Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application

CHUANPENG YANG, Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences, China
YAO ZHU*, Zhejiang University, China
WANG LU*, Tsinghua University, China
YIDONG WANG, Peking University, China
QIAN CHEN, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China
CHENLONG GAO, Institute of Computing Technology, Chinese Academy of Sciences, China
BINGJIE YAN, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China
YIQIANG CHEN, Institute of Computing Technology, Chinese Academy of Sciences, China

Large Language Models (LLMs) have showcased exceptional capabilities in various domains, attracting significant interest from both academia and industry. Despite their impressive performance, the substantial size and computational demands of LLMs pose considerable challenges for practical deployment, particularly in environments with limited resources. The endeavor to compress language models while maintaining their accuracy has become a focal point of research. Among the various methods, knowledge distillation has emerged as an effective technique to enhance inference speed without greatly compromising performance. This paper presents a thorough survey from three aspects: method, evaluation, and application, exploring knowledge distillation techniques tailored specifically for LLMs. Specifically, we divide the methods into white-box KD and black-box KD to better illustrate their differences. Furthermore, we also explored the evaluation tasks and distillation effects between different distillation methods, and proposed directions for future research. Through in-depth understanding of the latest advancements and practical applications, this survey provides valuable resources for researchers, paving the way for sustained progress in this field.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **General and reference** → **Surveys and overviews**.

Additional Key Words and Phrases: Knowledge Distillation, Large Language Models, Evaluation

*Correspondence to: Yao Zhu and Wang Lu.

Authors' addresses: Chuanpeng Yang, Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, yangchuanpeng@iie.ac.cn; Yao Zhu, Zhejiang University, China, ee_zhuy@zju.edu.cn; Wang Lu, Tsinghua University, China, newlw230630@gmail.com; Yidong Wang, Peking University, Haidian Qu, Beijing Shi, China, yidongwang37@gmail.com; Qian Chen, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China, chenqian20b@ict.ac.cn; Chenlong Gao, Institute of Computing Technology, Chinese Academy of Sciences, China, gaochenlong@ict.ac.cn; Bingjie Yan, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China, bj.yan@ieee.org; Yiqiang Chen, Institute of Computing Technology, Chinese Academy of Sciences, China, yqchen@ict.ac.cn.

ACM 2157-6912/2024/10-ART
https://doi.org/10.1145/3699518

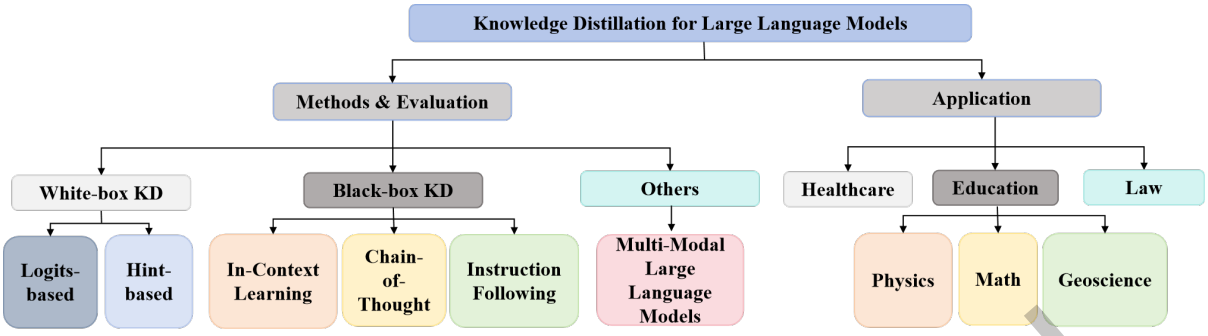

Fig. 1. The general taxonomy framework of this survey.

1 INTRODUCTION

The emergence of Large Language Models (LLMs) [2, 17, 130, 146, 166] has significantly improved text generation quality in various generative tasks, becoming a pivotal and widely discussed topic in the field of artificial intelligence. These models, compared to their predecessors, show superior generalization to unseen data. Moreover, they exhibit capabilities that smaller models lack, such as multi-step reasoning [47, 69, 83] and instruction-following [103, 144, 154]. The success of LLMs is often attributed to increased training data and a larger number of model parameters (e.g., GPT-3 with 175 billion parameters [12]). However, the expansion in parameter size brings significant drawbacks, particularly in terms of high inference costs and substantial memory requirements, making practical deployment challenging. For example, GPT-3 requires around 350GB of model storage (float16) and at least 5 A100 GPUs with 80GB of memory each for inference, contributing significantly to carbon emissions. To mitigate these challenges, model compression [30, 40] has emerged as a viable solution. Model compression aims to transform large, resource-heavy models into more compact versions that are suitable for storage on constrained mobile devices. This process may involve optimizing for reduced latency to achieve faster execution or balancing between minimal latency and model performance. Thus, a key goal in applying these high-capacity models in real-world scenarios is to compress them, reducing the number of parameters while preserving maximum performance.

As the necessity to reduce computational resource demands becomes increasingly crucial, Knowledge Distillation (KD) [43] emerges as a promising technique. KD is a machine learning method focused on compressing and speeding up models by transferring knowledge from a large, complex model to a smaller, more efficient one. This technique is frequently employed to condense the knowledge stored in large deep neural network models into smaller counterparts, thus reducing computational resource requirements and improving inference speed without substantial performance sacrifices. Fundamentally, knowledge distillation leverages the extensive knowledge acquired by a large model on a substantial dataset to guide the training of a smaller model. This knowledge typically includes the output probability distribution, intermediate layer representations, and loss function of the large model. During training, the smaller models aim not only to match the original data labels but also to mimic the behavior of the larger models. For advanced models like GPT-4 [2], which are accessible only through APIs, the generated instructions and explanations can aid in the training of student models [54].

With recent advancements in knowledge distillation, several studies have synthesized the latest progress in various distillation techniques. Specifically, Gou *et al.*[37] provide an extensive review of knowledge distillation, addressing six critical aspects: knowledge categories, training schemes, teacher-student architectures, distillation algorithms, performance comparisons, and applications. Similarly, Wang *et al.*[141] summarize the research progress and technical details of knowledge distillation techniques related to visual tasks comprehensively.

Alkhulaifi *et al.*[4] introduce an innovative metric known as the distillation metric, which they employ to evaluate different knowledge compression methods. Additionally, Hu *et al.*[48] explore various teacher-student architectures across multiple distillation objectives, presenting different knowledge representations and their corresponding optimization goals. They also provide a systematic overview of teacher-student architectures, incorporating representative learning algorithms and effective distillation schemes.

Existing reviews on knowledge distillation have laid a crucial foundation and offered valuable insights into model compression [13, 51, 64]. However, the emergence of LLMs has brought several new challenges to KD: 1) Large language models are designed not for single tasks like text generation but for broad generality across various tasks and unseen data, including emergent capabilities. Therefore, assessing the generalization of compressed LLMs requires careful and thorough evaluation. 2) The existing review is only a summary of existing work, without providing specific examples of KD technology applied to compress and deploy LLMs in real-world scenarios. This case study can help readers choose the best KD scheme for LLMs of different scales.

To tackle these challenges, a variety of knowledge distillation algorithms specifically designed for LLMs have been developed. This paper aims to provide a comprehensive and insightful guide to these methods. The overarching classification framework of our survey is depicted in Figure 1, which examines the distillation algorithm of LLMs from three aspects: method, evaluation, and application. To clearly explain these methods, we categorize them into white-box KD and black-box KD. White-box KD includes two distinct types: Logits-based methods [43], which transfer knowledge at the logits level, and Hint-based methods [109], which transmit knowledge through intermediate features. Black-box KD involves an API-based approach where only the outputs from the teacher model are accessible. This category typically includes three methods: In-Context Learning [52], Chain-of-Thought [69], and Instruction Following [144]. In addition, we simultaneously evaluate the effectiveness of the above two types of distillation algorithms on robustness benchmarks [94, 128, 138]. Finally, we discuss the relationships and application scenarios among different distillation methods and propose directions for future research.

The rest of this paper is organized as follows. Sec.2 briefly reviews the definition of knowledge distillation methods. Next, Sec.3 delves into the distillation and evaluation methods in the field of LLMs. Sec.4 presents applications while Sec.5 summarizes the challenges of knowledge distillation and explores future research directions. Lastly, Sec.6 concludes the paper.

2 OVERVIEW OF KNOWLEDGE DISTILLATION

In this section, we summarized the optimization objectives of each knowledge distillation algorithm.

2.1 Logits-based KD

As implied by its name, logic-based KD [43] is a distillation paradigm that employs logic within teacher models for knowledge transfer. We can formulate the general knowledge distillation loss function as follows:

$$\mathcal{L}_{\text{logits}} = KL\left(p^{\mathrm{t}} \| p^{\mathrm{s}}\right) = \sum_{j=1}^{C} p_j^{\mathrm{t}} \log\left(\frac{p_j^{\mathrm{t}}}{p_j^{\mathrm{s}}}\right), \tag{1}$$

$$p_i^{\mathrm{s}} = \frac{\exp\left(z_i^{\mathrm{s}}/\tau\right)}{\sum_{j=1}^{C} \exp\left(z_j^{\mathrm{s}}/\tau\right)}, \quad p_i^{\mathrm{t}} = \frac{\exp\left(z_i^{\mathrm{t}}/\tau\right)}{\sum_{j=1}^{C} \exp\left(z_i^{\mathrm{t}}/\tau\right)}, \tag{2}$$

where z^{s}, $z^{\mathrm{t}} \in \mathbb{R}^C$ denote the logits output of the student and teacher network, respectively. τ is a temperature parameter that adjusts the smoothness of the logits. C represents the number of classes. The Kullback-Leibler divergence (KLD) [43] loss can also be replaced with other functions, such as Reverse Kullback–Leibler (RKL) [20, 53, 65, 96] distillation, Jenson–Shannon (JS) [129] distillation, etc.

