SEEK: A Stacked Ensemble for Expert Knowledge

Kevin Kim
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
kkim170@jhu.edu

Sara Ren
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
sren16@jhu.edu

Camden A. Shultz
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
cshultz3@jhu.edu

Abstract

Large language models (LLMs) have seen recent spikes of popularity due to increasing quantities of training data and powerful compute. However, LLMs continue to hallucinate and are by no means memory efficient. We attempt to mitigate these issues through domain-specific knowledge distillation. Domain specific distillation requires reducing both the parameter size and vocabulary alterations while fine tuning to restricted domains of knowledge, such as law, medicine, math, or computer science, which allows models to perform more accurately in domain-specific answering tasks than domain-agnostic distillation. Through careful distillation, we propose a mixture-of-experts (MoE) of smaller and more efficient domain-specific and domain-agnostic models that in aggregate perform similarly to a larger model on domain-specific tasks.

1 Introduction & Motivation

Large language models (LLMs) have caught on in popularity in the recent years as their capabilities have developed, especially with the release of ChatGPT and the announcement of GPT4 (OpenAI, 2023). As demand for more accessible and personalized LLMs increase, we foresee issues with the large memory footprint of LLMs. We wish to experiment with model distillation as a way to take advantage of the weights of existing LLMs to lower the overall memory footprint. Knowledge distillation is the process of training a smaller model *student* on the input-output pairs of the larger *teacher* model. By using fine-tuned domain-specific models as teacher models, higher accuracies are achieved in domain-specific tasks. (Yao et al., 2021). We hypothesize that when these domain-specific student models are mixed together, a model with an overall smaller memory footprint but with higher domain-specific accuracy can be constructed.

In this work, a mixture-of-experts model is built by stacking distilled BERT models for Multiple Choice Question Answering fine-tuned for the domains of arithmetic, science, and medicine. A mixture-of-experts model was chosen as it has been shown to achieve up to 1000x more improvements on model capacity (Shazeer et al., 2017), making it an appealing method when combining smaller models. Mixture-of-expert ensembles also take into account multiple experts, which allows for a more well-rounded response. We also aim to investigate whether multiple experts can help with answering questions across different domains.

1st Conference on Natural Language Processing - Self Supervised Models (JHU NLP-SSM 2023).

We introduce SEEK, a stacked ensemble of expert knowledge for multiple choice question answering. We show that smaller specialized learners in aggregate can perform nearly as well as larger fine-tuned teacher models can. Rather than exponentially increasing model size based on exponential scaling laws, we have a model that can grow linearly in size across domains. The sections of this paper are as follows: In section 2 we outline current related works in knowledge distillation and mixture-of-experts. In sections 3 and 4 we expand on the importance of knowledge distillation and mixture-of-experts as a valid path for compressed knowledge. In section 5 we introduce SEEK, followed by experiments and ablation studies in section 6. Results are outlined in section 7.

2 Related Work

Knowledge Distillation Knowledge distillation continues to be an active field of research. DistillBert (Sanh et al., 2020), one of the most utilized distilled networks demonstrated that first fine-tuning teacher models on the target task increased performance compared to first distilling and then fine-tuning student models. Many others, including Adapt-and-Distill (Yao et al., 2021), Domain-specific knowledge distillation (Howell et al., 2022), and Knowledge Distillation Transfer Sets (Peris et al., 2022), have shown that distilling models using domain-specific knowledge can improve performance in downstream tasks. Many works of knowledge distillation continue to focus on the computational efficiency of smaller models and application to on-the-edge device computing rather than expert knowledge retention.

Model Compression Techniques Model compression techniques have grown in popularity as large language models are frequently exceeding hundreds of billions of parameters. Model compression is typically divided into two sub-categories: quantization and pruning. Many, including Gupta et al., 2015, have explored quantization in an effort to reduce the memory footprint of large models. It's been shown that with little reductions in performance, moving from 32-bit to 16, and 8 bit models can result in significant boosts in efficiency. With similar motivations, Wang, et al., 2021 present structured pruning as a technique to reduce redundant model parameters. While not entirely orthogonal, in this paper we demonstrate that similar strides towards model efficiency can be achieved through expert routing to distilled domain-specific models while retaining similar and/or better performance on specific downstream tasks.

