A Survey of Large Language Models

Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen

Abstract—Ever since the Turing Test was proposed in the 1950s, humans have explored the mastering of language intelligence by machine. Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable artificial intelligence (AI) algorithms for comprehending and grasping a language. As a major approach, *language modeling* has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various natural language processing (NLP) tasks. Since the researchers have found that model scaling can lead to an improved model capacity, they further investigate the scaling effect by increasing the parameter scale to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement, but also exhibit some special abilities (*e.g.,* in-context learning) that are not present in small-scale language models (*e.g.,* BERT). To discriminate the language models in different parameter scales, the research community has coined the term *large language models (LLM)* for the PLMs of significant size (*e.g.,* containing tens or hundreds of billions of parameters). Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT (a powerful AI chatbot developed based on LLMs), which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. Considering this rapid technical progress, in this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Besides, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions. This survey provides an up-to-date review of the literature on LLMs, which can be a useful resource for both researchers and engineers.

Index Terms—Large Language Models; Emergent Abilities; Adaptation Tuning; Utilization; Alignment; Capacity Evaluation

✦

1 INTRODUCTION

LANGUAGE is a prominent ability in human beings to express and communicate, which develops in early childhood and evolves over a lifetime [1, 2]. Machines, however, cannot naturally grasp the abilities of understanding and communicating in the form of human language, unless equipped with powerful artificial intelligence (AI) algorithms. It has been a longstanding research challenge to achieve this goal, to enable machines to read, write, and communicate like humans [3].

Technically, *language modeling (LM)* is one of the major approaches to advancing language intelligence of machines. In general, LM aims to model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens. The research of LM has received extensive attention in the literature, which can be divided into four major development stages:

• *Statistical language models (SLM)*. SLMs [4–7] are developed based on *statistical learning* methods that rose in the 1990s. The basic idea is to build the word prediction model based on the Markov assumption, *e.g.,* predicting the next word based on the most recent context. The SLMs with

a fixed context length n are also called n-gram language models, *e.g.,* bigram and trigram language models. SLMs have been widely applied to enhance task performance in information retrieval (IR) [8, 9] and natural language processing (NLP) [10–12]. However, they often suffer from the curse of dimensionality: it is difficult to accurately estimate high-order language models since an exponential number of transition probabilities need to be estimated. Thus, specially designed smoothing strategies such as back-off estimation [13] and Good–Turing estimation [14] have been introduced to alleviate the data sparsity problem.

• *Neural language models (NLM)*. NLMs [15–17] characterize the probability of word sequences by neural networks, *e.g.,* recurrent neural networks (RNNs). As a remarkable contribution, the work in [15] introduced the concept of *distributed representation* of words and built the word prediction function conditioned on the aggregated context features (*i.e.,* the distributed word vectors). By extending the idea of learning effective features for words or sentences, a general neural network approach was developed to build a unified solution for various NLP tasks [18]. Further, word2vec [19, 20] was proposed to build a simplified shallow neural network for learning distributed word representations, which were demonstrated to be very effective across a variety of NLP tasks. These studies have initiated the use of language models for representation learning (beyond word sequence modeling), having an important impact on the field of NLP.

• *Pre-trained language models (PLM).* As an early attempt, ELMo [21] was proposed to capture context-aware word representations by first pre-training a bidirectional LSTM (biLSTM) network (instead of learning fixed word representations) and then fine-tuning the biLSTM network according to specific downstream tasks. Further, based on the highly parallelizable Transformer architecture [22] with self-attention mechanisms, BERT [23] was proposed by pre-training bidirectional language models with specially designed pre-training tasks on large-scale unlabeled corpora. These pre-trained context-aware word representations are very effective as general-purpose semantic features, which have largely raised the performance bar of NLP tasks. This study has inspired a large number of follow-up work, which sets the "*pre-training* and *fine-tuning*" learning paradigm. Following this paradigm, a great number of studies on PLMs have been developed, introducing either different architectures [24, 25] (*e.g.,* GPT-2 [26] and BART [24]) or improved pre-training strategies [27–29]. In this paradigm, it often requires fine-tuning the PLM for adapting to different downstream tasks.

• *Large language models (LLM).* Researchers find that scaling PLM (*e.g.,* scaling model size or data size) often leads to an improved model capacity on downstream tasks (*i.e.,* following the scaling law [30]). A number of studies have explored the performance limit by training an ever larger PLM (*e.g.,* the 175B-parameter GPT-3 and the 540B-parameter PaLM). Although scaling is mainly conducted in model size (with similar architectures and pre-training tasks), these large-sized PLMs display different behaviors from smaller PLMs (*e.g.,* 330M-parameter BERT and 1.5B-parameter GPT-2) and show surprising abilities (called *emergent abilities* [31]) in solving a series of complex tasks. For example, GPT-3 can solve few-shot tasks through *in-context learning*, whereas GPT-2 cannot do well. Thus, the research community coins the term "*large language models (LLM)*"[1] for these large-sized PLMs [32–35]. A remarkable application of LLMs is *ChatGPT*[2] that adapts the LLMs from the GPT series for dialogue, which presents an amazing conversation ability with humans.

In the existing literature, PLMs have been widely discussed and surveyed [36–39], while LLMs are seldom reviewed in a systematic way. To motivate our survey, we first highlight three major differences between LLMs and PLMs. First, LLMs display some surprising emergent abilities that may not be observed in previous smaller PLMs. These abilities are key to the performance of language models on complex tasks, making AI algorithms unprecedently powerful and effective. Second, LLMs would revolutionize the way that humans develop and use AI algorithms. Unlike small PLMs, the major approach to accessing LLMs is through the prompting interface (*e.g.,* GPT-4 API). Humans have to understand how LLMs work and format their tasks in a way that LLMs can follow. Third, the development of LLMs no longer draws a clear distinction between research and engineering. The training of LLMs requires extensive practical experiences in large-scale data processing and distributed parallel training. To develop capable LLMs, researchers have to solve complicated engineering issues, working with engineers or being engineers.

Nowadays, LLMs are posing a significant impact on the AI community, and the advent of ChatGPT and GPT-4 leads to the rethinking of the possibilities of artificial general intelligence (AGI). OpenAI has published a technical article entitled "*Planning for AGI and beyond*", which discusses the short-term and long-term plans to approach AGI [40], and a more recent paper has argued that GPT-4 might be considered as an early version of an AGI system [41]. The research areas of AI are being revolutionized by the rapid progress of LLMs. In the field of NLP, LLMs can serve as a general-purpose language task solver (to some extent), and the research paradigm has been shifting towards the use of LLMs. In the field of IR, traditional search engines are challenged by the new information seeking way through AI chatbots (*i.e.,* ChatGPT), and *New Bing*[3] presents an initial attempt that enhances the search results based on LLMs. In the field of CV, the researchers try to develop ChatGPT-like vision-language models that can better serve multimodal dialogues [42–45], and GPT-4 [46] has supported multimodal input by integrating the visual information. This new wave of technology would potentially lead to a prosperous ecosystem of real-world applications based on LLMs. For instance, Microsoft 365 is being empowered by LLMs (*i.e.,* Copilot) to automate the office work, and OpenAI supports the use of plugins in ChatGPT for implementing special functions.

Despite the progress and impact, the underlying principles of LLMs are still not well explored. Firstly, it is mysterious why emergent abilities occur in LLMs, instead of smaller PLMs. As a more general issue, there lacks a deep, detailed investigation of the key factors that contribute to the superior abilities of LLMs. It is important to study when and how LLMs obtain such abilities [47]. Although there are some meaningful discussions about this problem [31, 47], more principled investigations are needed to uncover the "*secrets*" of LLMs. Secondly, it is difficult for the research community to train capable LLMs. Due to the huge demand of computation resources, it is very costly to carry out repetitive, ablating studies for investigating the effect of various strategies for training LLMs. Indeed, LLMs are mainly trained by industry, where many important training details (*e.g.,* data collection and cleaning) are not revealed to the public. Thirdly, it is challenging to align LLMs with human values or preferences. Despite the capacities, LLMs are also likely to produce toxic, fictitious, or harmful contents. It requires effective and efficient control approaches to eliminating the potential risk of the use of LLMs [46].

Faced with both opportunities and challenges, it needs more attention on the research and development of LLMs. In order to provide a basic understanding of LLMs, this survey conducts a literature review of the recent advances in LLMs from four major aspects, including *pre-training* (how to pre-train a capable LLM), *adaptation tuning* (how to effectively tune pre-trained LLMs from the two perspectives of effectiveness and safety), *utilization* (how to use LLMs for solving various downstream tasks) and *capability eval-*

---

1. Note that a LLM is not necessarily more capable than a small PLM, and emergent abilities may not occur in some LLMs.

2. https://openai.com/blog/chatgpt/

3. https://www.bing.com/new

uation (how to evaluate the abilities of LLMs and existing empirical findings). We thoroughly comb the literature and summarize the key findings, techniques, and methods of LLMs. For this survey, we also create a GitHub project website by collecting the supporting resources for LLMs, at the link https://github.com/RUCAIBox/LLMSurvey. We are also aware of several related review articles on PLMs or LLMs [32, 36, 38, 39, 43, 48–54]. These papers either discuss PLMs or some specific (or general) aspects of LLMs. Compared with them, we focus on the techniques and methods to develop and use LLMs and provide a relatively comprehensive reference to important aspects of LLMs.

The remainder of this survey is organized as follows: Section 2 introduces the background for LLMs, with the terminology, settings, resources, and organization outline, followed by the summarization of available resources for developing LLMs in Section 3. Sections 4, 5, 6, and 7 review and summarize the recent progress from the four aspects of pre-training, adaptation tuning, utilization, and capacity evaluation, respectively. Finally, we conclude the survey in Section 8 by summarizing the major findings and discuss the remaining issues for future work.

2 OVERVIEW

In this section, we present an overview about the background of LLMs and then summarize the technical evolution of the GPT-series models.

2.1 Background for LLMs

Typically, *large language models* (LLMs) refer to Transformer language models that contain hundreds of billions (or more) of parameters[4], which are trained on massive text data [32], such as GPT-3 [55], PaLM [56], Galactica [35], and LLaMA [57]. LLMs exhibit strong capacities to understand natural language and solve complex tasks (via text generation). To have a quick understanding of how LLMs work, this part introduces the basic background for LLMs, including scaling laws, emergent abilities and key techniques.

Scaling Laws for LLMs. Currently, LLMs are mainly built upon the Transformer architecture [22], where multi-head attention layers are stacked in a very deep neural network. Existing LLMs adopt similar Transformer architectures and pre-training objectives (*e.g.,* language modeling) as small language models. While, LLMs largely scale the model size, data size, and total compute (orders of magnification). Extensive research has shown that scaling can largely improve the model capacity of LLMs [26, 55, 56]. Thus, it is useful to establish a quantitative approach to characterizing the scaling effect. Next, we introduce two representative scaling laws for Transformer language models [30, 34].

● *KM scaling law*[5]. In 2020, Kaplan et al. [30] (the OpenAI team) firstly proposed to model the power-law relationship

of model performance with respective to three major factors, namely model size (N), dataset size (D), and the amount of training compute (C), for neural language models. Given a compute budget c, they empirically presented three basic formulas for the scaling law[6]:

$$
\begin{aligned}
L(N) &= \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13} \quad (1) \\
L(D) &= \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13} \\
L(C) &= \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}
\end{aligned}
$$

where $L(\cdot)$ denotes the cross entropy loss in nats. The three laws were derived by fitting the model performance with varied data sizes (22M to 23B tokens), model sizes (768M to 1.5B non-embedding parameters) and training compute, under some assumptions (*e.g.,* the analysis of one factor should be not bottlenecked by the other two factors). They showed that the model performance has a strong dependence relation on the three factors.

● *Chinchilla scaling law*. As another representative study, Hoffmann et al. [34] (the Google DeepMind team) proposed an alternative form for scaling laws to instruct the compute-optimal training for LLMs. They conducted rigorous experiments by varying a larger range of model sizes (70M to 16B) and data sizes (5B to 500B tokens), and fitted a similar scaling law yet with different coefficients as below [34]:

$$
L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}, \quad (2)
$$

where $E = 1.69, A = 406.4, B = 410.7, \alpha = 0.34$ and $\beta = 0.28$. By optimizing the loss $L(N, D)$ under the constraint $C \approx 6ND$, they showed that the optimal allocation of compute budget to model size and data size can be derived as follows:

$$
N_{opt}(C) = G\left(\frac{C}{6}\right)^a, \quad D_{opt}(C) = G^{-1}\left(\frac{C}{6}\right)^b, \quad (3)
$$

where $a = \frac{\alpha}{\alpha+\beta}$, $b = \frac{\beta}{\alpha+\beta}$ and G is a scaling coefficient that can be computed by A, B, α and β. As analyzed in [34], given an increase in compute budget, the KM scaling law favors a larger budget allocation in model size than the data size, while the Chinchilla scaling law argues that the two sizes should be increased in equal scales, *i.e.,* having similar values for a and b in Equation (3).

Though with some restricted assumptions, these scaling laws provide an intuitive understanding of the scaling effect, making it feasible to predict the performance of LLMs during training [46]. However, some abilities (*e.g.,* in-context learning [55]) are unpredictable according to the scaling law, which can be observed only when the model size exceeds a certain level (as discussed below).

---

4. In existing literature, there is no formal consensus on the minimum parameter scale for LLMs, since the model capacity is also related to data size and total compute. In this survey, we take a slightly loose definition of LLMs, and mainly focus on discussing language models with a model size larger than 10B.

5. Since there was not a model trained following this law in the original paper, we took the last names of the two co-first authors to name this scaling law.