2.2 Hint-based KD

Given the restricted ability of students to extract knowledge in logit-based knowledge distillation, researchers strive to more precisely replicate the behavior of teachers. Consequently, intermediate feature-based knowledge distillation [46, 122] was introduced. This technique involves matching the outputs of the intermediate layers between student and teacher models. This approach requires students to understand both the results and the processes leading to those results. The general form of the feature-based knowledge distillation loss function is outlined below:

$$\mathcal{L}_{\text{hint}} = \mathcal{H}\left(F^{\mathbf{s}}, F^{\mathbf{t}}\right) = \left\|F^{\mathbf{t}} - \phi\left(F^{\mathbf{s}}\right)\right\|^2, \tag{3}$$

where $F^{\mathbf{s}}, F^{\mathbf{t}} \in \mathbb{R}^{H\times W\times C}$ denote the intermediate features of the student and teacher networks, respectively. The function ϕ is used to ensure that the student features match the dimensions of the teacher features. The metric function is represented by $\mathcal{H}$, and as an example, we use mean square error.

2.3 In-Context Learning

ICL [12, 52] utilizes a natural language prompt composed of task descriptions or several task examples as demonstrations. Formally, let $D_K = \{f(x_1, y_1), \ldots, f(x_k, y_k)\}$ represent a set of k examples, where $f(x_k, y_k)$ is a function that converts the k-th task example into a natural language prompt. Given the task description I, the demonstration set D_k, and a new input query x_{k+1}, the predicted output $\hat{y}_{k+1}$ generated by the LLM can be described by the following formula:

$$\text{LLM}(I, \underbrace{f(x_1, y_1), \ldots, f(x_k, y_k)}_{\text{demonstrations}}, f(\underbrace{x_{k+1}}_{\text{input}}, \underbrace{\quad}_{\text{answer}})) \rightarrow \hat{y}_{k+1}, \tag{4}$$

where the answer y_{k+1} is left blank for the LLM to predict. The student model is used to predict the results generated by the LLM.

2.4 Chain-of-Thought

CoT [47, 69, 83, 136] integrates intermediate reasoning steps into prompts, rather than relying solely on simple input-output pairs as done in ICL.

$$\text{LLM}(I, \underbrace{f(x_1, r_1, y_1), \ldots, f(x_k, r_k, y_k)}_{\text{demonstrations}}, f(\underbrace{x_{k+1}}_{\text{input}}, \underbrace{\quad}_{\text{rational}}, \underbrace{\quad}_{\text{answer}})) \rightarrow \hat{r}_{k+1}, \hat{y}_{k+1}. \tag{5}$$

where r_k represents the rationale provided by the user that explains why the answer to x_k is y_k. At this point, the student model not only needs to predict the labels of the teacher model, but also needs to emulate the reasons generated by the teacher.

2.5 Instruction Following

By fine-tuning on a structured multitask dataset that utilizes natural language descriptions, LLMs exhibit proficiency on unseen tasks that are similarly expressed in instructional formats [98, 112, 149]. Through instruction tuning, LLMs can follow task guidelines for new assignments without needing explicit examples, thus improving their generalization abilities. The process of distilling instruction-following skills involves generating task-specific instructions with the LLM and then fine-tuning the student model using this instruction dataset.

3 KNOWLEDGE DISTILLATION IN LARGE LANGUAGE MODELS

The Transformer architecture is highly scalable, allowing for the creation of extremely large models with billions or even trillions of parameters. It underpins many of the most prominent large-scale models in NLP, CV, and

Table 1. Comparison of representative white-box KD methods on large language models. The compression ratio indicates the proportion of the original uncompressed model size to the compressed model size.

Models	Distillation Type	Teacher Model	Compression Rate	Evaluation Task	Comparison with Teacher Model
DistillBiLSTM [126]	logits-based	BERT_{base}	114×	GLUE[137]: SST-2/QQP/MNLI-m/MNLI-mm	81.1/87.7 (92% performance)
DistillBERT [111]	logits-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI/QNLI/RTE/MRPC/CoLA/STS-B/WNLI	77.0/79.5 (97% performance)
MixKD [75]	logits-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI-m/MNLI-mm/QNLI/RTE/MRPC	77.2/82.6 (93% performance)
ReAugKD [163]	logits-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/QNLI/RTE/MRPC/CoLA	81.8/82.3 (99% performance)
PD [134]	logits-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI/QNLI/RTE/MRPC	82.1/81.7 (100.5% performance)
MINILLM [38]	logits-based	$\text{GPT-2}_{1.5B}$	12×	Dolly[38]/SelfInst[144]/Vicuna[24]/S-NI[145]/UnNI[45]	22.0 ~25.5 / 24.1 (91% ~106% performance)
		OPT_{13B}	10×		24.3 ~26.9 / 29.4 (92% ~102% performance)
		LLaMA_{13B}	2×		29.7 / 29.4 (101% performance)
GKD [3]	logits-based	T5_{XL}	39×	XSum[93]/WMT14EN-DE[10]/GSM8K[27]/MMLU[41]/BBH[119]	14.5/26.0 (56% performance)
MiniMA [161]	logits-based	LLaMA-2_{7B}	2×	MMLU[41]/CEval[133]/DROP[33]/BBH[119]/GSM8K[27] HumanEval[21]	21.7/28.5 (76% performance)
PKD [122]	hint-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI-m/MNLI-mm/QNLI/RTE/MRPC	77.7/84.9 (92% performance)
MetaDistil [168]	hint-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI/QNLI/RTE/MRPC/CoLA/STS-B	80.4/80.7 (99% performance)
AD-KD [155]	hint-based	BERT_{base}	2×	GLUE[137]: SST-2/QQP/MNLI/QNLI/RTE/MRPC/CoLA/STS-B	83.4/84.1 (99% performance)
XtremeDistil [91]	hint-based	mBERT_{base}	35×	Multilingual NER[100]/IMDB[82]/SST-2[117]/Elec[85]/DbPedia[165]/Ag News[165]	88.6/92.7 (95% performance)
TinyBERT [55]	hint-based	BERT_{base}	7×	GLUE[137]: SST-2/QQP/MNLI-m/MNLI-mm/QNLI/RTE/MRPC/CoLA/STS-B	77.0/79.5 (97% performance)
MobileBERT [123]	hint-based	IB-BERT_{large}	4×	GLUE[137]: SST-2/QQP/MNLI-m/MNLI-mm/QNLI/RTE/MRPC/CoLA/STS-B	77.7/78.3 (99% performance)
MiniLM [143]	hint-based	BERT_{base}	2×	SQuAD2[106]/ GLUE[137]: SST-2/MNLI-m	80.4/81.5 (99% performance)
TED [74]	hint-based	DeBERTaV3_{base}	2×	GLUE[137]:SST-2/QQP/MNLI-m/MNLI-mm/QNLI/RTE/MRPC/CoLA/STS-B	87.5/88.9 (98% performance)
HomoDistil [73]	hint-based	BERT_{base}	7×	GLUE[137]: SST-2/QQP/MNLI/QNLI/RTE/MRPC/CoLA/STS-B	79.0/84.6 (93% performance)

multimodal domains. For example, notable large language models like the GPT series [2, 12], LLaMA [130], and Qwen [7] are based on its decoder-only configuration. Before 2023, research on Transformer-based NLP distillation [111, 126] mainly centered around the BERT architecture. However, with the rise of pre-trained large language models [2, 98], there has been increasing interest in distilling Transformers with billion-scale parameters and in developing more efficient distillation methods for scenarios with limited data and high computational costs [44, 47]. The existing distillation algorithms are mainly divided into two categories: white-box KD and black-box KD.

3.1 White-box Knowledge Distillation

White-box distillation depends on methods that require access to the teacher model's internal data during training, utilizing the accessible internal information of the teacher model. In the following discussion, we explore two distinct types of white-box knowledge distillation. Firstly, logits-based methods, introduced by Hinton *et al.*[43], transfer knowledge at the logits level, where the knowledge is conveyed using the teacher model's logits. Given the limited knowledge acquired by students in logits-based knowledge distillation, researchers aim to more accurately replicate the teacher's behavior. To this end, Romero *et al.*[109] propose hint-based knowledge distillation, which involves aligning the feature outputs of intermediate layers between the student and teacher models. This approach requires the student to understand not only the final results but also the processes leading to those results. In the following section, we analyze in detail the characteristics of each method from the perspective of evaluation tasks (as shown in Table 1). Furthermore, we evaluate the strengths and weaknesses of the two types of distillation algorithms based on robustness, providing certain guidance in the applicable scenarios of the algorithms.

3.1.1 Logits-based KD. The distillation of Bidirectional Long Short-Term Memory Networks (BiLSTM) [126] marks the earliest attempt to apply knowledge distillation to BERT [60]. The distillation objective is to minimize the mean squared error loss between the logits of the student network and those of the teacher. This approach has been tested on three tasks: sentence classification and sentence matching. Experimental results show that the shallow BiLSTM-based model achieves performance comparable to the ELMo language model [104], but with approximately 100 times fewer parameters and a 15-fold increase in inference speed. Similarly, DistillBERT [111] initializes a shallower student model using the teacher's parameters and minimizes the difference in soft target probabilities between the teacher and student, a technique known as word-level knowledge distillation. It

introduced a triple loss that combines language modeling, distillation, and cosine distance loss to leverage the inductive bias learned by the pre-trained model. DistilBERT achieved performance equivalent to or exceeding the ELMo baseline in nine tasks. Compared to BERT, DistilBERT maintains 97% of the performance while reducing the number of parameters by 40%. MixKD [75] extends the concept of encouraging students to mimic teachers' logits by using linear interpolation of example pairs. It improves the effectiveness of knowledge distillation by using data augmentation to create additional samples from the available task-specific data. This approach mirrors students learning more effectively from teachers by asking further questions to explore their answers and concepts in depth, providing more data for student models to extract insights from large-scale language models. Evaluation results across six datasets show that MixKD significantly outperforms traditional knowledge distillation and previous methods in compressing large language models. ReAugKD [163] includes both an inference stage and a training stage. In the inference stage, it aggregates soft labels generated by teachers that closely resemble student embeddings. During the training phase, a novel relationship KD loss is used to minimize the differences between teacher-student embeddings and their distributions. Evaluation results on six datasets demonstrated that ReAugKD achieved superior performance compared to the baseline, with a latency overhead of less than 3% of the baseline, highlighting that integrating retrieval information can significantly improve generalization ability. Turc *et al.*[134] proposed a pre-training distillation (PD) method, which is a universal yet straightforward algorithm for building compact models. It consists of three standard training operation sequences and can be applied to any architecture choice. The method also explores transferring task knowledge from large fine-tuned models using traditional logits-based KD and evaluates its performance on six datasets. On average, this pre-training distillation method performs best and even surpasses the corresponding teacher model. The above distillation algorithms are all based on BERT as the teacher model and GLUE as the evaluation benchmark. With the increasing size of the model, existing distillation algorithms and evaluation standards can no longer meet the requirements

