Beyond Answers: Transferring Reasoning Capabilities to Smaller LLMs Using Multi-Teacher Knowledge Distillation

Yijun Tian*
yijun.tian@nd.edu
University of Notre Dame
USA

Yikun Han*
yikunhan@umich.edu
University of Michigan
USA

Xiusi Chen*
xchen@cs.ucla.edu
University of California, Los Angeles
USA

Wei Wang
weiwang@cs.ucla.edu
University of California, Los Angeles
USA

Nitesh V. Chawla
nchawla@nd.edu
University of Notre Dame
USA

Abstract

Transferring the reasoning capability from stronger large language models (LLMs) to smaller ones has been quite appealing, as smaller LLMs are more flexible to deploy with less expense. Among the existing solutions, knowledge distillation stands out due to its outstanding efficiency and generalization. However, existing methods suffer from several drawbacks, including limited knowledge diversity and the lack of rich contextual information. To solve the problems and facilitate the learning of compact language models, we propose TinyLLM, a new knowledge distillation paradigm to learn a small student LLM from multiple large teacher LLMs. In particular, we encourage the student LLM to not only generate the correct answers but also understand the rationales behind these answers. Given that different LLMs possess diverse reasoning skills, we guide the student model to assimilate knowledge from various teacher LLMs. We further introduce an in-context example generator and a teacher-forcing Chain-of-Thought strategy to ensure that the rationales are accurate and grounded in contextually appropriate scenarios. Extensive experiments on six datasets across two reasoning tasks demonstrate the superiority of our method. Results show that TinyLLM can outperform large teacher LLMs significantly, despite a considerably smaller model size. The source code is available at: https://github.com/YikunHan42/TinyLLM.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**.

Keywords

Knowledge Distillation, Large Language Models, Knowledge Reasoning

*Equally contributed.

WSDM '2025, March 10-14, 2025, Hannover, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V. Chawla. 2024. Beyond Answers: Transferring Reasoning Capabilities to Smaller LLMs Using Multi-Teacher Knowledge Distillation. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have recently taken over various domains and web applications, including society [46], education [72], and recommendations [69]. Although cutting-edge language models like GPT-4 and Claude-2 have shown remarkable capability in producing coherent and contextually appropriate text, their smaller counterparts often fall short, especially in tasks that demand sophisticated reasoning and a deep level of understanding [64]. This discrepancy has been unveiled as the well-known scaling law of LLMs, which suggests a correlation between model size and reasoning, linguistic, and generalization capabilities [29]. However, deploying these colossal models in a real-world setting poses significant challenges due to their computational requirements and resource demands, underscoring the importance of building efficient, smaller models that retain the power of their larger counterparts. Previous studies have shown that knowledge distillation is an instrumental tool in mitigating the performance gap between larger LLMs and smaller ones [22, 60]. Examples of effective distillation methods include DistilBERT [48], Alpaca [52] and Vicuna [77].

However, existing methods suffer from two major drawbacks: (1) **Limited Knowledge Diversity**: Current research predominantly employs a single-teacher approach, which confines the learning scope of the student model to the knowledge derived from its own training and architecture designs [21, 32, 40, 62]. This restricts the student model to a single perspective, potentially overlooking the diverse problem-solving strategies and reasoning capabilities exhibited by different models, limiting its breadth and depth of understanding. (2) **Lack of Rich Contextual Information**: While rationales play a vital role in effective reasoning [30, 65], current research primarily focuses on leveraging ground truth labels, which indicate the correct answer but do not provide insights into the reasoning and thought process behind that answer. In other words, learning the ground truth labels exclusively failed to capture the nuanced decision-making processes of the teachers, which are crucial for tasks requiring complex reasoning and interpretation.

To solve these issues, we propose TinyLLM, a paradigm that facilitates the learning of a small student LLM by distilling knowledge from multiple large teacher LLMs with rationale guidance. Specifically, TinyLLM mitigates the limited knowledge diversity issue by involving multiple teacher models as *co-advisors*, which introduces a richer, varied knowledge source for the student to learn from. To fully exploit each teacher model and mitigate the lack of rich contextual information problem, TinyLLM asks the teacher for the credible rationales to support the answers, thereby providing the student with a deeper understanding of the problem-solving process. By learning from multiple teachers, the student model can inherit a broader range of skills and knowledge, leading to better generalization capabilities. In addition, to ensure the rationales are grounded in contextually appropriate scenarios and reflect the true underlying reasoning procedure, TinyLLM features an in-context example generator and a teacher-forcing Chain-of-Thought strategy, making the teachers understand the task through demonstrations and therefore generate the accurate rationales.

To thoroughly evaluate our approach, we conduct experiments on six datasets in commonsense and biomedical reasoning tasks. The results show that the usage of our paradigm enhances performance by **+5.07%** to **+15.69%** compared to full fine-tuning. Compared to the teacher models, TinyLLM achieve superior performance improvement, e.g., up to **+23.40%** with significantly smaller model size, e.g., **1.1%** to **26.0%**. Furthermore, compared to the state-of-the-art distillation method, we improve the performance by **+10.00%** to **+11.79%** across different model sizes. In addition, we perform efficiency analyses, ablation studies, parameter sensitivities, and case studies to demonstrate and validate the effectiveness of the proposed method.

To summarize, our main contributions are as follows:

- We identify two critical limitations in the existing knowledge distillation landscape for LLMs: 1) limited knowledge diversity and 2) lack of rich contextual information.
- To solve these two problems, we propose TinyLLM, a novel knowledge distillation paradigm to learn a small student LLM from multiple large teacher LLMs.
- TinyLLM encompasses several innovative designs including an in-context example generator, a teacher-forcing Chain-of-Thought strategy, and a joint learning objective from various teachers.
- Extensive experiments validate the superiority of TinyLLM across six datasets and two reasoning tasks, with performance improving by up to **+15.69%** compared to full fine-tuning, up to **+23.40%** compared to teacher models, and up to **+11.79%** compared to state-of-the-art. In addition, TinyLLM holds a significantly smaller model size, e.g., **1.1%** to **26.0%** compared to the teachers.