6. Here, N_c, D_c and C_c are measured in the number of non-embedding parameters, the number of training tokens and the number of FP-days, respectively. According to the original paper [30], C_c and C should be denoted by C_c^{min} and C_{min}, corresponding to the optimal use of compute. While, we use the simplified notations for ease of discussions.

Emergent Abilities of LLMs. In the literature [31], *emergent abilities* of LLMs are formally defined as "the abilities that are not present in small models but arise in large models", which is one of the most prominent features that distinguish LLMs from previous PLMs. It further introduces a notable characteristic when emergent abilities occur [31]: performance rises significantly above random when the scale reaches a certain level. By analogy, such an emergent pattern has close connections with the phenomenon of *phase transition* in physics [31, 58]. In principle, emergent abilities can be defined in relation to some complex tasks [31, 59], while we are more concerned with general abilities that can be applied to solve a variety of tasks. Here, we briefly introduce three typical emergent abilities for LLMs and representative models that possess such an ability[7].

• *In-context learning.* The in-context learning (ICL) ability is formally introduced by GPT-3 [55]: assuming that the language model has been provided with a natural language instruction and/or several task demonstrations, it can generate the expected output for the test instances by completing the word sequence of input text, without requiring additional training or gradient update[8]. Among the GPT-series models, the 175B GPT-3 model exhibited a strong ICL ability in general, but not the GPT-1 and GPT-2 models. While, such an ability also depends on the specific downstream task. For example, the ICL ability can emerge on the arithmetic tasks (*e.g.,* the 3-digit addition and subtraction) for the 13B GPT-3, but 175B GPT-3 even cannot work well on the Persian QA task [31].

• *Instruction following.* By fine-tuning with a mixture of multi-task datasets formatted via natural language descriptions (called *instruction tuning*), LLMs are shown to perform well on unseen tasks that are also described in the form of instructions [28, 61, 62]. With instruction tuning, LLMs are enabled to follow the task instructions for new tasks without using explicit examples, thus having an improved generalization ability. According to the experiments in [62], instruction-tuned LaMDA-PT [63] started to significantly outperform the untuned one on unseen tasks when the model size reached 68B, but not for 8B or smaller model sizes. A recent study [64] found that a model size of 62B is at least required for PaLM to perform well on various tasks in four evaluation benchmarks (*i.e.,* MMLU, BBH, TyDiQA and MGSM), though a much smaller size might suffice for some specific tasks (*e.g.,* MMLU).

• *Step-by-step reasoning.* For small language models, it is usually difficult to solve complex tasks that involve multiple reasoning steps, *e.g.,* mathematical word problems. While, with the chain-of-thought (CoT) prompting strategy [33], LLMs can solve such tasks by utilizing the prompting mechanism that involves intermediate reasoning steps for deriving the final answer. This ability is speculated to be potentially obtained by training on code [33, 47]. An empirical study [33] has shown that CoT prompting can bring performance gains (on arithmetic reasoning benchmarks) when applied to PaLM and LaMDA variants with a model size larger than 60B, while its advantage over the standard prompting becomes more evident when the model size exceeds 100B. Besides, the performance improvement with CoT prompting seems to be also varied for different tasks, *e.g.,* GSM8K > MAWPS > SWAMP for PaLM [33].

Key Techniques for LLMs. It has been a long way that LLMs evolve into the current state: *general* and *capable* learners. In the development process, a number of important techniques are proposed, which largely improve the capacity of LLMs. Here, we briefly list several important techniques that (potentially) lead to the success of LLMs, as follows.

• *Scaling*. As discussed in previous parts, there exists an evident scaling effect in Transformer language models: larger model/data sizes and more training compute typically lead to an improved model capacity [30, 34]. As two representative models, GPT-3 and PaLM explored the scaling limits by increasing the model size to 175B and 540B, respectively. Furthermore, since compute budget is usually limited, scaling laws can be employed to conduct a more compute-efficient allocation of the compute resources. For example, Chinchilla (with more training tokens) outperforms its counterpart model Gopher (with a larger model size) by increasing the data scale with the same compute budget [34]. While, it should be noted that data scaling should be with careful cleaning process, since the quality of pre-training data plays a key role in the model capacity.

• *Training*. Due to the huge model size, it is very challenging to successfully train a capable LLM. Distributed training algorithms are needed to learn the network parameters of LLMs, in which various parallel strategies are often jointly utilized. To support distributed training, several optimization frameworks have been released to facilitate the implementation and deployment of parallel algorithms, such as DeepSpeed [65] and Megatron-LM [66–68]. Besides, optimization tricks are also important for training stability and model performance, *e.g.,* restart to overcome training loss spike [56] and mixed precision training [69]. More recently, GPT-4 [46] proposes to develop special infrastructure and optimization methods that reliably predict the performance of large models with much smaller models.

• *Ability eliciting*. After being pre-trained on large-scale corpora, LLMs are endowed with potential abilities as general-purpose task solvers. While, these abilities might not be explicitly exhibited when LLMs perform some specific tasks. As the technical approach, it is useful to design suitable task instructions or specific in-context learning strategies to elicit such abilities. For instance, chain-of-thought prompting has been shown to be useful to solve complex reasoning tasks by including intermediate reasoning steps. Besides, we can further perform instruction tuning on LLMs with task descriptions expressed in natural language, for improving the generalizability of LLMs on unseen tasks. While, these techniques mainly correspond to the emergent abilities of LLMs, which may not show the same effect on small language models.

• *Alignment tuning*. Since LLMs are trained to capture the data characteristics of pre-training corpora (including

---

7. It is difficult to accurately examine the critical size for emergent abilities of LLMs (*i.e.,* the minimum size to possess an ability), since it might vary for different models or tasks. Besides, existing studies often test emergent abilities on very limited model sizes for a specific LLM. For example, PaLM is often tested with three sizes of 8B, 62B and 540B. It is unclear about the model performance of the untested sizes.

8. In a recent study [60], it also shows that in-context learning implicitly performs meta-optimization through the attention mechanism.

both high-quality and low-quality data), they are likely to generate toxic, biased, or even harmful content for humans. It is necessary to align LLMs with human values, *e.g., helpful, honest*, and *harmless*. For this purpose, InstructGPT [61] designs an effective tuning approach that enables LLMs to follow the expected instructions, which utilizes the technique of *reinforcement learning with human feedback* [61, 70]. It incorporates human in the training loop with elaborately designed labeling strategies. ChatGPT is indeed developed on a similar technique to InstructGPT, which shows a strong alignment capacity in producing high-quality, harmless responses, *e.g.*, rejecting to answer insulting questions.

• *Tools manipulation*. In essence, LLMs are trained as text generators over massive plain text corpora, thus performing less well on the tasks that are not best expressed in the form of text (*e.g.*, numerical computation). Besides, their capacities are also limited to the pre-training data, *e.g.*, the inability to capture up-to-date information. To tackle these issues, a recently proposed technique is to employ external tools to compensate for the deficiencies of LLMs [71, 72]. For example, LLMs can utilize the calculator for accurate computation [71] and employ search engines to retrieve unknown information [72]. More recently, ChatGPT has enabled the mechanism of using external plugins (existing or newly created apps)[9], which are by analogy with the "*eyes and ears*" of LLMs. Such a mechanism can broadly expand the scope of capacities for LLMs.

Besides, many other factors (*e.g.*, the upgrade of hardware) also contribute to the success of LLMs. While, we limit our discussion to the major technical approaches and key findings for developing LLMs.

2.2 Technical Evolution of GPT-series Models

Due to the excellent capacity in communicating with humans, ChatGPT has ignited the excitement of the AI community since its release. ChatGPT is developed based on the powerful GPT model with specially optimized conversation capacities. Considering the ever-growing interest in ChatGPT and GPT models, we add a special discussion about the technical evolution of the GPT-series models, to briefly summarize the progress how they have been developed in the past years. Overall, the research of OpenAI on LLMs can be roughly divided into the following stages[10].

Early Explorations. According to one interview with Ilya Sutskever[11] (a co-founder and chief scientist of OpenAI), the idea of approaching intelligent systems with language models was already explored in the early days of OpenAI, while it was attempted with recurrent neural networks (RNN) [104]. With the advent of Transformer, OpenAI developed two initial GPT models, namely GPT-1 [105] and GPT-2 [26], which can considered as the foundation to more powerful models subsequently *i.e.*, GPT-3 and GPT-4.

• *GPT-1*. In 2017, the Transformer model [22] was introduced by Google, and the OpenAI team quickly adapted their language modeling work to this new neural network architecture. They released the first GPT model in 2018, *i.e.*, GPT-1 [105], and coined the abbreviation term *GPT* as the model name, standing for *Generative Pre-Training*. GPT-1 was developed based on a generative, decoder-only Transformer architecture, and adopted a hybrid approach of unsupervised pretraining and supervised fine-tuning. GPT-1 has set up the core architecture for the GPT-series models and established the underlying principle to model natural language text, *i.e.*, predicting the next word.

• *GPT-2*. Following a similar architecture of GPT-1, GPT-2 [26] increased the parameter scale to 1.5B, which was trained with a large webpage dataset WebText. As claimed in the paper of GPT-2, it sought to perform tasks via unsupervised language modeling, without explicit fine-tuning using labeled data. To motivate the approach, they introduced a probabilistic form for multi-task solving, *i.e.*, $p(output|input, task)$ (similar approaches have been adopted in [106]), which predicts the output conditioned on the input and task information. To model this conditional probability, language text can be naturally employed as a unified way to format input, output and task information. In this way, the process of solving a task can be cast as a word prediction problem for generating the solution text. Further, they introduced a more formal claim for this idea: "Since the (task-specific) supervised objective is the same as the unsupervised (language modeling) objective but only evaluated on a subset of the sequence, the global minimum of the unsupervised objective is also the global minimum of the supervised objective (for various tasks)" [26][12]. A basic understanding of this claim is that each (NLP) task can be considered as the word prediction problem based on a subset of the world text. Thus, unsupervised language modeling could be capable in solving various tasks, if it was trained to have sufficient capacity in recovering the world text. These early discussion in GPT-2's paper echoed in the interview of Ilya Sutskever by Jensen Huang: "What the neural network learns is some representation of the process that produced the text. This text is actually a projection of the world...the more accurate you are in predicting the next word, the higher the fidelity, the more resolution you get in this process..."[13].

Capacity Leap. Although GPT-2 is intended to be an "unsupervised multitask learner", it overall has an inferior performance compared with supervised fine-tuning state-of-the-art methods. While, it has a relatively small model size, it has widely fine-tuned in downstream tasks, especially the dialog tasks [107, 108]. Based on GPT-2, GPT-3 demonstrates a key capacity leap by scaling of the (nearly same) generative pre-training architecture.

• *GPT-3*. GPT-3 [55] was released in 2020, which scaled the model parameters to an ever larger size of 175B. In the GPT-3's paper, it formally introduced the concept of

---

9. https://openai.com/blog/chatgpt-plugins
10. Note that the discussion of this part can be somewhat subjective. The overall viewpoints and summaries are made based on the understanding of the authors by surveying the papers, blog articles, interview reports and APIs released by OpenAI.
11. https://hackernoon.com/an-interview-with-ilya-sutskever-co-founder-of-openai

12. To better understand this sentence, we put some explanation words in parentheses.
13. https://lifearchitect.ai/ilya/

Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.,* the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

in-context learning (ICL)[14], which utilizes LLMs in a few-shot or zero-shot way. ICL can teach (or instruct) LLMs to understand the tasks in the form of natural language text. With ICL, the pre-training and utilization of LLMs converge to the same language modeling paradigm: pre-training predicts the following text sequence conditioned on the context, while ICL predicts the correct task solution, which can be also formatted as a text sequence, given the task description and demonstrations. GPT-3 not only demonstrates very excellent performance in a variety of NLP tasks, but also on a number of specially designed tasks that require the abilities of reasoning or domain adaptation. Although the GPT-3's paper does not explicitly discuss the emergent abilities of LLMs, we can observe large performance leap that might transcend the basic scaling law [30], *e.g.,* larger models have significantly stronger ICL ability (illustrated in the original Figure 1.2 of the GPT-3's paper [55]). Overall, GPT-3 can be viewed as a remarkable landmark in the journey evolving from PLMs to LLMs. It has empirically proved that scaling the neural networks to a significant size can lead to a huge increase in model capacity.

Capacity Enhancement. Due to the strong capacities, GPT-3 has been the base model to develop even more capable LLMs for OpenAI. Overall, OpenAI has explored two major approaches to further improving the GPT-3 model, *i.e.,* training on code data and alignment with human preference, which are detailed as follows.

- *Training on code data.* A major limitation of the original

GPT-3 model (pre-trained on plain text) lies in the lack of the reasoning ability on complex tasks, *e.g.,* completing the code and solving math problems. To enhance this ability, Codex [89] was introduced by OpenAI in July 2021, which was a GPT model fine-tuned on a large corpus of GitHub code. It demonstrated that Codex can solve very difficult programming problems, and also lead to a significant performance improvement in solving math problems [109]. Further, a contrastive approach [110] to training text and code embedding was reported in January 2022, which was shown to improve a series of related tasks (*i.e.,* linear-probe classification, text search and code search). Actually, the GPT-3.5 models are developed based on a code-based GPT model (*i.e.,* `code-davinci-002`), which indicates that training on code data is a very useful practice to improve the model capacity of GPT models, especially the reasoning ability. Besides, there is also a speculation that training on code data can greatly increase the chain-of-thought prompting abilities of LLMs [47], while it is still worth further investigation with more thorough verification.