MINILLM [38] addresses the limitations of traditional logits-based Knowledge Distillation methods by proposing an innovative approach to distill large language models (LLMs) into smaller ones, focusing on minimizing the forward Kullback-Leibler divergence during free-running generation. This method replaces the standard KD method's forward KLD target with a reverse KLD, which is more suitable for generating KD on language models and aims to prevent student models from overestimating the low probability distribution of teacher distributions. To further stabilize and accelerate training, an effective optimization method is introduced, comprising three key steps: 1) single-step decomposition to reduce variance, 2) teacher mixed sampling to mitigate reward hacking, and 3) length normalization to counteract length bias. MINILLM is applied to models ranging in size from 120M to 13B parameters. Experimental evaluations on five datasets using Rouge-L [77], human judgment, and GPT-4 feedback consistently demonstrate that this approach outperforms the standard KD baseline. Further research and analysis indicate that MINILLM can reduce exposure bias and improve long-response generation performance. Similar to MINILLM, GKD [3] moves beyond relying solely on a fixed set of output sequences, training student models to generate their own sequences with feedback from the teacher model. Unlike supervised KD methods, GKD allows for the use of alternative loss functions between the student and teacher, which is advantageous when student models lack the expressive capability to effectively mimic teacher distributions. Additionally, GKD enables the seamless integration of distillation and Reinforcement Learning (RL) fine-tuning for language models. By providing flexibility to optimize alternative divergence measures such as reverse KL and generalized JSD, GKD allows limited student capacity to focus on generating samples similar to those produced under teacher supervision. It has been demonstrated that on-policy GKD facilitates the integration of distillation with RL [98] fine-tuning of language models, a combination not previously explored. Regarding performance enhancement for initial students, on average, GKD yielded a relative gain of 2.1 times for abstracts, 1.7 times for machine translation, and 1.9 times for arithmetic reasoning tasks across different sizes of T5 student models, underscoring the effectiveness of GKD. In terms of performance enhancement for initial students, GKD showed average relative gains of 2.1 times for abstracts, 1.7 times for machine translation, and 1.9 times for arithmetic reasoning tasks

across various sizes of T5 student models, highlighting the effectiveness of GKD. Wen *et al.*[151] proposed the f-DISTILL framework, which formulates sequence-level knowledge distillation by minimizing a generalized f-divergence function. This framework introduces four distillation variants, demonstrating that existing SeqKD [63] and ENGINE [132] methods are approximations of KL and reverse KL distillation. Furthermore, the f-DISTILL method includes step-wise decomposition to convert the complex sequence-level divergence into a more manageable word-level loss. This facilitates easier calculation. This method was evaluated on four datasets: DART for data-to-text generation [92], XSum for summarization [93], WMT16 EN-RO for machine translation [11], and Commonsense Dialogue [167]. The experiments demonstrated that f-DISTILL variants outperformed existing distribution-matching KD methods, leading to performance improvements when combined with representation-matching KD methods.Additionally, the results indicated that symmetric distillation loss is superior to asymmetric distillation loss, confirming that extreme mode averaging or collapse is suboptimal. MiniMA [161] found that the optimal distillation effect occurs when the student model is approximately 40% the size of the teacher model. It combines structured pruning with logit-based knowledge distillation, using LLaMA2-7B [130] as the teacher model to train the 3B MiniMA model. The results showed that MiniMA achieved impressive performance in knowledge, reasoning, and encoding, while using a similar or even fewer number of tokens than the teacher model.

3.1.2 Hint-based KD. The feature-based knowledge distillation methods [46, 122] extract knowledge from the embedding space, transformer layers, and prediction layers, allowing the student model to learn various aspects of the teacher model comprehensively. For instance, Sun *et al.*[122] proposed a patient knowledge distillation (PKD) method aimed at compressing a large-scale teacher model into an equally effective lightweight student model. They proposed two distinct distillation strategies: 1) PKD-Last: The student model learns from the last k layers of the teacher model, based on the assumption that the top layers contain the most informative knowledge. 2) PKD-Skip: The student learns from every k-layer of the teacher, suggesting that the lower layers also contain essential information that should be gradually transferred during distillation. Experiments conducted on seven datasets across four tasks—sentiment classification, paraphrase similarity matching, natural language inference, and machine reading comprehension—showed that the PKD method outperformed standard knowledge distillation methods. It achieved superior performance and better generalization, significantly enhancing training efficiency and reducing storage requirements while maintaining accuracy comparable to the original large-scale model. MetaDistill [168] offers a simple and efficient alternative to traditional KD methods by keeping the teacher model fixed during training. Within the meta-learning framework, teacher networks enhance knowledge transfer to student networks by distilling feedback on student performance. Additionally, a pilot update mechanism is introduced to improve the alignment between internal learners and meta-learners, focusing on enhancing internal learners' performance. Extensive experiments have validated the effectiveness and versatility of this method across text and image classification tasks. Furthermore, experiments on the GLUE benchmark have shown that MetaDistill significantly outperforms traditional knowledge distillation, achieving state-of-the-art performance compression. AD-KD [155] addresses two key limitations of existing knowledge distillation methods. First, student models often merely mimic the teacher's behavior without developing their own reasoning capabilities. Second, these methods typically focus on transferring knowledge specific to complex models while neglecting data-specific knowledge. To overcome these issues, AD-KD introduces an innovative attribution-driven knowledge distillation method, which calculates the importance score of each input token using a gradient-based attribution approach [124]. To minimize the impact of less significant dimensions in the teacher's input embeddings, a top-K strategy filters out dimensions with lower attribution scores. The remaining scores are aggregated and normalized to reflect the importance of individual tokens. Additionally, this method extracts all potential predicted attribution knowledge, not just the highest probability prediction. To improve knowledge transfer for reasoning and generalization, AD-KD explores multi-view attribution distillation of all potential decisions made by the

teacher. Experimental results on the GLUE benchmark indicate that this method surpasses several state-of-the-art approaches in performance.

Mukherjee *et al.*[91] present XtremeDistil, a distillation method leveraging internal representations and parameter projections that are independent of the teacher's architecture. Unlike previous approaches focused on single-language GLUE tasks, this method distills multilingual Named Entity Recognition (NER) across 41 languages, using the multilingual bidirectional encoder representation from Transformers (mBERT) [131] as the teacher model. Experimental results indicate that XtremeDistil achieves higher compression and faster inference speeds. Additionally, the study explored several previously unexamined aspects of distillation, including the effects of unlabeled transmission data and annotation resources, the selection of multilingual word embeddings, architectural modifications, and inference delays. This method significantly compressed the teacher model by up to 35 times in terms of parameters and reduced batch inference delay by 51 times while maintaining 95% of the performance in large-scale multilingual NER and either matching or surpassing it in classification tasks. TinyBERT [55] integrates pre-trained distillation with fine-tuning distillation to capture both general domain and task-specific knowledge from BERT. It extracts multiple types of knowledge from different layers, including the embedding layer, hidden states, attention matrices, and transformation layers. During the GLUE benchmark evaluation, its teacher model BERT_{base} achieved a performance exceeding 96.8%, while offering inference speeds that were 7.5 to 9.4 times faster. MiniLM [143] introduced a depth self-attention distillation framework for task-agnostic Transformer-based language model (LM) distillation. This method isolates the self-attention module of the teacher model's final Transformer layer and uses the scaled dot-product between values within this module as a novel form of depth self-attention knowledge. This technique addresses the challenge of layer alignment between teacher and student models by transforming various dimensional representations of both models into a relation matrix of matching dimensionality, without requiring additional parameters for transforming student representations. This enhances the depth flexibility of the student model. MiniLM retained over 99% accuracy on the SQuAD 2.0 [106] and various GLUE benchmark tasks while using only 50% of the Transformer parameters and computational resources of the teacher model. This demonstrates the effectiveness of employing a teacher assistant [90] in distilling large pre-trained Transformer-based models. TED [74] introduces an innovative task-aware layout distillation method designed to combat underfitting in student models and remove unnecessary information from teachers' hidden representations. This method aligns the hidden representations of students and teachers at each level, employing task-aware filters to extract relevant knowledge for the target task. By doing so, it narrows the knowledge gap between the models and enhances the student's ability to adapt to the target task. MobileBERT [123] and HomoBERT [73] primarily focus on adjusting the model's width while maintaining its depth. This contrasts with Turc *et al.*[134], who found that altering model depth significantly impacts performance. MobileBERT introduces bottlenecks and inverted bottlenecks to both teacher and student models to modify hidden dimensions. However, this approach can disrupt the parameter balance between the multi-head attention and feed-forward networks, which is mitigated by using a stacked Feed-Forward Network (FFN) approach. Knowledge extraction is then carried out through the attention and hidden states of the transformer layers. HomoBERT, on the other hand, employs pruning. It starts by initializing the student model with the teacher model to ensure minimal initial divergence. It then targets input embeddings, hidden states, attention matrices, and output logits for pruning to create the distillation loss function. In each iteration, the most significant neurons are pruned based on importance scores, and the student model is trained using the distillation loss. This iterative process continues until the student model achieves the desired size. While white-box distillation is limited by the proprietary nature of LLMs, restricting its applicability, the rise of diverse open-source LLMs like Alpaca [127] and Vicuna [24] offers promising prospects for the future of white-box distillation.