2 Related Work

In this section, we review existing work including large language models, chain of thought, and knowledge distillation.

Large Language Models. Recent advancements have seen the proposal of various Large Language Models (LLMs) [5, 13, 44, 56, 57], which have showcased remarkable performance across a spectrum of tasks [37, 38, 49, 51, 66]. Central to these developments is question answering, a task that necessitates intricate reasoning and comprehensive understanding skills for text interpretation and generating suitable responses to queries [8, 39, 54, 78]. Despite their formidable learning capabilities, LLMs encounter limitations in accurately capturing factual knowledge and are prone to producing unsubstantiated responses [3, 27, 76]. Moreover, the extensive number of parameters within LLMs complicates their adaptation for downstream tasks [50, 68]. To mitigate these challenges, several approaches aim to lessen the dependency on intensive training and reduce computational costs [23, 31, 33]. For example, Prompt Tuning [17, 31, 36, 61] employs soft prompts to adapt pre-trained LLMs for specific tasks.

Chain of Thought. Recently, the use of rationales generated by LLMs has become a popular trend, setting itself apart from the traditional reliance on human-generated rationales [19, 24]. Previously, human rationales have been used for model regularization [47], as additional inputs for predictions [45], and to improve model performance [6, 15, 18, 25, 43, 71, 74, 75]. They also serve as gold standard labels for generating similar rationales to enhance interpretability [14, 42, 59, 67]. However, the cost of human rationales limits their widespread use. On the other hand, modern LLMs can generate high-quality reasoning steps to explain their predictions [30, 65], improving performance in few-shot or zero-shot learning [30, 62, 65] and serving as self-improvement data [26, 73]. However, LLMs' size hinders their deployment in practice. Correspondingly, recent research explores leveraging generated rationales for training smaller, task-specific models with minimal computational and memory overhead [21, 32, 40, 63]. For example, PINTO [62] presents an LLM pipeline that rationalizes via prompt-based learning. However, they still rely on an LLM for rationale generation at test-time, not fully addressing deployment challenges. In this work, we propose a multi-task learning paradigm with superior chain-of-thought reasoning capabilities, avoiding the dependence on teacher models during the test phase.

Knowledge Distillation. Recent LLMs such as PaLM 540b [2, 10] present formidable challenges in terms of inference and fine-tuning, primarily attributable to the extensive computational resources they necessitate. This dependency and requirement on computation underscore the pivotal role of knowledge distillation [1, 9, 16, 20, 53, 55, 70], which has proved to alleviate the resource limitation by training a smaller model to mimic the large teacher model. In addition, the employment of the Chain-of-Thought paradigm has facilitated the generation of deliberative reasoning samples from the teacher models [21], allowing student models to concurrently grasp both the answers and the intricate reasoning of the teachers. This process strengthens the student model through multi-task learning endeavors [22]. Correspondingly, efforts have been made to generate various rationales for each inquiry, seeking to ensure consistency in the predictions [7, 35]. However, depending on the rationales from a single teacher model introduces bias and compromises thoroughness, thus not fully tapping into the capabilities of multi-teacher learning paradigms. These paradigms, by their very nature, hold the potential to enhance knowledge diversity, an aspect that remains largely unexplored.

3 Method

In this section, we formally present TinyLLM to resolve the challenges described in the Introduction. In particular, we start by describing the preliminary. Next, we introduce the details of TinyLLM by first obtaining rationales from multiple teachers, and then learning a small student using the obtained rationales. The TinyLLM pipeline is shown in Figure 1.

3.1 Preliminary

Multiple Choice Question Answering. A k-way multiple choice question answering (MCQA) is defined as follows: Given a question Q_i, a set of candidate answer options $O_i = \{O_{i1}, O_{i2}, ..., O_{ik}\}$, the model is tasked with selecting the correct answer from the set O_i, such that the selected answer aligns the ground truth label A_i.

Knowledge Distillation. The knowledge distillation process begins with the teacher model, denoted as T parameterized by θ_T, which has been pre-trained on a large corpus. Later, the student model, S, with parameter θ_S, is tasked with distilling knowledge directly from T, leveraging the strong capabilities of T. Correspondingly, the objective function can be formulated as: $\mathcal{L} = \ell(S, T)$, where ℓ indicates the learning function, e.g., cross-entropy loss between the prediction output of the student and the target output generated by the teacher.

3.2 Obtaining Rationales from Teachers

In-context Example Generator. To enable that rationales generated by teacher models are grounded in contextually appropriate scenarios, we introduce an optional in-context example generator. This tool provides additional context by generating examples that include both questions and corresponding rationales in a zero-shot setting, enhancing the teacher models' comprehension of task-specific nuances. Each generated example, covering diverse aspects of the dataset, is concatenated with the actual question to form a comprehensive input. By including a range of in-context examples, we create a richer input structure that aids teacher LLMs in generating higher-quality rationales. This approach allows the student model to learn not only from correct answers but also from the underlying reasoning, thereby enhancing both the accuracy and interpretability of the distilled model.

Teacher-forcing Chain-of-Thought. In addition, we design a teacher-forcing strategy to ensure the validity of the rationales. Compared to existing methods that simply employ regular chain-of-thought (CoT) mechanisms [30, 65], wherein an LLM is prompted with sets of questions and options $\{Q_i, O_i\}$ to elicit rationales R_i directly, TinyLLM posits a distinct advantage in integrating the correct answer A_i into the input. We hypothesize that the inclusion of A_i alongside Q_i and O_i facilitates a more nuanced understanding of the input context and the correct logical rationales leading to the answer, thereby facilitating a more informed and accurate generation process. Specifically, we consider the concatenation of questions, options, and answers $\{Q_i, O_i, A_i\}$ as the input to LLMs.