Expert Routing Machine learning models are inherently limited by their parameter counts, and consequently their capacity to absorb information. Mixture-of-Experts, a technique where subsets of a network are activated conditioned on the given example have proven useful in enabling much larger parameters without the computational overhead of computing inference over the entire set of weights. A trainable Sparsely-Gated Mixture-of-Experts layer is introduced that learns to attend over the sub-networks (Shazeer et al., 2017). GShard (Lepikhin et al., 2020) shows that similar layers can be scaled up to MoE Transformers with 600 billion parameters. Switch Transformer (Fedus et al., 2021) grows this to trillions of parameters by only routing to 1 sub-network. More recently, Zhou et al., 2022 describes the effect of expert choice routing, where each expert picks top-k tokens instead of conditioning expert assignment from based on input tokens. It was noted that this method achieves better load balancing across experts and better training efficiency compared to previous methods. Inspired by this, we plan to develop an attention mechanism across the ensemble of student models, attending to each model's strengths ad hoc.

3 Expert Knowledge Distillation

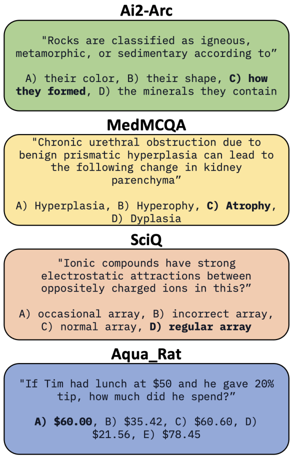
Domain-specific knowledge distillation is a technique in machine learning that involves transferring knowledge from a larger or more complex model, often referred to as the teacher model, to a smaller or simpler model, referred to as the student model, which is specialized to perform well on a specific domain or task. The process of knowledge transfer involves training the teacher model on a large dataset covering multiple domains or tasks and using it to generate soft targets or labels that contain additional information beyond the ground truth labels.

The soft targets, which typically include the probabilities associated with each label, are then used to train the student model on a smaller dataset specific to the target domain or task. The student model is trained to mimic the behavior of the teacher model in generating soft targets, as well as in producing accurate predictions on the specific domain or task. This allows the student model to learn from the knowledge of the larger teacher model, without requiring access to the entire training dataset used to train the teacher model. Moreover, it has been shown that effective knowledge distillation outperforms finetuning smaller models since soft targets have more entropy than single labels, effectively allowing smaller models to learn more complex parameters than they could otherwise.

Domain-specific knowledge distillation is particularly useful in scenarios where the larger teacher model is too resource-intensive to be deployed in a real-world application or where the smaller student model is required to operate with limited computational resources. By distilling the knowledge of the teacher model into the student model, the resulting model can be both smaller and more accurate, making it well-suited for deployment in real-world applications with limited computational resources.

Student models can be well initialized by selecting a subset of attention layers from the teacher model for use in the student. This works in part due to the residual connections between attention layers, resulting in modest modifications between multiple layer (Sanh, et al., 2020).

SEEK's expert models are formed by distilling fine-tuned teacher models on each of the specific domains. The data used for distillation is consistent with that for fine-tuning. The largest reduction in model parameters is made by reducing attention layers. Our expert models contain 6 attention layers compared to the 12 in teacher models, approximately halving each expert to 66 million parameters from the 109 million in standard BERT models.



Ai2-Arc

"Rocks are classified as igneous, metamorphic, or sedimentary according to”

A) their color, B) their shape, **C) how they formed**, D) the minerals they contain

MedMCQA

"Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following change in kidney parenchyma”

A) Hyperplasia, B) Hyperophy, **C) Atrophy**, D) Dyplasia

SciQ

"Ionic compounds have strong electrostatic attractions between oppositely charged ions in this?”

A) occasional array, B) incorrect array, C) normal array, **D) regular array**

Aqua_Rat

"If Tim had lunch at $50 and he gave 20% tip, how much did he spend?”

A) $60.00, B) $35.42, C) $60.60, D) $21.56, E) $78.45

Figure 1: List of expert datasets with examples of questions and answers

Expert Datasets Four datasets were used to fine tune models in this experiment.