Table 2. The evaluation results of GPT-2. The optimal score for each model size is displayed in bold, and the score for student models outperforming teachers is marked with an ∗.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
1.5B	Teacher	54.73	96.15	86.78	93.92	80.25	87.65	87.67	38.55	0.40
120M	SFT	57.43	96.15	91.74	86.49	61.73	88.89	96.17	14.26	0
	KD	83.78	**94.87**∗	89.26	92.57	69.14	**82.72**∗	94.08	12.96	0
	SeqKD	70.27	96.15	90.08	87.84	75.31	83.95	**88.83**	12.61	0
	RKL	66.89	98.72	**87.60**	85.81	65.43	93.83	98.08	7.46	0
	JS	**54.05**∗	100	94.21	**77.70**∗	74.07	95.06	97.83	20.25	0
	MINILLM	64.19	100	89.26	90.54	**53.09**∗	84.57	95.50	**23.32**	**1.25**∗
340M	SFT	70.27	97.44	92.56	95.27	**55.56**∗	83.33	94.00	46.36	0
	KD	63.51	98.72	84.30	90.54	72.84	80.86	92.42	52.12	0
	SeqKD	66.89	97.44	81.82	93.24	72.84	79.01	97.08	47.87	0
	RKL	62.16	96.15	95.04	95.27	70.37	92.59	96.92	33.64	0
	JS	62.16	96.15	95.04	95.27	70.37	92.59	96.92	32.29	0
	MINILLM	**54.73**∗	**92.31**∗	**78.51**∗	**87.16**∗	69.14	**75.31**∗	**88.75**	**52.45**∗	**0.68**∗
760M	SFT	56.76	97.44	90.91	92.57	80.25	91.36	91.75	28.52	0
	KD	**53.38**∗	**94.87**∗	90.91	95.27	64.20	87.65	95.58	30.77	0
	SeqKD	55.41	96.15	90.91	97.30	80.25	89.51	94.17	27.60	0
	RKL	64.86	97.44	90.91	92.57	85.19	95.06	98.67	20.53	0
	JS	60.14	98.72	95.04	94.59	77.78	96.30	98.58	19.90	0
	MINILLM	54.05	96.15	**81.82**∗	**89.19**∗	**60.49**∗	**72.22**∗	**85.33**∗	**47.00**∗	0

3.2 Robustness Evaluation of White-box KD

There are various evaluation standards for existing white-box KD algorithms, most of which utilize BERT as the teacher model. However, the effectiveness of these distillation algorithms in the context of LLMs remains unclear. Building on the work presented in [139], we conducted a unified evaluation of these algorithms from a robustness perspective, specifically focusing on adversarial robustness and out-of-distribution (OOD) robustness. Both types of robustness pertain to performance under input disturbances, which is particularly critical for safety-sensitive applications. Adversarial robustness examines the stability of models against adversarial and imperceptible disturbances, while OOD robustness assesses performance on unseen data that differs from the training data distribution. To evaluate adversarial robustness, we employed the AdvGLUE [138] and ANLI [95] benchmarks, using Attack Success Rate (ASR) as the metric. For OOD robustness, we used the Flipkart [135] review and DDXPlus [128] medical diagnostic datasets, with F1-score (F1) as the indicator. Inspired by the work on MINILLM [38], we utilized the Dolly [1] dataset for distillation, fine-tuning both student and teacher models. We evaluated five distillation algorithms and four models concurrently to assess their robustness.

The evaluation results are shown in Tables 2-4. Firstly, we observed that MINILLM demonstrated superior overall distillation performance in GPT-2. Notably, for the 340M-sized GPT-2, it achieved state-of-the-art results on both adversarial and out-of-distribution datasets when compared to the other four distillation algorithms. Furthermore, MINILLM outperformed the other algorithms on the Flipkart and DDXPlus datasets for GPT-2 of

[1] https://github.com/databrickslabs/dolly/tree/master

Table 3. The evaluation results of OPT.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
13B	Teacher	52.70	94.87	76.03	91.89	61.73	70.99	99.58	54.81	0.18
1.3B	SFT	54.05	94.87	78.51	94.59	61.73	77.78	98.00	51.06	**1.84***
	KD	52.03	**65.38***	**77.69**	**68.92***	50.62	**69.14***	87.67	44.45	0
	SeqKD	48.65	93.59	80.17	85.14	83.95	75.31	**78.17***	28.85	0.30
	RKL	**45.95***	94.87	92.56	97.30	74.07	90.12	99.00	46.41	0
	JS	47.97	82.05	80.99	89.86	48.15	89.51	94.67	45.15	0
	MINILLM	47.30	94.87	78.51	79.73	**49.38***	71.60	89.75	**55.88***	0.13
2.7B	SFT	51.35	96.15	92.56	97.30	69.14	93.83	98.17	47.86	0
	KD	**44.59***	84.62	**73.55***	83.78	64.20	**66.05***	**70.67***	41.83	0
	SeqKD	47.97	**75.64***	77.69	**77.03***	**61.73**	72.22	88.50	34.85	0
	RKL	56.08	98.72	90.91	99.32	100	95.68	96.83	53.01	**0.29***
	JS	67.57	96.15	86.78	97.97	75.31	89.51	92.25	46.55	0
	MINILLM	47.30	94.87	**73.55***	91.89	66.67	72.84	81.17	**58.09***	0.09
6.7B	SFT	49.32	89.74	85.12	81.76	77.78	79.63	**75.58***	43.05	0.06
	KD	**45.27***	94.87	80.17	85.14	58.02	**73.46**	81.33	50.99	0
	SeqKD	50.68	92.31	85.95	77.70	86.42	80.86	78.17	32.01	**0.65***
	RKL	60.81	91.03	95.87	88.51	**49.38***	93.21	97.92	24.16	0
	JS	63.51	**88.46***	94.21	**69.59***	62.96	88.27	98.58	26.05	0
	MINILLM	50.68	92.31	**74.38***	95.27	80.25	75.93	86.75	**56.43***	0

Table 4. The evaluation results of LLaMA and LLaMA2.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
13B	Teacher	41.89	65.38	76.03	52.70	56.79	64.20	66.42	56.01	9.09
LLaMA_7B	SFT	47.97	67.95	71.90	48.65	46.91	66.67	67.33	52.65	2.81
	KD	43.24	60.26	73.55	54.73	56.79	**64.20**	**66.58**	**65.16***	4.32
	SeqKD	**40.54***	65.38	**67.77***	49.32	**41.98***	64.81	66.67	56.58	**5.66**
	RKL	**40.54***	76.92	72.73	**45.27***	50.62	64.81	67.75	52.74	0.43
	JS	43.92	66.67	73.55	52.70	56.79	**64.20**	67.00	57.35	3.44
	MINILLM	42.57	**48.72***	71.90	51.35	56.79	64.81	66.92	52.27	5.21
13B	Teacher	54.05	47.44	65.29	51.35	46.91	62.96	71.00	29.76	0
LLaMA2_7B	SFT	**43.92***	69.23	75.21	63.51	61.73	66.05	69.92	50.76	0.17
	KD	47.97	70.51	74.38	52.70	56.79	64.20	83.67	55.51	1.75
	SeqKD	49.32	66.67	75.21	52.70	56.79	63.59	84.50	**55.63***	1.67
	RKL	52.03	61.54	**73.55**	**47.97***	58.02	**63.58**	72.33	55.42	4.37
	JS	51.35	**57.69**	78.51	56.76	**37.04***	66.67	**69.00***	50.38	**12.19***
	MINILLM	50.00	75.64	75.21	58.78	48.15	69.75	83.67	31.94	0

Table 5. Comparison of representative black-box KD methods on large language models.

Models	Distillation Type	Teacher Model	Compression Rate	Evaluation Task	Comparison with Teacher Model
ILD [52]	ICL	BERT_{large}	13×	LAMA[105]	52.4/57.3 (91% performance)
		GPT-2_{large}	6×	CrossFit[159]	61.2/66.2 (92% performance)
LLM-R [140]	ICL	LLaMA_{13B}	2×	Commonsense/Coreference/NLI/Paraphrase/Sentiment/Data-to-text/Summarize/etc.(30)	68.8/64.6 (107% performance)
MT-CoT [75]	CoT	$\text{GPT-3}_{text-davinci-002}$	58×	CommonsenseQA[125]/StrategyQA[35]/OpenbookQA[87]	80.5/82.1 (98% performance)
Distilling step-by-step [47]	CoT	PaLM_{540B}	2455×	e-SNLI[14]/ANLI[95]/CQA[125]/SVAMP[102]	58.4/72.3 (81% performance)
Fine-tune-CoT [44]	CoT	$\text{InstructGPT 175B}_{text-davinci-002}$	26×	SingleEq/AddSub/MultiArith/GSM8K/AQUA-RAT/SVAMP/StrategyQA/etc.(12)	42.2/65.5 (64% performance)
MCC-KD [19]	CoT	GPT-3.5_{Turbo}	/	GSM8K[27]/ASDiv[86]/SVAMP[102]/CommonsenseQA[125]	66.2/75.8 (87% performance)
SCOTT [142]	CoT	GPT-neox_{20B}	7×	CSQA[125]/CREAK[97]/QASC[61]/StrategyQA[35]	69.6/72.3 (96% performance)
LTD [56]	CoT	$\text{Codex}_{code-davinci-002}$	29×	SVAMP[102]/GSM8K[27]/MathQA[5]	45.9/56.8 (81% performance)
PaD [170]	CoT	PaLM_{60B}	78×	SVAMP[102]/GSM8K[27]/ASDiv[86]/MultiArith[110]/BBH[119]	43.9/50.2 (88% performance)
Learn-to-Reason [16]	CoT	ChatGPT_{175B}	29×	GSM8K[27]/MultiArith[110]/SVAMP[102]/CSQA[125]/StrategyQA[35]	62.1/76.1 (82% performance)
LaMini-LM [154]	IF	Alpaca_{7B}	9×	Multiple-Choice QA/Extractive QA/Sentiment Analysis/Paraphrase Identification/etc.(15)	60.8/62.3 (98% performance)
Lion [54]	IF	ChatGPT_{175B}	25×	BBH[119]	32.0/48.9 (65% performance)
UniNER [169]	IF	$\text{GPT-3.5}_{turbo-0301}$	/	UNIVERSAL NER BENCHMARK [169]	41.7/34.9 (119% performance)

any size, highlighting its exceptional generalization capability to out-of-distribution data. Secondly, for the OPT model, we discovered that the most straightforward KD algorithm, which employs the teacher distribution as supervision for each token step to fine-tune the student model, achieved the best overall performance. Likewise, MINILLM outperformed other distillation algorithms and even exceeded the performance of teacher models for OPTs of any size on the Flipkart dataset. Finally, for LLaMA, SeqKD demonstrated a comparatively better distillation effect, whereas for LLaMA2, JS showed a relatively superior performance. This suggests that even when the model size is identical and the model structure is similar, the effectiveness of the same distillation algorithm can vary significantly.