- *Human alignment.* The related research of human alignment can be dated back to the year 2017 (or earlier) for OpenAI: a blog article entitled "learning from human preferences"[15] was posted on the OpenAI blog describing a work that applied reinforcement learning (RL) to learn from the *preference comparisons* annotated by humans [70] (similar to the *reward training* step in the aligning algorithm of InstructGPT in Figure 6). Shortly after the release of this RL paper [70], the paper of the Proximal Policy Optimiza-

14. GPT-2 essentially used ICL for unsupervised task learning, though it wasn't called ICL at that time.

15. https://openai.com/research/learning-from-human-preferences

TABLE 1
Statistics of large language models (having a size larger than 10B in this survey) in recent years, including the capacity evaluation, pre-training data scale (either in the number of tokens or storage size) and hardware resource costs. In this table, we only include LLMs with a public paper about the technical details. Here, "Release Time" indicates the date when the corresponding paper was officially released. "Publicly Available" means that the model checkpoints can be publicly accessible while "Closed Source" means the opposite. "Adaptation" indicates whether the model has been with subsequent fine-tuning: IT denotes instruction tuning and RLHF denotes reinforcement learning with human feedback. "Evaluation" indicates whether the model has been evaluated with corresponding abilities in their original paper: ICL denotes in-context learning and CoT denotes chain-of-thought. "*" denotes the largest publicly available version.

| | Model | Release Time | Size (B) | Base Model | Adaptation | | Pre-train Data Scale | Latest Data Timestamp | Hardware (GPUs / TPUs) | Training Time | Evaluation | |
					IT	RLHF					ICL	CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu-α [75]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [88]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [89]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [90]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [91]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [92]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA-PT	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [93]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [94]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [72]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [95]	Dec-2021	260	-	-	-	-	-	-	-	✓	-
	GLaM [96]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [63]	Jan-2022	137	-	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [64]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [64]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [103]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

tion (PPO) [111] was published in July 2017, which now has been the foundational RL algorithm for learning from human preferences [61]. Later in January 2020, GPT-2 was fine-tuned using the aforementioned RL algorithms [70, 111], which leveraged human preferences to improve the capacities of GPT-2 on NLP tasks. In the same year, another work [112] trained a summarization model for optimizing human preferences in a similar way. Based on these prior work, InstructGPT [61] was proposed in January 2022 to improve the GPT-3 model for human alignment, which formally established a three-stage *reinforcement learning from human feedback (RLHF)* algorithm. Note that it seems that the wording of *"instruction tuning"* has seldom been used in OpenAI's paper and documentation, which is substituted by *supervised fine-tuning on human demonstrations* (*i.e.,* the first step of the RLHF algorithm [61]). In addition to improving the instruction following capacity, the RLHF algorithm is particularly useful to mitigate the issues of generating harm or toxic content for LLMs, which is key to the safe deployment of LLMs in practice. OpenAI describes their approach to alignment research in a technical article [113], which has summarized three promising directions: "training AI

systems to use human feedback, to assist human evaluation and to do alignment research".

These enhancement techniques lead to the improved GPT-3 models with stronger capacities, which are called GPT-3.5 models by OpenAI (see the discussion about the OpenAI API in Section 3.1).

The Milestones of Language Models. Based on all the exploration efforts, two major milestones have been achieved by OpenAI, namely ChatGPT [114] and GPT-4 [46], which have largely raised the capacity bar of existing AI systems.

• *ChatGPT*. In November 2022, OpenAI released the conversation model ChatGPT, based on the GPT models (GPT-3.5 and GPT-4). As the official blog article introduced [114], ChatGPT was trained in a similar way as InstructGPT (called "a sibling model to InstructGPT" in the original post), while specially optimized for dialogue. They reported a difference between the training of ChatGPT and InstructGPT in the data collection setup: human-generated conversations (playing both the roles of user and AI) are combined with the InstructGPT dataset in a dialogue format for training ChatGPT. ChatGPT exhibited superior capacities in communicating with humans: possessing a vast store of knowledge, skill at reasoning on mathematical problems, tracing the context accurately in multi-turn dialogues, and aligning well with human values for safe use. Later on, the plugin mechanism has been supported in ChatGPT, which further extends the capacities of ChatGPT with existing tools or apps. So far, it seems to be the ever most powerful chatbot in the AI history. The launch of ChatGPT has a significant impact on the AI research in the future, which sheds light on the exploration of human-like AI systems.

• *GPT-4*. As another remarkable progress, GPT-4 [46] was released in March 2023, which extended the text input to multimodal signals. Overall, GPT-4 has stronger capacities in solving complex tasks than GPT-3.5, showing a large performance improvement on many evaluation tasks. A recent study [41] investigated the capacities of GPT-4 by conducting qualitative tests with human-generated problems, spanning a diverse range of difficult tasks, and showed that GPT-4 can achieve more superior performance than prior GPT models such as ChatGPT. Furthermore, GPT-4 responds more safely to malicious or provocative queries, due to a six-month iterative alignment (with an additional safety reward signal in the RLHF training). In the technical report, OpenAI has emphasized how to safely develop GPT-4 and applied a number of intervention strategies to mitigate the possible issues of LLMs, such as hallucinations, privacy and overreliance. For example, they introduced the mechanism called *read teaming* [115] to reduce the harm or toxic content generation. As another important aspect, GPT-4 has been developed on a well-established deep learning infrastructure with improved optimization methods. They introduced a new mechanism called *predictable scaling* that can accurately predict the final performance with a small proportion of compute during model training.

Despite the huge progress, there are still limitations with these superior LLMs, *e.g.,* generating hallucinations with factual errors or potentially risky response within some specific context [46]. More limitations or issues of LLMs will be discussed in Section 7. It poses long-standing research challenges to develop more capable, safer LLMs. From the perspective of engineering, OpenAI has adopted an iterative deployment strategy [116] to develop the models and products by following a five-stage development and deployment life-cycle, which aims to effectively reduce the potential risks of using the models. In the following, we will dive into the technical details in order to have a specific understanding of how they have been developed.

3 RESOURCES OF LLMs

It is by no means an easy job to develop or reproduce LLMs, considering the challenging technical issues and huge demands of computation resources. A feasible way is to learn experiences from existing LLMs and reuse publicly available resources for incremental development or experimental study. In this section, we briefly summarize the publicly available resources for developing LLMs, including model checkpoints (or APIs), corpora and libraries.

3.1 Publicly Available Model Checkpoints or APIs

Given the huge cost of model pre-training, well-trained model checkpoints are critical to the study and development of LLMs for the research community. Since the parameter scale is a key factor to consider for using LLMs, we categorize these public models into two scale levels (*i.e., tens of billions of parameters* and *hundreds of billions of parameters*), which is useful for users to identify the suitable resources according to their resource budget. Besides, for inference, we can directly employ public APIs to perform our tasks, without running the model locally. Next, we introduce the publicly available model checkpoints and APIs.

Models with Tens of Billions of Parameters. Most of the models in this category have a parameter scale ranging from 10B to 20B, except LLaMA [57] (containing 65B parameters in the largest version) and NLLB [82] (containing 54.5B parameters in the largest version). Other models within this range include mT5 [74], PanGu-α [75], T0 [28], GPT-NeoX-20B [78], CodeGen [77], UL2 [80], Flan-T5 [64], and mT0 [84]. Among them, Flan-T5 (11B version) can serve as a premier model for research on instruction tuning, since it explores the instruction tuning from three aspects [64]: increasing the number of tasks, scaling the model size, and fine-tuning with chain-of-thought prompting data. Besides, CodeGen (11B version), as an autoregressive language model designed for generating code, can be considered as a good candidate for exploring the code generation ability. It also introduces a new benchmark MTPB [77] specially for multi-turn program synthesis, which is composed by 115 expert-generated problems. To solve these problems, it requires LLMs to acquire sufficient programming knowledge (*e.g.,* math, array operations, and algorithms). As for multilingual tasks, mT0 (13B version) might be a good candidate model, which has been fine-tuned on multilingual tasks with multilingual prompts. Furthermore, PanGu-α [75] shows good performance in Chinese downstream tasks in zero-shot or few-shot settings, which is developed based on the deep learning framework MindSpore [117]. Note that PanGu-α [75] holds multiple versions of models (up to 200B parameters), while the largest public version

has 13B parameters. As a more recent release, LLaMA (65B version) [57], which contains approximately five times as many parameters as other models, has exhibited superior performance in tasks related to instruction following. Due to the openness and effectiveness, LLaMA has attracted significant attention from the research community, and many efforts [118–121] have been devoted to fine-tuning or continually pre-training its different model versions for implementing new models or tools. Typically, pre-training models at this scale require hundreds or even thousands of GPUs or TPUs. For instance, GPT-NeoX-20B uses 12 supermicro servers, each equipped with 8 NVIDIA A100-SXM4-40GB GPUs, while LLaMA utilizes 2,048 A100-80G GPUs as reported in their original publications. To accurately estimate the computation resources needed, it is suggested to use the metrics measuring the number of involved computations such as *FLOPS* (*i.e.,* FLoating point number Operations Per Second) [30].

Models with Hundreds of Billions of Parameters. For models in this category, only a handful of models have been publicly released. For example, OPT [81], OPT-IML [85], BLOOM [69], and BLOOMZ [84] have nearly the same number of parameters as GPT-3 (175B version), while GLM [83] and Galactica [35] have 130B and 120B parameters, respectively. Among them, OPT (175B version) has been specially motivated for open sharing, which aims to enable researchers to carry out reproducible research at scale. For research in cross-lingual generalization, BLOOM (176B version) and BLOOMZ (176B version) can be used as base models, due to the competence in multilingual language modeling tasks. Among these models, OPT-IML have been tuned with instructions, which might be good candidates for studying the effect of instruction tuning. Models of this scale typically require thousands of GPUs or TPUs to train. For instance, OPT (175B version) used 992 A100-80GB GPUs, while GLM (130B version) used a cluster of 96 NVIDIA DGX-A100 (8x40G) GPU nodes.

Public API of LLMs. Instead of directly using the model copies, APIs provide a more convenient way for common users to use LLMs, without the need of running the model locally. As a representative interface for using LLMs, the APIs for the GPT-series models [46, 55, 61, 89] have been widely used for both academia and industry[16]. OpenAI has provided seven major interfaces to the models in GPT-3 series: `ada`, `babbage`, `curie`, `davinci` (the most powerful version in GPT-3 series), `text-ada-001`, `text-babbage-001`, and `text-curie-001`. Among them, the first four interfaces can be further fine-tuned on the host server of OpenAI. In particular, `babbage`, `curie`, and `davinci` correspond to the GPT-3 (1B), GPT-3 (6.7B), and GPT-3 (175B) models, respectively [55]. Besides, there are also two APIs related to Codex [89], called `code-cushman-001` (a powerful and multilingual version of the Codex (12B) [89]) and `code-davinci-002`. Further, GPT-3.5 series include one base model `code-davinci-002` and three enhanced versions, namely `text-davinci-002`, `text-davinci-003`, and `gpt-3.5-turbo-0301`. It is worth noting that

`gpt-3.5-turbo-0301` is the interface to invoke Chat-GPT. More recently, OpenAI has also released the corresponding APIs for GPT-4, including `gpt-4`, `gpt-4-0314`, `gpt-4-32k`, and `gpt-4-32k-0314`. Overall, the choice of API interfaces depends on the specific application scenarios and response requirements. The detailed usage can be found on their project websites[17].

TABLE 2
Statistics of commonly-used data sources.

Corpora	Size	Source	Latest Update Time
BookCorpus [122]	5GB	Books	Dec-2015
Gutenberg [123]	-	Books	Dec-2021
C4 [73]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [124]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [125]	120GB	CommonCrawl	Apr-2019
OpenWebText [126]	38GB	Reddit links	Mar-2023
Pushift.io [127]	2TB	Reddit links	Mar-2023
Wikipedia [128]	21GB	Wikipedia	Mar-2023
BigQuery [129]	-	Codes	Mar-2023
the Pile [130]	800GB	Other	Dec-2020
ROOTS [131]	1.6TB	Other	Jun-2022

3.2 Commonly Used Corpora

In contrast to earlier PLMs, LLMs which consist of a significantly larger number of parameters require a higher volume of training data that covers a broad range of content. For this need, there are increasingly more accessible training datasets that have been released for research. In this section, we will briefly summarize several widely used corpora for training LLMs. Based on their content types, we categorize these corpora into six groups: Books, CommonCrawl, Reddit links, Wikipedia, Code, and others.

Books. BookCorpus [122] is a commonly used dataset in previous small-scale models (*e.g.,* GPT [105] and GPT-2 [26]), consisting of over 11,000 books covering a wide range of topics and genres (*e.g.,* novels and biographies). Another large-scale book corpus is Project Gutenberg [123], consisting of over 70,000 literary books including novels, essays, poetry, drama, history, science, philosophy, and other types of works in the public domain. It is currently one of the largest open-source book collections, which is used in training of MT-NLG [97] and LLaMA [57]. As for Books1 [55] and Books2 [55] used in GPT-3 [55], they are much larger than BookCorpus but have not been publicly released so far.

CommonCrawl. CommonCrawl [132] is one of the largest open-source web crawling databases, containing a petabyte-scale data volume, which has been widely used as training data for existing LLMs. As the whole dataset is very large, existing studies mainly extract subsets of web pages from it within a specific period. However, due to the widespread existence of noisy and low-quality information in web data, it is necessary to perform data preprocessing before usage. Based on CommonCrawl, there are four filtered datasets that are commonly used in existing work: C4 [73], CC-Stories [124], CC-News [27], and RealNews [125]. The Colossal Clean Crawled Corpus (C4) includes five variants[18],

16. https://platform.openai.com/docs/api-reference/introduction

17. https://platform.openai.com/docs/models/overview
18. https://www.tensorflow.org/datasets/catalog/c4

namely en (806G), en.noclean (6T), realnewslike (36G), web-textlike (17G), and multilingual (38T). The *en* version has been utilized for pre-training T5 [73], LaMDA [63], Gopher [59], and UL2 [80]. The multilingual C4, also called mC4, has been used in mT5 [74]. CC-Stories (31G) is composed of a subset of CommonCrawl data, in which the contents are made in a story-like way. While, the original source of CC-Stories is not available now, so a reproduction version, *CC-Stories-R* [133], has been included in Table 2. Moreover, two news corpora extracted from Common-Crawl, *i.e.,* REALNEWS (120G) and CC-News (76G), are also commonly used as the pre-training data.