Rationales from Multiple Teachers. Given M teachers, TinyLLM pioneers the usage of a multi-teacher architecture in which each teacher T^m is an LLM. In particular, the rationale R_i^m produced by a specific teacher model θ_{T^m} for the ith question is derived using the question Q_i, options O_i, correct answer A_i, and in-context examples P_i. The process is formalized as follows:

$$R_i^m = T^m(Q_i, O_i, A_i, P_i; \theta_{T^m}). \tag{1}$$

3.3 Learning a Small Student

A straightforward strategy to incorporate rationales as supervision is to append each rationale R_i^m generated by the teacher models as supplementary input to the student model, along with the question Q_i and options O_i. However, this method faces challenges due to limitations in computational resources at the inference stage, especially because rationales must be pre-generated for every data sample in both training and test sets [62]. To overcome this issue, we employ rationales as a form of supervisory signal during the training process to develop a model that is adept at generating its explanations. Subsequently, this trained model can be utilized on the test set, eliminating the need for pre-generated rationales to facilitate accurate reasoning. Specifically, TinyLLM integrates rationales from multiple teacher models into a unified multi-task instruction tuning framework. This necessitates the assignment of a unique prefix p for distinguishing between learning tasks from different teachers. The student model is trained not only to predict labels but also to generate rationales akin to those produced by the teachers. Accordingly, the overall loss function $\mathcal{L}$ is as follows:

$$\mathcal{L} = \mathcal{L}_A + \sum_{m=1}^{M} \alpha^m \mathcal{L}_{T^m}, \tag{2}$$

where $\mathcal{L}_A$ denotes the objective of learning from ground truth answers, $\mathcal{L}_{T^m}$ indicates the objective of learning from m-th teacher, α^m is the importance weight for T^m, and M is the number of teacher LLMs. Formally, $\mathcal{L}_A$ and $\mathcal{L}_{T^m}$ are defined as follows:

$$\mathcal{L}_A = \frac{1}{N} \sum_{i=1}^{N} \ell(S(Q_i, O_i, p_A; \theta_S), A_i), \tag{3}$$

$$\mathcal{L}_{T^m} = \frac{1}{N} \sum_{i=1}^{N} \ell(S(Q_i, O_i, p_m; \theta_S), R_i^m), \tag{4}$$

where N is the number of data samples, ℓ indicates the cross-entropy loss between the predicted and target tokens. Here $\mathcal{L}_A$ encourages the student S to generate ground truth answer A_i by minimizing the difference between it and the student output given the question Q_i, options O_i, and instruction prefix p_A for generating answers. On the other hand, $\mathcal{L}_T^m$ facilitates the student S to mimic the reasoning capability of teacher T^m by learning from its rationale R_i^m, with the guidance of instruction prefix p_m for T^m.

4 Experiments

In this section, we rigorously test TinyLLM against a series of empirical benchmarks across varied datasets and reasoning tasks. In addition, we conduct efficiency analyses, ablation studies, parameter sensitivity tests, and case studies to demonstrate the effectiveness and superiority of our method.

4.1 Experimental Setup

Datasets. For the task of commonsense reasoning, we use OpenBookQA (OBQA) [41], The AI2 Reasoning Challenge (ARC) [12],

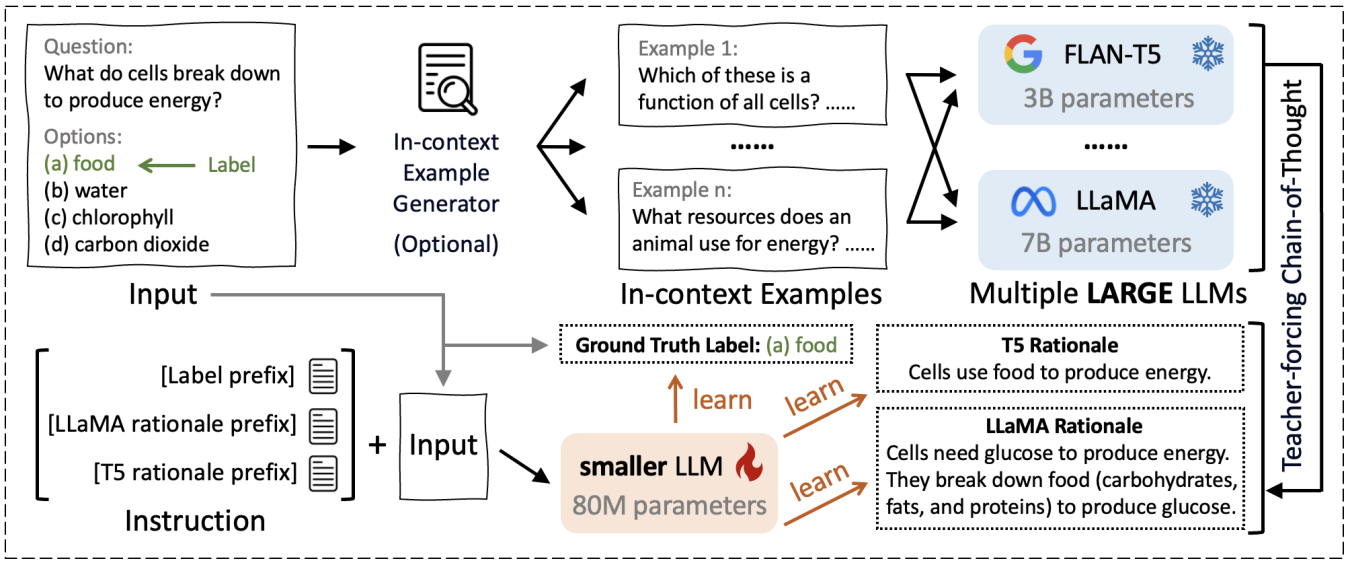

Figure 1: Pipeline of TinyLLM: Given an input question, we first generate in-context examples and obtain rationales from multiple large LLMs via a teacher-forcing Chain-of-Thought strategy. Later, a small student LLM is trained to integrate rationales from different teachers via multi-task instruction tuning, along with the ground truth label.

