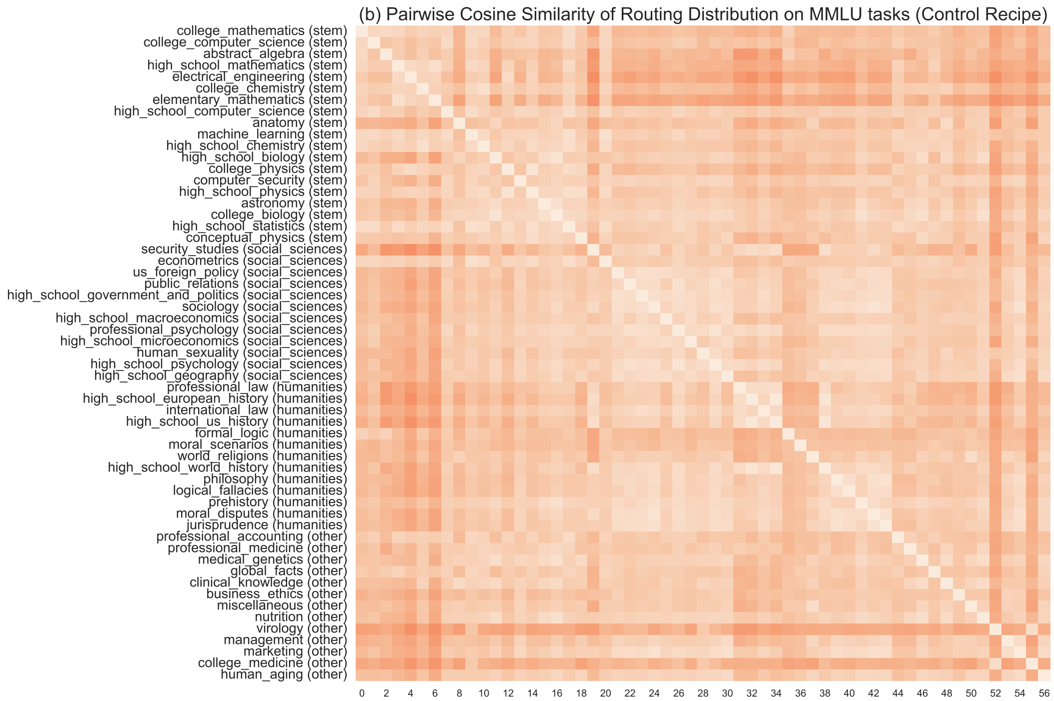

Figure 6 (b): MoE Routing distribution similarity across MMLU 57 tasks for the control recipe. This figure shares the same setting and the same color bar with Figure 6 (a).

the routing distribution can vary significantly among different tasks. We also visualize the model trained with the control recipe (as in Figure 6 (b)), in which the routing distributions are similar across different tasks.

A case study of MoE routing. This analysis uses the MoE model trained with the main recipe. As shown in Figure 7, the routing distributions are different in different layers. The bottom (shallow) layer has the most balanced expert distribution. In the middle layer, the experts 10 and 15 are selected more other than the other experts. The final layer (deep) comes to be more balanced than the middle layer. These findings reveal that MoE routing distributions are related to such information as context, word, position, etc.

Our study seems to verify our hypothesis that expert networks in GRIN MoE have developed highly-specialized and heterogeneous expertise. As pointed out in Wei et al. (2024), such experts are likely to improve models' capacity.

6 Conclusion

In this paper, we describe in detail a new MoE model, known as GRIN MoE, and the model training techniques (i.e., sparse backpropgationa and model parallelism configuration) used to train the model. Compared to dense models trained on the same pretraining corpus, GRIN MoE demonstrates a remarkable scaling potential of MoE. We also provided a summary of our observations and insights gained during GRIN MoE 's development, aiming to deepen our understanding of MoE training. Through controlled and semi-controlled experiments we have demonstrated how gradient estimation methods and model parallelism strategies, along with corresponding auxiliary adaptations, significantly improves model training.

Looking ahead, many important open questions remain. For example, the training and inference of MoE models present challenges to both algorithms and engineering implementations. Also, since softmax is originally designed to approximate the argmax operation, it presents new challenges to approximate topk as sampling. We plan to further

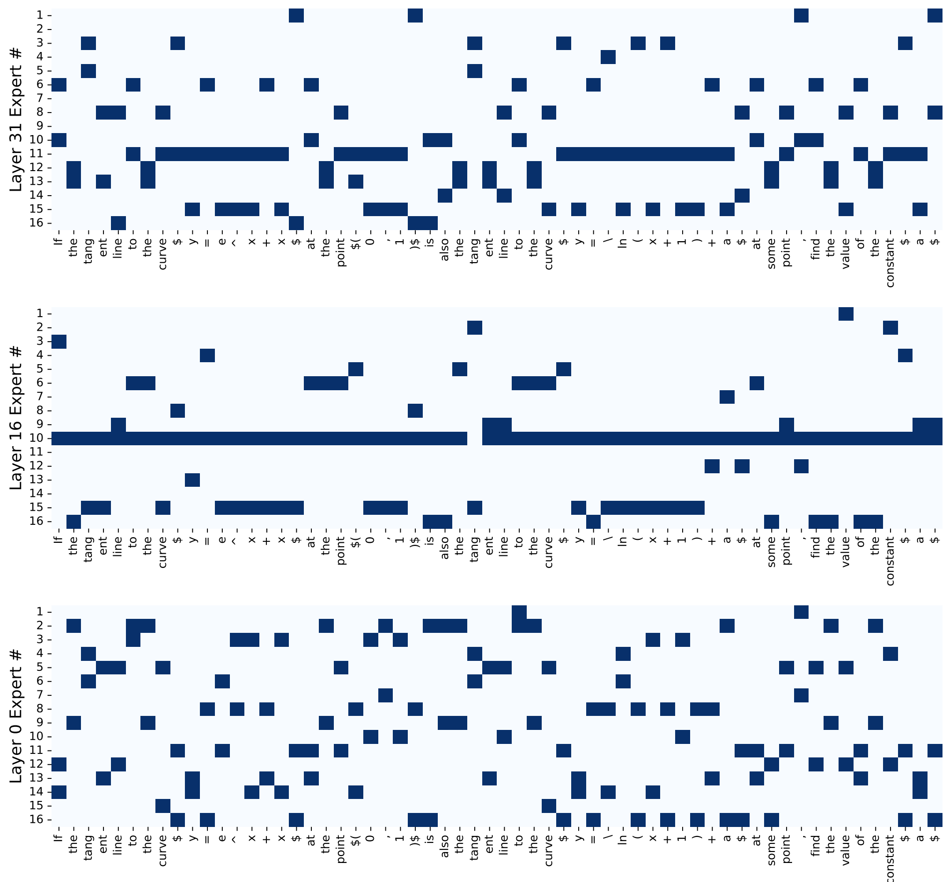

Figure 7: Case study of MoE trained by main recipe. It visualizes how experts are selected at different position in different layers. We select one question from newly released College Entrance Exam for the case study and more information is in Appendix D.15.

explore solutions to these challenges, with a focus on enhancing sparsity and developing efficient computing and scaling methods to advance state-of-the-art MoE modeling.

Acaknowledgement

We would like to thank the engineering team at Microsoft for providing computation infrastructure support. Also, we would like to thank Mojan Javaheripi, Suriya Gunasekar, and Sébastien Bubeck for the pre-training data. Additionally, we would like to thank Abhishek Goswami, Alessandro Sordoni, Amin Saied, Ammar Ahmad Awan, Anh Nguyen, Benjamin Sorscher, Caio César Teodoro Mendes, Chandan Singh, Colin White, Eric Xihui Lin, Gustavo de Rosa, Hanxiao Liu, Joe Mayer, Liliang Ren, Marko Radmilac, Michael Santacroce, Michael Wyatt, Olatunji Ruwase, Philipp Witte, Russell Hewett, Swadheen Shukla, Xia Song, Yadong Lu, Yang Liu, Yuanzhi Li, Yunan Zhang, Zichong Li, Zhen Zheng for their assistance and helpful discussions.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024.

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebron, F., and Sanghai, S. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Bouamor, H., Pino, J., and Bali, K. (eds.), *EMNLP*, 2023.

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

Bengio, Y., Léonard, N., and Courville, A. C. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*, 2013.

Bisk, Y., Zellers, R., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, 2019.

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., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*, 2019.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457*, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv:2307.08691*, 2023.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 2022.

Gale, T., Narayanan, D., Young, C., and Zaharia, M. A. Megablocks: Efficient sparse training with mixture-of-experts. In *MLSys*, 2023.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Haider, E., Perez-Becker, D., Portet, T., Madan, P., Garg, A., Majercak, D., Wen, W., Kim, D., Yang, Z., Zhang, J., Sharma, H., Bullwinkel, B., Pouliot, M., Minnich, A., Chawla, S., Herrera, S., Warreth, S., Engler, M., Lopez, G., Chikanov, N., Dheekonda, R. S. R., Jagdagdorj, B.-E., Lutz, R., Lundeen, R., Westerhoff, T., Bryan, P., Seifert, C., Kumar, R. S. S., Berkley, A., and Kessler, A. Phi-3 safety post-training: Aligning language models with a "break-fix" cycle. *arXiv:2407.13833*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *ACL Findings*, 2023.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 1991.

Jain, N., yeh Chiang, P., Wen, Y., Kirchenbauer, J., Chu, H.-M., Somepalli, G., Bartoldson, B. R., Kailkhura, B., Schwarzschild, A., Saha, A., Goldblum, M., Geiping, J., and Goldstein, T. NEFTune: Noisy embeddings improve instruction finetuning. In *ICLR*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *arXiv:2310.06825*, 2023a.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *arXiv:2310.06825*, 2023b.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv:2401.04088*, 2024.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv:2009.13081*, 2020.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 1994.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Kim, Y. J., Awan, A. A., Muzio, A., Salinas, A. F. C., Lu, L., Hendy, A., Rajbhandari, S., He, Y., and Awadalla, H. H. Scalable and efficient moe training for multitask multilingual models. *arXiv:2109.10465*, 2021.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N. M., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. In *ICLR*, 2021.

Lewis, M., Bhosale, S., Dettmers, T., Goyal, N., and Zettlemoyer, L. Base layers: Simplifying training of large, sparse models. In *ICML*, 2021.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*, 2021.

Liu, L., Liu, X., Gao, J., Chen, W., and Han, J. Understanding the difficulty of training transformers. In *EMNLP*, 2020.

Liu, L., Dong, C., Liu, X., Yu, B., and Gao, J. Bridging discrete and backpropagation: Straight-through and beyond. In *NeurIPS 2023*, 2023a.

Liu, L., Gao, J., and Chen, W. Sparse backpropagation for moe training. In *WANT@NeurIPS 2023*, 2023b.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv:2306.02707*, 2023.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., and Zaharia, M. A. Efficient large-scale language model training on gpu clusters using megatron-lm. In *SC21*, 2021.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. In *ACL*, 2020.

OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2024.

Rosenblatt, F. *The perceptron, a perceiving and recognizing automaton*. Cornell Aeronautical Laboratory, 1957.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020.

Shazeer, N. M., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., vinay uday prabhu, Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*, 2023.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 2024.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL Findings*, 2023.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*, 2019.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Wei, T., Zhu, B., Zhao, L., Cheng, C., Li, B., Lü, W., Cheng, P., Zhang, J., Zhang, X., Zeng, L., Wang, X., Ma, Y., Hu, R., Yan, S., Fang, H., and Zhou, Y. Skywork-moe: A deep dive into training techniques for mixture-of-experts language models. *arXiv:2406.06563*, 2024.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv:2406.19314*, 2024.

Yang, G., Hu, J. E., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J. W., Chen, W., and Gao, J. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. In *NeurIPS*, 2021.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *ACL*, 2019.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. Agieval: A human-centric benchmark for evaluating foundation models. In *NAACL-HLT Findings*, 2024.

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

Appendix I

SparseMixer-v2 and Setting

Table of Contents

A SparseMixer-v2

SparseMixer-v2 has two important components, i.e., approximating `TopK` with discrete variable sampling, and estimating gradients in a scalable manner. Here, we first introduce SparseMixer-v2 for `Top1` MoE, then extend it for `TopK` MoE.

A.1 Discrete Variable Sampling for Approximating `Top1`

In conventional MoE training, it is a standard practice to add jitter noise to expert routing, which would lead to implicit expert sampling. With SparseMixer-v2, we try to replace the implicit expert sampling with explicit expert sampling.