Reddit Links. Reddit is a social media platform that enables users to submit links and text posts, which can be voted on by others through "upvotes" or "downvotes". Highly up-voted posts are often considered useful, and can be utilized to create high-quality datasets. WebText [26] is a well-known corpus composed of highly upvoted links from Reddit, but it is not publicly available. As a surrogate, there is a readily accessible open-source alternative called OpenWebText [126]. Another corpus extracted from Reddit is PushShift.io [127], a real-time updated dataset that consists of historical data from Reddit since its creation day. Pushshift provides not only monthly data dumps but also useful utility tools to support users in searching, summarizing, and conducting preliminary investigations on the entire dataset. This makes it easy for users to collect and process Reddit data.

Wikipedia. Wikipedia [128] is an online encyclopedia containing a large volume of high-quality articles on diverse topics. Most of these articles are composed in an expository style of writing (with supporting references), covering a wide range of languages and fields. Typically, the English-only filtered versions of Wikipedia are widely used in most LLMs (*e.g.,* GPT-3 [55], LaMDA [63], and LLaMA [57]). Wikipedia is available in multiple languages, so it can be used in multilingual settings.

Code. To collect code data, existing work mainly crawls open-source licensed codes from the Internet. Two major sources are public code repositories under open-source licenses (*e.g.,* GitHub) and code-related question-answering platforms (*e.g.,* StackOverflow). Google has publicly released the BigQuery dataset [129], which includes a substantial number of open-source licensed code snippets in various programming languages, serving as a representative code dataset. CodeGen has utilized BIGQUERY [77], a subset of the BigQuery dataset, for training the multilingual version of CodeGen (CodeGen-Multi).

Others. The Pile [130] is a large-scale, diverse, and open-source text dataset consisting of over 800GB of data from multiple sources, including books, websites, codes, scientific papers, and social media platforms. It is constructed from 22 diverse high-quality subsets. The Pile dataset is widely used in models with different parameter scales, such as GPT-J (6B) [134], CodeGen (16B) [77], and Megatron-Turing NLG (530B) [97]. Besides, ROOTS [131] is composed of various smaller datasets (totally 1.61 TB of text) and covers 59 different languages (containing natural languages and programming languages), which have been used for training BLOOM [69].

In practice, it commonly requires a mixture of different data sources for pre-training LLMs (see Figure 2), instead of a single corpus. Therefore, existing studies commonly mix several ready-made datasets (*e.g.,* C4, OpenWebText, and the Pile), and then perform further processing to obtain the pre-training corpus. Besides, to train the LLMs that are adaptive to specific applications, it is also important to extract data from relevant sources (*e.g.,* Wikipedia and BigQuery) for enriching the corresponding information in pre-training data. To have a quick reference of the data sources used in existing LLMs, we present the pre-training corpora of three representative LLMs:

• *GPT-3* (175B) [55] was trained on a mixed dataset of 300B tokens, including CommonCrawl [132], WebText2 [55], Books1 [55], Books2 [55], and Wikipedia [128].

• *PaLM* (540B) [56] uses a pre-training dataset of 780B tokens, which is sourced from social media conversations, filtered webpages, books, Github, multilingual Wikipedia, and news.

• *LLaMA* [57] extracts training data from various sources, including CommonCrawl, C4 [73], Github, Wikipedia, books, ArXiv, and StackExchange. The training data size for LLaMA (6B) and LLaMA (13B) is 1.0T tokens, while 1.4T tokens are used for LLaMA (32B) and LLaMA (65B).

3.3 Library Resource

In this part, we briefly introduce a series of available libraries for developing LLMs.

• **Transformers** [135] is an open-source Python library for building models using the Transformer architecture, which is developed and maintained by Hugging Face. It has a simple and user-friendly API, making it easy to use and customize various pre-trained models. It is a powerful library with a large and active community of users and developers who regularly update and improve the models and algorithms.

• **DeepSpeed** [65] is a deep learning optimization library (compatible with PyTorch) developed by Microsoft, which has been used to train a number of LLMs, such as MT-NLG [97] and BLOOM [69]. It provides the support of various optimization techniques for distributed training, such as memory optimization (ZeRO technique, gradient checkpointing), and pipeline parallelism.

• **Megatron-LM** [66–68] is a deep learning library developed by NVIDIA for training large-scale language models. It also provides rich optimization techniques for distributed training, including model and data parallelism, mixed-precision training, and FlashAttention. These optimization techniques can largely improve the training efficiency and speed, enabling efficient distributed training across GPUs.

• **JAX** [136] is a Python library for high-performance machine learning algorithms developed by Google, allowing users to easily perform computations on arrays with hardware acceleration (*e.g.,* GPU or TPU). It enables efficient computation on various devices and also supports several featured functions, such as automatic differentiation and just-in-time compilation.

• **Colossal-AI** [137] is a deep learning library developed by HPC-AI Tech for training large-scale AI models. It is

implemented based on PyTorch and supports a rich collection of parallel training strategies. Furthermore, it can also optimize heterogeneous memory management with methods proposed by PatrickStar [138]. Recently, a ChatGPT-like model called ColossalChat [121] has been publicly released with two versions (7B and 13B), which are developed using Colossal-AI based on LLaMA [57].

• **BMTrain** [139] is an efficient library developed by OpenBMB for training models with large-scale parameters in a distributed manner, which emphasizes code simplicity, low resource, and high availability. BMTrain has already incorporated several common LLMs (*e.g.,* Flan-T5 [64] and GLM [83]) into its ModelCenter, where developers can use these models directly.

• **FastMoE** [140] is a specialized training library for MoE (*i.e.,* mixture-of-experts) models. It is developed based on PyTorch, prioritizing both efficiency and user-friendliness in its design. FastMoE simplifies the process of transferring Transformer models to MoE models and supports both data parallelism and model parallelism during training.

Besides the above library resources, existing deep learning frameworks (*e.g.,* PyTorch [141], TensorFlow [142], MXNet [143], PaddlePaddle [144], MindSpore [117] and OneFlow [145]) have also provided the support for parallel algorithms, which are commonly used for training large-scale models.

4 PRE-TRAINING

Pre-training establishes the basis of the abilities of LLMs. By pre-training on large-scale corpora, LLMs can acquire essential language understanding and generation skills [55, 56]. In this process, the scale and quality of the pre-training corpus are critical for LLMs to attain powerful capabilities. Besides, to effectively pre-train LLMs, model architectures, acceleration methods, and optimization techniques need to be well designed. In what follows, we first discuss the data collection and processing in Section 4.1, then introduce the commonly used model architectures in Section 4.2, and finally present the training techniques to stably and efficiently optimize LLMs in Section 4.3.

4.1 Data Collection

Compared with small-scale language models, LLMs have a stronger demand for high-quality data for model pre-training, and their model capacities largely rely on the pre-training corpus and how it has been preprocessed. In this part, we discuss the collection and processing of pre-training data, including data sources, preprocessing methods, and important analysis of how pre-training data affects the performance of LLMs.

4.1.1 Data Source

To develop a capable LLM, it is key to collect a large amount of natural language corpus from various data sources. Existing LLMs mainly leverage a mixture of diverse public textual datasets as the pre-training corpus. Figure 2 shows the distribution of the sources of pre-training data for a number of representative LLMs.

The source of pre-training corpus can be broadly categorized into two types: general data and specialized data.

General data, such as webpages, books, and conversational text, is utilized by most LLMs [55, 56, 81] due to its large, diverse, and accessible nature, which can enhance the language modeling and generalization abilities of LLMs. In light of the impressive generalization capabilities exhibited by LLMs, there are also studies that extend their pre-training corpus to more specialized datasets, such as multilingual data, scientific data, and code, endowing LLMs with specific task-solving capabilities [35, 56, 77]. In what follows, we describe these two types of pre-training data sources and their effects on LLMs. For a detailed introduction to the commonly used corpus, one can refer to Section 3.2.

General Text Data. As we can see in Figure 2, the vast majority of LLMs adopt general-purpose pre-training data, such as webpages, books, and conversational text, which provides rich text sources on a variety of topics. Next, we briefly summarize three important kinds of general data.

• *Webpages.* Owing to the proliferation of the Internet, various types of data have been created, which enables LLMs to gain diverse linguistic knowledge and enhance their generalization capabilities [26, 73]. For convenient use of these data resources, a large amount of data is crawled from the web in previous work, such as CommonCrawl [132]. However, the crawled web data tends to contain both high-quality text, such as Wikipedia and low-quality text, like spam mail, thus it is important to filter and process webpages for improving the data quality.

• *Conversation text.* Conversation data can enhance the conversational competence of LLMs [81] and potentially improve their performance on a range of question-answering tasks [56]. Researchers can utilize subsets of public conversation corpus (*e.g.,* PushShift.io Reddit corpus) [127, 146] or collect conversation data from online social media. Since online conversational data often involves discussions among multiple participants, an effective processing way is to transform a conversation into a tree structure, where the utterance is linked to the one it responds to. In this way, the multi-party conversation tree can be divided into multiple sub-conversations, which can be collected in the pre-training corpus. Furthermore, a potential risk is that the excessive integration of dialogue data into LLMs may result in a side effect [81]: declarative instructions and direct interrogatives are erroneously perceived as the beginning of conversations, thus leading to a decline in the efficacy of the instructions.

• *Books.* Compared to other corpus, books provide an important source of formal long texts, which are potentially beneficial for LLMs to learn linguistic knowledge, model long-term dependency, and generate narrative and coherent texts. To obtain open-source book data, existing studies usually adopt the Books3 and Bookcorpus2 datasets, which are available in the Pile dataset [130].

Specialized Text Data. Specialized datasets are useful to improve the specific capabilities of LLMs on downstream tasks. Next, we introduce three kinds of specialized data.

• *Multilingual text.* Besides the text in the target language, integrating a multilingual corpus can enhance the multilingual abilities of language understanding and generation. For example, BLOOM [69] and PaLM [56] have curated multilingual data covering 46 and 122 languages, respectively, within their pre-training corpora. These models
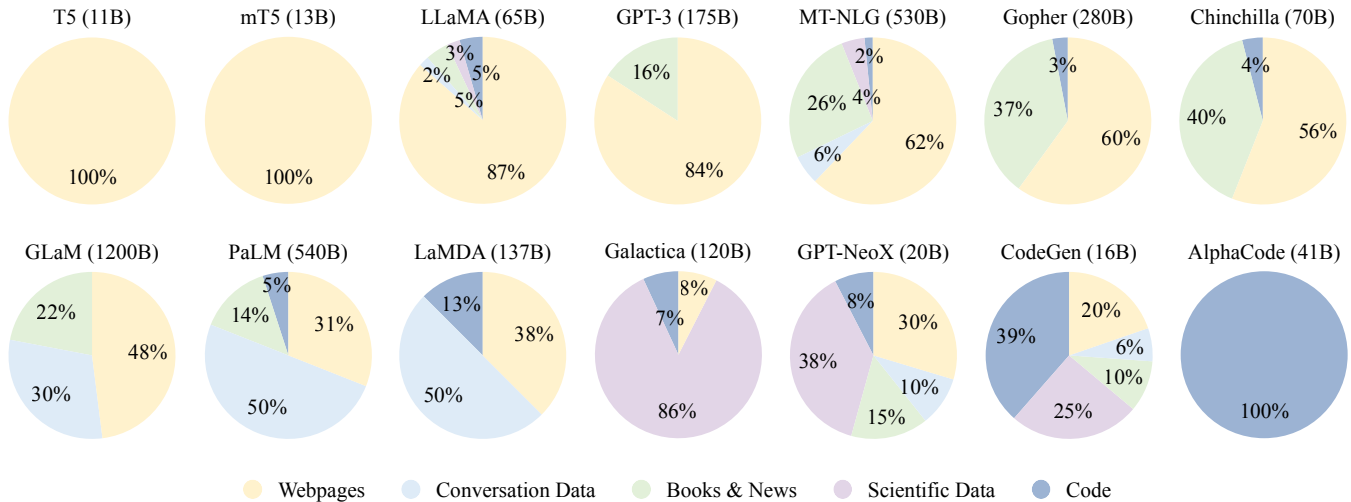
Fig. 2. Ratios of various data sources in the pre-training data for existing LLMs.

demonstrate impressive performance in multilingual tasks, such as translation, multilingual summarization, and multilingual question answering, and achieve comparable or superior performance to the state-of-the-art models that are fine-tuned on the corpus in the target language(s).

• *Scientific text.* The exploration of science by humans has been witnessed by the increasing growth of scientific publications. In order to enhance the understanding of scientific knowledge for LLMs [35, 147], it is useful to incorporate a scientific corpus for model pre-training [35, 147]. By pre-training on a vast amount of scientific text, LLMs can achieve impressive performance in scientific and reasoning tasks [148]. To construct the scientific corpus, existing efforts mainly collect arXiv papers, scientific textbooks, math webpages, and other related scientific resources. Due to the complex nature of data in scientific fields, such as mathematical symbols and protein sequences, specific tokenization and preprocessing techniques are usually required to transform these different formats of data into a unified form that can be processed by language models.