Physical Interaction Question Answering (PIQA) [4], and RiddleSense (Riddle) [34]. For the task of biomedical reasoning, we consider PubMedQA (PQA) [28] and BioASQ [58].

Baselines. We benchmark TinyLLM against the teacher's performance and various baseline methods, including Inference-only that only leverage the pre-trained model for evaluation without training, and multiple fine-tuning methods that provide further adaptation. In particular, we consider LoRA [23], full fine-tuning, and the PINTO method [62] for the fine-tuning methods. We also compare TinyLLM with various knowledge distillation strategies. To illustrate, we include standard KD [20] that enforces the student to mimic the teacher's labels and the Distill-step-by-step method [22] that leverage rationales.

Implementation Details. For all distillation baselines and TinyLLM, we set the learning rate to 5×10^{-5}, batch size to 8, maximum input length to 1024, and epoch to 1. For Distill-step-by-step and TinyLLM, the trade-off weights α_{T_n} are explored within {0.01, 0.1, 0.5, 1, 2, 3}. We report the best result for Distill-step-by-step by leveraging different teacher models. For the choice of LLMs, we use FLAN-T5 [11] small (80M), base (250M), and large (780M) as the student, and FLAN-T5 xlarge (3B) and LLaMA 2-chat [57] (7B) as teachers. Experiments are conducted on four NVIDIA H100 Tensor Core GPUs.

4.2 Performance Comparison

Comparison to Baselines Methods. We conducted a thorough evaluation of our method, TinyLLM, across six diverse datasets spanning two distinct reasoning tasks: commonsense reasoning and biomedical reasoning. The detailed results are presented in Table 1, offering a comprehensive view of the performance landscape across different model sizes and methodologies.

From the data, several key insights emerge. Notably, while full fine-tuning is theoretically capable of maximizing parameter adjustments and should, in principle, yield the best results, it does not consistently outperform more parameter-efficient methods such as LoRA. This outcome suggests that simply having a larger number of adjustable parameters does not guarantee improved generalization or performance, especially when the fine-tuning process may inadvertently overfit to the training data or fail to capture nuanced task-specific knowledge.

In stark contrast, TinyLLM demonstrates consistent and substantial performance improvements across all datasets and model sizes, underscoring the robustness and adaptability of our approach. Specifically, the quantitative gains achieved by TinyLLM are noteworthy: we observe an average performance boost of **+15.69%**, **+11.55%**, and **+5.07%** for student models with 80M, 250M, and 780M parameters, respectively, compared to full fine-tuning. These improvements are significant, particularly considering that larger models often exhibit diminishing returns on performance gains, making these results even more compelling.

Furthermore, when compared to the state-of-the-art distillation method, Distill-step-by-step, TinyLLM achieves impressive relative improvements of **+10.00%**, **+10.32%**, and **+11.79%** for the 80M, 250M, and 780M student models, respectively. These results highlight the effectiveness of our method in distilling knowledge from large models into smaller, more efficient student models. The consistent superiority of TinyLLM across different model sizes and datasets demonstrates the advantages of our design choices, including the integration of intermediate rationale guidance and the strategic balance between mimicking teacher outputs and adapting to new data.

Comparison to Teachers. In addition to outperforming baseline methods, TinyLLM also exhibits superior performance compared to

Table 1: Overall experimental results. The best results across different datasets and LLM sizes are highlighted in bold. Δ_{FF} and $\Delta_{Distill}$ represent the relative performance improvement of TinyLLM to Full Fine-Tuning and Distill-step-by-step, respectively. Accuracy is used as the evaluation metric.

Setting	Method	Commonsense Reasoning				Biomedical Reasoning		Total
		OBQA	ARC	PIQA	Riddle	PQA	BioASQ	
3B/7B Teacher	FLAN-T5 xlarge	69.20	68.24	58.43	53.73	71.50	65.85	64.49
	LLaMA 2	58.60	45.90	78.80	47.65	54.50	73.75	59.87
80M Student Size: 2.7%/1.1%	Inference	16.60	19.31	20.78	13.33	38.00	47.97	26.00
	PINTO	46.40	26.87	48.10	25.29	60.00	80.49	47.86
	LoRA	37.80	27.12	39.93	39.80	53.75	78.05	46.08
	Full Fine-tuning	41.60	27.47	42.33	42.75	56.25	78.86	48.21
	Standard KD	45.80	29.53	49.29	36.27	58.00	81.30	49.43
	Distill-step-by-step	46.40	30.47	50.38	36.67	59.00	81.30	50.70
	TinyLLM	**49.40**	**33.05**	**53.65**	**51.18**	**62.00**	**85.37**	**55.78**
	Δ_{FF}	↑ 18.75%	↑ 20.31%	↑ 26.74%	↑ 19.72%	↑ 10.22%	↑ 8.26%	↑ 15.69%
	$\Delta_{Distill}$	↑ 6.47%	↑ 8.47%	↑ 6.49%	↑ 39.57%	↑ 5.08%	↑ 5.01%	↑ 10.00%
250M Student Size: 8.3%/3.6%	Inference	31.00	23.00	30.47	30.78	48.00	57.72	36.83
	PINTO	50.40	38.63	52.12	34.90	61.75	82.93	53.46
	LoRA	51.40	37.25	47.66	53.14	62.00	82.93	55.73
	Full Fine-tuning	56.60	38.88	47.55	52.55	64.75	89.43	58.29
	Standard KD	55.40	43.69	55.93	42.94	64.25	86.18	58.07
	Distill-step-by-step	56.80	43.86	56.37	45.69	64.75	86.18	58.94
	TinyLLM	**64.20**	**48.50**	**60.17**	**60.78**	**66.25**	**90.24**	**65.02**
	Δ_{FF}	↑ 13.43%	↑ 24.74%	↑ 26.54%	↑ 15.66%	↑ 2.32%	↑ 0.91%	↑ 11.55%
	$\Delta_{Distill}$	↑ 13.03%	↑ 10.58%	↑ 6.74%	↑ 33.03%	↑ 2.32%	↑ 4.71%	↑ 10.32%
780M Student Size: 26.0%/11.1%	Inference	50.40	51.07	51.90	39.80	64.25	63.41	53.47
	PINTO	62.20	52.10	57.13	42.94	70.00	84.55	61.49
	LoRA	64.00	57.77	57.02	68.63	70.25	86.18	67.31
	Full Fine-tuning	71.20	62.92	58.43	68.82	70.25	90.24	70.31
	Standard KD	65.80	56.05	60.72	52.94	70.00	86.99	65.42
	Distill-step-by-step	66.80	57.42	61.37	53.92	70.00	86.99	66.08
	TinyLLM	**74.40**	**64.29**	**67.90**	**70.98**	**73.00**	**92.68**	**73.88**
	Δ_{FF}	↑ 4.49%	↑ 2.18%	↑ 16.21%	↑ 3.14%	↑ 3.91%	↑ 2.70%	↑ 5.07%
	$\Delta_{Distill}$	↑ 11.38%	↑ 11.96%	↑ 10.64%	↑ 31.64%	↑ 4.29%	↑ 6.54%	↑ 11.79%