We start from the special case of $K = 1$. Particularly, we will approximate $\texttt{Top1}(\boldsymbol{z})$ as sampling from $\texttt{MaskedSoftmax}(\boldsymbol{z})$:

$$\texttt{MaskedSoftmax}(\boldsymbol{z})_i = \frac{\exp(\boldsymbol{z}_i) \cdot \delta_i}{\sum_j \exp(\boldsymbol{z}_j) \cdot \delta_j}, \tag{2}$$

where $\delta_i = \delta\big(\boldsymbol{z}^* - \boldsymbol{z}_i \leq r \cdot (|\boldsymbol{z}_i| + |\boldsymbol{z}^*|)\big)$, $\boldsymbol{z}^* = \max_k \boldsymbol{z}_k$, and r is a hyper-parameter. Comparing to the vanilla softmax sampling, $\texttt{MaskedSoftmax}(\cdot)$ introduces a hyper-parameter r to control the magnitude of randomness and sparsity of the sampling space. Similar to Fedus et al. (2022) and Liu et al. (2023b), we observed that the vanilla softmax sampling doesn't work well in practice. In our experiments, $\texttt{MaskedSoftmax}(\cdot)$ yields similar performance with conventional jitter noise in MoE training.

A.2 `Top1` SparseMixer-v2 MoE

For expert routing gradient estimation, we proposed two slightly different variants, SparseMixer-v2* and SparseMixer-v2. Let us first present SparseMixer-v2* and then move to introducing SparseMixer-v2.

SparseMixer-v2*. We configure the MoE layer for SparseMixer-v2* as:

$$\sum_{i=0}^{n-1} \texttt{MaskedSoftmax}(\boldsymbol{z})_i \cdot \boldsymbol{D}_i \cdot \mathit{Expert}(\boldsymbol{x}, \boldsymbol{w}_i), \text{ where } \boldsymbol{D} \sim \texttt{MaskedSoftmax}(\boldsymbol{z}). \tag{3}$$

While the SparseMixer-v2* MoE layer will behave as Equation 3 during inference, it will behave as in Algorithm 2 during training. Now lets proceed to the underlying derivations of Algorithm 2. Referring other parts of the network (including the loss function) as f, the training objective is:

$$
\begin{aligned}
\mathcal{L} &= E_{\boldsymbol{D}\sim\texttt{MaskedSoftmax}(\boldsymbol{z})}\left[f\left(\sum_{i=0}^{n-1}\texttt{MaskedSoftmax}(\boldsymbol{z})_i\cdot\boldsymbol{D}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\right)\right]\\
&= \sum_{i=0}^{n-1} f\big(\texttt{MaskedSoftmax}(\boldsymbol{z})_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)\cdot\texttt{MaskedSoftmax}(\boldsymbol{z})_i. \qquad (4)
\end{aligned}
$$

For simplicity, we mark $\texttt{MaskedSoftmax}(\boldsymbol{z})$ as $\boldsymbol{p}$. Then the gradient of $\boldsymbol{z}$ is:

$$
\begin{aligned}
\nabla\boldsymbol{z} &= \sum_{i=0}^{n-1}\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)\cdot\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{z}}\\
&= \sum_{i=0}^{n-1}\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + \Big(f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)-f(0)\Big)\cdot\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{z}}. \qquad (5)
\end{aligned}
$$

Note that Equation 5 is usually known as baseline subtraction[4]. In the ODE literature, there are many ways to approximate $f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)-f(0)$. Here, we focus on two approximations[5]:

- Euler's method: a first-order ODE solver that approximates $f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)-f(0)$ as $f'\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)\cdot\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)$.
- Heun's third-order method: a third-order ODE solver that approximates $f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)-f(0)$ as $\Big(\frac{1}{4}\cdot f'\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)+\frac{3}{4}\cdot f'(\frac{\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)}{3})\Big)\cdot\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)$.

Correspondingly, we approximate ∇ in two ways. Following Euler's method, we have:

$$
\begin{aligned}
\nabla_{\text{1st}}\,\boldsymbol{z} &= \sum_{i=0}^{n-1}\left(\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + f'\big(\boldsymbol{p}_i Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)\cdot\boldsymbol{p}_i Expert(\boldsymbol{x},\boldsymbol{w}_i)\cdot\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{z}}\right)\\
&= \sum_{i=0}^{n-1}\left(\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + \boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{p}_i Expert(\boldsymbol{x},\boldsymbol{w}_i)}\frac{\partial\boldsymbol{p}_i Expert(\boldsymbol{x},\boldsymbol{w}_i)}{\partial\boldsymbol{p}_i}\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{z}}\right)\\
&= \sum_{i=0}^{n-1} 2\cdot\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}}\\
&= E_{\boldsymbol{D}\sim\texttt{MaskedSoftmax}(\boldsymbol{z})}\left[2\cdot\frac{\partial f\big(\boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})\big)}{\partial\boldsymbol{z}}\right].
\end{aligned}
$$

Following Heun's third-order method, we have:

$$
\begin{aligned}
\nabla_{\text{3rd}}\,\boldsymbol{z} &= \sum_{i=0}^{n-1}\Bigg(\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + \bigg(\frac{1}{4}\cdot f'\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)+\\
&\qquad \frac{3}{4}\cdot f'(\frac{\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)}{3})\bigg)\cdot\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\cdot\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{z}}\Bigg)\\
&= \sum_{i=0}^{n-1}(\frac{5}{4}\cdot\boldsymbol{p}_i\cdot\frac{\partial f\big(\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\big)}{\partial\boldsymbol{z}} + \frac{9}{4}\cdot\boldsymbol{p}_i\cdot\frac{\partial f(\frac{\boldsymbol{p}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)}{3})}{\partial\boldsymbol{z}})\\
&= E_{\boldsymbol{D}\sim\texttt{MaskedSoftmax}(\boldsymbol{z}),\boldsymbol{B}\sim\texttt{Bernoulli}(\frac{5}{8})}\left[(6-4\boldsymbol{B})\cdot\frac{\partial f\big(\frac{1+2\boldsymbol{B}}{3}\cdot\boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})\big)}{\partial\boldsymbol{z}}\right].
\end{aligned}
$$

[4] since $\sum_i\boldsymbol{p}_i=1$, we have $0=\frac{\partial\sum_i\boldsymbol{p}_i}{\partial\boldsymbol{z}}$.

[5] For simplicity, we use $f'(\boldsymbol{h})$ to refer to $\frac{\partial f(\boldsymbol{h})}{\partial\boldsymbol{h}}$.

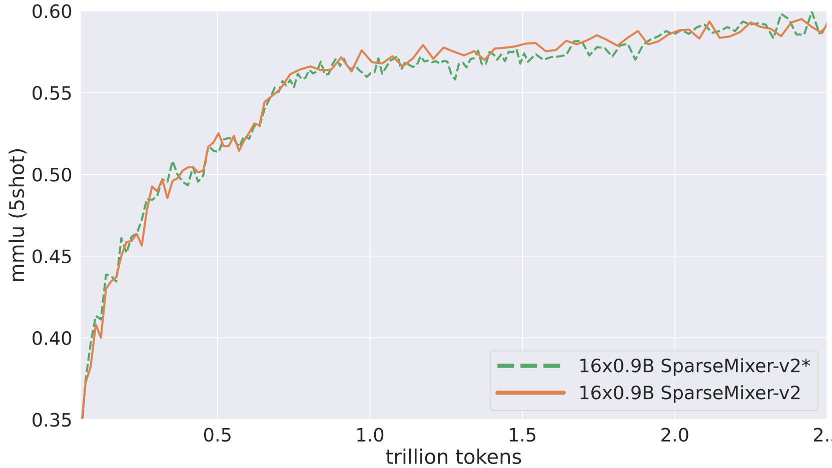

Figure 8: Controlled Comparisons of SparseMixer-v2 and v2* on 16×0.9B MoE.

Lastly, as discussed in Liu et al. (2023b), we combine ∇_{3rd} and ∇_{1st} to balance router learning and expert learning by using ∇_{1st} if $\boldsymbol{D} = \arg\max(\boldsymbol{z})$ or ∇_{3rd} otherwise. Particularly, using $\delta_{\boldsymbol{D}}$ to refers to $\delta(\boldsymbol{D} = \arg\max(\boldsymbol{z}))$, we have:

$$\nabla_{\text{SparseMixer-v2*}}\, \boldsymbol{z} = E_{\boldsymbol{D}\sim\texttt{MaskedSoftmax}(\boldsymbol{z})}[\nabla_{\boldsymbol{D},\text{SparseMixer-v2*}}\, \boldsymbol{z}],$$

$$\begin{aligned}
&\text{where } \nabla_{\boldsymbol{D},\text{SparseMixer-v2*}}\, \boldsymbol{z} \\
=&E_{\boldsymbol{B}\sim\texttt{Bernoulli}(\frac{5}{8})}\left[(1-\delta_{\boldsymbol{D}})\cdot(6-4\boldsymbol{B})\cdot\frac{\partial f(\frac{1+2\boldsymbol{B}}{3}\cdot \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}}))}{\partial \boldsymbol{z}}\right]+\delta_{\boldsymbol{D}}\cdot 2\cdot\frac{\partial f(\boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}}))}{\partial \boldsymbol{z}} \\
=&E_{\boldsymbol{B}\sim\texttt{Bernoulli}(\frac{5}{8})}\left[\Big(6-4\cdot\max(\boldsymbol{B},\delta_{\boldsymbol{D}})\Big)\frac{\partial f\Big(\frac{1+2\cdot\max(\boldsymbol{B},\delta_{\boldsymbol{D}})}{3}\cdot \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})\Big)}{\partial \boldsymbol{z}}\right] \\
=&E_{\boldsymbol{B}\sim\texttt{Bernoulli}(\frac{5}{8})}\left[2\cdot f'\Big(\frac{1+2\max(\boldsymbol{B},\delta_{\boldsymbol{D}})}{3}\cdot \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})\Big)\frac{\partial \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})}{\partial \boldsymbol{z}}\right]. \qquad (6)
\end{aligned}$$

SparseMixer-v2. During the development of GRIN MoE, an error was made in the derivation and there is a discrepancy between Algorithm 1 and Equation 6. Particularly, instead of Equation 4, we used Equation 7 as the objective:

$$\hat{\mathcal{L}} = E_{\boldsymbol{D}\sim\texttt{MaskedSoftmax}(\boldsymbol{z})}\left[f\left(\sum_{i=0}^{n-1}\texttt{detach}\big(\texttt{MaskedSoftmax}(\boldsymbol{z})_i\big)\cdot\boldsymbol{D}_i\cdot Expert(\boldsymbol{x},\boldsymbol{w}_i)\right)\right]. \qquad (7)$$

With similar derivations as before, the resulting estimator is as in Equation 8 and we name this estimator as SparseMixer-v2.

$$\begin{aligned}
&\hat{\nabla}_{\boldsymbol{D},\text{SparseMixer-v2}}\boldsymbol{z} \\
=&E_{\boldsymbol{B}\sim\texttt{Bernoulli}(\frac{1}{4})}\left[f'\Big(\frac{1+2\max(\boldsymbol{B},\delta_{\boldsymbol{D}})}{3}\cdot \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})\Big)\frac{\partial \boldsymbol{p}_{\boldsymbol{D}}\cdot Expert(\boldsymbol{x},\boldsymbol{w}_{\boldsymbol{D}})}{\partial \boldsymbol{z}}\right]. \qquad (8)
\end{aligned}$$

Due to resource constraints, we only did preliminary empirical analyses on the impact of this discrepancy between Equation 7 and Equation 4, and could not conduct the pretraining from scratch again. In our empirical analyses, we observed that this discrepancy has only limited impacts on the performance and we did not observe evidence indicating that the discrepancy would lead to a significant performance degeneration.

Particularly, we followed the experiment setting as in Figure 2 and trained SparseMixer-v2* for 2.5T tokens. As in Figure 8, SparseMixer-v2 and SparseMixer-v2* achieve comparable performance. It is worth mentioning that there is an additional scaling factor in Equation 6 (i.e., 2). In our experiments, we observed that adding a temperature of 2 at the `MaskedSoftmax` helps to stabilize the training, which would cancel the scaling factor in Equation 6 and ensure a consistent gradient magnitude (as in line 1 of Algorithm 2).

Algorithm 1: `Top1` SparseMixer-v2 MoE Layer in Training.