• *Code.* Program synthesis has been widely studied in the research community [89, 149–152], especially the use of PLMs trained on code [134, 153]. However, it remains challenging for these PLMs (*e.g.,* GPT-J [134]) to generate high-quality and accurate programs. Recent studies [89, 152] have found that training LLMs on a vast code corpus can lead to a substantial improvement in the quality of the synthesized programs. The generated programs can successfully pass expert-designed unit-test cases [89] or solve competitive programming questions [98]. In general, two types of code corpora are commonly used for pre-training LLMs. The first source is from programming question answering communities like Stack Exchange [154, 155]. The second source is from public software repositories such as GitHub [77, 89, 152], where code data (including comments and docstrings) are collected for utilization. Compared to natural language text, code is in the format of a programming language, corresponding to long-range dependencies and accurate execution logic [156]. A recent study [47] also speculates that training on code might be a source of complex reasoning

abilities (*e.g.,* chain-of-thought ability [33]). Besides, it has been shown that formatting reasoning tasks into code can help LLMs generate more accurate results [156, 157].

4.1.2 Data Preprocessing

After collecting a large amount of text data, it is essential to preprocess the data for constructing the pre-training corpus, especially removing noisy, redundant, irrelevant, and potentially toxic data [56, 59], which may largely affect the capacity and performance of LLMs. In this part, we review the detailed data preprocessing strategies to improve the quality of the collected data [59, 69, 96]. A typical pipeline of preprocessing the pre-training data for LLMs has been illustrated in Figure 3.

Quality Filtering. To remove low-quality data from the collected corpus, existing work generally adopts two approaches: (1) classifier-based, and (2) heuristic-based. The former approach trains a selection classifier based on high-quality texts and leverages it to identify and filter out low-quality data. Typically, these methods [55, 56, 96] train a binary classifier with well-curated data (*e.g.,* Wikipedia pages) as positive instances and sample candidate data as negative instances, and predict the score that measures the quality of each data example. However, several studies [59, 96] also find that a classifier-based approach may result in the unintentional removal of high-quality texts in dialectal, colloquial, and sociolectal languages, which potentially leads to bias in the pre-training corpus and diminishes the corpus diversity. As the second approach, several studies, such as BLOOM [69] and Gopher [59], employ heuristic-based approaches to eliminate low-quality texts through a set of well-designed rules, which can be summarized as follows:

• *Language based filtering.* If a LLM would be mainly used in the tasks of certain languages, the text in other languages can be filtered.

• *Metric based filtering.* Evaluation metrics about the generated texts, *e.g.,* perplexity, can be employed to detect and remove unnatural sentences.
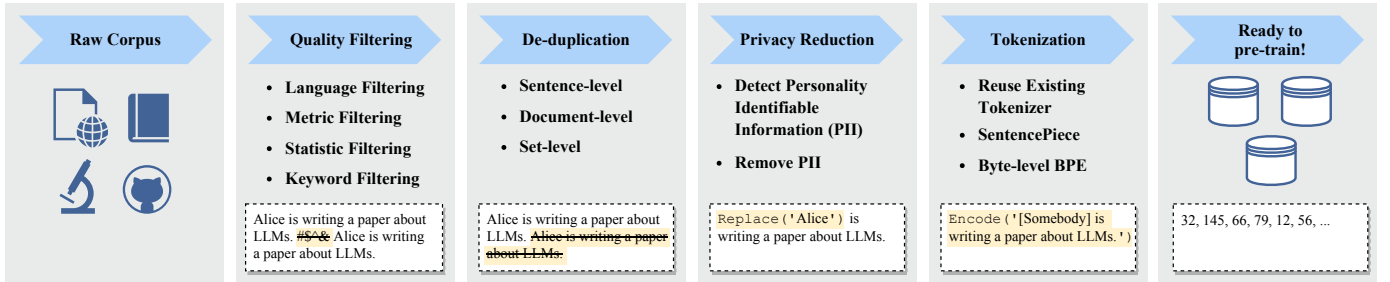
Raw Corpus	Quality Filtering	De-duplication	Privacy Reduction	Tokenization	Ready to pre-train!
	• Language Filtering • Metric Filtering • Statistic Filtering • Keyword Filtering	• Sentence-level • Document-level • Set-level	• Detect Personality Identifiable Information (PII) • Remove PII	• Reuse Existing Tokenizer • SentencePiece • Byte-level BPE	
	Alice is writing a paper about LLMs. #$^& Alice is writing a paper about LLMs.	Alice is writing a paper about LLMs. Alice is writing a paper about LLMs.	Replace('Alice') is writing a paper about LLMs.	Encode('[Somebody] is writing a paper about LLMs.')	32, 145, 66, 79, 12, 56, ...

Fig. 3. An illustration of a typical data preprocessing pipeline for pre-training large language models.

- *Statistic based filtering.* Statistical features of a corpus, *e.g.,* the punctuation distribution, symbol-to-word ratio, and sentence length, can be utilized to measure the text quality and filter the low-quality data.

- *Keyword based filtering.* Based on specific keyword set, the noisy or unuseful elements in the text, such as HTML tags, hyperlinks, boilerplates, and offensive words, can be identified and removed.

De-duplication. Existing work [158] has found that duplicate data in a corpus would reduce the diversity of language models, which may cause the training process to become unstable and thus affect the model performance. Therefore, it is necessary to de-duplicate the pre-training corpus. Specially, de-duplication can be performed at different granularities, including sentence-level, document-level, and dataset-level de-duplication. First, low-quality sentences that contain repeated words and phrases should be removed, as they may introduce repetitive patterns in language modeling [159]. At the document level, existing studies mostly rely on the overlap ratio of surface features (*e.g.,* words and n-grams overlap) between documents to detect and remove duplicate documents containing similar contents [57, 59, 69, 160]. Furthermore, to avoid the dataset contamination problem, it is also crucial to prevent the overlap between the training and evaluation sets [56], by removing the possible duplicate texts from the training set. It has been shown that the three levels of de-duplication are useful to improve the training of LLMs [56, 161], which should be jointly used in practice.

Privacy Redaction. The majority of pre-training text data is obtained from web sources, including user-generated content involving sensitive or personal information, which may increase the risk of privacy breaches [162]. Thus, it is necessary to remove the *personally identifiable information (PII)* from the pre-training corpus. One direct and effective approach is to employ rule-based methods, such as keyword spotting, to detect and remove PII such as names, addresses, and phone numbers [131]. Furthermore, researchers also find that the vulnerability of LLMs under privacy attacks can be attributed to the presence of duplicate PII data in the pre-training corpus [163]. Therefore, de-duplication can also reduce privacy risks to some extent.

Tokenization. Tokenization is also a crucial step for data preprocessing. It aims to segment raw text into sequences of individual tokens, which are subsequently used as the inputs of LLMs. Although it is expedient to leverage an existing tokenizer (*e.g.,* OPT [81] and GPT-3 [55] utilize the tokenizer of GPT-2 [26]), using a tokenizer specially designed for the pre-training corpus can be highly beneficial [69], especially for the corpus that consists of diverse domains, languages, and formats. Therefore, several recent LLMs train the customized tokenizers specially for the pre-training corpus with SentencePiece [164]. The byte-level *Byte Pair Encoding (BPE)* algorithm [165] is utilized to ensure that the information after tokenization is lossless [56, 59]. While, normalization techniques in BPE, such as NFKC [166], may degrade the tokenization performance [34, 59, 69].

4.1.3 Effect of Pre-training Data on LLMs

Unlike small-scale PLMs, it is usually infeasible to iterate the pre-training of LLMs multiple times, due to the huge demand for computational resources. Thus, it is particularly important to construct a well-prepared pre-training corpus before training a LLM. In this part, we discuss how the quality and distribution of the pre-training corpus potentially influence the performance of LLMs.

Mixture of Sources. As discussed before, pre-training data from different domains or scenarios has distinct linguistic characteristics or semantic knowledge. By pre-training on a mixture of text data from diverse sources, LLMs can acquire a broad scope of knowledge and may exhibit a strong generalization capacity. When mixing different sources, one needs to carefully set the distribution of pre-training data, since it is also likely to affect the performance of LLMs on downstream tasks [59]. Gopher [59] conducts the ablation experiment on data distribution to examine the impact of mixed sources on downstream tasks. Experimental results on the LAMBADA dataset [167] show that increasing the proportion of books data can improve the capacity of the model in capturing long-term dependencies from text, and increasing the proportion of the C4 dataset [73] leads to performance improvement on the C4 validation dataset [59]. While, as a side effect, training on excessive data about a certain domain would affect the generalization capability of LLMs on other domains [35, 59]. Therefore, it is suggested that researchers should carefully determine the proportion of data from different domains in the pre-training corpus, in order to develop LLMs that better meet their specific needs. The readers can refer to Figure 2 for a comparison of the data sources for different LLMs.

Amount of Pre-training Data. For pre-training an effective LLM, it is important to collect sufficient high-quality data

that satisfies the data quantity demand of the LLM. Existing studies have found that with the increasing parameter scale in the LLM, more data is also required to train the model [34, 57]: a similar scaling law as model size is also observed in data size, with respect to model performance. A recent study has shown that a number of existing LLMs suffer from sub-optimal training due to inadequate pretraining data [34]. By conducting extensive experiments, it further demonstrates increasing the model size and data size in equal scales can lead to a more compute-efficient model (*i.e.,* the Chinchilla model), for a given compute budget. More recently, LLaMA [57] shows that with more data and longer training, smaller models can also achieve good performance. Overall, it is suggested that researchers should pay more attention to the amount of high-quality data for adequately training the model, especially when scaling the model parameters.

Quality of Pre-training Data. Existing work has shown that pre-training on the low-quality corpus, such as noisy, toxic, and duplicate data, may hurt the performance of models [59, 158, 160, 163]. For developing a well-performing LLM, it is crucial to consider both the quantity ant the quality of the collected training data. Recent studies, such as T5 [73], GLaM [96], and Gopher [59], have investigated the influence of data quality on the performance of downstream tasks. By comparing the performance of models trained on the filtered and unfiltered corpus, they reach the same conclusion that pre-training LLMs on cleaned data can improve the performance. More specifically, the duplication of data may result in "*double descent*" (referring to the phenomenon of performance initially deteriorating and subsequently improving) [158, 168], or even overwhelm the training process [158]. Besides, it has been shown that duplicate data degrades the ability of LLMs to copy from the context, which might further affect the generalization capacity of LLMs using in-context learning [158]. Therefore, as suggested in [56, 59, 69], it is essential to incorporate preprocessing methods on the pre-training corpus carefully (as illustrated in Section 4.1.2), to improve stability of the training process and avoid affecting the model performance.

4.2 Architecture

In this section, we review the architecture design of LLMs, *i.e.,* mainstream architecture, pre-training objective, and detailed configuration. Table 3 presents the model cards of several representative LLMs with public details.

4.2.1 Mainstream Architectures

Due to the excellent parallelizability and capacity, the Transformer architecture [22] has become the de facto backbone to develop various LLMs, making it possible to scale language models to hundreds or thousands of billions of parameters. In general, the mainstream architectures of existing LLMs can be roughly categorized into three major types, namely encoder-decoder, causal decoder, and prefix decoder, as shown in Figure 4.

Encoder-decoder Architecture. The vanilla Transformer model is built on the encoder-decoder architecture [22], which consists of two stacks of Transformer blocks as the encoder and decoder, respectively. The encoder adopts stacked multi-head self-attention layers to encode the input sequence for generating its latent representations, while the decoder performs cross-attention on these representations and autoregressively generates the target sequence. Encoder-decoder PLMs (*e.g.,* T5 [73] and BART [24]) have shown effectiveness on a variety of NLP tasks. So far, there are only a small number of LLMs that are built based on the encoder-decoder architecture, *e.g.,* Flan-T5 [64]. We leave a detailed discussion about the architecture selection in Section 4.2.4.

Causal Decoder Architecture. The causal decoder architecture incorporates the unidirectional attention mask, to guarantee that each input token can only attend to the past tokens and itself. The input and output tokens are processed in the same fashion through the decoder. As representative language models of this architecture, the GPT-series models [26, 55, 105] are developed based on the causal-decoder architecture. In particular, GPT-3 [55] has successfully demonstrated the effectiveness of this architecture, also showing an amazing in-context learning capability of LLMs. Interestingly, GPT-1 [105] and GPT-2 [26] do not exhibit such superior abilities as those in GPT-3, and it seems that scaling plays an important role in increasing the model capacity of this model architecture. So far, the causal decoders have been widely adopted as the architecture of LLMs by various existing LLMs, such as OPT [81], BLOOM [69], and Gopher [59]. Note that both the causal decoder and prefix decoder discussed next belong to decoder-only architectures. While, when mentioning "decoder-only architecture", it mainly refers to the causal decoder architecture in existing literature, unless specified.

Prefix Decoder Architecture. The prefix decoder architecture (*a.k.a.,* non-causal decoder [169]) revises the masking mechanism of causal decoders, to enable performing bidirectional attention over the prefix tokens [170] and unidirectional attention only on generated tokens. In this way, like the encoder-decoder architecture, the prefix decoders can bidirectionally encode the prefix sequence and autoregressively predict the output tokens one by one, where the same parameters are shared during encoding and decoding. Instead of pre-training from scratch, a practical suggestion is to continually train causal decoders and then convert them into prefix decoders for accelerating convergence [29], *e.g.,* U-PaLM [102] is derived from PaLM [56]. Existing representative LLMs based on prefix decoders include GLM-130B [83] and U-PaLM [102].

For the three types of architectures, we can also consider extending them via the mixture-of-experts (MoE) scaling, in which a subset of neural network weights for each input are sparsely activated, *e.g.,* Switch Transformer [25] and GLaM [96]. It has been shown that substantial performance improvement can be observed by increasing either the number of experts or the total parameter size [171].