the original teacher models. This is particularly remarkable given the significant difference in model size. For instance, a 780M parameter student model trained with TinyLLM achieves an impressive average performance score of 73.88 across various datasets. This score represents a substantial improvement of **+14.56%** over the performance of the much larger 3B parameter teacher model and an even more striking **+23.40%** improvement over the 7B parameter teacher.

These results not only demonstrate the efficacy of TinyLLM in transferring knowledge but also suggest that our approach enables smaller models to generalize better and perform tasks more effectively than their larger counterparts. Moreover, the efficiency gains are particularly pronounced when considering smaller student models, such as the 250M parameter model, which manages to surpass both the 3B and 7B parameter teachers with relative improvements of **+0.82%** and **+8.60%**, respectively. This achievement is all the more noteworthy given that the 250M model operates with only **8.3%** and **3.6%** of the teacher models' parameters, showcasing the remarkable efficiency of TinyLLM in compressing and enhancing the performance of smaller models without compromising accuracy.

4.3 Efficiency Analysis of Training Set Size in Knowledge Transfer

Advantage Over Standard Knowledge Distillation. To thoroughly assess the efficiency and effectiveness of our proposed method, TinyLLM, we conducted a series of experiments that evaluate its performance across varying training set sizes, particularly in comparison to the state-of-the-art Distill-step-by-step method. This analysis is crucial for understanding how well our model performs not only with full datasets but also under conditions of limited training data, which is often a real-world constraint.

As illustrated in Figure 2, TinyLLM consistently outperforms the Distill-step-by-step method across all tested ratios of the training data, demonstrating its robustness and efficiency. This superior

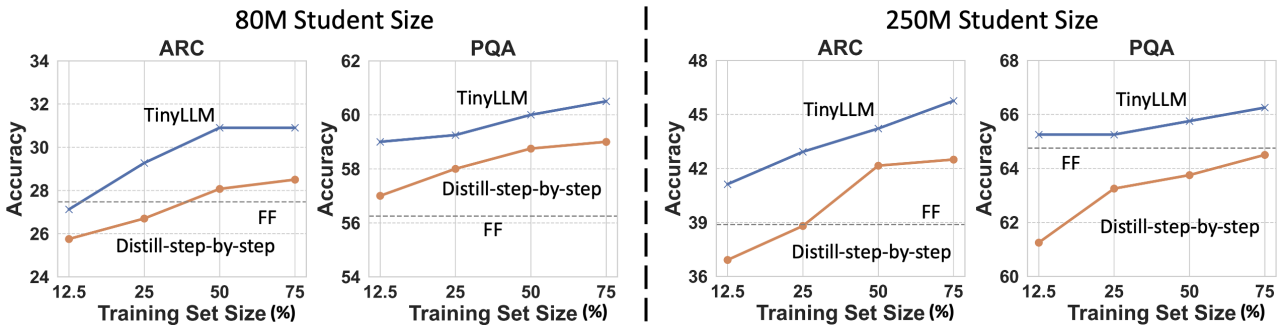

Figure 2: A comparative analysis of TinyLLM against the state-of-the-art Distill-step-by-step method using 80M and 250M FLAN-T5 model architectures across various training set sizes. Dotted line indicates the full fine-tuning (FF) using 100% dataset. It is evident that TinyLLM consistently surpasses the performance of both Distill-step-by-step and full fine-tuning. Notably, TinyLLM achieves this superior accuracy while employing substantially fewer training examples.

performance is particularly evident even as we reduce the size of the training set. The trend indicates that TinyLLM can effectively leverage smaller amounts of data to achieve comparable or better results than methods that require larger training datasets.

A striking example of this efficiency is observed in the context of the PQA dataset. Here, TinyLLM, when trained with only 12.5% of the available training data, not only meets but often exceeds the performance levels achieved by Distill-step-by-step, which utilizes the full training set. This significant reduction in required training data, without sacrificing performance, underscores the practical advantages of TinyLLM, particularly in scenarios where data is scarce or expensive to obtain.

This finding holds across different model sizes, including both 80M and 250M parameter student models. The consistent outperformance of TinyLLM in these cases suggests that our method is highly effective in transferring knowledge efficiently, making it a versatile tool for various applications, regardless of the model size. This efficiency could be particularly beneficial in resource-constrained environments, where computational resources and training time are limited.

Outperforming Full Fine-Tuning. Beyond its advantages over standard knowledge distillation methods, TinyLLM also demonstrates significant improvements compared to traditional full fine-tuning approaches, even when using the entire dataset. This comparison further highlights the efficiency of our method in terms of both data usage and computational resources.

In particular, when training a 250M parameter model with TinyLLM on the ARC and PQA datasets, as well as training an 80M parameter model on the PQA dataset, only 12.5% of the full training data is necessary to surpass the benchmarks established by full fine-tuning. This result is remarkable, as it shows that TinyLLM can achieve superior performance with a fraction of the data typically required for full fine-tuning, significantly reducing the computational cost and time required for training.