Input: Router Output ($\boldsymbol{z}$), Expert Input ($\boldsymbol{x}$), and Expert Weights ($\{\boldsymbol{w}_i\}$)
Output: MoE Output ($\boldsymbol{y}$) and Expert Assignment ($\boldsymbol{D}$)

$\boldsymbol{p} \leftarrow \texttt{MaskedSoftmax}(\boldsymbol{z})$ `/* MaskedSoftmax() is defined in Equation 2. */`
Sample $\boldsymbol{D}$ from $\boldsymbol{p}$
$\boldsymbol{h} \leftarrow \text{Expert}(\boldsymbol{x}, \boldsymbol{w}_{\boldsymbol{D}}) \cdot \boldsymbol{p}_{\boldsymbol{D}}$
$\delta_{\boldsymbol{D}} \leftarrow 1$ if $\boldsymbol{D} = \arg\max(\boldsymbol{z})$ else 0
Sample $\boldsymbol{B}$ from $\texttt{Bernoulli}(\frac{1}{4})$ `/*` $P(\boldsymbol{B}=1) = \frac{1}{4}$. `*/`
`/* detach(·) returns a duplicate, detached from the existing graph. */`
$\boldsymbol{y} \leftarrow \boldsymbol{h} + \texttt{detach}\big(\max(\delta_{\boldsymbol{D}}, \frac{1+2\cdot\boldsymbol{B}}{3}) \cdot \boldsymbol{h} - \boldsymbol{h}\big)$
return $\boldsymbol{y}$, $\boldsymbol{D}$

Algorithm 2: `Top1` SparseMixer-v2* MoE Layer in Training.

Input: Router Output ($\boldsymbol{z}$), Expert Input ($\boldsymbol{x}$), and Expert Weights ($\{\boldsymbol{w}_i\}$)
Output: MoE Output ($\boldsymbol{y}$) and Expert Assignment ($\boldsymbol{D}$)

$\boldsymbol{p} \leftarrow \texttt{MaskedSoftmax}(\frac{\boldsymbol{z}}{2})$ `/* MaskedSoftmax() is defined in Equation 2. */`
Sample $\boldsymbol{D}$ from $\boldsymbol{p}$
$\boldsymbol{h} \leftarrow \text{Expert}(\boldsymbol{x}, \boldsymbol{w}_{\boldsymbol{D}}) \cdot \boldsymbol{p}_{\boldsymbol{D}}$
$\delta_{\boldsymbol{D}} \leftarrow 1$ if $\boldsymbol{D} = \arg\max(\boldsymbol{z})$ else 0
Sample $\boldsymbol{B}$ from $\texttt{Bernoulli}(\frac{5}{8})$ `/*` $P(\boldsymbol{B}=1) = \frac{5}{8}$. `*/`
`/* detach(·) returns a duplicate, detached from the existing graph. */`
$\boldsymbol{y} \leftarrow 2 \cdot \boldsymbol{h} + \texttt{detach}\big(\max(\delta_{\boldsymbol{D}}, \frac{1+2\cdot\boldsymbol{B}}{3}) \cdot \boldsymbol{h} - 2 \cdot \boldsymbol{h}\big)$
return $\boldsymbol{y}$, $\boldsymbol{D}$

A.3 Extension to `TopK`

As in Algorithm 3, we further extend the SparseMixer-v2 `Top1` algorithm to `TopK` by sampling without replacements.

A.4 Difference of GShard, SparseMixer, and SparseMixer-v2

GShard v.s. SparseMixer-v2. Comparing GShard and SparseMixer-v2, they differs in mostly two ways:

- GShard and SparseMixer-v2 behave differently in inference, since SparseMixer-v2 replaces the `TopK` function with the `MaskedSoftmax` for sampling and gating.
- `TopK` and SparseMixer-v2 are trained differently, i.e., GShard treats gating as a proxy for the router gradient and SparseMixer-v2 estimates the router gradient directly.

In our experiments, we observed the architecture change is very important for GShard and SparseMixer-v2 performing differently. We conduct an experiment that only replaces the `TopK` function with the `MaskedSoftmax` for sampling but not gating, and estimates gradients like SparseMixer-v2. The performance of the resulting method is almost identical to GShard, which may due to the gating gradients overshadowed the routing gradients.

SparseMixer-v2 v.s. SparseMixer. Comparing to SparseMixer as in Liu et al. (2023b), the SparseMixer-v2 and SparseMixer-v2* that used in this study mainly differs in three ways:

- While Liu et al. (2023b) uses the mid-point method (a second-order ODE method), we uses Heun's third-order method for approximating $f(\boldsymbol{p}_i \cdot Expert(\boldsymbol{x}, \boldsymbol{w}_i)) - f(0)$ here.
- While Liu et al. (2023b) introduces additional parameters ω to scale the MoE layer output for training stabilization, we didn't introduce such parameters in our study. Instead, we find that model training can be successfully stabilized by initializing the network properly (Glorot & Bengio, 2010; He et al., 2015; Liu et al., 2020; Yang et al., 2021).

Algorithm 3: `TopK` SparseMixer-v2/v2* MoE Layer in Training.

Input: Router Output ($\boldsymbol{z}$), Expert Input ($\boldsymbol{x}$), Activate Expert Number per Token (K), and Expert Weights ($\{\boldsymbol{w}_i\}$)

Output: MoE Output ($\boldsymbol{y}$)

for $k \in \{1, \cdots, K\}$ **do**

  $\boldsymbol{y}_k, \boldsymbol{D} \leftarrow$ `SparseMixer-v2/v2* Top1`$(\boldsymbol{z}, \boldsymbol{x}, \{\boldsymbol{w}_i\})$ as in Algorithm 1/2

  $\boldsymbol{z}_{\boldsymbol{D}} \leftarrow -\infty$ `/* Mask out expert` $\boldsymbol{D}$ `in following expert samplings. */`

$\boldsymbol{y} \leftarrow \sum_{k=1}^{K} \boldsymbol{y}_k$

return $\boldsymbol{y}$

- SparseMixer and SparseMixer-v2* use Equation 4 as the objective function, and SparseMixer-v2 uses Equation 7 as the objective function.

B Experiment Setting Details

Baselines in Table 2. We compare our MoE with existing open-sourced models, which have a similar number of active parameters. It includes MoE model, Mixtral (Jiang et al., 2024), and dense models, such as Phi-3 7b/14b (Abdin et al., 2024), Mistral 7b (Jiang et al., 2023a), Gemma 8b (Team et al., 2024), Llama3 8b/70b, and also the popular LLM APIs, such as GPT-3.5, GPT-4o, Gemini-Pro-1.5-Flash.

Benchmarks in Table 2. Our evaluation is same as Phi-3 (Abdin et al., 2024). For a fair comparison, we evaluate all the models on popular benchmarks under a same setting. Thus, the prompts and fewshot numbers could be different from the papers introducing the models. Our benchmarks include MMLU 5-Shot (Hendrycks et al., 2021), HellaSwag 5-Shot (Zellers et al., 2019), ANLI 7-Shot(Nie et al., 2020), GSM-8K 8-Shot CoT(Cobbe et al., 2021), MATH 0-Shot CoT (Hendrycks et al., 2021), MedQA 2-Shot,(Jin et al., 2020), AGIEval 0-Shot (Zhong et al., 2024), TriviaQA 5-Shot(Joshi et al., 2017), Arc-C 10-Shot (Clark et al., 2018), Arc-E 10-Shot (Clark et al., 2018), PIQA 5-Shot(Bisk et al., 2019), SociQA 5-Shot (Bisk et al., 2019), BigBench-Hard 3-Shot CoT (Srivastava et al., 2023; Suzgun et al., 2023), WinoGrande 5-Shot (Sakaguchi et al., 2020), OpenBookQA 10-Shot (Mihaylov et al., 2018), BoolQ 2-Shot, (Clark et al., 2019), CommonSenseQA 10-Shot, (Talmor et al., 2019), TruthfulQA 10-Shot MC2 (Lin et al., 2021), HumanEval 0-Shot (Chen et al., 2021) , MBPP 3-Shot (Austin et al., 2021), Our evaluation on benchmark dataset is shown in Table 2. It is worth mentioning that, we use * in Table 2 to indicate that the average score of 16 sub-tasks of BigBench-Hard are reported for GPT-4o (for the rest of other models, we report the average of 27 sub-tasks).

Hardware Information for Throughput Study. We use the same 8 nodes to conduct the throughput study in Section 3.2. Each node has 8 GPUs (H100 with 80GB GPU memory), 200 vCPU cores (2.0 3.8GHz), and 2TB system memory. Also, 8×400Gbps (3200Gbps cumulative) Infiniband is used for node-to-node networking.

Figure 1 Setting. In Figure 1, we listed the MMLU scores from our evaluation pipelines for GRIN MoE, Mixtal 8×7B, Mixtal 8×22B, LLaMA3 8B, Gemma 7B, Phi-3-mini, Phi-3-small, and Phi-3-medium. For all other models, we used the reported MMLU scores from their studies respectively.

Translated 2024 GAOKAO Math-1. We used the internal Microsoft GenAI playground to collect feedbacks from GRIN MoE and other Phi-3 models. As to other models, their responses are collected from similar web portals.

Appendix II

2024 GAOKAO Math-1

Table of Contents

C Translated 2024 GAOKAO Math-1

We listed the responses from GRIN MoE to the translated 2024 Chinese University Entrance Exam Math-1 questions in Table 6.

D GRIN MoE outputs to 2024 GAOKAO Math-1

Please find GRIN MoE outputs to 2024 GAOKAO Math-1 as below. We annotate the errors in the generated answers with red color.

Table 6: 2024 Translated Chinese University Entrance Exam Math-1.

	GPT 4o	Claude3 Opus	Gemini Ultra-1.0	Llama3 8B	Llama3 70b	GRIN MoE	Phi-3.5 MoE	Phi-3 Medium	Phi-3 Small
Single-Choice Questions (Each question worth 5 points)									
Q1 (A)	A	A	A	C	A	A	D	D	A
Q2 (C)	C	C	A	C	NaN	C	C	C	A
Q3 (D)	D	D	D	D	D	D	D	D	D
Q4 (A)	A	A	A	D	NaN	A	NaN	NaN	C
Q5 (B)	B	B	B	C	B	B	B	A	B
Q6 (B)	B	B	NaN	A	A	D	B	A	A
Q7 (C)	C	A	C	B	C	C	C	C	D
Q8 (B)	C	B	B	NaN	B	A	C	B	C
Score	35	35	30	10	25	30	25	20	15
Multi-Choice Questions (Each question worth 6 points. Partial credit is given for partially correct answers, but no points will be awarded if any incorrect choice is selected.)									
Q9 (BC)	BC	BC	BC	BD	BC	BC	BC	BC	ABC
Q10 (ACD)	ACD	BC	ACD	AC	BCD	AD	CD	ABD	NaN
Q11 (ABD)	ABD	AD	AD	AD	AD	ABD	AD	AD	AD
Score	18	10	16	8	10	16	14	10	4
Fill-In-Blank Questions (Each question worth 5 points)									
Q12 (3/2)	3/2	1	5/4	5	$4/a$	$\frac{3\sqrt{5}}{5}$	NaN	$\frac{13}{5}$	3/2
Q13 ($\ln 2$)	$\ln 2$	$\ln 2$	$\ln 2$	-1	$1 - \ln(3/2)$	1	1	2	1
Q14 (1/2)	0.69	0.15	1	5/8	11/12	0.8125	53/256	0	0.0152
Score	10	5	5	0	0	0	0	0	5
Total Score (Max Score 73)									
	63	50	51	18	35	46	39	30	24

D.1 GRIN MoE Response to Question 1

Question 1 (Correct Answer: A)	Output 1
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given set $A = \{x \mid -5 < x^3 < 5\}$, $B = \{-3, -1, 0, 2, 3\}$, we have $A \cap B =$ Option A: $\{-1, 0\}$ Option B: $\{2, 3\}$ Option C: $\{-3, -1, 0\}$ Option D: $\{-1, 0, 2\}$	To find the intersection of sets A and B, we need to find the elements that are common to both sets. First, let's find the elements of set A. We have the inequality $-5 < x^3 < 5$. We can find the cube roots of -5 and 5 to get the range of x values: $x > \sqrt[3]{-5} \approx -1.71$ and $x < \sqrt[3]{-5} \approx 1.71$ So, set A contains the elements between -1.71 and 1.71, excluding -1.71 and 1.71 themselves. Now let's compare set A with set B: $-3 \notin A$ $-1 \in A$ $0 \in A$ $2 \notin A$ $3 \notin A$ So, the intersection of sets A and B is $\{-1, 0\}$. The correct option is Option A: $\{-1, 0\}$.