3.3 Discussion on White-box KD

Logits-based KD methods typically focus on aligning the output distributions between the teacher and student models. In contrast, hint-based KD methods can convey richer information by aligning the intermediate layers, leading to better results. However, implementing layer-to-layer knowledge distillation necessitates careful design of the layer mappings between the teacher and student models and requires a deep understanding of the model architecture. Both logits-based and hint-based KD methods demand substantial GPU memory during the distillation process. Even though the teacher network doesn't need backpropagation, the activation of intermediate features during forward propagation consumes a significant amount of GPU memory. Therefore, exploring ways to reduce training costs and shorten training times is crucial.

3.4 Black-box Knowledge Distillation

The two previously discussed distillation techniques rely on access to the internal data of the teacher model, categorizing them as white-box distillation methods, which require internal data during training. However, many modern large-scale closed-source models do not provide access to internal data, limiting us to using only model predictions. Distillation where knowledge is transferred solely through the teacher model's predictions is known as black-box knowledge distillation. Researchers have found that when model parameters are sufficiently large, the models exhibit remarkable versatility, enabling them to handle complex tasks. Many black-box distillation methods take advantage of this capability, typically utilizing three techniques: In-Context Learning, Chain-of-Thought, and Instruction Following. In this section, we further categorize black-box KD methods based on the use of emergent capabilities.

3.4.1 In-Context Learning. ICL was initially introduced in GPT-3 [12], where it employs a natural language prompt that includes both task descriptions and multiple task examples as demonstrations. The process begins with the task description, followed by selecting specific instances from the task dataset to serve as examples.

These instances are then formatted into natural language prompts using a predefined template and arranged in a particular order. Finally, the test samples are incorporated into the input of the LLM to produce the output.

Expanding on this concept, Huang et al. [52] propose In-Context Learning Distillation, which aims to enhance the few-shot learning capabilities of multitask models by effectively extracting and transferring knowledge through context learning and language modeling objectives. This approach introduces two paradigms for few-shot learning: Meta In-context Tuning and Multitask In-context Tuning. In Meta-ICT [22, 88], the language model undergoes meta-training across a broad spectrum of tasks using in-context learning objectives. Subsequently, it adapts to unseen target tasks through in-context learning. However, the efficacy of in-context learning heavily relies on the knowledge accumulated during pretraining [108], potentially limiting its ability to fully leverage the input-label correspondence provided in the training data [89]. To address this limitation, an alternative few-shot learning paradigm called Multitask In-Context Tuning is proposed. While Meta-ICT enables the student model to adapt to new tasks via context learning and teacher guidance, Multitask-ICT treats all target tasks as training tasks and utilizes examples directly from these tasks for in-context learning distillation. These two paradigms for few-shot learning involve a trade-off between performance and computational efficiency. Results across tasks such as classification, natural language inference, and question answering indicate that Multitask-ICT achieves a reduction in model size by 93% while retaining 91.4% of the teacher's performance. Therefore, Multitask-ICT proves to be more effective, albeit with higher computational costs. LLM-R [140] utilizes a pre-trained frozen LLM to retrieve high-quality contextual examples, which are then ranked to generate training data. Subsequently, it constructs a reward model using a cross-encoder to capture ranking preferences. Finally, knowledge distillation is applied to train a dense retriever based on dual encoders. Our comprehensive evaluation of LLM-R across diverse tasks consistently demonstrates superior performance compared to several robust baselines. Furthermore, our model exhibits scalability across different task sizes and LLM architectures. Detailed analysis indicates that our approach enhances context learning performance by an average of 7.8%, with consistent improvements observed across various sizes of LLMs.

3.4.2 Chain-of-Thought. Chain-of-Thought (CoT) [47, 69, 83, 136] represents an advanced prompting strategy aimed at enhancing LLMs' ability to tackle complex reasoning tasks. Unlike the input-output pair approach used in ICL for prompt formulation, CoT integrates intermediate inference steps that incorporate final outputs into the prompts. Typically, CoT distillation [16, 19, 44, 56, 59, 68, 69, 116, 142, 147, 170] involves leveraging large-scale models to construct enriched datasets focused on reasoning tasks, which are then utilized for fine-tuning student models. Thus, the primary focus is on generating high-quality rationales for training and ensuring effective utilization of these rationales by students [47, 59, 69, 116, 142].

Li *et al.*[69] pioneered the use of explanations generated by LLMs to enhance the training of smaller inference machines. They systematically explored three methods for deriving interpretations from LLMs and integrated them into a multitask learning framework to empower compact models with robust reasoning and interpretative capabilities. Across multiple inference tasks, experiments consistently demonstrated that their approach outperforms baseline fine-tuning methods under various conditions. Notably, it achieved up to a 9.5% accuracy improvement over GPT-3 (175B) after 60 rounds of fine-tuning on Commonsense QA. The high-quality explanations generated by their method elucidate the rationale behind AI's interpretable predictions. Hsieh *et al.*[47] introduced step-by-step distilling, a novel and straightforward approach aimed at reducing the amount of training data required to refine and fine-tune LLMs into smaller models. Central to their method is a paradigm shift: LLMs are not merely sources of noisy labels but proxies capable of providing natural language reasoning to justify their predictions. Empirical findings across four NLP benchmark tests yielded three notable outcomes. Firstly, compared to fine-tuning and traditional distillation methods, their model reduced the average number of training samples required by over 50% (with some reductions exceeding 85%), leading to improved performance. Secondly, their model achieved superior performance to LLMs while being significantly smaller in size, thereby reducing

computational resources for deployment. Thirdly, their method concurrently reduced model size and required data to outperform LLMs. For example, their final iteration of the 770M T5 model surpassed the performance of a 540B parameter LLM, utilizing only 80% of the labeled dataset.

Moreover, Ho *et al.*[44] propose fine-tuning CoT, a method harnessing LLMs' reasoning capabilities to guide smaller models in solving complex tasks. By generating multiple inference solutions from the teacher model through random sampling, they enrich the training data of the student model. Evaluation across 12 tasks using widely accessible models demonstrates that fine-tuning CoT achieves significant inference performance in smaller models while preserving much of the generality of hint-based CoT inference, previously reliant on models with over 100 billion parameters. Consequently, models with as few as 0.3 billion parameters can outperform larger counterparts in specific tasks, even surpassing the performance of the teacher model with 175 billion parameters. Similarly, Chen *et al.*[19] introduced Multi-CoT Consistent Knowledge Distillation (MCC-KD) to efficiently capture the diversity and coherence of reasoning capabilities. In MCC-KD, multiple fundamental principles are generated for each question, and the consistency between corresponding predictions is strengthened by minimizing bidirectional KL divergence between answer distributions. MCC-KD's efficacy is evaluated on mathematical reasoning and common sense reasoning benchmarks across various model architectures. Empirical findings not only confirm MCC-KD's superior performance on in-distribution datasets but also highlight its robust generalization ability on out-of-distribution datasets. Fu *et al.*[34] Fu *et al.*[34] apply CoT to specialize smaller language models for multi-step mathematical reasoning tasks. The SOCRATIC CoT method, as detailed by Shridhar *et al.*[116], decomposes the original problem into a series of sub-problems and employs a pair of compact distillation models: a problem decomposer and a sub-problem solver. These models collaborate to break down and resolve complex problems presented in new tasks. Evaluation across various inference datasets, including GSM8K, StrategyQA, and SVAMP, demonstrates that this distillation approach significantly enhances the performance of smaller models by over 70% compared to the baseline. On the other hand, SCOTT [142] introduces a core principle of leveraging a LLM to guide the correct answer through comparative decoding. This method encourages the teacher model to generate tokens that align closely with the correct answer, thereby improving the fidelity of the distillation process. Jie *et al.*[56] and Zhu *et al.*[170] enhance mathematical reasoning capabilities through program distillation. Chae *et al.*[16] and Wang *et al.*[147] propose an interactive multi-loop learning framework. In this framework, the former focuses on training students using multi-hop reasoning, while the latter actively communicates their learning status to the LLM teacher. Subsequently, the teacher offers customized explanations for the students' feedback, guiding them to reflect on their errors.

3.4.3 Instruction Following. The instruction following capability aims to enhance the language model's ability to perform new tasks without heavy reliance on limited examples. Through fine-tuning across various tasks specified by instructions, the language model demonstrates its proficiency in accurately executing tasks described in previously unseen instructions. However, in black-box distillation, knowledge transfer relies solely on datasets, making the availability of a sufficiently large dataset crucial. Therefore, collaborative efforts in these approaches [54, 103, 144, 154] involve creating a comprehensive dataset comprising instructions, inputs, and outputs. This dataset enables the student model to acquire extensive knowledge from the teacher model.