Moreover, in the case of the 80M parameter model trained on the ARC dataset, TinyLLM achieves higher accuracy than full fine-tuning even with a 75% reduction in training samples. This result is particularly important, as it demonstrates that TinyLLM not only reduces the need for extensive data but also maintains or improves model performance, making it an attractive alternative to full fine-tuning.

Table 2: Impact of in-context examples, the teacher-forcing strategy and contributions of various teachers.

Variant	Commonsense				Biomedical	
	OBQA	**ARC**	**PIQA**	**Riddle**	**PQA**	**BioASQ**
w/o in-context	73.20	63.09	66.27	69.22	70.75	86.99
w/o LLaMA	73.00	62.32	66.70	68.82	69.25	87.81
w/o T5	73.80	61.80	66.49	68.63	69.50	88.62
w/o diverse teachers	73.80	62.49	66.81	68.82	70.00	89.43
w/o teacher-forcing	73.80	60.94	65.94	69.02	70.25	90.24
TinyLLM	**74.40**	**64.29**	**67.90**	**70.98**	**73.00**	**92.68**

4.4 Ablation Study

To provide a comprehensive evaluation of our proposed method, TinyLLM, we conducted an ablation study in Table 2 to validate the contributions of key components in enhancing the reasoning capabilities of the distilled LLM. Specifically, we focused on assessing the impact of the in-context example generator, the use of rationales from multiple teacher models, and the teacher-forcing strategy. By isolating these components, we aim to understand their individual and collective contributions to the overall performance of TinyLLM. We designed four ablation variants of TinyLLM, each purposefully modified to test the significance of specific components:

- **w/o in-context** removes the use of in-context examples during rationale generation. This variant is crucial for evaluating the role of in-context examples in guiding the student model to produce more accurate, relevant, and contextually appropriate rationales.
- **w/o LLaMa** and **w/o T5** exclude the rationale supervision provided by the respective teacher models during the distillation process. By removing the influence of either LLaMa or T5, these variants help us understand the individual contributions of each teacher's rationales.
- **w/o diverse teachers** excludes the weaker teacher model, instead generating multiple rationales from the stronger teacher model. This variant is designed to test the effectiveness of using a diverse set of teachers, as opposed to relying on a single, potentially more robust teacher.

- **w/o teacher-forcing** eliminates the teacher-forcing strategy during rationale generation. Teacher-forcing is a technique where the model is trained using the correct output (from the teacher) as input for the next step, rather than its own previous prediction. By removing this strategy, we aim to assess its effectiveness in helping the student model generate higher-quality rationales.

Table 2 provides a comparative analysis between each ablation variant and the complete TinyLLM model. From the data presented in Table 2, several important observations emerge: (1) TinyLLM consistently outperforms all ablation variants, demonstrating that the integration of all components—rationales from multiple teachers, in-context examples, and the teacher-forcing strategy—collectively contributes to the model's superior performance. This result highlights the synergistic effect of these components, where their combination leads to significant improvements in the reasoning capabilities of the distilled LLM. (2) There is no substantial performance gap between the different ablation variants, suggesting that while each component contributes to the model's performance, none of them alone is solely responsible for the observed improvements. This finding implies a balanced importance among the components, with each playing a complementary role in refining the model's reasoning abilities.

In terms of computational overhead, while TinyLLM demonstrates significant improvements in reasoning capabilities, it is essential to consider the costs introduced by using multiple teacher models and the in-context example generator. Although all variant methods in Table 2 take the same time for student model distillation, inference times for generating rationales differ among teacher models, impacting overall computational cost. Specifically, LLaMA 2 7B requires more inference time than FLAN-T5 xlarge on the same dataset, indicating that the choice of teacher model affects scalability and practicality. For resource-constrained applications, selecting teacher models with lower computational demands—such as APIs like OpenAI's GPT-3.5 Turbo or GPT-4—offers a balanced trade-off between inference speed and performance, making TinyLLM adaptable to various deployment environments.

4.5 Parameter Sensitivity

To thoroughly evaluate the robustness and adaptability of our proposed model, we conducted parameter sensitivity experiments on two distinct datasets: ARC for commonsense reasoning and PQA for biomedical reasoning. These experiments are critical for understanding how different parameter settings, particularly the trade-off weights α^{T5} and α^{LLaMA}, influence the model's performance across various tasks. The results of these sensitivity analyses are depicted in Figure 3.

In this analysis, our primary focus is on exploring the effects of varying the trade-off weights α^{T5} and α^{LLaMA}, which balance the influence of rationales provided by the T5 and LLaMA teacher models, respectively. This exploration reveals the model's adaptability and how it responds to different parameter configurations, offering insights into the optimal settings for maximizing performance across different tasks and datasets. From the results shown in Figure 3, our key observations are as follows.

Optimal Parameter Variability Across Datasets and Tasks. It is evident that the optimal parameter settings for α^{T5} and α^{LLaMA}

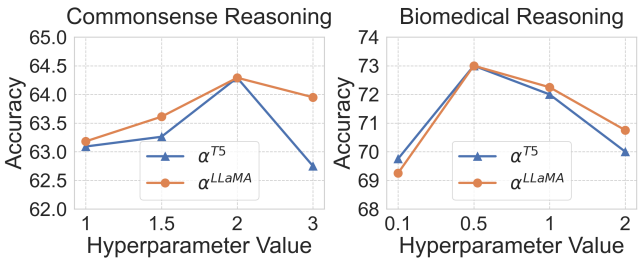

Figure 3: Performance w.r.t. different values of weight α.

vary depending on the dataset and the nature of the reasoning task. For biomedical reasoning tasks, such as those found in the PQA dataset, questions tend to be lengthy and complex, often requiring deep comprehension and synthesis of detailed information. In these cases, the impact of rationales from teacher models is somewhat diminished, as the complexity of the content overshadows the direct utility of rationale guidance. As a result, a smaller value of α is sufficient to achieve optimal performance. On the other hand, commonsense reasoning tasks, exemplified by the ARC dataset, typically involve more concise and straightforward questions. In these scenarios, the rationales provided by teacher models play a more critical role in guiding the student model's reasoning process, leading to the need for a larger α value to fully leverage this guidance.