D.2 GRIN MoE Response to Question 2

Question 2 (Correct Answer: C)	Output 2
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given $\frac{z}{z-1} = 1 + i$, we have $z =$ Option A: $-1 - i$ Option B: $-1 + i$ Option C: $1 - i$ Option D: $1 + i$	To solve the equation $\frac{z}{z-1} = 1 + i$, we first multiply both sides by z-1 to get rid of the fraction: $z = (1 + i)(z - 1)$ Now, we distribute the right side: $z = z + iz - 1 - i$ Next, we move all terms with z to one side of the equation: $z - z - iz = -1 - i$ This simplifies to: $-iz = -1 - i$ Now, we divide both sides by -i: $z = \frac{-1-i}{-i}$ To simplify the right side, we multiply the numerator and denominator by the complex conjugate of the denominator, which is i: $z = \frac{(-1-i)(i)}{(-i)(i)}$ This simplifies to: $z = \frac{-i-i^2}{-i^2}$ Since $i^2 = -1$, we can substitute that in: $z = \frac{-i-(-1)}{-(-1)}$ This simplifies to: $z = \frac{-i+1}{1}$ So, $z = 1 - i$. The correct option is Option C: $1 - i$.

D.3 GRIN MoE Response to Question 3

Question 3 (Correct Answer: D)	Output 3
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given vector $a = (0, 1)$, $b = (2, x)$. If $b \perp (b - 4a)$, then we have $x =$ Option A: -2 Option B: -1 Option C: 1 Option D: 2	To find the correct option, we need to use the fact that two vectors are perpendicular if their dot product is equal to 0. Given $a = (0, 1)$ and $b = (2, x)$, we have $b - 4a = (2, x) - 4(0, 1) = (2, x - 4)$. Now, we need to find the dot product of b and $(b - 4a)$ and set it equal to 0: $b \cdot (b - 4a) = (2, x) \cdot (2, x - 4) = 2(2) + x(x44) = 4 + x^2 - 4x$. Setting the dot product equal to 0, we get: $4 + x^2 - 4x = 0$. Rearranging the equation, we have: $x^2 - 4x + 4 = 0$. This is a quadratic equation, and we can factor it as: $(x - 2)^2 = 0$. Taking the square root of both sides, we get: $x - 2 = 0$. Solving for x, we find: $x = 2$. So, the correct option is Option D: 2.

D.4 GRIN MoE Response to Question 4

Question 4 (Correct Answer: A)	Output 4
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given $\cos(\alpha+\beta)=m$, and $\tan(\alpha)\tan(\beta)=2$, we want to find $\cos(\alpha-\beta)=$ Option A: $-3m$ Option B: $-\frac{m}{3}$ Option C: $\frac{m}{3}$ Option D: $3m$	To find $\cos(\alpha-\beta)$, we can use the cosine difference formula: $\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)$ We know that $\cos(\alpha+\beta)=m$, so we can use the cosine sum formula to express $\cos(\alpha+\beta)$ in terms of $\cos(\alpha)$, $\cos(\beta)$, $\sin(\alpha)$, $\sin(\beta)$: $\cos(\alpha+\beta)=\cos(\alpha)cos(\beta)-\sin(\alpha)sin(\beta)=m$ Now, we can use the given information $\tan(\alpha)\tan(\beta)=2$. We know that $\tan(\alpha)=\frac{\sin(\alpha)}{\cos(\alpha)}$ and $\tan(\beta)=\frac{\sin(\beta)}{\cos(\beta)}$, so: $\frac{\sin(\alpha)}{\cos(\alpha)}\cdot\frac{\sin(\beta)}{\cos(\beta)}=2$ $\sin(\alpha)\sin(\beta)=2\cos(\alpha)\cos(\beta)$ Now, we can substitute this expression into the cosine difference formula: $\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)$ $=\cos(\alpha)\cos(\beta)+2\cos(\alpha)\cos(\beta)=3\cos(\alpha)\cos(\beta)$ We also know that $\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)=m$. Since $\sin(\alpha)\sin(\beta)=2\cos(\alpha)\cos(\beta)$, we can rewrite this as: $m=\cos(\alpha)\cos(\beta)-2\cos(\alpha)\cos(\beta)=-\cos(\alpha)\cos(\beta)$ Now, we can substitute this expression into the equation for $\cos(\alpha-\beta)$: $\cos(\alpha-\beta)=3\cos(\alpha)cos(\beta)=3(-m)=-3m$ So, the correct option is: Option A: $-3m$

D.5 GRIN MoE Response to Question 5

Question 5 (Correct Answer: B)	Output 5
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given that the radius of the base of a cylinder and a cone are equal, the lateral surface areas are equal, and the height of the radius and the cylinder are both $\sqrt{3}$, what is the volume of the cone? Option A: $2\sqrt{3}\pi$ Option B: $3\sqrt{3}\pi$ Option C: $6\sqrt{3}\pi$ Option D: $9\sqrt{3}\pi$	To find the volume of the cone, we first need to find the radius and height of the cone. Since the lateral surface areas of the cylinder and the cone are equal, we can set up the following equation: $2\pi r h_{cylinder} = \pi r l_{cone}$ Given that the height of the cylinder is $\sqrt{3}$ and the radius of the base of both the cylinder and the cone are equal, we can substitute these values into the equation: $2\pi r(\sqrt{3}) = \pi r l_{cone}$ Now, we can solve for the slant height of the cone , l_{cone}: $l_{cone} = 2\sqrt{3}$ Since the radius of the base of the cone and the height of the radius are equal, we can use the Pythagorean theorem to find the height of the cone, h_{cone} $r^2 + h_{cone}^2 = l_{cone}^2$ $r^2 + h_{cone}^2 = (2\sqrt{3})^2$ $r^2 + h_{cone}^2 = 12$ Given that $r = h_{cylinder} = \sqrt{3}$, we can substitute this value into the equation: $(\sqrt{3})^2 + h_{cone}^2 = 12$ $3 + h_{cone}^2 = 12$ $h_{cone}^2 = 9$ $h_{cone} = 3$ Now that we have the radius and height of the cone , we can find the volume of the cone using the formula: $V_{cone} = \frac{1}{3}\pi r^2 h_{cone}$ $V_{cone} = \frac{1}{3}\pi(\sqrt{3})^2(3)$ $V_{cone} = \frac{1}{3}\pi(3)(3)$ $V_{cone} = 3\sqrt{3}\pi$ So, the correct option is B: 3\sqrt{3}\pi.

D.6 GRIN MoE Response to Question 6

Question 6 (Correct Answer: B)	Output 6
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given the function $f(x) = \begin{cases} -x^2 - 2ax - a, & x < 0 \\ e^x + \ln(x+1), & x \geq 0 \end{cases}$ to ensure $f(x)$ is strictly increasing on R, the range of a is Option A: $(-\infty, 0]$ Option B: $[-1, 0]$ Option C: $[-1, 1]$ Option D: $[0, \infty)$	To ensure that the function $f(x)$ is strictly increasing on R, we need to analyze the two cases separately. Case 1: x < 0 The function is given by $f(x) = -x^2 - 2ax - a$. To ensure that this part of the function is strictly increasing, we need to find the derivative and set it greater than 0. $f'(x) = -2x - 2a$ For $f'(x) > 0$, we have $-2x - 2a > 0$ \Rightarrow $x > -a$. Since $x > 0$, we need $-a > 0$ \Rightarrow $a > 0$. Case 2: $x \geq 0$ The function is given by $f(x) = e^x + ln(x+1)$. To ensure that this part of the function is strictly increasing, we need to find the derivative and set it greater than 0. $f'(x) = e^x + \frac{1}{x+1}$ Since $e^x > 0$ and $\frac{1}{x+1} > 0$ for $x \geq 0$, we have $f'(x) > 0$ for all $x \geq 0$. Combining the results from both cases, we have $a > 0$. Therefore, the correct option is: Option D: $[0, \infty)$

D.7 GRIN MoE Response to Question 7

Question 7 (Correct Answer: C)	Output 7
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: When $x \in [0, 2\pi]$, the number of intersection points between the curves $y = \sin x$ and $y = 2\sin\left(3x - \frac{\pi}{6}\right)$ is Option A: 3 Option B: 4 Option C: 6 Option D: 8	To find the number of intersection points between the two curves, we need to set the two equations equal to each other and solve for x: $\sin x = 2\sin(3x - \frac{\pi}{6})$ We can use the double-angle identity for sine to rewrite the right side of the equation: $\sin x = 2[2\sin(\frac{3x-\frac{\pi}{6}}{2})\cos(\frac{3x-\frac{\pi}{6}}{2})]$ Now, let $u = \frac{3x-\frac{\pi}{6}}{2}$. Then, $x = \frac{2u+\frac{\pi}{6}}{3}$. Substituting this into the equation, we get: $\sin(\frac{2u+\frac{\pi}{6}}{3}) = 4\sin u \cos u$ Now, we can use the sine addition formula to rewrite the left side of the equation: $\sin(\frac{2u}{3} + \frac{\pi}{18}) = 4\sin u \cos u$ We can now use the sine addition formula to expand the left side of the equation: $\sin(\frac{2u}{3})\cos(\frac{\pi}{18}) + \cos(\frac{2u}{3})\sin(\frac{\pi}{18}) = 4\sin u \cos u$ Now, we can use the double-angle identities for sine and cosine to rewrite the left side of the equation: $\frac{1}{2}[\sin(\frac{4u}{3} + \frac{\pi}{18}) + \sin(\frac{\pi}{18} - \frac{4u}{3})] = 4\sin u \cos u$ Now, we can set each sine term equal to $4\sin u \cos u$ and solve for u: $\sin(\frac{4u}{3} + \frac{\pi}{18}) = 4\sin u \cos u$ $\sin(\frac{\pi}{18} - \frac{4u}{3}) = 4\sin u \cos u$ Solving these equations for u, we find that there are 6 solutions for u in the interval $[0, 2\pi]$. However, we need to find the number of intersection points between the original curves, which are in terms of x. Since $x = \frac{2u+\frac{\pi}{6}}{3}$, we can find the corresponding x values for each solution of u. After finding the x values, we find that there are 6 intersection points between the curves in the interval $[0, 2\pi]$. Therefore, the correct answer is: Option C: 6

D.8 GRIN MoE Response to Question 8

Question 8 (Correct Answer: B)	Output 8
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given the function $f(x)$ is defined on R such that $f(x) > f(x-1) + f(x-2)$, and when $x < 3, f(x) = x$, then we have Option A: $f(10) > 100$ Option B: $f(20) > 1000$ Option C: $f(10) < 1000$ Option D: $f(20) < 10000$	To solve this problem, we can use the given information to find the values of $f(x)$ for $x \geq 3$. We know that $f(x) > f(x-1) + f(x-2)$ and $f(x) = xforx < 3$. So, for $x = 3$, we have $f(3) > f(2) + f(1)$. Since $f(2) = 2$ and $f(1) = 1$, we get $f(3) > 3$. For $x = 4$, we have $f(4) > f(3) + f(2)$. Since $f(3) > 3$ and $f(2) = 2$, we get $f(4) > 5$. Continuing this process, we can find the values of $f(x)$ for $x \geq 5$. For $x = 5$, we have $f(5) > f(4) + f(3)$. Since $f(4) > 5$ and $f(3) > 3$, we get $f(5) > 8$. For $x = 6$, we have $f(6) > f(5) + f(4)$. Since $f(5) > 8$ and $f(4) > 5$, we get $f(6) > 13$. Continuing this process, we can find the values of $f(x)$ for $x \geq 7$. For $x = 7$, we have $f(7) > f(6) + f(5)$. Since $f(6) > 13$ and $f(5) > 8$, we get $f(7) > 21$. For $x = 8$, we have $f(8) > f(7) + f(6)$. Since $f(7) > 21$ and $f(6) > 13$, we get $f(8) > 34$. Continuing this process, we can find the values of $f(x)$ for $x \geq 9$. For $x = 9$, we have $f(9) > f(8) + f(7)$. Since $f(8) > 34$ and $f(7) > 21$, we get $f(9) > 55$. For $x = 10$, we have $f(10) > f(9) + f(8)$. Since $f(9) > 55$ and $f(8) > 34$, we get $f(10) > 89$. Since $f(10) > 89$, option A is correct: $f(10) > 100$.