- **Aqua_Rat:** The first dataset used was the AquaRat dataset, a math dataset that is a superset of the MathQ dataset. It consists of 100,000 algebraic word problems (Ling et al., 2017).These questions had 5 multiple choice selections, so all questions that had the fifth choice as the correct choice were filtered out to fit our superset.
- **Ai2Arc:** The second dataset was Ai2Arc, a collection of 7,787 grade school level multiple choice science questions (Clark et al., 2018).
- **MedMCQA:** MedMCQA is a corpus of multiple choice questions that are reflective of medical school entrance examinations. MedMCQA has more than 194k AIIMS and NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects (Pal et al., 2022). The dataset also consists of expert reasoning for correct answers, making it a interesting task for future open-domain quesiton answering. Some of the questions in MedMCQA were multiple-select, which we filtered out before using for our models.
- **SciQ:** The fourth dataset used was SciQ, a crowdsourced dataset of a wide span of science exam questions about Physics, Chemistry and Biology. All 3,679 questions were multiple choice with four answer options (Welbl et al., 2017).

Distillation Loss To train our student model, we use a weighted combination of Kullback-Leibler Divergence and cross entropy loss across our teacher output and ground truth, respectively.

$$L(x; W) = \alpha(KLdiv(y_s, y_t)) + (1 - \alpha)(CE(y_s, y))$$

To distill our student models, alpha was kept at 0.9. Previous works (Sanh et al., 2020), have also experimented with soft targets over model parameters themselves adding L2 loss for comparable attention layers between the student and teacher.

4 Mixture of Experts Layer

In most Mixture-of-Expert (MoE) models, input data is first fed into a gating network, which takes as input a data point and outputs a set of probabilities over the experts. A wide variety of gating networks exist, ranging from simple feed-forward neural networks taking in the input to more complex models that consider the interactions between input features. Sparsely gated MoEs enable much larger models as only subsets of the network are trained simultaneously. Typically the top *k* experts scored by the gating mechanism are selected.

Each expert is a sub-model that specializes in a particular aspect of the problem. The experts do not need to be homogeneous, but should receive and produce the same structures in vector space. Each expert receives a subset of the input data and produces its own output. The outputs of the experts are then combined to produce the final output. During training, the gating network learns to allocate the input data to the appropriate experts, and the experts learn to produce accurate outputs for their assigned data points. It has been shown that MoE models can learn to specialize each of their experts. There are several ways to combine the outputs of the experts. The most common approach is to use a weighted sum, where the weights are determined by the gating network probabilities, however non-linear combinations can also be learned via a neural network.

MoE has several advantages over single-model approaches. First, it allows the model to capture complex, multi-modal distributions in the input data by learning a mixture of models. Second, it can handle non-stationary data distributions by adapting the gating network probabilities as the data changes. Finally, it can improve the overall accuracy and robustness of the model by combining the strengths of multiple models for inference.

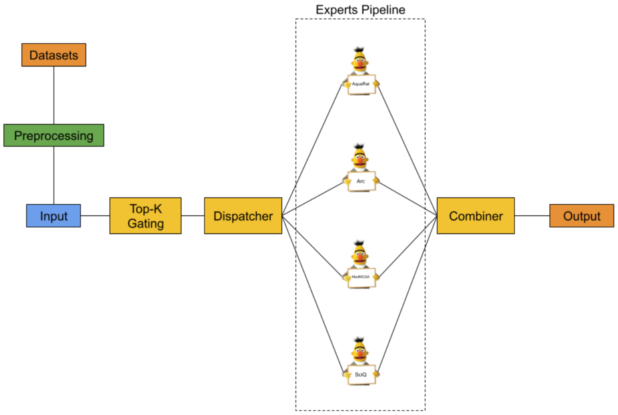


Figure 2: Our Mixture of Experts (MoE) architecture, which utilizes Top-K Gating to dispatch the input to the correct expert(s). The output is a weighted sum of these experts, which is then used as the label for multiple choice question answering.

5 A Stacked Ensemble for Expert Knowledge

Fine-tuned Teacher Models Four standard size BERT models are used as teacher models, covering three specific domains and one semi-domain-agnostic dataset. They are the BERT base model[1], SciBERT[2], MathBERT[3], and OntoMedQA[4]. All of the pretrained models were fine-tuned on in-domain multiple choice questions, using ai2_arc , sciq, aqua_rat, and MedMCQA, respectively. These large models for multiple choice question answering serve as the teacher models for distillation.