Effect of Increasing α Values on Performance. Increasing the values of α^{T5} and α^{LLaMA} generally leads to improved performance, which can be attributed to the enhanced Chain-of-Thought reasoning capabilities of the student model. By placing greater emphasis on the rationales from teacher models, the student model benefits from a richer multi-task learning experience, which enhances its prediction capabilities. However, this improvement has its limits. Excessively high values of α can degrade performance by causing the model to overly focus on reasoning processes at the expense of prediction accuracy. This shift in focus may lead to a situation where the model becomes too reliant on rationales, potentially overfitting to the rationale structure rather than learning to generalize effectively from the data.

Differential Sensitivity of α^{T5} and α^{LLaMA} Across Datasets. The sensitivity of the parameters α^{T5} and α^{LLaMA} varies between different datasets, reflecting the diverse contributions of the T5 and LLaMA teacher models to the reasoning process. For commonsense reasoning tasks, we observe that α^{T5} is more sensitive than α^{LLaMA}. This suggests that the rationales generated by the T5 model are particularly valuable for tasks involving straightforward reasoning, where the logical flow and concise reasoning provided by T5 are more impactful. Conversely, in biomedical reasoning tasks, the sensitivities of α^{T5} and α^{LLaMA} are more balanced, indicating that both teacher models offer valuable, albeit different, insights that complement each other in processing the complex and detailed content typical of biomedical texts. This balance highlights the importance of leveraging diverse teacher models to capture the full range of reasoning required for different tasks.

Question: Which tool would be best to use to determine how long it takes a cup of water to boil? Options: (a) balance (b) hot plate (c) stopwatch (d) thermometer		**Question:** When cold air is at the tops of mountains, the air Options: (a) will flow toward areas of lower pressure in valleys. (b) will flow toward areas of higher pressure in valleys. (c) is considered to be less dense than warmer air. (d) is considered to be free of oxygen atoms.	
T5 Rationale: Boiling water is a state of water in which the temperature is above 100°C. Thermometer is a tool that measures temperature. **LLaMA Rationale:** What is the purpose of determining how long it takes a cup of water to boil? To measure the time it takes for water to reach its boiling point. Which tool would be best suited for this task? A stopwatch would be the best tool to use because it allows you to measure the time it takes for the water to reach its boiling point.		**T5 Rationale:** Cold air is at the tops of mountains, so it has less density. Denser air is less likely to flow. Cold air is at the tops of mountains, so it has less density. **LLaMA Rationale:** The cold air at the tops of mountains is denser than the warmer air in the valleys due to the lower temperature. As a result, the cold air will flow downward, toward areas of lower pressure, which are typically in the valleys.	
Answer	**Ground Truth:** (c) **T5:** (d), **LLaMA:** (c) **Full Fine-tuning:** (d) ✗ **Distill-step-by-step:** (d) ✗ **TinyLLM:** (c) ✓	**Answer**	**Ground Truth:** (a) **T5:** (c), **LLaMA:** (a) **Full Fine-tuning:** (b) ✗ **Distill-step-by-step:** (c) ✗ **TinyLLM:** (a) ✓

Figure 4: Case study of different models' prediction. Examples are selected from the ARC and PIQA datasets. In both cases, TinyLLM successfully generates the correct answer.

4.6 Case Study

To gain a more intuitive understanding of why TinyLLM consistently outperforms other models, we conducted case studies by comparing the predictions generated by different models. These case studies offer valuable insights into the specific scenarios where TinyLLM excels and highlight the advantages of our multi-teacher approach. Figure 4 presents two representative examples that we randomly selected from the ARC and PIQA datasets, illustrating the differences in model predictions and the underlying rationales.

In the first example, taken from the ARC dataset, we observe a significant discrepancy in the reasoning capabilities of the teacher models. The T5 model provides a completely incorrect rationale, leading to an incorrect prediction. In contrast, LLaMA generates a meaningful and accurate rationale, which correctly guides the prediction. Despite LLaMA's correct reasoning, the full fine-tuning method fails to produce the correct answer, likely due to overfitting or an inability to effectively integrate the rationale during the fine-tuning process. Similarly, the state-of-the-art Distill-step-by-step method also predicts incorrectly, likely because the distillation process, which leverages T5's reasoning, introduces noise that misguides the student model. This occurs even though T5 significantly outperforms LLaMA in overall accuracy on the ARC dataset (as shown in Table 1). However, TinyLLM demonstrates its robustness by correctly inferring the answer (a), effectively synthesizing the rationales from both T5 and LLaMA, and filtering out the noise introduced by the incorrect rationale from T5. This example highlights the strength of TinyLLM in balancing and integrating multiple sources of knowledge, leading to superior prediction accuracy.

5 Conclusion and Future Work

In this paper, we propose TinyLLM, a novel knowledge distillation paradigm to learn a small student LLM from multiple large teacher LLMs. TinyLLM involves several principled designs, such as learning contextually appropriate rationales using an in-context example generator, enabling the credibility of rationales with a teacher-forcing Chain-of-Thought strategy, and inheriting a wider range of knowledge from various teachers. Our extensive empirical evaluation and in-depth analysis, conducted across six datasets spanning two reasoning tasks, demonstrate that TinyLLM brings significant and consistent improvements by up to 15.69% over full fine-tuning, up to **+23.40%** over teacher models, and up to **+11.79%** over state-of-the-art. Moreover, TinyLLM holds a significantly smaller model size, e.g., **1.1%** to **26.0%** compared to the sizes of the teachers.