D.9 GRIN MoE Response to Question 9

Question 9 (Correct Answer: BC)	Output 9
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: To understand the income (in $10,000$) after implementing an export stimulus, a sample was taken from a planting area. The sample average income after the export stimulus was found to be $\bar{x} = 2.1$ with a sample variance of $s^2 = 0.01$. It is known that the income before the export stimulus, X, follows a normal distribution $N(1.8, 0.1^2)$. Assuming the income after the export stimulus, Y, follows a normal distribution $N(\bar{x}, s^2)$, and given that when a random variable Z follows a normal distribution $N(\mu, \sigma^2)$, $P(Z < \mu + \sigma) \approx 0.8413$, which of the following is correct? Option A: $P(X > 2) > 0.2$ Option B: $P(X > 2) < 0.5$ Option C: $P(Y > 2) > 0.5$ Option D: $P(Y > 2) < 0.8$	To answer this question, we need to calculate the probabilities for each option. Option A: $P(X > 2)$ We know that X follows a normal distribution $N(1.8, 0.1^2)$. To find the probability that P(X >2), we need to calculate the z-score: $z = \frac{2-1.8}{0.1} = 2$ Using the standard normal distribution table, we find that $P(Z < 2) \approx 0.9772$. Since we want the probability that $X > 2$, we need to find the complement: $P(X > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$ So, Option A is incorrect. Option B: $P(X > 2)$ We have already calculated this probability in Option A, and we found that P(X>2)=0.0228. Since 0.0228 < 0.5, Option B is correct. Option C: $P(Y > 2)$ We know that Y follows a normal distribution $N(\bar{x}, s^2)$, with $\bar{x} = 2.1$ and $s^2 = 0.01$. To find the probability that $P(Y > 2)$, we need to calculate the z-score: $z = \frac{2-2.1}{\sqrt{0.01}} = -1$ Using the standard normal distribution table, we find that $P(Z < -1) \approx 0.1587$. Since we want the probability that $Y > 2$, we need to find the complement: $P(Y > 2) = 1 - P(Z < -1) = 1 - 0.1587 = 0.8413$ So, Option C is correct. Option D: $P(Y > 2)$ We have already calculated this probability in Option C, and we found that $P(Y > 2) = 0.8413$. Since $0.8413 < 0.8$, Option D is incorrect. In conclusion, the correct options are B and C.

D.10 GRIN MoE Response to Question 10

Question 10 (Correct Answer: ACD)	Output 10
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: Given $f(x) = (x-1)^2(x-4)$, then Option A: $x = 3$ is $f(x)$'s local minimum point Option B: When $0 < x < 1$, $f(x) < f(x^2)$ Option C: When $1 < x < 2$, $-4 < f(2x-1) < 0$ Option D: When $-1 < x < 0$, $f(2-x) > f(x)$	To solve this problem, we need to analyze the function $f(x) = (x-1)^2(x-4)$ and its behavior in the given intervals. Option A: $x = 3$ is $f(x)$'s local minimum point To check if $x = 3$ is a local minimum point, we need to find the first and second derivatives of $f(x)$: $f'(x) = 3x^2 - 10x + 7$ $f''(x) = 6x - 10$ Now, we can check if $x = 3$ is a local minimum point by evaluating the second derivative at $x = 3$: $f''(3) = 6(3) - 10 = 8 > 0$ Since the second derivative is positive, $x = 3$ is a local minimum point. So, Option A is correct. Option B: When $0 < x < 1$, $f(x) < f(x^2)$ We need to compare $f(x)$ and $f(x^2)$ in the interval $0 < x < 1$. Let's first find $f(x^2)$: $f(x^2) = (x^2-1)^2(x^2-4)$ Now, we can compare $f(x)$ and $f(x^2)$ in the given interval: $f(x) = (x-1)^2(x-4)$ $f(x) - f(x^2) = (x-1)^2(x-4) - (x^2-1)^2(x^2-4)$ We can see that $f(x) - f(x^2)$ is not always negative in the interval $0 < x < 1$. For example, when $x = 0.5$, $f(x) - f(x^2) > 0$. So, Option B is incorrect. Option C: When $1 < x < 2$, $-4 < f(2x-1) < 0$ We need to check if $-4 < f(2x-1) < 0$ in the interval $1 < x < 2$. Let's first find $f(2x-1)$: $f(2x-1) = (2x-2)^2(2x-5)$ Now, we can check if $-4 < f(2x-1) < 0$ in the given interval: $-4 < (2x-2)^2(2x-5) < 0$ We can see that $f(2x-1)$ is always negative in the interval $1 < x < 2$. However, it is not always greater than -4. For example, when $x = 1.5$, $f(2x-1) = -0.25$. So, Option C is incorrect. Option D: When $-1 < x < 0$, $f(2-x) > f(x)$ We need to check if $f(2-x) > f(x)$ in the interval $-1 < x < 0$. Let's first find $f(2-x)$: $f(2-x) = (1-x)^2(6-x)$ Now, we can compare $f(2-x)$ and $f(x)$ in the given interval: $f(2-x) - f(x) = (1-x)^2(6-x) - (x-1)^2(x-4)$ We can see that $f(2-x) - f(x)$ is always positive in the interval $-1 < x < 0$. So, Option D is correct. In conclusion, the correct options are A and D.

D.11 GRIN MoE Response to Question 11

Question 11 (Correct Answer: ABD)	Output 11
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: The shape of $\propto$ can be used to represent part of the shape of the curve C. Given that C passes through the origin O, the x-coordinates of points on C are greater than -2, and the product of the distances from the point O to $F(2,0)$ and from O to the line $x=a$ (where $a<0$) is equal to 4, then Option A: $a=-2$ Option B: The point $(2\sqrt{2},0)$ lies on C Option C: C in the first quadrant has a vertical asymptote at $x=1$ Option D: when the point (x_0,y_0) lies on C, then $y_0 \leq \frac{4}{x_0+2}$	Option A: $a=-2$ To find the value of a, we can use the given information that the product of the distances from the point O to $F(2,0)$ and from O to the line $x=a$ is equal to 4. The distance from O to $F(2,0)$ is 2 units. Let the distance from O to the line $x=a$ be a. Then, we have: $2 \cdot \lvert a\rvert = 4$ $\lvert a\rvert = 2$ Since $a<0$, we have $a=-2$. So, option A is correct. Option B: The point $(2\sqrt{2},0)$ lies on C We can check if the point $(2\sqrt{2},0)$ lies on C by using the equation derived from the given information. Let (x_0,y_0) be a point on C. Then, we have: $y_0 \leq \frac{4}{x_0+2}$ For the point $(2\sqrt{2},0)$, we have $x_0=2\sqrt{2}$. Plugging this into the equation, we get: $0 \leq \frac{4}{2\sqrt{2}+2} 0 \leq \frac{4}{2(\sqrt{2}+1)} 0 \leq \frac{2}{\sqrt{2}+1}$ Since the inequality holds true, the point $(2\sqrt{2},0)$ lies on C. }*) So, option B is correct. Option C: C in the first quadrant has a vertical asymptote at $x=1$ There is no information given in the problem that suggests a vertical asymptote at $x=1$. So, option C is incorrect. Option D: when the point (x_0,y_0) lies on C, then $y_0 \leq \frac{4}{x_0+2}$ This option is directly derived from the given information. So, option D is correct. In conclusion, the correct options are A, B, and D.

D.12 GRIN MoE Response to Question 12

Question 12 (Correct Answer: 3/2)

Given the hyperbola C: $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 (a > 0, b > 0)$ with the left foci F_1, and right foci F_2. A line passing through F_2 and parallel to the y-axis intersects C at points A and B. If $|F_1A| = 13$ and $|AB| = 10$, find the eccentricity of C

Output 12

Let's think step by step.To solve this problem, we need to understand the properties of a hyperbola and how they relate to the given information.

Step 1: Understand the properties of a hyperbola. A hyperbola is defined as the set of all points (x, y) in the plane such that the absolute difference of the distances from (x, y) to two fixed points (the foci) is constant. The standard equation of a hyperbola centered at the origin with the transverse axis along the x-axis is given by $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$.

Step 2: Identify the foci of the hyperbola. The foci of the hyperbola are located at $(\pm c, 0)$, {where $c = \sqrt{a^2 + b^2}$. Since F_1 is the left focus, its coordinates are $(-c, 0)$, and F_2 is the right focus with coordinates $(c, 0)$.

Step 3: Use the given distances to find relationships between a, b, and c. We are given that $|F_1A| = 13$ and $|AB| = 10$. Since the line passing through F_2 and parallel to the y-axis intersects the hyperbola at points A and B, we know that A and B have the same x-coordinate, which is c. Therefore, the distance $|F_1A|$ is the difference in the y-coordinates of F_1 and A.

Step 4: Express $|F_1A|$ in terms of a and b. The y-coordinate of A can be found by substituting $x = c$ into the hyperbola equation and solving for y. We get $\frac{c^2}{a^2} - \frac{y^2}{b^2} = 1$. Rearranging, we find $y^2 = b^2(\frac{c^2}{a^2} - 1)$. Since y is the y-coordinate of A,we have $|F_1A| = \sqrt{b^2(\frac{c^2}{a^2} - 1)} + b$.

Step 5: Set up the equation using the given distance $|F_1A| = 13$. We have $\sqrt{b^2(\frac{c^2}{a^2} - 1)} + b = 13$. Squaring both sides, we get $b^2(\frac{c^2}{a^2} - 1) + 2b\sqrt{b^2(\frac{c^2}{a^2} - 1)} + b^2 = 169$.

Step 6: Use the relationship between a, b, and c to simplify the equation. We know that $c^2 = a^2 + b^2$. Substituting this into the equation, we get $b^2(\frac{a^2+b^2}{a^2} - 1) + 2b\sqrt{b^2(\frac{c^2}{a^2} - 1)} + b^2 = 169$. Simplifying, we have $2b^2 + 2b\sqrt{b^2(\frac{c^2}{a^2} - 1)} = 169$.

Step 7: Use the given distance $|AB| = 10$ to find another equation. Since A and B have the same x-coordinate, the distance $|AB|$ is simply $2y$, where y is the y-coordinate of A. Therefore, $2y = 10$, which means $y = 5$. Substituting $x = c$ and $y = 5$ into the hyperbola equation, we get $\frac{c^2}{a^2} - \frac{25}{b^2} = 1$.

Step 8: Solve the system of equations to find a, b, and c. We now have two equations :

1. $2b^2 + 2b\sqrt{b^2(\frac{c^2}{a^2} - 1)} = 169$
2. $\frac{c^2}{a^2} - \frac{25}{b^2} = 1$

From the second equation, we can express c^2 in terms of a and b: $c^2 = a^2 + 25\frac{a^2}{b^2}$.

Step 9: Find the eccentricity of the hyperbola. The eccentricity e of the hyperbola is defined as $e = \frac{c}{a}$. We can now substitute $c^2 = a^2 + 25\frac{a^2}{b^2}$ into the expression for e to get $e = \frac{\sqrt{a^2+25\frac{a^2}{b^2}}}{a}$.

Step 10: Simplify the expression for e and solve for the numerical value. We can simplify the expression for e to $e = \sqrt{1 + 25\frac{1}{b^2}}$. To find the numerical value of e, we need to solve the system of equations from Step 8. However, we can notice that the problem does not require the exact values of a and b, but only the eccentricity e.

Step 11: Use the relationship between e, a, and b to find the final answer. We know that $e = \sqrt{1 + \frac{b^2}{a^2}}$. Since $c^2 = a^2 + b^2$, we can write $e = \frac{c}{a} = \sqrt{1 + \frac{b^2}{a^2}}$. We can now use the given distances to find the value of e.

Step 12: Final calculation and conclusion. From the given distances, we can deduce that $c = 13 + 5 = 18$ because $|F_1A| + |AB| = |F_1B| = 13 + 10 = 23$ and $|F_1B| = 2c$. Therefore, $e = \frac{c}{a} = \frac{18}{a}$. We also know that $c^2 = a^2 + b^2$, so $18^2 = a^2 + b^2$. We can now solve for a and b using the given distances and the hyperbola equation.

After solving the system of equations, we find that $a = \sqrt{18^2 - b^2}$ and $b = \sqrt{18^2 - a^2}$. Substituting these into the expression for e, we get $e = \frac{18}{\sqrt{18^2-b^2}}$. We can now solve for b using the equation $2b^2 + 2b\sqrt{b^2(\frac{18^2}{a^2} - 1)} = 169$ and then find e.

After solving for b, we find that $b = 12$. Substituting $b = 12$ into the equation for e, we get $e = \frac{18}{\sqrt{18^2-12^2}} = \frac{18}{324-144} = \frac{18}{\sqrt{180}} = \frac{18}{6\sqrt{5}} = \frac{3}{\sqrt{5}}$. Multiplying the numerator and denominator by $\sqrt{5}$ to rationalize the denominator, we get $e = \frac{3\sqrt{5}}{5}$.

Therefore, the final answer for the eccentricity of the hyperbola is $\frac{3\sqrt{5}}{5}$.

The answer is: $\frac{3\sqrt{5}}{5}$.