4.2.2 Detailed Configuration

Since the launch of Transformer [22], various improvements have been proposed to enhance its training stability, performance, and computational efficiency. In this part, we will discuss the corresponding configurations for four major

TABLE 3
Model cards of several selected LLMs with public configuration details. Here, PE denotes position embedding, #L denotes the number of layers, #H denotes the number of attention heads, d_{model} denotes the size of hidden states, and MCL denotes the maximum context length during training.

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Causal decoder	175B	Pre Layer Norm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [75]	Causal decoder	207B	Pre Layer Norm	Learned	GeLU	✓	64	128	16384	1024
OPT [81]	Causal decoder	175B	Pre Layer Norm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Causal decoder	540B	Pre Layer Norm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [69]	Causal decoder	176B	Pre Layer Norm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [97]	Causal decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Causal decoder	280B	Pre RMS Norm	Relative	-	-	80	128	16384	2048
Chinchilla [34]	Causal decoder	70B	Pre RMS Norm	Relative	-	-	80	64	8192	-
Galactica [35]	Causal decoder	120B	Pre Layer Norm	Learned	GeLU	×	96	80	10240	2048
LaMDA [63]	Causal decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [91]	Causal decoder	178B	Pre Layer Norm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Causal decoder	65B	Pre RMS Norm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [83]	Prefix decoder	130B	Post Deep Norm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [73]	Encoder-decoder	11B	Pre RMS Norm	Relative	ReLU	×	24	128	1024	512
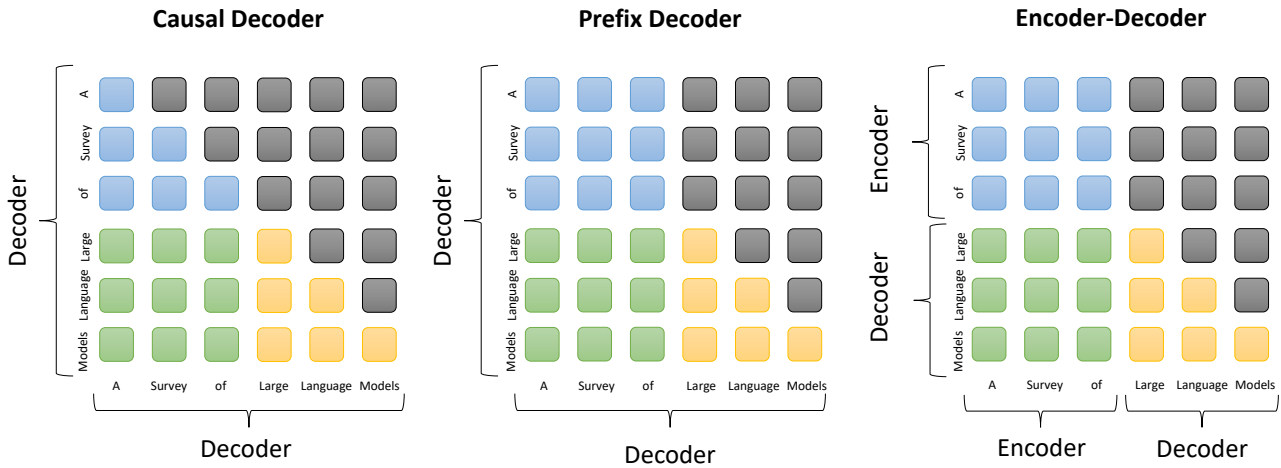
Fig. 4. A comparison of the attention patterns in three mainstream architectures. Here, the blue, green, yellow and grey rounded rectangles indicate the attention between prefix tokens, attention between prefix and target tokens, attention between target tokens, and masked attention respectively.

parts of the Transformer, including normalization, position embeddings, activation functions, and attention and bias. To make this survey more self-contained, we present the detailed formulations for these configurations in Table 4.

Normalization. Training instability is a challenging issue for pre-training LLMs. To alleviate this problem, layer normalization (Layer Norm, LN) [173] is widely employed in Transformer architectures. The position of LN is vital to the performance of LLMs. While the initial Transformer [22] uses post-LN, most LLMs employ pre-LN for more stable training in spite of decreasing performance [182]. Based on pre-LN, Sandwich-LN [172] adds extra LN before the residual connections to avoid value explosion. However, it has been found that Sandwich-LN sometimes fails to stabilize the training of LLMs and may lead to the collapse of training [83]. Recently, several advanced normalization techniques have been proposed as alternatives to LN. In Gopher [59] and Chinchilla [34], RMS Norm [174] is employed due to its superiority in training speed and performance [183]. Compared with LN, DeepNorm [175] has shown a better capability to ensure the stability in training, which has been adopted by GLM-130B with post normaliza-

tion. In addition, adding an extra LN after the embedding layer can also stabilize the training of LLMs. However, it tends to incur a significant performance drop [184], which has been removed in several recent LLMs [69].

Activation Functions. To obtain good performance, activation functions also need to be properly set in feed-forward networks. In existing LLMs, GeLU activations [185] are widely used. Besides, in the latest LLMs (*e.g.,* PaLM and LaMDA), variants of GLU activation [179, 186] have also been utilized, especially the SwiGLU and GeGLU variants, which often achieve better performance in practice [183]. However, compared with GeLU, they require extra parameters (about 50%) in the feed-forward networks [184].

Position Embeddings. Since the self-attention modules in Transformer are permutation equivariant, position embeddings are employed to inject absolute or relative position information for modeling sequences. There are two variants of absolute position embeddings in the vanilla Transformer [22], *i.e.,* sinusoids and learned position embeddings, where the latter is commonly employed in LLMs. Unlike absolute position embeddings, relative positional encodings

TABLE 4
Detailed formulations for the network configurations. Here, Sublayer denotes a FFN or a self-attention module in a Transformer layer, d denotes the size of hidden states, $\mathbf{p}_i$ denotes position embedding at position i, A_{ij} denotes the attention score between a query and a key, r_{i-j} denotes a learnable scalar based on the offset between the query and the key, and $\mathbf{R}_{\theta,t}$ denotes a rotary matrix with rotation degree $t \cdot \theta$.

Configuration	Method	Equation
Normalization position	Post Norm [22]	$\text{Norm}(\mathbf{x}+\text{Sulayerb}(\mathbf{x}))$
	Pre Norm [26]	$\mathbf{x} + \text{Sublayer}(\text{Norm}(\mathbf{x}))$
	Sandwich Norm [172]	$\mathbf{x} + \text{Norm}(\text{Sublayer}(\text{Norm}(\mathbf{x})))$
Normalization method	LayerNorm [173]	$\frac{\mathbf{x}-\mu}{\sqrt{\sigma}} \cdot \gamma + \beta, \quad \mu = \frac{1}{d}\sum_{i=1}^{d} x_i, \quad \sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu))^2}$
	RMSNorm [174]	$\frac{\mathbf{x}}{\text{RMS}(\mathbf{x})} \cdot \gamma, \quad \text{RMS}(\mathbf{x}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}$
	DeepNorm [175]	$\text{LayerNorm}(\alpha \cdot \mathbf{x} + \text{Sublayer}(\mathbf{x}))$
Activation function	ReLU [176]	$\text{ReLU}(\mathbf{x}) = \max(\mathbf{x}, \mathbf{0})$
	GeLU [177]	$\text{GeLU}(\mathbf{x}) = 0.5\mathbf{x} \otimes [1 + \text{erf}(\mathbf{x}/\sqrt{2})], \quad \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt$
	Swish [178]	$\text{Swish}(\mathbf{x}) = \mathbf{x} \otimes \text{sigmoid}(\mathbf{x})$
	SwiGLU [179]	$\text{SwiGLU}(\mathbf{x_1}, \mathbf{x_2}) = \text{Swish}(\mathbf{x_1}) \otimes \mathbf{x_2}$
	GeGLU [179]	$\text{GeGLU}(\mathbf{x_1}, \mathbf{x_2}) = \text{GeLU}(\mathbf{x_1}) \otimes \mathbf{x_2}$
Position embedding	Absolute [22]	$\mathbf{x}_i = \mathbf{x}_i + \mathbf{p}_i$
	Relative [73]	$A_{ij} = \mathbf{W}_q\mathbf{x}_i\mathbf{x}_j^T\mathbf{W}_k^T + r_{i-j}$
	RoPE [180]	$A_{ij} = \mathbf{W}_q\mathbf{x}_i\mathbf{R}_{\theta,i-j}\mathbf{x}_j^T\mathbf{W}_k^T$
	Alibi [181]	$A_{ij} = \mathbf{W}_q\mathbf{x}_i\mathbf{R}_{\theta,i-j}\mathbf{x}_j^T\mathbf{W}_k^T A_{ij} = \mathbf{W}_q\mathbf{x}_i\mathbf{x}_j^T\mathbf{W}_k^T - m(i-j)$

generate embeddings according to the offsets between keys and queries [73], so it can perform well on sequences longer than those it has seen during training, *i.e.*, extrapolation [181]. ALiBi [181] biases attention scores using a penalty based on the distance between keys and queries. Empirical results have shown that it has better zero-shot generalization with a stronger extrapolation capacity than other position embeddings [29]. Besides, by setting specific rotatory matrices based on the absolute position, the scores between keys and queries in RoPE [180] can be computed with relative position information, which is useful to model long sequences. As a result, RoPE has been widely adopted in several latest LLMs [56, 57, 83]

Attention and Bias. Beyond the full self-attention in the original Transformer [22], sparse attention with lower computation complexity is employed in GPT-3 (*i.e.*, Factorized Attention [55, 187]). In order to effectively and efficiently model longer sequences, more attempts have been made by either introducing special attention patterns [188, 189] or considering GPU memory access (*i.e.*, FlashAttention [190]). Besides, following the original Transformer, most LLMs keep the biases in each dense kernel and Layer Norm. However, in PaLM [56] and Galactica [35], biases are removed. It demonstrates that no biases can enhance training stability for LLMs [56].

To put all these discussions together, we summarize the suggestions from existing literature for detailed configuration. For stronger generalization and training stability, it is suggested to choose the pre RMS Norm for layer normalization, and SwiGLU or GeGLU as the activation function. While, LN may not be used immediately after embedding layers, which is likely to incur performance degradation. Besides, as for position embeddings, RoPE or ALiBi is a better choice since it performs better on long sequences.

4.2.3 Pre-training Tasks

Pre-training plays a key role that encodes general knowledge from large-scale corpus into the massive model parameters. For training LLMs, there are two commonly used pre-training tasks, namely language modeling and denoising autoencoding.

Language Modeling. The language modeling task (LM) is the most commonly used objective to pre-train decoder-only LLMs, *e.g.*, GPT3 [55] and PaLM [56]. Given a sequence of tokens $\mathbf{x} = \{x_1, \ldots, x_n\}$, the LM task aims to autoregressively predict the target tokens x_i based on the preceding tokens $x_{<i}$ in a sequence. A general training objective is to maximize the following likelihood:

$$\mathcal{L}_{LM}(\mathbf{x}) = \sum_{i=1}^{n} \log P(x_i|x_{<i}). \tag{4}$$

Since most language tasks can be cast as the prediction problem based on the input, these decoder-only LLMs might be potentially advantageous to implicitly learn how to accomplish these tasks in a unified LM way. Some studies have also revealed that decoder-only LLMs can be naturally transferred to certain tasks by autoregressively predicting the next tokens [26, 55], without fine-tuning. An important variant of LM is the *prefix language modeling* task, which is designed for pre-training models with the prefix decoder architecture. The tokens within a randomly selected prefix would not be used in computing the loss of prefix language modeling. With the same amount of tokens seen during pre-training, prefix language modeling performs slightly worse than language modeling, since fewer tokens in the sequence are involved for model pre-training [29].

Denoising Autoencoding. Besides conventional LM, the denoising autoencoding task (DAE) has also been widely used to pre-train language models [24, 73]. The inputs $\mathbf{x}_{\backslash \tilde{\mathbf{x}}}$ for DAE task are corrupted text with randomly replaced spans. Then, the language models are trained to recover the

replaced tokens $\tilde{\mathbf{x}}$. Formally, the training objective of DAE is denoted as follows:

$$\mathcal{L}_{DAE}(\mathbf{x}) = \log P(\tilde{\mathbf{x}}|\mathbf{x}_{\backslash \tilde{\mathbf{x}}}). \tag{5}$$

However, the DAE task seems to be more complicated in implementation than LM task. As a result, it has not been widely used to pre-train large language models. Existing LLMs that take DAE as pre-training objectives include T5 [73] and GLM-130B [83]. These models are mainly trained to recover the replaced spans in an autoregressive way.

4.2.4   Summary and Discussion

The choice of architecture and pre-training tasks may incur different inductive biases for LLMs, which would lead to different model capacities. In this part, we summarize some important findings or discussions in the existing literature on this issue.

• By pre-training with the LM objective, it seems that causal decoder architecture can achieve a more superior zero-shot and few-shot generalization capacity. Existing research has shown that without multi-task fine-tuning, the causal decoder has better zero-shot performance than other architectures [29]. The success of GPT-3 [55] has demonstrated that the large causal decoder model can be a good few-shot learner. In addition, instruction tuning and alignment tuning discussed in Section 5 have been proven to further enhance the capability of large causal decoder models [61, 62, 64].

• Scaling law has been widely observed in causal decoders. By scaling the model size, the dataset size, and the total computation, the performance of causal decoders can be substantially improved [30, 55]. Thus, it has become an important strategy to increase the model capacity of the causal decoder via scaling. However, more detailed investigation on encoder-decoder models is still lacking, and more efforts are needed to investigate the performance of encoder-decoder models at a large scale.

More research efforts about the discussions on architectures and pre-training objectives are in need to analyze how the choices of the architecture and pre-training tasks affect the capacity of LLMs, especially for encoder-decoder architectures. Besides the major architecture, the detailed configuration of LLM is also worth attention, which has been discussed in Section 4.2.2.