Table 1: List of pretrained models and accuracies on domain-specific datasets after fine-tuning

Model	Dataset	Finetuned Accuracy
MathBert	aqua_rat	0.3684
BERT-base	ai2_arc	0.5121
OntoMedQA	medmcqa	0.3008
SciBert	sciq	.754

Student Models To create the student model, we initialize another BertForMultipleChoice model with half (6) of the amount of hidden layers. Reducing attention layers is the most effective way to reduce parameter counts. We then copied the weights of every other hidden layer in the teacher model. We then trained the student models on the domain-specific datasets again, trying to match the answer of the teacher model using soft targets of the answer classes as well as the ground truth answer provided from our dataset. An alpha value of 0.9 was used to balance the importance of the two optimization tasks.

Mixture of Experts Similar to Shazeer et al., 2017, the MoE network proposed here consist of $n = 4$ expert models $E_1, \ldots, E_n$ with a sparse gating network G that produces a sparse probability distribution over the experts given input x. Top $k = 2$ experts are chosen. In order to reduce the size of the gating network, input data is parsed in the following ways. x_e, denoting Question-answer pairs of the form $[$*question + answer A*$,$ *question + answer B*$,$ *question + answer C*$,$ *question + answer D*$]$ are sent to each of the selected experts. A summary of the question and answers, x_g, is given to the gating network as $[Question + AnswerA, AnswerB, AnswerC, AnswerD]$. The embeddings of x_g are encoded by a BERT model and fed to a feed forward network to produce conditional probabilities over the top k experts. Specifically, a softmax gating function, a trivial choice for non-sparse gating, is augmented with trainable weights and random Gaussian noise. The feature-extraction portion of the network is not backpropagated over during training, and thus training times are rather efficient. It is expected, however, that a smaller encoding model can be used while retaining similar performance.

$$G(x) = Softmax\big(TopK\big(x_g W_{gate} + \mathcal{N}(\mu,\, \sigma^2) \cdot Softplus(x_g W_{noise})\big)\big) \tag{1}$$

That is the conditional probability that a value is in the top k given random noise is computed during training, providing a differentiable loss that is also capable of helping to balance the load of experts. Details of this loss are discussed in appendix A of Shazeer, et al. 2017, however it can be summarized as the square of the coefficient of variation of the load vector. At inference, no noise is given, the prediction is deterministic and losses cannot be back propagated.

Furthermore, a weighted sum of the expert outputs is chosen as a aggregator. Therefore the output of the MoE can be described as:

$$y = \sum_{i=1}^{n} G(x)_i E_i(x) \tag{2}$$

The authors Shazeer et al. 2017 note that if the number of experts is very large, the branching factor can be reduced by stacking multiple MoE gating networks, creating an hierarchical MoE. Due to the specific architecture of chosen domain-specific experts here, we found this to not be necessary.

[1] https://huggingface.co/bertbaseuncased

[2] https://huggingface.co/allenai/scibert_scivocab_uncased

[3] https://huggingface.co/tbs17/MathBERT

[4] https://huggingface.co/sahillihas/OntoMedQA

In our implementation, we pass the question into a Hugging Face Pipeline instead of the model directly. This allows us to standardize the data to the format mentioned above and abstract model vocabulary and technical niches into the pipeline. The pipeline takes in the input from the dispatcher, then preprocesses the data to fit the model, and the model's inference is called. This also makes attaching other experts in the future much easier.

6 Experiments

Experiments

Evaluation Strategy To evaluate our ensemble model, we combined 800 multiple choice questions. 200 questions were selected from each test split, or a portion of the validation split if the test splits did not contain ground truth answers. Accuracy was used as the evaluation metric. The distribution of questions are balanced across the datasets.

Experiments An ensemble of finetuned teacher models is first evaluated as a baseline. As outlined in Table 2, an accuracy of 44% accuracy was achieved on our evaluation bench-set. To directly compare, an ensemble of distilled student models is evaluated. 44% accuracy is achieved with only 68.9% of the parameters. It is important to note that a significant portion of model parameters originates from the BERT encoder described above (109 million).