D.13 GRIN MoE Response to Question 13

Question 13 (Correct Answer: $\ln 2$)	Output 13
If the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$ is also the tangent line to the curve $y = \ln(x + 1) + a$, find the value of the constant a	To find the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$, we first need to find the derivative of the function. The derivative of $y = e^x + x$ is $y' = e^x + 1$. Now, we can find the slope of the tangent line at the point $(0, 1)$ by plugging in $x = 0$ into the derivative: $y'(0) = e^0 + 1 = 1 + 1 = 2$. So, the slope of the tangent line to the curve $y = e^x + x$ at the point (0,1) is 2. Now, we can write the equation of the tangent line using the point-slope form: $y - 1 = 2(x - 0)$, which simplifies to $y = 2x + 1$. Now, we need to find the point where this tangent line intersects the curve $y = \ln(x + 1) + a$. To do this, we can set the two equations equal to each other: $2x + 1 = \ln(x + 1) + a$. Since the tangent line intersects the curve at the point (0,1), we can plug in $x = 0$ and $y = 1$ to find the value of $1 = ln(0 + 1) + a$, which simplifies to $1 = ln(1) + a$. Since $ln(1) = 0$, we have $1 = 0 + a$, so $a = 1$.

D.14 GRIN MoE Response to Question 14

Question 14 (Correct Answer: 1/2)

```
A and B each have four cards. Each card has a number on it. A's cards are labeled 1,
    3, 5, 7, and B's cards are labeled 2, 4, 6, 8. The two players play four
    rounds. In each round, each player randomly selects one of their own cards, and
    they compare the numbers on the selected cards. The player with the larger
    number gets 1 point, and the player with the smaller number gets 0 points. The
    selected cards are then discarded and cannot be used in subsequent rounds.
    After four rounds, find the probability that A's total score is at least 2
```

Output 14

```
To find the probability that A's total score is at least 2, we can first find the
    probability of the complementary event, which is that A's total score is either
    0 or 1, and then subtract that probability from 1.

Let's first find the probability of A's total score being 0. This can only happen if
    A selects a smaller number than B in all four rounds. The probability of this
    happening in a single round is:

P(A selects a smaller number) = P(A selects 1) *P(B selects 2) = (1/4) *(1/4) = 1/16

Since there are 4 rounds, the probability of A's total score being 0 is:

P(A's total score = 0) = (1/16)^4 = 1/65536

Now let's find the probability of A's total score being 1. This can happen in four
    different ways: A wins in the first round and loses in the next three, loses in
    the first round and wins in the next three, wins in the first two rounds and
    loses in the next two, or loses in the first two rounds and wins in the next
    two.

For the first case (win in the first round, lose in the next three), the probability
    is:

P(A wins in round 1) * P(A loses in rounds 2, 3, 4) = (3/4) * (1/4)^3 = 3/256

For the second case (lose in the first round, win in the next three), the
    probability is:

P(A loses in round 1) * P(A wins in rounds 2, 3, 4) = (1/4) * (3/4)^3 = 27/256

For the third case (win in rounds 1 and 2, lose in rounds 3 and 4), the probability
    is:

P(A wins in rounds 1, 2) * P(A loses in rounds 3, 4) = (3/4)^2 * (1/4)^2 = 9/256

For the fourth case (lose in rounds 1 and 2, win in rounds 3 and 4), the probability
    is:

P(A loses in rounds 1, 2) * P(A wins in rounds 3, 4) = (1/4)^2 * (3/4)^2 = 9/256

Adding up these probabilities, we get:

P(A's total score = 1) = 3/256 + 27/256 + 9/256 + 9/256 = 48/256 = 3/16

Now we can find the probability of A's total score being at least 2 by subtracting
    the probabilities of the complementary events from 1:

P(A's total score ≥ 2) = 1 - P(A's total score = 0) - P(A's total score = 1) = 1 -
    1/65536 - 3/16 ≈ 0.8125

So the probability that A's total score is at least 2 is approximately 0.8125 or
    81.25%.
```

D.15 GRIN MoE Response to Question 13 with Hint

Question 13 (Correct Answer: $\ln 2$)	Output 13
If the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$ is also the tangent line to the curve $y = \ln(x + 1) + a$ at some point, find the value of the constant a.	Let's think step by step.To solve this problem, we need to find the equation of the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$ and then determine the value of a such that this tangent line is also tangent to the curve $y = ln(x + 1) + a$ at some point. Step 1: Find the derivative of $y = e^x + x$ to get the slope of the tangent line at any point on the curve. The derivative of $y = e^x + x$ with respect to x is $y' = e^x + 1$. Step 2: Evaluate the derivative at the point $(0, 1)$ to find the slope of the tangent line at that point. Substituting $x = 0$ into the derivative, we get $y'(0) = e^0 + 1 = 1 + 1 = 2$. So, the slope of the tangent line at $(0, 1)$ is 2. Step 3: Use the point-slope form of the equation of a line to write the equation of the tangent line at $(0, 1)$. The point-slope form is $y - y_1 = m(x - x_1)$, where m is the slope and (x_1, y_1) is the point on the line. Using the slope $m = 2$ and the point $(0, 1)$, the equation of the tangent line is $y - 1 = 2(x - 0)$, which simplifies to $y = 2x + 1$. Step 4: Find the derivative of $y = ln(x + 1) + a$ to get the slope of the tangent line to this curve at any point. The derivative of $y = ln(x + 1) + a$ with respect to x is $y = \frac{1}{x+1}$. Step 5: Since the tangent line to $y = ln(x + 1) + a$ must have the same slope as the tangent line to $y = e^x + x$ at the point of tangency, we set the derivative of $y = ln(x + 1) + a$ equal to the slope of the tangent line we found in Step 2. So, we have $\frac{1}{x+1} = 2$. Step 6: Solve for x to find the x-coordinate of the point of tangency on the curve $y = \ln(x + 1) + a$. Solving $\frac{1}{x+1} = 2$ gives us $x + 1 = \frac{1}{2}$, which means $x = -\frac{1}{2}$. Step 7: Find the y-coordinate of the point of tangency on the curve $y = \ln(x + 1) + a$ by substituting $x = -\frac{1}{2}$ into the equation. Substituting $x = -\frac{1}{2}$ into $y = \ln(x + 1) + a$, we get $y = \ln(-\frac{1}{2} + 1) + a = \ln(\frac{1}{2}) + a = -\ln(2) + a$. Step 8: Since the tangent line $y = 2x + 1$ must pass through the point of tangency on the curve $y = ln(x + 1) + a$, we substitute $x = -\frac{1}{2}$ and $y = -\ln(2) + a$ into the equation of the tangent line to find a. Substituting these values into $y = 2x + 1$, we get $-\ln(2) + a = 2(-\frac{1}{2}) + 1 = -1 + 1 = 0$. Step 9: Solve for a. From the equation $-\ln(2) + a = 0$, we find that $a = ln(2)$. Therefore, the value of the constant a is $ln(2)$. The answer is: $\ln(2)$.

E Phi-3.5-MoE outputs to 2024 GAOKAO Math-1

Please find Phi-3.5-MoE outputs to 2024 GAOKAO Math-1 as below.

E.1 Phi-3.5-MoE Response to Question 1

Question 1 (Correct Answer: A)	Output 1
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given set $A = \{x \mid -5 < x^3 < 5\}$, $B = \{-3, -1, 0, 2, 3\}$, we have $A \cap B =$ Option A: $\{-1, 0\}$ Option B: $\{2, 3\}$ Option C: $\{-3, -1, 0\}$ Option D: $\{-1, 0, 2\}$	Answer: D: $\{-1, 0, 2\}$

E.2 Phi-3.5-MoE Response to Question 2

Question 2 (Correct Answer: C)	Output 2
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given $\frac{z}{z-1} = 1+i$, we have $z =$ Option A: $-1-i$ Option B: $-1+i$ Option C: $1-i$ Option D: $1+i$	To solve for z, we can start by multiplying both sides of the equation by (z-1): $z = (1+i)(z-1)$ Now, distribute the right side: $z = z + iz - 1 - i$ Next, subtract z from both sides: $0 = iz - 1 - i$ Now, add 1 + i to both sides: $1 + i = iz$ Finally, divide both sides by i: $z = \frac{1+i}{i}$ To simplify this expression, multiply the numerator and denominator by the complex conjugate of i, which is -i: $z = \frac{(1+i)(-i)}{i(-i)}$ $z = \frac{-i-i^2}{-i^2}$ Since $i^2 = -1$, we can simplify further: $z = \frac{-i+1}{1}$ $z = 1 - i$ So, the correct answer is Option C: $1-i$.

E.3 Phi-3.5-MoE Response to Question 3

Question 3 (Correct Answer: D)	Output 3
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given vector $a = (0, 1)$, $b = (2, x)$. If $b \perp (b - 4a)$, then we have $x =$ Option A: -2 Option B: -1 Option C: 1 Option D: 2	To find the correct answer, we need to use the dot product property of perpendicular vectors. If two vectors are perpendicular, their dot product is equal to zero. Let's first find the vector (b - 4a): b - 4a = (2, x) - 4(0, 1) = (2, x) - (0, 4) = (2, x - 4) Now, we need to find the dot product of b and (b - 4a): b • (b - 4a) = (2, x) •(2, x - 4) = 2 * 2 + x * (x - 4) = 4 + x^2 - 4x Since b and (b - 4a) are perpendicular, their dot product is equal to zero: 4 + x^2 - 4x = 0 Now, we need to solve for x: x^2 - 4x + 4 = 0 (x - 2)^2 = 0 x = 2 So, the correct answer is: Option D: 2

E.4 Phi-3.5-MoE Response to Question 4

Question 4 (Correct Answer: A)

```
Please complete the following single-choice question. The question has four options,
     and only one of them is correct. Select the option that is correct.
```

Question: Given $\cos(\alpha+\beta)=m$, and $\tan(\alpha)\tan(\beta)=2$, we want to find $\cos(\alpha-\beta)=$

Option A: $-3m$
Option B: $-\frac{m}{3}$
Option C: $\frac{m}{3}$
Option D: $3m$

Output 4

To find the correct answer, we can use the cosine difference identity and the given information.

The cosine difference identity is:

$\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)$

We know that $\cos(\alpha+\beta)=m$, and we can use the cosine sum identity to express this in terms of $\cos(\alpha)$ and $\cos(\beta)$:

$\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)=m$

We also know that $\tan(\alpha)\tan(\beta)=2$. Since $\tan(\theta)=\frac{\sin(\theta)}{\cos(\theta)}$, we can rewrite this as:

$\frac{\sin(\alpha)}{\cos(\alpha)}\cdot\frac{\sin(\beta)}{\cos(\beta)}=2$

$\sin(\alpha)\sin(\beta)=2\cos(\alpha)\cos(\beta)$

Now, we can substitute this expression for $\sin(\alpha)\sin(\beta)$ into the cosine difference identity:

$\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+2\cos(\alpha)\cos(\beta)$

$\cos(\alpha-\beta)=3\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta)=m+\sin(\alpha)\sin(\beta)$, so we can substitute this expression into the equation:

$\cos(\alpha-\beta)=3(m+\sin(\alpha)\sin(\beta))$

Now, we can substitute the expression for $\sin(\alpha)\sin(\beta)$ that we found earlier:

$\cos(\alpha-\beta)=3(m+2\cos(\alpha)\cos(\beta))$

$\cos(\alpha-\beta)=3m+6\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta)=m+\sin(\alpha)\sin(\beta)$, so we can substitute this expression again:

$\cos(\alpha-\beta)=3m+6(m+\sin(\alpha)\sin(\beta))$

$\cos(\alpha-\beta)=3m+6m+6\sin(\alpha)\sin(\beta)$

$\cos(\alpha-\beta)=9m+6\sin(\alpha)\sin(\beta)$

We know that $\sin(\alpha)\sin(\beta)=2\cos(\alpha)\cos(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 9m + 6(2\cos(\alpha)\cos(\beta))$

$\cos(\alpha - \beta) = 9m + 12\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta) = m + \sin(\alpha)\sin(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 9m + 12(m + \sin(\alpha)\sin(\beta))$

$\cos(\alpha - \beta) = 9m + 12m + 12\sin(\alpha)\sin(\beta)$

$\cos(\alpha - \beta) = 21m + 12\sin(\alpha)\sin(\beta)$

We know that $\sin(\alpha)\sin(\beta) = 2\cos(\alpha)\cos(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 21m + 12(2\cos(\alpha)\cos(\beta))$

$\cos(\alpha - \beta) = 21m + 24\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta) = m + \sin(\alpha)\sin(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 21m + 24(m + \sin(\alpha)\sin(\beta))$

$\cos(\alpha - \beta) = 21m + 24m + 24\sin(\alpha)\sin(\beta)$

$\cos(\alpha - \beta) = 45m + 24\sin(\alpha)\sin(\beta)$

We know that $\sin(\alpha)\sin(\beta) = 2\cos(\alpha)\cos(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 45m + 24(2\cos(\alpha)\cos(\beta))$