4.3   Model Training

In this part, we review the important settings, techniques, or tricks for training LLMs.

4.3.1   Optimization Setting

For parameter optimization of LLMs, we present the commonly used settings for batch training, learning rate, optimizer, and training stability.

Batch Training. For language model pre-training, existing work generally sets the batch size to a large number (*e.g.,* 8,196 examples or 16M tokens) to improve the training stability and throughput. For LLMs such as GPT-3 and PaLM, they have introduced a new strategy that dynamically increases the batch size during training, ultimately reaching a million scale. Specifically, the batch size of GPT-3 is gradually increasing from 32K to 3.2M tokens. Empirical results have demonstrated that the dynamic schedule of batch size can effectively stabilize the training process of LLMs [56].

Learning Rate. Existing LLMs usually adopt a similar learning rate schedule with the warm-up and decay strategies during pre-training. Specifically, in the initial 0.1% to 0.5% of the training steps, a linear warm-up schedule is employed for gradually increasing the learning rate to the maximum value that ranges from approximately 5×10^{-5} to 1×10^{-4} (*e.g.,* 6×10^{-5} for GPT-3). Then, a cosine decay strategy is adopted in the subsequent steps, gradually reducing the learning rate to approximately 10% of its maximum value, until the convergence of the training loss.

Optimizer. The Adam optimizer [191] and AdamW optimizer [192] are widely utilized for training LLMs (*e.g.,* GPT-3), which are based on adaptive estimates of lower-order moments for first-order gradient-based optimization. Commonly, its hyper-parameters are set as follows: $\beta_1 = 0.9$, $\beta_2 = 0.95$ and $\epsilon = 10^{-8}$. Meanwhile, the Adafactor optimizer [193] has also been utilized in training LLMs (*e.g.,* PaLM and T5), which is a variant of the Adam optimizer specially designed for conserving GPU memory during training. The hyper-parameters of the Adafactor optimizer are set as: $\beta_1 = 0.9$ and $\beta_2 = 1.0 - k^{-0.8}$, where k denotes the number of training steps.

Stabilizing the Training. During the pre-training of LLMs, it often suffers from the training instability issue, which may cause the model collapse. To address this issue, weight decay and gradient clipping have been widely utilized, where existing studies [55, 69, 81, 83, 97] commonly set the threshold of gradient clipping to 1.0 and weight decay rate to 0.1. However, with the scaling of LLMs, the training loss spike is also more likely to occur, leading to unstable training. To mitigate this problem, PaLM [56] and OPT [81] use a simple strategy that restarts the training process from an earlier checkpoint before the occurrence of the spike and skips over the data that may have caused the problem. Further, GLM [83] finds that the abnormal gradients of the embedding layer usually lead to spikes, and proposes to shrink the embedding layer gradients to alleviate it.

4.3.2   Scalable Training Techniques

As the model and data sizes increase, it has become challenging to efficiently train LLMs under a limited computational resource. Especially, two primary technical issues are required to be resolved, *i.e.,* increasing training throughput and loading larger models into GPU memory. In this part, we review several widely used approaches in existing work to address the above two challenges, namely 3D parallelism [66, 194, 195], ZeRO [196], and mixed precision training [197], and also give general suggestions about how to utilize them for training.

3D Parallelism. 3D parallelism is actually a combination of three commonly used parallel training techniques, namely data parallelism, pipeline parallelism [194, 195], and tensor

TABLE 5
Detailed optimization settings of several existing LLMs.

Model	Batch Size (#tokens)	Learning Rate	Warmup	Decay Method	Optimizer	Precision Type	Weight Decay	Grad Clip	Dropout
GPT3 (175B)	32K→3.2M	6×10^{-5}	yes	cosine decay to 10%	Adam	FP16	0.1	1.0	-
PanGu-α (200B)	-	2×10^{-5}	-	-	Adam	-	0.1	-	-
OPT (175B)	2M	1.2×10^{-4}	yes	manual decay	AdamW	FP16	0.1	-	0.1
PaLM (540B)	1M→4M	1×10^{-2}	no	inverse square root	Adafactor	BF16	lr^2	1.0	0.1
BLOOM (176B)	4M	6×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	0.0
MT-NLG (530B)	64 K→3.75M	5×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	-
Gopher (280B)	3M→6M	4×10^{-5}	yes	cosine decay to 10%	Adam	BF16	-	1.0	-
Chinchilla (70B)	1.5M→3M	1×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	-	-	-
Galactica (120B)	2M	7×10^{-6}	yes	linear decay to 10%	AdamW	-	0.1	1.0	0.1
LaMDA (137B)	256K	-	-	-	-	BF16	-	-	-
Jurassic-1 (178B)	32 K→3.2M	6×10^{-5}	yes	-	-	-	-	-	-
LLaMA (65B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
GLM (130B)	0.4M→8.25M	8×10^{-5}	yes	cosine decay to 10%	AdamW	FP16	0.1	1.0	0.1
T5 (11B)	64K	1×10^{-2}	no	inverse square root	AdaFactor	-	-	-	0.1
ERNIE 3.0 Titan (260B)	-	1×10^{-4}	-	-	Adam	FP16	0.1	1.0	-
PanGu-Σ (1.085T)	0.5M	2×10^{-5}	yes	-	Adam	FP16	-	-	-

parallelism [66][19]. We next introduce the three parallel training techniques.

• *Data parallelism.* Data parallelism is one of the most fundamental approaches to improving the training throughput. It replicates the model parameters and optimizer states across multiple GPUs and then distributes the whole training corpus into these GPUs. In this way, each GPU only needs to process the assigned data for it, and performs the forward and backward propagation to obtain the gradients. The computed gradients on different GPUs will be further aggregated to obtain the gradients of the entire batch for updating the models in all GPUs. In this way, as the calculations of gradients are independently performed on different GPUs, the data parallelism mechanism is highly scalable, enabling the way that increases the number of GPUs to improve training throughput. Furthermore, this technique is simple in implementation, and most of existing popular deep learning libraries have already implemented data parallelism, such as TensorFlow and PyTorch.

• *Pipeline parallelism.* Pipeline parallelism aims to distribute the different layers of a LLM into multiple GPUs. Especially, in the case of a Transformer model, pipeline parallelism loads consecutive layers onto the same GPU, to reduce the cost of transmitting the computed hidden states or gradients between GPUs. However, a naive implementation of pipeline parallelism may result in a lower GPU utilization rate as each GPU has to wait for the previous one to complete the computation, leading to the unnecessary cost of *bubbles overhead* [194]. To reduce these bubbles in pipeline parallelism, GPipe [194] and PipeDream [195] propose the techniques of padding multiple batches of data and asynchronous gradient update to improve the pipeline efficiency.

• *Tensor parallelism.* Tensor parallelism is also a commonly used technique that aims to decompose the LLM for multi-GPU loading. Unlike pipeline parallelism, tensor parallelism focuses on decomposing the tensors (the parameter matrices) of LLMs. For a matrix multiplication operation $Y = XA$ in the LLM, the parameter matrix A can be split

into two submatrices, A_1 and A_2, by column, which can be expressed as $Y = [XA_1, XA_2]$. By placing matrices A_1 and A_2 on different GPUs, the matrix multiplication operation would be invoked at two GPUs in parallel, and the final result can be obtained by combining the outputs from the two GPUs through across-GPU communication. Currently, tensor parallelism has been supported in several open-source libraries, *e.g.,* Megatron-LM [66], and can be extended to higher-dimensional tensors. Besides, Colossal-AI has also implemented tensor parallelism for higher-dimensional tensors [198–200] and proposed sequence parallelism [201] especially for sequence data, which can further decompose the attention operation of the Transformer model.

ZeRO. ZeRO [196] technique, proposed by the DeepSpeed [65] library, focuses on the issue of memory redundancy in data parallelism. As mentioned before, data parallelism requires each GPU to store the same copy of a LLM, including model parameters, model gradients, and optimizer parameters. Whereas, not all of the above data is necessary to be retained on each GPU, which would cause a memory redundancy problem. To resolve it, the ZeRO technique aims to retain only a fraction of data on each GPU, while the rest data can be retrieved from other GPUs when required. Specifically, ZeRO provides three solutions, depending on how the three parts of the data are stored, namely optimizer state partitioning, gradient partitioning, and parameter partitioning. Empirical results indicate that the first two solutions do not increase the communication overhead, and the third solution increases about 50% communication overhead but saves memory proportional to the number of GPUs. PyTorch has implemented a similar technique as ZeRO, called FSDP [202].

Mixed Precision Training. In previous PLMs (*e.g.,* BERT [23]), 32-bit floating-point numbers, also known as FP32, have been predominantly used for pre-training. In recent years, to pre-train extremely large language models, some studies [197] have started to utilize 16-bit floating-point numbers (FP16), which reduces memory usage and communication overhead. Additionally, as popular NVIDIA GPUs (*e.g.,* A100) have twice the amount of FP16 computa-

19. Model parallelism is a more broader term that includes tensor parallelism and pipeline parallelism in some work [66].

tion units as FP32, the computational efficiency of FP16 can be further improved. However, existing work has found that FP16 may lead to the loss of computational accuracy [59, 69], which affects the final model performance. To alleviate it, an alternative called *Brain Floating Point (BF16)* has been used for training, which allocates more exponent bits and fewer significant bits than FP16. For pre-training, BF16 generally performs better than FP16 on representation accuracy [69].

Overall Training Suggestion. In practice, the above training techniques, especially 3D parallelism, are often jointly used to improve the training throughput and large model loading. For instance, researchers have incorporated 8-way data parallelism, 4-way tensor parallelism, and 12-way pipeline parallelism, enabling the training of BLOOM [69] on 384 A100 GPUs. Currently, open-source libraries like DeepSpeed [65], Colossal-AI [137], and Alpa [203] can well support the three parallel training methods. To reduce the memory redundancy, ZeRO, FSDP, and activation recomputation techniques [68, 204] can be also employed for training LLMs, which have already been integrated into DeepSpeed, PyTorch, and Megatron-LM. Besides, the mixed precision training technique such as BF16 can be also leveraged to improve the training efficiency and reduce GPU memory usage, while it requires necessary support on hardware (*e.g.,* A100 GPU). Because training large models is a time-intensive process, it would be useful to forecast the model performance and detect abnormal issues at an early stage. For this purpose, GPT-4 [46] has recently introduced a new mechanism called *predictable scaling* built on a deep learning stack, enabling the performance prediction of large models with a much smaller model, which might be quite useful for developing LLMs. In practice, one can further leverage the supporting training techniques of mainstream deep learning frameworks. For instance, PyTorch supports the data parallel training algorithm FSDP [202] (*i.e.,* fully sharded data parallel), which allows for partial offloading of training computations to CPUs if desired.

Besides the above training strategies, it is also important to improve the inference speed for using LLMs. Typically, quantization techniques are widely used to reduce both the time and space costs of LLMs during the inference stage [205]. With some loss in model performance, quantized language models have smaller model sizes and can achieve faster inference speed [83, 206, 207]. For model quantization, a popular choice is INT8-quantization [206]. Further, some research work attempts to develop more aggressive INT4-quantization methods [83]. Recently, quantized model copies of several publicly available language models have been released on Hugging Face, including BLOOM[20], GPT-J[21], and ChatGLM[22].

5 ADAPTATION TUNING OF LLMs

After pre-training, LLMs can acquire the general abilities for solving various tasks. However, increasing studies have shown that LLM's abilities can be further adapted according

to specific goals. In this section, we introduce two major approaches to adapting pre-trained LLMs, namely instruction tuning and alignment tuning. The former approach mainly aims to enhance (or unlock) the abilities of LLMs, while the latter approach aims to align the behaviors of LLMs with human values or preferences. Further, we will also discuss efficient tuning for rapid model adaptation. In what follows, we will introduce the three parts in detail.

TABLE 6
A detailed list of available task collections for instruction tuning. Note that OIG is a large collection consisting of existing collections.

Collections	Time	#Task types	#Tasks	#Examples
Nat. Inst. [208]	Apr-2021	6	61	193K
CrossFit [209]	Apr-2021	13	160	7.1M
FLAN [62]	Sep-2021	12	62	4.4M
P3 [210]	Oct-2021	13	267	12.1M
ExMix [211]	Nov-2021	11	107	18M
UnifiedSKG [212]	Jan-2022	6	21	812K
Super. Inst. [79]	Apr-2022	76	1616	5M
MVPCorpus [213]	Jun-2022	11	77	41M
xP3 [84]	Nov-2022	17	85	81M
OIG[23]	Mar-2023	-	-	43M

5.1 Instruction Tuning

In essence, instruction tuning is the approach to fine-tuning pre-trained LLMs on a collection of formatted instances in the form of natural language [62], which is highly related to supervised fine-tuning [61] and multi-task prompted training [28]. In order to perform instruction tuning, we first need to collect or construct instruction-formatted instances. Then, we employ these formatted instances to fine-tune LLMs in a supervised learning way (*e.g.,* training with the sequence-to-sequence loss). After instruction tuning, LLMs can demonstrate superior abilities to generalize to unseen tasks [28, 62, 64], even in a multilingual setting [84].

A recent survey [214] presents a systematic overview of the research on instruction tuning. In comparison to that, we mainly focus on the effect of instruction tuning on LLMs and provide detailed guidelines or strategies for instance collection and tuning. Besides, we also discuss the use of instruction tuning for satisfying the real needs of users, which has been widely applied in existing LLMs, *e.g.,* InstructGPT [61] and GPT-4 [46].