Ablation Studies Further ablation studies over the MoE experts is performed. We examined leave one out ensembles, in which a single expert is removed from the ensemble. This allows us to evaluate the impact a particular expert has in overall performance. We show robustness between experts with similar domains.

7 Results & Discussion

Table 2: Results of various MoE ensembles, including various leaveoneout experiments. Model parameters are provided to compare accuracy loss with parameter count.

Model	**Training Accuracy**	**Testing Accuracy**	**Model Parameters (M)**
Teacher Models	82%	44%	548.3
Distilled Student Models	84%	40%	378.2
Leave-One-Out (Arc)	67%	35%	311.3
Leave-One-Out (Medical)	67%	37%	311.3
Leave-One-Out (Science)	84%	40%	311.3
Leave-One-Out (Math)	67%	35%	311.3

Though the ensemble of student models is 4% worse on testing data, it is 170.1 million parameters less. While it is unfortunate that the student models perform worse than the teacher models, this was to be expected since the models are much smaller. Note that just BERT (BERT-based-uncased) show an accuracy of 23% on our test dataset, which is no different from just random chance. Thus, we are able to still be better than a base BERT, but no better than a teacher.

While the goal of this paper is not to propose a new metric, to further quantify our results, we propose one simple method. The metric will be the accuracy divided by the log of the number of parameters. This metric will reward accuracy but punish the number of parameters, while taking into account the exponential scaling laws of large-language models (Kaplan et al., 2020). This way, we are measuring the accuracy versus the potential knowledge available from the model size.

$$score = \frac{accuracy}{\log(\#Params)}$$

From the size-adjusted scores (Table 3), we can see that our student models do perform worse. This is most likely due to a loss of information during distillation.

Table 3: Size-adjusted scores

Model	Training Score	Testing Score	Model Parameters (M)
Teacher Models	9.38	5.03	548.3
Distilled Student Models	9.79	4.66	378.2
Leave-One-Out (Arc)	7.89	4.12	311.3
Leave-One-Out (Medical)	7.89	4.36	311.3
Leave-One-Out (Science)	9.89	4.71	311.3
Leave-One-Out (Math)	7.89	4.12	311.3

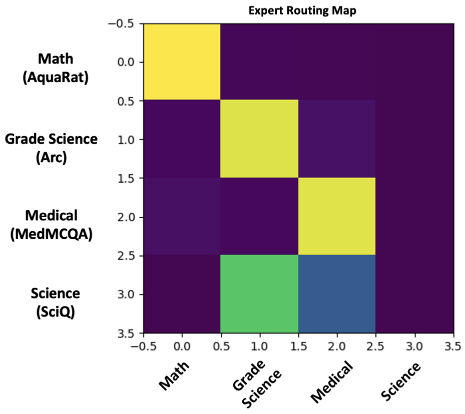
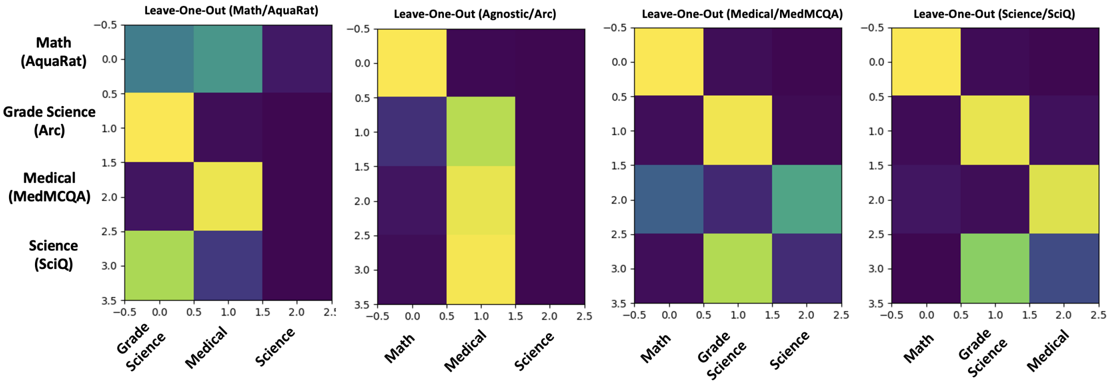


Figure 3: Expert Routing Maps for SEEK

In Figures 3 and 4 we illustrate expert routing maps between question sources (the datasets they origionate from) and the experts that they are routed to. We show that for most datasets, there exists a one-to-one correspondence. In the case that two experts cover similar domains, it has been shown that one predominately takes over. This can be seen where Arc (grade-school science questions) is preferred over SciQ models.