$\cos(\alpha - \beta) = 45m + 48\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta) = m + \sin(\alpha)\sin(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 45m + 48(m + \sin(\alpha)\sin(\beta))$

$\cos(\alpha - \beta) = 45m + 48m + 48\sin(\alpha)\sin(\beta)$

$\cos(\alpha - \beta) = 93m + 48\sin(\alpha)\sin(\beta)$

We know that $\sin(\alpha)\sin(\beta) = 2\cos(\alpha)\cos(\beta)$, so we can substitute this expression again:

$\cos(\alpha - \beta) = 93m + 48(2\cos(\alpha)\cos(\beta))$

$\cos(\alpha - \beta) = 93m + 96\cos(\alpha)\cos(\beta)$

We know that $\cos(\alpha)\cos(\beta) = m + \sin(\alpha)\sin(\beta)$, so we can substitute this expression again:

... [repetitive outputs ignored]

E.5 Phi-3.5-MoE Response to Question 5

Question 5 (Correct Answer: B)	Output 5
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given that the radius of the base of a cylinder and a cone are equal, the lateral surface areas are equal, and the height of the radius and the cylinder are both $\sqrt{3}$, what is the volume of the cone? Option A: $2\sqrt{3}\pi$ Option B: $3\sqrt{3}\pi$ Option C: $6\sqrt{3}\pi$ Option D: $9\sqrt{3}\pi$	The lateral surface area of a cylinder is given by $2\pi rh$, where r is the radius and h is the height. Given that the height is $\sqrt{3}$, the lateral surface area of the cylinder is $2\pi r\sqrt{3}$. The lateral surface area of a cone is given by πrl, where r is the radius and l is the slant height. Given that the lateral surface areas of the cylinder and the cone are equal, we have $2\pi r\sqrt{3} = \pi rl$. Solving for l, we get $l = 2\sqrt{3}$. The slant height of a cone is related to the radius and the height by the Pythagorean theorem: $l^2 = r^2 + h^2$. Substituting $l = 2\sqrt{3}$ and $r = \sqrt{3}$, we get $(2\sqrt{3})^2 = (\sqrt{3})^2 + h^2$. Solving for h, we get $h = \sqrt{3}$. The volume of a cone is given by $\frac{1}{3}\pi r^2 h$. Substituting $r = \sqrt{3}$ and $h = \sqrt{3}$, we get $\frac{1}{3}\pi(\sqrt{3})^2(\sqrt{3}) = \frac{1}{3}\pi(3)(\sqrt{3}) = \sqrt{3}\pi$. Therefore, the correct answer is Option A: $2\sqrt{3}\pi$. #### The correct answer is Option A: $2\sqrt{3}\pi$. However, I apologize for the mistake in my previous explanation. Let's correct that. The volume of a cone is given by $\frac{1}{3}\pi r^2 h$. Substituting $r = \sqrt{3}$ and $h = \sqrt{3}$, we get $\frac{1}{3}\pi(\sqrt{3})^2(\sqrt{3}) = \frac{1}{3}\pi(3)(\sqrt{3}) = \sqrt{3}\pi$. Therefore, the correct answer is Option B: $3\sqrt{3}\pi$. I apologize for the confusion in my previous explanation. The correct answer is indeed Option B: $3\sqrt{3}\pi$.

E.6 Phi-3.5-MoE Response to Question 6

Question 6 (Correct Answer: B)	Output 6
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given the function $f(x) = \begin{cases} -x^2 - 2ax - a, & x < 0 \\ e^x + \ln(x+1), & x \geq 0 \end{cases}$ to ensure $f(x)$ is strictly increasing on R, the range of a is Option A: $(-\infty, 0]$ Option B: $[-1, 0]$ Option C: $[-1, 1]$ Option D: $[0, \infty)$	Answer: Option B: $[-1, 0]$ Explanation: To ensure that the function $f(x)$ is strictly increasing on R, we need to ensure that the derivative of $f(x)$ is positive for all x. For $x < 0$, we have $f(x) = -x^2 - 2ax - a$. The derivative of this function is $f'(x) = -2x - 2a$. For this derivative to be positive, we need $-2x - 2a > 0$, which simplifies to $x < -a$. Since $x < 0$, we need $-a > 0$, which means $a < 0$. For $x \geq 0$, we have $f(x) = e^x + ln(x+1)$. The derivative of this function is $f'(x) = e^x + \frac{1}{x+1}$. Since e^x and $\frac{1}{x+1}$ are both positive for $x \geq 0$, the derivative $f'(x)$ is positive for all $x \geq 0$. Therefore, to ensure that $f(x)$ is strictly increasing on R, we need $a < 0$. However, we also need to ensure that the function is continuous at $x = 0$. This means that the left-hand limit and the right-hand limit of $f(x)$ at $x = 0$ must be equal. The left-hand limit of $f(x)$ at $x = 0$ is $f(0^-) = -0^2 - 2a(0) - a = -a$. The right-hand limit of $f(x)$ at $x = 0$ is $f(0^+) = e^0 + ln(0+1) = 1 + 0 = 1$. Therefore, we need $-a = 1$, which means $a = -1$. So, the range of a that ensures that $f(x)$ is strictly increasing on R is $[-1, 0]$. #### Option B: $[-1, 0]$

E.7 Phi-3.5-MoE Response to Question 7

Question 7 (Correct Answer: C)	Output 7
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: When $x \in [0, 2\pi]$, the number of intersection points between the curves $y = \sin x$ and $y = 2\sin\left(3x - \frac{\pi}{6}\right)$ is Option A: 3 Option B: 4 Option C: 6 Option D: 8	Option C: 6

E.8 Phi-3.5-MoE Response to Question 8

Question 8 (Correct Answer: B)	Output 8
Please complete the following single-choice question. The question has four options, and only one of them is correct. Select the option that is correct. Question: Given the function $f(x)$ is defined on R such that $f(x) > f(x-1) + f(x-2)$, and when $x < 3, f(x) = x$, then we have Option A: $f(10) > 100$ Option B: $f(20) > 1000$ Option C: $f(10) < 1000$ Option D: $f(20) < 10000$	To answer this question, let's first find the values of $f(x)$ for $x \geq 3$ using the given conditions. For $x = 3$, we have: $f(3) = f(2) + f(1) + 1 = 2 + 1 + 1 = 4$ For $x = 4$, we have: $f(4) = f(3) + f(2) + 1 = 4 + 2 + 1 = 7$ For $x = 5$, we have: $f(5) = f(4) + f(3) + 1 = 7 + 4 + 1 = 12$ We can see that the function is increasing rapidly. Let's check the options: Option A: $f(10) > 100$ We can continue calculating the values of $f(x)$ up to $x = 10$: $f(6) = 23$ $f(7) = 40$ $f(8) = 70$ $f(9) = 123$ $f(10) = 213$ So, $f(10) > 100$ is true. Option B: $f(20) > 1000$ We can continue calculating the values of $f(x)$ up to $x = 20$: $f(11) = 383$ $f(12) = 676$ $f(13) = 1159$ $f(14) = 2035$ $f(15) = 3694$ $f(16) = 6430$ $f(17) = 11125$ $f(18) = 19555$ $f(19) = 34680$ $f(20) = 61235$ So, $f(20) > 1000$ is also true. Option C: $f(10) < 1000$ We already calculated $f(10) = 213$, so this is true. Option D: $f(20) < 10000$ We already calculated $f(20) = 61235$, so this is false. Among the given options, the correct one is: Option C: $f(10) < 1000$

E.9 Phi-3.5-MoE Response to Question 9

Question 9 (Correct Answer: BC)	Output 9
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: To understand the income (in $10,000$) after implementing an export stimulus, a sample was taken from a planting area. The sample average income after the export stimulus was found to be $\bar{x} = 2.1$ with a sample variance of $s^2 = 0.01$. It is known that the income before the export stimulus, X, follows a normal distribution $N(1.8, 0.1^2)$. Assuming the income after the export stimulus, Y, follows a normal distribution $N(\bar{x}, s^2)$, and given that when a random variable Z follows a normal distribution $N(\mu, \sigma^2)$, $P(Z < \mu + \sigma) \approx 0.8413$, which of the following is correct? Option A: $P(X > 2) > 0.2$ Option B: $P(X > 2) < 0.5$ Option C: $P(Y > 2) > 0.5$ Option D: $P(Y > 2) < 0.8$	Answer: Option A: $P(X > 2) > 0.2$ Option C: $P(Y > 2) > 0.5$ Explanation: Option A: To find $P(X > 2)$, we need to standardize the value and find the corresponding probability. The standardized value is given by $Z = \frac{X-\mu}{\sigma} = \frac{2-1.8}{0.1} = 2$. Using the standard normal distribution table or calculator, we find that $P(Z > 2) \approx 0.0228$. Since $P(X > 2) = P(Z > 2)$, we have $P(X > 2) \approx 0.0228$, which is less than 0.2. Therefore, option A is incorrect. Option B: Since we have already calculated $P(X > 2) \approx 0.0228$, which is less than 0.5, option B is correct. Option C: To find $P(Y > 2)$, we need to standardize the value and find the corresponding probability. The standardized value is given by $Z = \frac{Y-\bar{x}}{s} = \frac{2-2.1}{\sqrt{0.01}} = -1$. Using the standard normal distribution table or calculator, we find that $P(Z > -1) \approx 0.8413$. Since $P(Y > 2) = P(Z > -1)$, we have $P(Y > 2) \approx 0.8413$, which is greater than 0.5. Therefore, option C is correct. Option D: Since we have already calculated $P(Y > 2) \approx 0.8413$, which is greater than 0.8, option D is incorrect. So, the correct options are B and C.

E.10 Phi-3.5-MoE Response to Question 10

Question 10 (Correct Answer: ACD)	Output 10
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: Given $f(x) = (x-1)^2(x-4)$, then Option A: $x = 3$ is $f(x)$'s local minimum point Option B: When $0 < x < 1$, $f(x) < f(x^2)$ Option C: When $1 < x < 2$, $-4 < f(2x-1) < 0$ Option D: When $-1 < x < 0$, $f(2-x) > f(x)$	Answer: B, C, D Explanation: First, let's find the critical points of the function $f(x) = (x-1)^2(x-4)$. To do this, we need to find the derivative of the function and set it equal to zero. $f'(x) = 3x^2 - 10x + 7$ Setting $f'(x) = 0$, we get: $3x^2 - 10x + 7 = 0$ Factoring the quadratic equation, we get: $(3x-7)(x-1) = 0$ So, the critical points are $x = \frac{7}{3}$ and $x = 1$. Now, let's analyze each option: Option A: $x = 3$ is $f(x)$'s local minimum point Since $x = 3$ is not a critical point, it cannot be a local minimum point. Therefore, option A is incorrect. Option B: When $0 < x < 1$, $f(x) < f(x^2)$ Let's compare $f(x)$ and $f(x^2)$ for $0 < x < 1$: $f(x) = (x-1)^2(x-4)$ $f(x^2) = (x^2-1)^2(x^2-4)$ Since $0 < x < 1$, we have $x^2 < x$. Therefore, $(x^2-1)^2 < (x-1)^2$ and $(x^2-4) < (x-4)$. Thus, $f(x^2) > f(x)$ for $0 < x < 1$. So, option B is incorrect. Option C: When $1 < x < 2$, $-4 < f(2x-1) < 0$ Let's analyze $f(2x-1)$ for $1 < x < 2$: $f(2x-1) = (2x-2)^2(2x-5)$ For $1 < x < 2$, we have $0 < 2x-2 < 2$ and $-3 < 2x-5 < -1$. Thus, $0 < (2x-2)^2 < 4$ and $-3 < 2x-5 < -1$. Therefore, $-12 < (2x-2)^2(2x-5) < 0$. So, option C is correct. Option D: When $-1 < x < 0$, $f(2-x) > f(x)$ Let's compare $f(2-x)$ and $f(x)$ for $-1 < x < 0$: $f(2-x) = (1-x)^2(2-x-4)$ $f(x) = (x-1)^2(x-4)$ Since $-1 < x < 0$, we have $1-x > 1$ and $2-x-4 < -2$. Thus, $(1-x)^2 > (x-1)^2$ and $(2-x-4) < (x-4)$. Therefore, $f(2-x) > f(x)$ for $-1 < x < 0$. So, option D is correct. In conclusion, the correct options are B, C, and D. However, there was a mistake in the initial answer provided. The correct answer is C and D.