Specifically, Wang *et al.*[144] propose self-instruction, a semi-automatic process that utilizes indicator signals from the model itself to refine the language model's instructions. The process begins with a constrained seed set of manually crafted tasks, such as the 175 tasks used in our study, to guide the overall generation process. Initially, the prompt model uses this initial set of instructions to generate a broader array of task descriptions. Furthermore, for newly generated sets of instructions, the framework creates input-output instances that can be used for supervised instruction tuning in the future. Finally, various heuristic methods are employed to automatically filter out low-quality or duplicate instructions before incorporating the remaining valid tasks into the task pool. This iterative process can be repeated multiple times until a significant number of tasks

are obtained. This method has influenced subsequent research, leading to adjustments in the 13B open-source models like Alpaca [127], Vicuna [24], and GPT4All [6] following this paradigm. Expanding on these ideas, Peng *et al.*[103] explore the use of GPT-4 to generate instruction-following data for fine-tuning LLMs. They curated a dataset of 52,000 instruction-following examples in both English and Chinese, along with feedback datasets generated by GPT-4. Using these datasets, they fine-tuned two student models, LLaMA-GPT4 and LLaMA-GPT4-CN. Additionally, they developed a feedback model to evaluate the quality of model responses. Wu *et al.*[154] meticulously compiled a dataset comprising 2.58 million instructions, ensuring coverage of diverse topics. These instructions were used as input to generate responses using GPT-3.5 Turbo. They fine-tuned a range of models under the LaMini-LM, including both encoder-decoder and decoder-only architectures. Evaluation of the LaMini-LM models' performance involved applying automatic metrics across 15 benchmarks, alongside manual assessment. Results illustrate that the proposed LaMini-LM model achieves comparable performance to competitive baselines despite being only one-tenth the size.

However, existing methodologies have predominantly concentrated on one-way knowledge distillation, where student model responses align with those of teacher models to generate instructions without incorporating a "feedback" mechanism. To address this limitation, Jiang *et al.*[54] introduce an innovative adversarial distillation framework consisting of three stages: imitation, discrimination, and generation. Leveraging the adaptable nature of LLMs, this framework incentivizes teacher models to identify "challenging" instructions and generate new instructions for student models, thereby enhancing the effectiveness of knowledge transfer. This approach achieves open-generation capability comparable to ChatGPT using only 70,000 training samples, surpassing traditional state-of-the-art instruction adjustment models (such as Vicuna-13B) by 55.4% and 16.7% on the zero-shot inference BBH and AGIEval tasks, respectively. In efforts to provide task-specific guidance, Chen *et al.*[18] propose a fine-tuning dataset for code generation instructions and develop a multi-round personalized distillation approach. This approach enables student models to first attempt solving tasks independently, followed by adaptive refinements provided by the teacher to enhance their performance through executive feedback. Unlike traditional knowledge transfer methods where the teacher's prior knowledge is directly imparted to students, personalized refinement offers individualized learning experiences by learning solely from examples of mistakes and iteratively improving their solutions. Meanwhile, UniversalNER [169] has conducted extensive research on named entity recognition tasks. Unlike the aforementioned methods that aim to increase instruction diversity, UniversalNER focuses on augmenting input diversity to enhance the model's generalization capabilities across various domains.

3.5 Robustness Evaluation of Black-box KD

Inspired by the work in [47], we conducted a unified evaluation of the step-by-step distillation algorithm based on CoT from a robustness perspective. Due to the closed-source nature of the PaLM 540B model, we adhered to the experimental setup in [47] and used the generated CoT interpretations to fine-tune the student model. The experimental results are presented in Tables 6-8. For GPT-2 models with 120M and 340M parameters, distillation using the interpretations from the ANLI and e-SNLI datasets produced better results. However, as the model size increases, the explanatory power of these two datasets diminishes, and a similar trend is observed in OPT models. For OPT models of various sizes, the explanatory distillation effects generated by ANLI and e-SNLI were suboptimal. This suggests that commonsense data (CQA) and mathematical data (SVAMP) are more conducive to CoT distillation in OPT models. Regardless of whether it is LLaMA or OPT, the distillation of CoT using CQA and SVAMP outperforms the distillation using the other two datasets on Flipkart and DDXPlus. This indicates that distillation of mathematical abilities and commonsense knowledge enhances the model's ability to generalize to out-of-distribution.

Table 6. The evaluation results of GPT-2.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
1.5B	Teacher	62.84	94.87	76.03	75.68	56.79	74.07	89.42	10.72	0
120M	ANLI	**56.76***	74.36	73.55	**57.43***	**45.68***	**64.81***	92.75	**25.32***	0
	CQA	85.81	96.15	95.87	96.62	98.77	93.83	94.17	8.56	0
	e-SNLI	99.32	**66.67***	**65.29***	60.81	58.02	71.60	**68.50***	1.38	0
	SVAMP	80.41	93.59	85.95	90.54	93.83	82.72	95.42	3.98	**1.02***
340M	ANLI	58.78	**50.00***	**78.51**	**61.49***	46.91	68.52	87.08	27.39	0
	CQA	52.70	87.18	88.43	94.59	97.53	91.36	94.08	**31.82***	0
	e-SNLI	99.32	71.79	80.99	65.54	**41.98***	**66.67***	**72.83***	4.61	0
	SVAMP	**50.00***	69.23	80.17	77.03	64.20	77.78	76.08	25.60	0
760M	ANLI	89.19	89.74	88.43	75.68	93.83	86.42	**92.83**	6.48	0
	CQA	**54.73***	**83.33***	90.08	75.00	**83.95**	94.44	98.83	**33.03***	0
	e-SNLI	100	100	100	100	100	100	100	1.38	0
	SVAMP	64.19	70.51	**87.60**	**70.27***	85.19	**85.19**	96.75	9.84	0

Table 7. The evaluation results of OPT.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
13B	Teacher	68.24	80.77	74.38	50.68	56.79	69.75	72.33	32.04	0
1.3B	ANLI	100	100	100	100	100	100	100	0	0
	CQA	50.00	**85.90**	**57.85***	**66.89**	**43.21***	**64.81***	**96.00**	36.00	0
	e-SNLI	75.00	93.59	81.82	83.78	67.90	79.63	97.33	2.93	0
	SVAMP	**49.32***	92.31	78.51	78.38	53.09	75.93	99.33	**52.44***	**0.85***
2.7B	ANLI	100	100	100	100	100	97.53	98.50	0	0
	CQA	**50.68***	87.18	80.99	75.00	**38.27***	**61.11***	81.92	31.15	**1.33***
	e-SNLI	87.84	79.49	90.91	87.84	86.42	86.42	96.50	3.48	0
	SVAMP	51.35	**65.38***	**68.60***	**61.49**	66.67	64.20	**70.33***	**43.51***	0.08
6.7B	ANLI	100	100	99.17	100	100	100	99.83	0	0
	CQA	**52.03***	98.72	85.12	**83.78**	**64.20**	93.83	97.08	**44.62***	0
	e-SNLI	100	**67.95***	91.74	84.46	70.37	85.80	94.75	0.87	0
	SVAMP	58.78	84.62	**76.03**	94.59	86.42	**72.84**	**82.83**	31.77	**0.15***

3.6 Discussion on Black-box KD

The black-box based KD method is typically used by LLMs to generate explanations or instruction pairs to fine-tune the student model. In this approach, only the teacher model generates data, and only the student model is involved in training, making it memory-efficient. However, most current methods rely on closed-source teacher

Table 8. The evaluation results of LLaMA and LLaMA2.

Params	Method	Adversarial Robustness(ASR↓)							OOD Robustness(F1↑)	
		SST-2	QQP	MNLI	QNLI	RTE	MNLI-MM	ANLI	Flipkart	DDXPlus
13B	Teacher	50.68	69.23	62.81	58.78	46.91	77.78	69.83	49.16	5.74
LLaMA_7B	ANLI	**41.22***	60.26	**53.72***	**47.97***	54.32	**61.11***	71.08	2.16	4.80
	CQA	46.62	**46.15***	66.12	53.38	**46.91**	64.81	77.42	**41.37**	**8.33***
	e-SNLI	49.32	84.62	57.85	58.11	61.73	63.58	**70.33**	1.38	1.46
	SVAMP	46.62	74.36	72.73	52.70	61.73	67.90	77.75	32.44	1.46
13B	Teacher	53.38	50.00	71.07	45.95	50.62	62.35	71.83	33.76	18.92
LLaMA2_7B	ANLI	64.19	61.54	**61.98***	52.70	56.79	61.11	**62.08***	23.88	1.02
	CQA	**44.59***	48.72	66.94	50.00	43.21	66.05	66.50	**35.10***	8.19
	e-SNLI	61.49	**44.87***	74.38	**40.54***	**41.98***	**58.64***	67.33	5.48	4.71
	SVAMP	58.11	53.85	78.51	56.08	61.73	69.14	70.42	27.01	**11.67**

models, and generating additional data can be costly. Additionally, many methods do not have open-source data generation techniques or involve closed-source generated data, posing challenges for the fair evaluation of these black-box based distillation algorithms.

3.7 Others

As large language models have advanced significantly, their inherent limitation lies in their inability to comprehend visual information, as they are primarily designed for processing discrete texts. Consequently, researchers are increasingly exploring ways to transfer the capabilities of language models into multimodal domains, where text and image data are integrated to enable a wider range of tasks [32, 36, 158]. Extracting knowledge from pre-trained multimodal models to enhance the performance and generalization of compact multimodal language models has become a focal point of interest in this field.

3.7.1 Muiti-Modal Large Language Models. Knowledge distillation for multimodal large models is still in its nascent stages, focusing primarily on refining instruction-following capabilities. Li *et al.*[70] have pioneered a novel framework featuring two stages for distilling knowledge in multimodal large models. The initial stage involves multimodal pre-training to align multimodal features through a projection layer. The second stage, termed multimodal competitive distillation, establishes a bidirectional feedback loop encompassing: 1) Multimodal instruction adjustment to ensure student responses align with teacher-provided multimodal instructions. 2) Multimodal evaluation to identify challenging multimodal instructions. 3) Multimodal augmentation, where new instructions are generated and combined with original images to create a new multimodal instruction dataset for training student models. Evaluation on datasets like ScienceQA [80], SEED-Bench [66], and LLaVA Test Set [78] demonstrates that CoMD surpasses existing models in inference tasks and zero-shot settings. Park *et al.*[101] developed a localized visual commonsense model by sampling localized commonsense knowledge from LLMs. Users can specify regions as inputs, and a separately trained critic model selects high-quality examples. Empirical results and human evaluations in the zero-shot setting indicate that this distillation method produces a more accurate VL inference model compared to simply passing generated reference expressions to baseline LLMs. Similarly, Hu *et al.*[49] introduced Instruction Tuning for Visual Program Distillation (VPD). VPD leverages LLMs' inference capability by sampling multiple candidate programs, executing and verifying them, and translating

Table 9. Applications of LLM distillation.