Figure 4: Leave-One-Out Ablation: Routing between sources and experts are visualized.



A) SEEK without Aqua-Rat expert. B) SEEK without Ai2-Arc expert. C) SEEK without MedMCQA expert. Note that more weight is put on the science expert instead. D) SEEK with SciQ expert.

8 Conclusion & Future Work

We introduced SEEK, a modular domain-specific mixture of expert models for multiple choice question answering. We demonstrated that smaller specialized learners in aggregate can retain similar performance while drastically reducing parameter counts. However, the performance is worse, even when adjusted for the log-size of the models. However, since the model is smaller (70% of its teacher size), the computational cost will also be much less. Thus, it may be important to quantify how much of an accuracy drop is permissible after distillation and ensembling the models.

The results above show that the pipeline laid out for this paper is effective at distilling then routing to a mixture of experts. Unsurprisingly, experts were able to handle questions from other similar sources in routing. This was shown in our ablation studies. Further exploration is necessary in the fine-tuning of various hyperparameters, especially those related to distillation, like the alpha value in the combined loss, and the fraction of layers copied.

It is expected that future work and more precise training experiments can yield even larger gains in the accuracy vs parameter trade off. Experiments on other models, whether other fine-tuned expert models or different model types, would allow this type of distill-and-ensemble work to apply to general question-answering and reasoning past basic multiple choice questions. For example, the GPT or T5 could yield more interesting results. However, due to compute and time limitations, we were only able to use BERT.

References

[1] Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-Distill: Developing Small, Fast and Effective Pretrained Language Models for Domains. arXiv preprint arXiv:2106.13474.

[2] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022. Mixture-of-Experts with Expert Choice Routing. arXiv preprint arXiv:2202.09368.

[3] Kristen Howell, Jian Wang, Akshay Hazare, Joseph Bradley, Chris Brew, Xi Chen, Matthew Dunn, Beth Ann Hockey, Andrew Maurer, and Dominic Widdows. 2022. Domain-specific knowledge distillation yields smaller and better models for conversational commerce. In Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 151-160, Online. Association for Computational Linguistics.

[4] Charith Peris, Lizhen Tan, Thomas Gueudre, Turan Gojayev, Pan Wei, and Gokmen Oz. Knowledge distillation transfer sets and their impact on downstream NLU tasks. 2022. In EMNLP 2022.

[5] OpenAI. GPT-4 Technical Report. 2023. arXiv preprint arXiv:2303.08774

[6] Victor Sanh, et al. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108

[7] Noam Shazeer, et al. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, arXiv preprint arXiv:1701.06538

[8] Dmitry Lepikhin, et al. 2020. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, arXiv preprint arXiv:2006.16668

[9] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. 2021. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

[10] William Fedus, et al. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint arXiv:2101.03961.

[11] Ziheng Wang, et al. 2021. Structured Pruning of Large Language Models. arXiv preprint arXiv:1910.

[12] Suyog Gupta, et al. 2015. Deep learning with limited numerical precision. ICML.

[13] Johannes Welbl, Nelson F. Liu. Matt Gardner. 2017. SciQ: Crowdsourcing Multiple Choice Science Questions. arXiv:1707.06209v1

[14] Pal, Ankit and Umapathi, Logesh Kumar and Sankarasubbu, Malaikannan. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. Proceedings of the Conference on Health, Inference, and Learning. 248–260.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457v

[16] Ling, Wang and Yogatama, Dani and Dyer, Chris and Blunsom, Phil. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. ACL.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805

[18] Beltagy, Iz and Lo, Kyle and Cohan, Arman. 2019. SciBERT: A Pretrained Language Model for Scientific Text. Association for Computational Linguistics

[19] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding. ArXiv:2105.00377

[20] Kaplan et al., 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361