As future work, we envision several directions to further enhance TinyLLM. One potential avenue is to integrate a broader set of teacher LLMs to examine whether the student model can effectively learn from a wider, possibly incoherent, range of knowledge and reasoning styles, managing any conflicting rationales that may arise. Another promising direction involves testing API-based teacher models, such as OpenAI's GPT-3.5 Turbo or GPT-4, to explore the feasibility and impact of using accessible, high-quality models in the distillation process without requiring local computational resources. Additionally, we plan to use open-source embeddings to represent questions in each dataset, allowing us to select in-context examples based on the top nearest neighbors. This strategy could lead to more relevant examples, enabling us to assess whether a targeted in-context example selection improves the overall performance of the student model.

References

[1] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. Generalized Knowledge Distillation for Auto-regressive Language Models. In *The Twelfth International Conference on Learning Representations*.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *ACL*.

[4] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

[6] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *NeurIPS*.

[7] Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. MCC-KD: Multi-CoT Consistent Knowledge Distillation. In *EMNLP Findings*.

[8] Xiusi Chen, Jyun-Yu Jiang, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Wei Wang. 2024. MinPrompt: Graph-based Minimal Prompt Data Augmentation for Few-shot Question Answering. In *ACL*.

[9] Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *ICCV*.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PALM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research* (2023).

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416* (2022).

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

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[14] Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. 2022. Honest Students from Untrusted Teachers: Learning an Interpretable Question-Answering Pipeline from a Pretrained Language Model. *arXiv preprint arXiv:2210.02498* (2022).

[15] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. In *ICML*.

[16] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* (2021).

[17] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332* (2021).

[18] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In *ACL*.

[19] Peter Hase and Mohit Bansal. 2022. When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data. In *ACL Workshop*.

[20] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[21] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large Language Models Are Reasoning Teachers. *arXiv preprint arXiv:2212.10071* (2022).

[22] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *ACL Findings*.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

[24] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

[25] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large Language Models Can Self-Improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 1051–1068.

[26] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610* (2022).

[27] Z Ji, N Lee, R Frieske, T Yu, D Su, Y Xu, E Ishii, Y J Bang, A Madotto, and P Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* (2023).

[28] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *EMNLP*.

[29] J Kaplan, S McCandlish, T Henighan, et al. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[30] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *NeurIPS*.

[31] B Lester, R Al-Rfou, and N Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*.

[32] Liunian Harold Li et al. 2023. Symbolic Chain-of-Thought Distillation: Small Models Can Also "Think" Step-by-Step. *arXiv preprint arXiv:2306.14050* (2023).

[33] X L Li and P Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL*.

[34] Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge. In *ACL Findings*.

[35] Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2023. Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models. *arXiv preprint arXiv:2311.09214* (2023).

[36] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602* (2021).

[37] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards Safer Large Language Models through Machine Unlearning. *arXiv preprint arXiv:2402.10058* (2024).

[38] Zheyuan Liu, Xiaoxin He, Yijun Tian, and Nitesh V Chawla. 2024. Can we soft prompt LLMs for graph learning tasks?. In *WWW*.

[39] P Lu, S Mishra, T Xia, L Qiu, K-W Chang, S-C Zhu, O Tafjord, P Clark, and A Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *NeurIPS*.

[40] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching Small Language Models to Reason. *arXiv preprint arXiv:2212.08410* (2022).

[41] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

[42] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain Their Predictions. *arXiv preprint arXiv:2004.14546* (2020).

[43] Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students?. In *ACL*.

[44] C Raffel, N Shazeer, A Roberts, et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* (2020).

[45] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

[46] Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations. In *EMNLP Findings*.

[47] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations. *arXiv preprint arXiv:1703.03717* (2017).

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[49] Yucheng Shi, Shaochen Xu, et al. 2023. Mededit: Model Editing for Medical Question Answering with External Knowledge Bases. *arXiv preprint arXiv:2309.16035* (2023).

[50] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, a Large-Scale Generative Language Model. *arXiv preprint arXiv:2201.11990* (2022).

[51] Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning. *arXiv preprint arXiv:2402.04401* (2024).

[52] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. *GitHub repository* (2023).

[53] Yijun Tian, Shichao Pei, Xiangliang Zhang, Chuxu Zhang, and Nitesh V Chawla. 2023. Knowledge Distillation on Graphs: A Survey. *arXiv preprint arXiv:2302.00219* (2023).

[54] Yijun Tian, Huan Song, Zichen Wang, Haozhu Wang, Ziqing Hu, Fang Wang, Nitesh V Chawla, and Panpan Xu. 2024. Graph neural prompting with large language models. In *AAAI*.

[55] Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V Chawla. 2023. Learning MLPs on Graphs: A Unified View of Effectiveness, Robustness, and Efficiency. In *ICLR*.

[56] Hugo Touvron, Thibaut Lavril, et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[57] Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[58] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* (2015).

[59] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).

[60] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863* 1 (2023).

[61] Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R Lyu. 2022. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM joint European software engineering conference and symposium on the foundations of software engineering*. 382–394.

[62] Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. PINTO: Faithful Language Reasoning Using Prompt-Generated Rationales. In *ICLR*.

[63] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *EMNLP Findings*.

[64] J Wei, Y Tay, R Bommasani, et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682* (2022).

[65] Jason Wei, Xuezhi Wang, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

[66] W Wei, X Ren, J Tang, Q Wang, L Su, S Cheng, J Wang, D Yin, and C Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *WSDM*.

[67] Sarah Wiegreffe, Ana Marasovic, and Noah A Smith. 2021. Measuring Association Between Labels and Free-Text Rationales. In *EMNLP*.

[68] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[69] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2023. A Survey on Large Language Models for Recommendation. *arXiv preprint arXiv:2305.19860* (2023).

[70] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116* (2024).

[71] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *NAACL*.

[72] Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. Generating and Evaluating Tests for K-12 Students with Language Model Simulations: A Case Study on Sentence Reading Efficiency. In *EMNLP*.

[73] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping Reasoning with Reasoning. In *NeurIPS*.

[74] Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *EMNLP*.

[75] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. [n. d.]. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research* ([n. d.]).

[76] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023).

[77] Lianmin Zheng, Wei-Lin Chiang, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

[78] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-S Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering. *arXiv preprint arXiv:2101.00774* (2021).