Models	Distillation Scenario	Teacher Model	Compression Rate	Evaluation Task	Comparison with Teacher Model
HuatuoGPT [162]	Healthcare	GPT-3.5$_{turbo}$	/	cMedQA2[164]/webMedQA[39]/Huatuo-26M[67]	25.1/18.6 (135% performance)
Chatdoctor [72]	Healthcare	GPT-3.5$_{turbo}$	/	HealthCareMagic100k[72]	84.5/84.1 (100% performance)
PMC-LLaMA [153]	Healthcare	ChatGPT	/	PubMedQA[58]/MedMCQA[99]/USMLE[57]	64.4/55.0 (117% performance)
DARWIN [156]	Education	GPT-3$_{175B}$	25×	SciQ[150]/FAIR[113]	93.6/82.7 (113% performance)
WizardMath [81]	Education	ChatGPT	/	GSM8k[27]/MATH[42]	52.2/57.5 (91% performance)
K2 [29]	Education	LLaMA$_{7B}$	1750×	GeoBench[29]	34.6/24.6 (141% performance)
LawyerLLaMA [50]	Law	GPT-3.5$_{turbo}$	/	C3[121]/CMNLI[157]/SciQ[150]/PIQA[8]	/
ChatLaw [28]	Law	Ziya-LLaMA$_{13B}$	/	Legal Multiple-choice Questions[28]	/

correct programs into language descriptions of inference steps for VLM distillation. Extensive experiments have shown that VPD enhances counting, spatial relationship understanding, and combinatorial reasoning abilities in VLMs, achieving state-of-the-art performance in challenging visual tasks such as MMBench [79], OK-VQA [84], A-OKVQA [114], TallyQA [1], POPE [71], and Hateful Memes [62].

4 APPLICATIONS

In this section, we briefly explore the applications of LLM distillation in various critical domains such as healthcare, education, and law.

4.1 Healthcare

Healthcare represents a critical domain deeply intertwined with human well-being. Since the inception of ChatGPT, numerous endeavors have endeavored to harness the prowess of ChatGPT and other LLMs in the realm of medicine. For example, Zhang *et al.*[162] introduced HuatuoGPT, a specialized LLM designed for medical consultations. By distilling data from ChatGPT and integrating real-world insights from physicians through supervised fine-tuning, HuatuoGPT incorporates a reward model aimed at synergizing the strengths derived from both datasets. Empirical results demonstrate that HuatuoGPT achieves state-of-the-art performance in medical consultations, outperforming GPT-3.5$_{turbo}$ across various metrics evaluated on GPT-4, including manual assessments and medical benchmark datasets. Li *et al.*[72] highlight the scarcity of LLMs specifically tailored to medical domains. Using LLaMA as a developmental and evaluative platform, they explored two enhancement strategies: model fine-tuning and knowledge integration to augment the efficacy of LLMs as medical chatbots. Fine-tuning the dialogue model on a dataset comprising 100K patient physiological dialogues sourced from online medical consultation platforms, their experiments demonstrate that the Chatdoctor model surpasses ChatGPT in terms of accuracy and F1 score. Furthermore, Wu *et al.*[153] introduced PMC-LLaMA, which amalgamates 4.8M biomedical academic papers and 30K medical textbooks to infuse data-centric knowledge, coupled with exhaustive fine-tuning tailored to specific domain directives. With a modest parameter count of 13B, PMC-LLaMA demonstrates outstanding performance, surpassing ChatGPT across various public medical question answering benchmarks.

4.2 Education

Education represents another critical domain where LLMs show significant promise. Current research demonstrates that LLMs can achieve proficiency comparable to students in standardized exams across various mathematical disciplines such as physics and computer science [2]. Xie *et al.*[156] introduced DARWIN, a framework aimed at enhancing natural sciences by accelerating and enriching the automation of discovery processes. This

approach incorporates the Scientific Instruction Generation (SIG) model, which integrates structured and unstructured scientific knowledge from public datasets and literature. By eliminating the need for manual extraction or domain-specific knowledge graphs, DARWIN achieves state-of-the-art performance across diverse scientific tasks. Luo *et al.*[81] proposed WizardMath, which utilizes the Reinforcement Learning from Evol-Instruct Feedback (RLEIF) technique to enhance the mathematical reasoning capabilities of LLaMA-2 [130]. This method employs math-specific Evol-Instruct to generate diverse mathematical instruction data, subsequently training the Instruction Reward Model (IRM) and the Process Supervised Reward Model (PRM) [160]. The IRM evaluates the quality of evolutionary instructions, while the PRM receives feedback at each step of the solution process. Through extensive experimentation on two mathematical reasoning benchmarks, GSM8k [27] and MATH [42], WizardMath significantly outperforms other open-source LLMs. Furthermore, Deng *et al.*[29] introduced K2, a LLM tailored for geoscience, and established the GeoBench, the first geoscience benchmark, to evaluate LLMs within this domain.

4.3 Law

Law, a domain rich in professional expertise, has recently adopted LLMs to address various legal tasks, such as legal document analysis [9] and legal document generation [25]. Huang *et al.*[50] integrated legal expertise into the continuous training phase of LLaMA by employing carefully designed supervised fine-tuning tasks. These tasks aimed to impart professional skills to the model while mitigating the issue of model-generated illusions. To enhance training, they introduced a retrieval module that extracts relevant legal articles before the model generates responses. Similarly, Cui *et al.*[28] integrated legal-specific data into LLaMA, resulting in the creation of ChatLaw. Concerned with the accuracy of reference retrieval from legal datasets, they developed a hybrid approach combining vector database retrieval and keyword-based retrieval. This approach addresses hallucination concerns and improves accuracy by implementing a self-attention mechanism. This mechanism enhances the ability of large models to correct errors within reference data, thereby improving coherence and augmenting problem-solving proficiency in legal contexts.

5 CHALLENGES AND FUTURE DIRECTIONS

5.1 Unified Evaluation Benchmark

The existing benchmark for evaluating knowledge distillation primarily falls into four categories: 1) General Language Understanding Evaluation (GLUE) Benchmark [137]: This benchmark consists of nine sentence-level classification tasks, including language acceptability [148], sentiment analysis [118], text similarity [15], entailment detection [31], and natural language inference [107]. It is commonly utilized to assess distillation methods employing BERT as the teacher model. 2) Multimodal Multitask Learning Understanding (MMLU) Benchmark [41]: This benchmark serves as a universal evaluation tool for assessing the multitasking knowledge comprehension abilities of LLMs. It covers various domains such as mathematics, computer science, humanities, and social sciences, featuring tasks of varying difficulty levels from basic to advanced. 3) BIG Bench [119]: A collaborative effort to create a comprehensive evaluation benchmark that explores the capabilities of existing LLMs across a diverse range of tasks. It includes 204 tasks spanning linguistics, child development, mathematics, common sense reasoning, biology, physics, social prejudice, software development, and more. 4) Human-Evaluated Language Models (HELM) Benchmark [76]: This is a holistic evaluation benchmark comprising 16 core scenarios and 7 indicator categories. It integrates various previously proposed evaluation benchmarks to provide a holistic assessment of LLM performance. These benchmarks collectively cover a wide array of mainstream LLM evaluation tasks. Additionally, there are specialized evaluation benchmarks tailored to specific tasks, such as TyDiQA [26] for evaluating multilingual knowledge utilization and MGSM [115] for assessing multilingual mathematical

reasoning. As large models continue to evolve, evaluation criteria are continually updated, and developing a unified evaluation standard for knowledge distillation remains a promising avenue of research.

5.2 Advanced Algorithms

Current methodologies primarily aim to equip student models with specific capabilities. For example, symbolic knowledge distillation [152] leverages LLMs to gather and filter data, extracting high-quality commonsense maps for training commonsense models. Similarly, DISCO [23] employs LLMs to acquire counterfactual data, which is then filtered using a large teacher Natural Language Inference model to improve students' proficiency in natural language reasoning tasks. As open-source LLMs continue to evolve, exploring white-box distillation algorithms for LLMs could prove to be an effective approach for integrating multiple capabilities. Furthermore, the current development pace of MLLMs distillation lags behind that of LLMs. Thus, investigating more advanced MLLMs distillation algorithms could facilitate the integration of multiple modalities more effectively.

5.3 Interpretability

Stanton *et al.*[120] explore the interpretability of knowledge distillation and introduce the concept of matching degree to enhance its reliability. Their study reveals several significant insights: 1) The relationship between student models' generalization performance and matching degree is not uniformly consistent. Excluding self-distillation, models with the best generalization performance do not always exhibit the highest fidelity. 2) There is a notable correlation between student models' fidelity and the calibration of the distillation process. Although the most faithful student model may not always achieve the highest accuracy, it consistently shows superior calibration. 3) Optimization during the knowledge distillation process is challenging, resulting in lower fidelity. Similarly, in the era of large language models, knowledge distillation faces comparable difficulties. For example, current methods struggle to elucidate how CoT-distillation imparts CoT capability to student language models or to determine the required amount of data for fine-tuning instructions. Therefore, integrating interpretability into the process is crucial for advancing LLM knowledge distillation. This integration not only aids in evaluating model distillation but also enhances the reliability and predictability of models in production

6 CONCLUSION

In this survey, we systematically investigate the knowledge distillation algorithms from three perspectives: methods, evaluation, and application. Compared to smaller models, distillation in larger models faces more challenges. Despite considerable efforts by existing algorithms to tackle these challenges, many still rely on frameworks initially tailored for compressing smaller models, while the challenge of compressing large models still exists. In the future, while ensuring the universality and generalization of LLMs, it becomes imperative to delve deeper into developing more efficient and effective compression algorithms. This survey aims to furnish valuable references, shed light on the current landscape, and advocate for ongoing exploration of this pivotal theme to enable the effective design, learning, and application of various distillation objectives within the teacher-student framework.