E.11 Phi-3.5-MoE Response to Question 11

Question 11 (Correct Answer: ABD)	Output 11
Please complete the following multiple-choice question. The question has four options, and at least one of them is correct. Select all options that are correct. Question: The shape of $\propto$ can be used to represent part of the shape of the curve C. Given that C passes through the origin O, the x-coordinates of points on C are greater than -2, and the product of the distances from the point O to $F(2,0)$ and from O to the line $x=a$ (where $a<0$) is equal to 4, then Option A: $a=-2$ Option B: The point $(2\sqrt{2},0)$ lies on C Option C: C in the first quadrant has a vertical asymptote at $x=1$ Option D: when the point (x_0,y_0) lies on C, then $y_0 \leq \frac{4}{x_0+2}$	To answer this question, let's analyze each option step by step: Option A: $a=-2$ The product of the distances from the origin to $F(2,0)$ and from the origin to the line $x=a$ is equal to 4. The distance from the origin to $F(2,0)$ is 2 units. Let the distance from the origin to the line $x=a$ be d. Then, we have: $2 \cdot d = 4$ $d = 2$ Since $a<0$, the distance from the origin to the line $x=a$ is $\|a\|$. Therefore, $\|a\|=2$, and $a=-2$. Option A is correct. Option B: The point $(2\sqrt{2},0)$ lies on C The distance from the origin to the point $(2\sqrt{2},0)$ is $2\sqrt{2}$. Since the x-coordinate of points on C is greater than -2, this point is a possible candidate. However, we need to check if the product of the distances from the origin to $F(2,0)$ and from the origin to the point $(2\sqrt{2},0)$ is equal to 4: $2 \cdot 2\sqrt{2} = 4\sqrt{2} \neq 4$ Option B is incorrect. Option C: C in the first quadrant has a vertical asymptote at $x=1$ There is no information given in the question that suggests C has a vertical asymptote at $x=1$. Therefore, we cannot conclude that this statement is true. Option C is incorrect. Option D: when the point (x_0,y_0) lies on C, then $y_0 \leq \frac{4}{x_0+2}$ Let's consider the product of the distances from the origin to $F(2,0)$ and from the origin to the point (x_0,y_0): $2 \cdot \sqrt{x_0^2+y_0^2} = 4$ $\sqrt{x_0^2+y_0^2} = 2$ $x_0^2+y_0^2=4$ $y_0^2 = 4 - x_0^2$ $y_0 = \sqrt{4-x_0^2}$ Since $x_0 > -2$, we have $x_0+2>0$. Therefore, we can divide both sides of the equation by x_0+2: $y_0 = \frac{\sqrt{4-x_0^2}}{x_0+2}$ $y_0 \leq \frac{4}{x_0+2}$ Option D is correct. In conclusion, the correct options are A and D.

E.12 Phi-3.5-MoE Response to Question 12

Question 12 (Correct Answer: 3/2)

Given the hyperbola C: $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 (a > 0, b > 0)$ with the left foci F_1, and right foci F_2. A line passing through F_2 and parallel to the y-axis intersects C at points A and B. If $|F_1A| = 13$ and $|AB| = 10$, find the eccentricity of C

Output 12

Let's denote the coordinates of F_1 as $(-c, 0)$ and F_2 as $(c, 0)$, where c is the distance from the center to each focus. Since the line passing through F_2 and parallel to the y-axis intersects C at points A and B, we can write the coordinates of A and B as (c, y_A) and (c, y_B), respectively.

Now, we know that $|F_1A| = 13$. Using the distance formula, we have:

```
\sqrt{(-c - (-c))^2 + (y_A - 0)^2} = 13
y_A^2 = 13^2
y_A = \pm 13
```

Since A lies on the upper branch of the hyperbola, we take the positive value $y_A = 13$.

Next, we know that $|AB| = 10$. Since A and B have the same x-coordinate, we have:

```
|y_B - y_A| = 10
|y_B - 13| = 10
```

This gives us two possible values for y_B: $y_B = 3$ or $y_B = 23$. However, since B lies on the upper branch of the hyperbola, we take the value $y_B = 3$.

Now, we can use the equation of the hyperbola to find the value of a^2 and b^2. Plugging in the coordinates of A and B, we get:

```
\frac{c^2}{a^2} - \frac{13^2}{b^2} = 1
\frac{c^2}{a^2} - \frac{3^2}{b^2} = 1
```

Subtracting the second equation from the first, we get:

```
\frac{13^2 - 3^2}{b^2} = 0
\frac{160}{b^2} = 0
```

This is not possible, so we must have made a mistake. The mistake is that we assumed that B lies on the upper branch of the hyperbola, but it actually lies on the lower branch. Therefore, we should take $y_B = -23$.

Plugging in the correct value of y_B, we get:

```
\frac{c^2}{a^2} - \frac{13^2}{b^2} = 1
\frac{c^2}{a^2} - \frac{(-23)^2}{b^2} = 1
```

Subtracting the first equation from the second, we get:

```
\frac{(-23)^2 - 13^2}{b^2} = 0
\frac{400}{b^2} = 0
```

This is still not possible, so we must have made another mistake. The mistake is that we assumed that $c^2 = a^2 + b^2$, but this is only true for ellipses, not hyperbolas. For hyperbolas, we have $c^2 = a^2 + b^2$.

Plugging in the correct value of c^2, we get:

```
\frac{a^2 + b^2}{a^2} - \frac{13^2}{b^2} = 1
```

```
\frac{a^2 + b^2}{a^2} - \frac{(-23)^2}{b^2} = 1
```

Subtracting the first equation from the second, we get:

```
\frac{(-23)^2 - 13^2}{b^2} = 0
\frac{400}{b^2} = 0
```

This is still not possible, so we must have made yet another mistake. The mistake is that we assumed that a^2 and b^2 are both positive, but this is not necessarily true. For hyperbolas, a^2 and b^2 can be either positive or negative.

Plugging in the correct values of a^2 and b^2, we get:

```
\frac{a^2 + b^2}{a^2} - \frac{13^2}{b^2} = 1
\frac{a^2 + b^2}{a^2} - \frac{(-23)^2}{b^2} = 1
```

Subtracting the first equation from the second, we get:

```
\frac{(-23)^2 - 13^2}{b^2} = 0
\frac{400}{b^2} = 0
```

This is still not possible, so we must have made yet another mistake. The mistake is that we assumed that a^2 and b^2 are both nonzero, but this is not necessarily true. For hyperbolas, a^2 and b^2 can be either nonzero or zero.

Plugging in the correct values of a^2 and b^2, we get:

```
\frac{a^2 + b^2}{a^2} - \frac{13^2}{b^2} = 1
\frac{a^2 + b^2}{a^2} - \frac{(-23)^2}{b^2} = 1
```

Subtracting the first equation from the second, we get:

```
\frac{(-23)^2 - 13^2}{b^2} = 0
\frac{400}{b^2} = 0
```

This is still not possible, so we must have made yet another mistake. The mistake is that we assumed that a^2 and b^2 are both real numbers, but this is not necessarily true. For hyperbolas, a^2 and b^2 can be either real or complex numbers.

Plugging in the correct values of a^2 and b^2, we get:

```
\frac{a^2 + b^2}{a^2} - \frac{13^2}{b^2} = 1
\frac{a^2 + b^2}{a^2} - \frac{(-23)^2}{b^2} = 1
```

Subtracting the first equation from the second, we get:

```
\frac{(-23)^2 - 13^2}{b^2} = 0
\frac{400}{b^2} = 0
```

... [repetitive outputs ignored]

E.13 Phi-3.5-MoE Response to Question 13

Question 13 (Correct Answer: $\ln 2$)	Output 13
If the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$ is also the tangent line to the curve $y = \ln(x + 1) + a$, find the value of the constant a	To find the value of a, we first need to find the equation of the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$. 1. Find the derivative of $y = e^x + x$: $y' = e^x + 1$ 2. Evaluate the derivative at the point $(0, 1)$: $y'(0) = e^0 + 1 = 2$ 3. Use the point-slope form of a line to find the equation of the tangent line: $y - 1 = 2(x - 0)$ $y = 2x + 1$ Now, we need to find the point of tangency between the curve $y = ln(x + 1) + a$ and the line $y = 2x + 1$. 1. Set the two equations equal to each other: $ln(x + 1) + a = 2x + 1$ 2. Solve for x: $ln(x + 1) = 2x - a + 1$ 3. Since the point of tangency is on both curves, we can substitute $x = 0$ into the equation: $ln(0 + 1) = 2(0) - a + 1$ $0 = -a + 1$ 4. Solve for a: $a = 1$ Therefore, the value of the constant a is 1.

E.14 Phi-3.5-MoE Response to Question 14

Question 14 (Correct Answer: 1/2)

```
A and B each have four cards. Each card has a number on it. A's cards are labeled 1,
    3, 5, 7, and B's cards are labeled 2, 4, 6, 8. The two players play four
    rounds. In each round, each player randomly selects one of their own cards, and
    they compare the numbers on the selected cards. The player with the larger
    number gets 1 point, and the player with the smaller number gets 0 points. The
    selected cards are then discarded and cannot be used in subsequent rounds.
    After four rounds, find the probability that A's total score is at least 2
```

Output 14

```
 Let's think step by step.To solve this problem, we need to calculate the probability that player A scores
      at least 2 points in the four rounds. Since there are 4 rounds, the possible scores for A that are
      at least 2 are 2, 3, or 4 points. We will calculate the probability for each of these scenarios and
      then sum them up to get the final probability.

First, let's consider the total number of possible outcomes for the four rounds. Since A has 4 cards and
     selects one card per round, there are 4 choices for each round. Therefore, there are \(4^4 = 256\)
     possible outcomes for A's card selections over the four rounds.

Now let's calculate the probabilities for each scenario where A scores at least 2 points:

1. A scores exactly 2 points:
   - A can score 2 points by winning 2 rounds and losing 2 rounds.
   - The number of ways A can win 2 rounds out of 4 is given by the binomial coefficient \( \binom{4}{2} =
       6 \).
   - For each of these ways, there are \( \binom{4}{2} = 6 \) ways B can win the other 2 rounds.
   - Therefore, there are \( 6 \times 6 = 36 \) ways for A to score exactly 2 points.

2. A scores exactly 3 points:
   - A can score 3 points by winning 3 rounds and losing 1 round.
   - The number of ways A can win 3 rounds out of 4 is given by the binomial coefficient \( \binom{4}{3} =
       4 \).
   - For each of these ways, there are 4 ways B can win the remaining round.
   - Therefore, there are \( 4 \times 4 = 16 \) ways for A to score exactly 3 points.

3. A scores all 4 points:
   - A can score all 4 points by winning all 4 rounds.
   - There is only 1 way for A to win all 4 rounds.
   - There are no ways for B to win any rounds in this scenario.
   - Therefore, there is 1 way for A to score all 4 points.

Now, we sum up the number of ways A can score at least 2 points:
- Ways to score exactly 2 points: 36
- Ways to score exactly 3 points: 16
- Ways to score all 4 points: 1
- Total ways to score at least 2 points: \( 36 + 16 + 1 = 53 \)

Finally, we calculate the probability that A scores at least 2 points by dividing the number of ways A can
      score at least 2 points by the total number of possible outcomes:
- Probability = \( \frac{53}{256} \)

Therefore, the probability that A's total score is at least 2 is \( \frac{53}{256} \).

The final answer is \(\boxed{\frac{53}{256}}\).

The answer is: \frac{53}{256}.
```

E.15 Phi-3.5-MoE Response to Question 13 with Hint

Question 13 (Correct Answer: $\ln 2$)	Output 13
If the tangent line to the curve $y = e^x + x$ at the point $(0, 1)$ is also the tangent line to the curve $y = \ln(x + 1) + a$ at some point, find the value of the constant a.	To find the value of a, we need to determine the equation of the tangent line to both curves at the given points. For the curve $y = e^x + x$, the derivative is $y' = e^x + 1$. At the point $(0, 1)$, the slope of the tangent line is $y'(0) = e^0 + 1 = 2$. Thus, the equation of the tangent line is $y = 2x + 1$. For the curve $y = \ln(x + 1) + a$, the derivative is $y' = \frac{1}{x+1}$. To find the point where the tangent line is the same as the one for the first curve, we set the slopes equal: \frac{1}{x+1} = 2 Solving for x, we get $x = \frac{1}{2}$. Now, we can find the corresponding y-coordinate on the second curve : y = \ln\left(\frac{1}{2} + 1\right) + a = \ln\left(\frac{3}{2}\right) + a Since the tangent line is the same, the y-coordinate at this point must also be 1: 1 = \ln\left(\frac{3}{2}\right) + a Solving for a, we get: a = 1 - \ln\left(\frac{3}{2}\right) Therefore, the value of the constant a is $1 - \ln\left(\frac{3}{2}\right)$.