REFERENCES

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: answering complex counting questions. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. 8076–8084.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2023. Generalized knowledge distillation for auto-regressive language models. *arXiv preprint arXiv:2306.13649* (2023).

[4] Abdolmaged Alkhulaifi, Fahad Alsahli, and Irfan Ahmad. 2021. Knowledge distillation in deep learning and its applications. *PeerJ Computer Science* 7 (2021), e474.

[5] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2357–2367.

[6] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. *GitHub* (2023).

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[8] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7432–7439.

[9] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can GPT-3 Perform Statutory Reasoning?. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23)*. 22–31.

[10] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*. 12–58.

[11] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*. Association for Computational Linguistics, 131–198.

[12] 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 1877–1901.

[13] Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. A Systematic Study of Knowledge Distillation for Natural Language Generation with Pseudo-Target Training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14632–14659.

[14] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 9560–9572.

[15] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1–14.

[16] Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue Chain-of-Thought Distillation for Commonsense-aware Conversational Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5606–5632.

[17] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* (2024).

[18] Hailin Chen, Amrita Saha, Steven Hoi, and Shafiq Joty. 2023. Personalized Distillation: Empowering Open-Sourced LLMs with Adaptive Learning for Code Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6737–6749.

[19] Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. MCC-KD: Multi-CoT Consistent Knowledge Distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 6805–6820.

[20] Liqun Chen, Shuyang Dai, Yunchen Pu, Erjin Zhou, Chunyuan Li, Qinliang Su, Changyou Chen, and Lawrence Carin. 2018. Symmetric variational autoencoder and connections to adversarial learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 661–669.

[21] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[22] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via Language Model In-context Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 719–730.

[23] Zeming Chen, Qiyue Gao, Kyle Richardson, Antoine Bosselut, and Ashish Sabharwal. 2022. DISCO: Distilling Phrasal Counterfactuals with Large Language Models. *arXiv preprint arXiv:2212.10534* (2022).

[24] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. (2023).

[25] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. ChatGPT goes to law school. *J. Legal Educ.* 71 (2021), 387.

[26] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics* 8 (2020), 454–470.
[27] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
[28] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092* (2023).
[29] Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2023. K2: A foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint arXiv:2306.05064* (2023).
[30] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proc. IEEE* 108, 4 (2020), 485–532.
[31] William B Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
[32] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).
[33] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2368–2378.
[34] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. *arXiv preprint arXiv:2301.12726* (2023).
[35] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* 9 (2021), 346–361.
[36] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).
[37] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.
[38] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.
[39] Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to Chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making* 19, 2 (2019), 91–100.
[40] Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. 2018. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European conference on computer vision (ECCV)*. 784–800.
[41] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.
[42] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).
[43] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[44] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071* (2022).
[45] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689* (2022).
[46] Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. DynaBERT: dynamic BERT with adaptive width and depth. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 9782–9793.
[47] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander 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. *arXiv preprint arXiv:2305.02301* (2023).
[48] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. 2023. Teacher-student architecture for knowledge distillation: A survey. *arXiv preprint arXiv:2308.04268* (2023).
[49] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models. *arXiv preprint arXiv:2312.03052* (2023).

[50] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer LLaMA Technical Report. *arXiv preprint arXiv:2305.15062* (2023).
[51] Tianyu Huang, Weisheng Dong, Fangfang Wu, Xin Li, and Guangming Shi. 2023. Uncertainty-Driven Knowledge Distillation for Language Model Compression. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* (2023), 2850–2858.
[52] Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. 2022. In-context Learning Distillation: Transferring Few-shot Learning Ability of Pre-trained Language Models. *arXiv preprint arXiv:2212.10670* (2022).
[53] Ferenc Huszár. 2015. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101* (2015).
[54] Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *arXiv preprint arXiv:2305.12870* (2023).
[55] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4163–4174.
[56] Zhanming Jie and Wei Lu. 2023. Leveraging training data in few-shot prompting for numerical reasoning. *arXiv preprint arXiv:2305.18170* (2023).
[57] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.
[58] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2567–2577.
[59] Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *Advances in Neural Information Processing Systems*, Vol. 36.
[60] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
[61] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A Dataset for Question Answering via Sentence Composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8082–8090.
[62] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2611–2624.
[63] Yoon Kim and Alexander M Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1317–1327.
[64] Hayeon Lee, Rui Hou, Jongpil Kim, Davis Liang, Sung Ju Hwang, and Alexander Min. 2023. A Study on Knowledge Distillation from Weak Teacher for Scaling Up Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11239–11246.
[65] Hyoje Lee, Yeachan Park, Hyun Seo, and Myungjoo Kang. 2023. Self-knowledge distillation via dropout. *Computer Vision and Image Understanding* 233 (2023), 103720.
[66] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).
[67] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. *arXiv preprint arXiv:2305.01526* (2023).
[68] Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic Chain-of-Thought Distillation: Small Models Can Also" Think" Step-by-Step. *arXiv preprint arXiv:2306.14050* (2023).
[69] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726* (2022).
[70] Xinwei Li, Li Lin, Shuai Wang, and Chen Qian. 2023. Unlock the power: Competitive distillation for multi-modal large language models. *arXiv preprint arXiv:2311.08213* (2023).
[71] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).
[72] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* 15, 6 (2023).
[73] Chen Liang, Haoming Jiang, Zheng Li, Xianfeng Tang, Bin Yin, and Tuo Zhao. 2023. Homodistil: Homotopic task-agnostic distillation of pre-trained transformers. *arXiv preprint arXiv:2302.09632* (2023).
[74] Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*. PMLR, 20852–20867.
[75] Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2020. MixKD: Towards Efficient Distillation of Large-scale Language Models. In *International Conference on Learning Representations*.

[76] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[77] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[78] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in neural information processing systems*, Vol. 36.
[79] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281* (2023).
[80] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, Vol. 35. 2507–2521.
[81] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583* (2023).
[82] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 142–150.
[83] 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).
[84] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.
[85] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text *(RecSys '13)*. 165–172.
[86] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 975–984.
[87] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2381–2391.
[88] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2791–2809.
[89] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 11048–11064.
[90] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 5191–5198.
[91] Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. XtremeDistil: Multi-stage Distillation for Massive Multilingual Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2221–2234.
[92] Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2021. DART: Open-Domain Structured Data Record to Text Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 432–447.
[93] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1797–1807.
[94] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4885–4901.
[95] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
[96] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-GAN: training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 271–279.
[97] Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. *arXiv preprint arXiv:2109.01653* (2021).
[98] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. 27730–27744.

[99] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*. PMLR, 248–260.

[100] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1946–1958.

[101] Jae Sung Park, Jack Hessel, Khyathi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Youngjae Yu, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, et al. 2023. Localized Symbolic Knowledge Distillation for Visual Commonsense Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[102] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2080–2094.

[103] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[104] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[105] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2463–2473.

[106] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *Cornell University - arXiv,Cornell University - arXiv* (Jun 2018).

[107] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[108] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[109] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[110] Subhro Roy and Dan Roth. 2015. Solving General Arithmetic Word Problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1743–1752.

[111] 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).

[112] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

[113] Matthias Scheffler, Martin Aeschlimann, Martin Albrecht, Tristan Bereau, Hans-Joachim Bungartz, Claudia Felser, Mark Greiner, Axel Groß, Christoph T Koch, Kurt Kremer, et al. 2022. FAIR data enabling new horizons for materials research. *Nature* 604, 7907 (2022), 635–642.

[114] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*. Springer, 146–162.

[115] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

[116] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. 7059–7073.

[117] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 455–465.

[118] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[119] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023).

[120] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does knowledge distillation really work?. In *Advances in Neural Information Processing Systems*, Vol. 34. 6906–6919.

[121] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *Transactions of the Association for Computational Linguistics* 8 (2020), 141–155.

[122] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4323–4332.

[123] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2158–2170.

[124] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. 3319–3328.

[125] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4149–4158.

[126] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136* (2019).

[127] 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).

[128] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[129] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1522–1531.

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

[131] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3632–3636.

[132] Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2819–2826.

[133] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).

[134] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962* (2019).

[135] Nirali Vaghani and Mansi Thummar. [n. d.]. Flipkart product reviews with sentiment dataset, 2023. *URL https://www. kaggle. com/dsv/4940809* ([n. d.]).

[136] Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003* (2023).

[137] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 353–355.

[138] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[139] Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

[140] Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164* (2023).

[141] Lin Wang and Kuk-Jin Yoon. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 3048–3068.

[142] Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879* (2023).

[143] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 5776–5788.

[144] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560* (2022).

[145] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv e-prints* (2022), arXiv–2204.

[146] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. *arXiv preprint arXiv:2306.05087* (2023).

[147] Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, et al. 2023. Democratizing Reasoning Ability: Tailored Learning from Large Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 1948–1966.

[148] Alex Warstadt, Amanpreet Singh, and Samuel Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641.

[149] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

[150] Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 94–106.

[151] Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-Divergence Minimization for Sequence-Level Knowledge Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10817–10834.

[152] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4602–4625.

[153] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *arXiv preprint arXiv:2305.10415* 6 (2023).

[154] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402* (2023).

[155] Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. 2023. AD-KD: Attribution-Driven Knowledge Distillation for Language Model Compression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[156] Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. 2023. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565* (2023).

[157] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4762–4772.

[158] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[159] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A Few-shot Learning Challenge for Cross-task Generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7163–7189.

[160] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825* (2023).

[161] Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. 2023. Towards the law of capacity gap in distilling language models. *arXiv preprint arXiv:2311.07052* (2023).

[162] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. HuatuoGPT, towards Taming Language Model to Be a Doctor. *arXiv preprint arXiv:2305.15075* (2023).

[163] Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Yiran Chen. 2023. ReAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 1128–1136.

[164] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access* 6 (2018), 74061–74071.

[165] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 649–657.

[166] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[167] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 121–132.

[168] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT Learns to Teach: Knowledge Distillation with Meta Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7037–7049.
[169] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279* (2023).
[170] Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xingwei Long, and Bowen Zhou. 2023. PaD: Program-aided Distillation Specializes Large Models in Reasoning. *arXiv preprint arXiv:2305.13888* (2023).

Received 28 February 2024; revised 24 June 2024; accepted 6 September 2024