5.1.1 Formatted Instance Construction

Generally, an instruction-formatted instance consists of a task description (called an *instruction*), an input-output pair, and a small number of demonstrations (optional). As important public resources, existing studies have released a large number of labeled data formatted in natural language (see the list of available resources in Table 6). Next, we introduce two major methods for constructing formatted instances (see an illustration in Figure 5) and then discuss several key factors for instance construction.

Formatting Existing Datasets. Before instruction tuning was proposed, several early studies [211, 213, 215, 216] collected the instances from a diverse range of tasks (*e.g.,* text
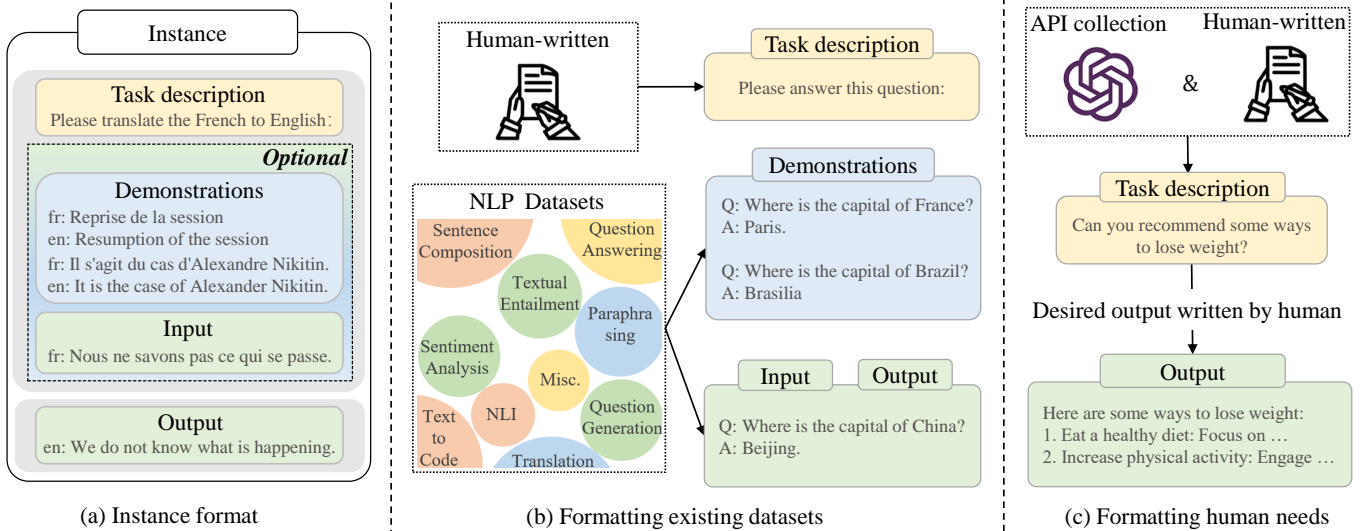
Fig. 5. An illustration of instance formatting and two different methods for constructing the instruction-formatted instances.

summarization, text classification, and translation) to create supervised multi-task training datasets. As a major source of instruction tuning instances, it is convenient to format these multi-task training datasets with natural language task descriptions. Specifically, recent work [28, 61, 62, 79] augments the labeled datasets with human-written task descriptions, which instructs LLMs to understand the tasks by explaining the task goal. For example, in Figure 5(b), a task description "*Please answer this question*" is added for each example in the question-answering task. After instruction tuning, LLMs can generalize well to other unseen tasks by following their task descriptions [28, 62, 64]. In particular, it has been shown that instructions are the crucial factor in task generalization ability for LLMs [62]: by fine-tuning the model on labeled datasets with the task descriptions removed, it results in a dramatic drop in model performance. To better generate labeled instances for instruction tuning, a crowd-sourcing platform, PromptSource [210] has been proposed to effectively create, share, and verify the task descriptions for different datasets. To enrich the training instances, several studies [28, 213, 217] also try to invert the input-output pairs of existing instances with specially designed task descriptions for instruction tuning. For instance, given a question-answer pair, we can create a new instance by predicting the question-conditioned answer (*e.g.,* "*Please generate a question based on the answer:*"). Besides, some work [218] also leverages heuristic task templates to convert massive unlabeled texts into labeled instances.

Formatting Human Needs. Despite that a large number of training instances have been formatted with instructions, they mainly come from public NLP datasets, either lacking instruction diversity or mismatching with real human needs [61]. To overcome this issue, InstructGPT [61] proposes to take the queries that real users have submitted to the OpenAI API as the task descriptions. User queries are expressed in natural languages, which are particularly suitable for eliciting the ability of instruction following for LLMs. Additionally, to enrich the task diversity, human labelers are also asked to compose the instructions for real-

life tasks, including open-ended generation, open question answering, brainstorming, and chatting. Then, they let another group of labelers directly answer these instructions as the output. Finally, they pair one instruction (*i.e.,* the collected user query) and the expected output (*i.e.,* the human-written answer) as a training instance. Note that Instruct-GPT also employs these real-world tasks formatted in natural language for alignment tuning (discussed in Section 5.2). Further, GPT-4 [46] has designed potentially high-risk instructions and guided the model to reject these instructions through supervised fine-tuning for safety concerns. Besides, to reduce the burden of human annotation, several semi-automated approaches [219–221] have also been proposed for constructing instances by feeding existing instances into LLMs to generate diverse task descriptions and instances.

Key Factors for Instance Construction. The quality of instruction instances has an important impact on the performance of the model. Here, we discuss some essential factors for instance construction.

• *Scaling the instructions.* It has been widely shown that scaling the number of tasks can largely enhance the generalization ability of LLMs [28, 62, 79]. With the increasing of the task number, the model performance initially shows a continuous growth pattern, while the gain becomes negligible when it reaches a certain level [64, 79]. A plausible speculation is that a certain number of representative tasks can provide relatively sufficient knowledge and adding more tasks may not bring additional gains [64]. Besides, it is also beneficial to enhance the diversity of the task descriptions in several aspects, such as length, structure, and creativity [28]. As for the number of instances per task, it has been found that a small number of instances can usually saturate the generalization performance of the model [62, 64]. Whereas, increasing the number of instances for some tasks to a large number (*e.g.,* a few hundreds) could potentially result in the overfitting issue and impair the model performance [79].

• *Formatting design.* As an important factor, the design of natural language format also highly impacts the generalization performance of LLMs [79]. Typically, we can add

task descriptions and optional demonstrations to the input-output pairs of existing datasets, where the task description is the most key part for LLMs to understand the task [79]. Further, it can lead to substantial improvements by using an appropriate number of exemplars as demonstrations [64], which also alleviates the model sensitivity to instruction engineering [62, 64]. However, incorporating other components (*e.g.,* things to avoid, reasons, and suggestions) into instructions may have a negligible or even adverse effect on the performance of LLMs [79, 208]. Recently, to elicit the step-by-step reasoning ability of LLMs, some work [64] proposes to include chain-of-thought (CoT) examples for some reasoning datasets, such as arithmetic reasoning. It has been shown that fine-tuning LLMs with both CoT and non-CoT examples can lead to a good performance across various reasoning tasks, including those that require multi-hop reasoning ability (*e.g.,* commonsense question answering and arithmetic reasoning) as well as those without the need for such a reasoning way (*e.g.,* sentiment analysis and extractive question answering) [64, 85].

To summarize, it seems that the diversity of instructions is more important than the number of instances since the well-performing InstructGPT [61] and Alpaca [221] utilize fewer but more diverse instructions (or instances) than the Flan-series LLMs [62, 64]. Further, it is more useful to invite labelers to compose human-need tasks than using dataset-specific tasks. While, it still lacks the guidelines to annotate human-need instances, making the task composition somehow heuristic. To reduce human efforts, we can either reuse existing formatted datasets (Table 6) or automatically construct the instructions using existing LLMs [219].

5.1.2 Instruction Tuning Strategies

Unlike pre-training, instruction tuning is often more efficient since only a moderate number of instances are used for training. Since instruction tuning can be considered as a supervised training process, its optimization is different from pre-training in several aspects [64], such as the training objective (*i.e.,* sequence-to-sequence loss) and optimization configuration (*e.g.,* smaller batch size and learning rate), which require special attention in practice. In addition to these optimization configurations, there are also two important aspects to consider for instruction tuning:

Balancing the Data Distribution. Since instruction tuning involves a mixture of different tasks, it is important to balance the proportion of different tasks during fine-tuning. A widely used method is the *examples-proportional mixing* strategy [73], *i.e.,* combining all the datasets and sampling each instance equally from the mixed datasets. Furthermore, increasing the sampling ratio of high-quality collections (*e.g.,* FLAN [62] and P3 [210]) can generally lead to performance improvement according to recent findings [64, 85]. While, it is common to set a *maximum cap* to control the maximum number of examples that a dataset can contain during instruction tuning [73], which is set to prevent larger datasets from overwhelming the entire distribution [73, 85]. In practice, the maximum cap is typically set to several thousands or tens of thousands according to different datasets [62, 64].

Combining Instruction Tuning and Pre-Training. To make the tuning process more effective and stable, OPT-IML [85] incorporates pre-training data during instruction tuning, which can be regarded as regularization for model tuning. Further, instead of using a separate two-stage process (*pre-training* then *instruction tuning*), some studies attempt to train a model from scratch with a mixture of pre-training data (*i.e.,* plain texts) and instruction tuning data (*i.e.,* formatted datasets) using multi-task learning [73, 211]. Specifically, GLM-130B [83] and Galactica [35] integrate instruction-formatted datasets as a small proportion of the pre-training corpora to pre-train LLMs, which potentially achieves the advantages of pre-training and instruction tuning at the same time.

5.1.3 The Effect of Instruction Tuning

In this part, we discuss the effect of instruction tuning on LLMs in two major aspects.

Performance Improvement. Despite being tuned on a moderate number of instances, instruction tuning has become an important way to improve or unlock the abilities of LLMs [64]. Recent studies have experimented with language models in multiple scales (ranging from 77M to 540B), showing that the models of different scales can all benefit from instruction tuning [64, 217], yielding improved performance as the parameter scale increases [84]. Further, smaller models with instruction tuning can even perform better than larger models without fine-tuning [28, 64]. Besides the model scale, instruction tuning demonstrates consistent improvements in various model architectures, pre-training objectives, and model adaptation methods [64]. In practice, instruction tuning offers a general approach to enhancing the abilities of existing language models [64] (including small-sized PLMs). Besides, it is also much less costly than pre-training, since the amount of instruction data required by LLMs is significantly smaller than pre-training data.

Task Generalization. Instruction tuning encourages the model to understand natural language instructions for task completion. It endows LLMs with the ability (often considered as an emergent ability) to follow human instructions [31] to perform specific tasks without demonstrations, even on unseen tasks [64]. A large number of studies have confirmed the effectiveness of instruction tuning to achieve superior performance on both seen and unseen tasks [85, 217]. Besides, instruction tuning has been shown to be useful in alleviating several weaknesses of LLMs (*e.g.,* repetitive generation or complementing the input without accomplishing a certain task) [61, 64], leading to a superior capacity to solve real-world tasks for LLMs. Furthermore, LLMs trained with instruction tuning can generalize to related tasks across languages. For example, BLOOMZ-P3 [84] is fine-tuned based on BLOOM [69] using English-only task collection P3 [210]. Interestingly, BLOOMZ-P3 can achieve a more than 50% improvement in multilingual sentence completion tasks compared to BLOOM, which shows that instruction tuning can help LLMs acquire general task skills from English-only datasets and transfer such skills into other languages [84]. In addition, it has been found that using English-only instructions can produce satisfactory

results on multilingual tasks [84], which helps reduce the effort of instruction engineering for a specific language.

5.2 Alignment Tuning

This part first presents the background of alignment with its definition and criteria, then focuses on the collection of human feedback data for aligning LLMs, and finally discusses the key technique of reinforcement learning from human feedback for alignment tuning.

5.2.1 Background and Criteria for Alignment

Background. LLMs have shown remarkable capabilities in a wide range of NLP tasks [55, 56, 62, 81]. However, these models may sometimes exhibit unintended behaviors, *e.g.*, fabricating false information, pursuing inaccurate objectives, and producing harmful, misleading, and biased expressions [61, 222]. For LLMs, the language modeling objective pre-trains the model parameters by word prediction while lacking the consideration of human values or preferences. To avert these unexpected behaviors, human alignment has been proposed to make LLMs act in line with human expectations [61, 100]. However, unlike the original pre-training and adaptation tuning (*e.g.*, instruction tuning), such an alignment requires considering very different criteria (*e.g.*, helpfulness, honesty, and harmlessness). It has been shown that alignment might harm the general abilities of LLMs to some extent, which is called *alignment tax* in related literature [61, 223, 224].

Alignment Criteria. Recently, there is increasing attention on developing multifarious criteria to regulate the behaviors of LLMs. Here, we take three representative alignment criteria (*i.e.*, helpful, honest, and harmless) as examples for discussion, which have been widely adopted in existing literature [61, 222, 223]. Besides, there are other alignment criteria for LLMs from different perspectives including behavior, intent, incentive, and inner aspects [222], which are essentially similar (or at least with similar alignment techniques) to the above three criteria. It is also feasible to modify the three criteria according to specific needs, *e.g.*, substituting honesty with correctness [100] or focusing on some specified criteria [224]. Next, we give brief explanations about the three representative alignment criteria:

• *Helpfulness.* To be helpful, the LLM should demonstrate a clear attempt to assist users in solving their tasks or answering questions in a concise and efficient manner as possible. At a higher level, when further clarification is needed, the LLM should demonstrate the capability of eliciting additional relevant information through pertinent inquiries and exhibit suitable levels of sensitivity, perceptiveness, and prudence [223]. Realizing the alignment of helpful behavior is challenging for LLMs since it is difficult to precisely define and measure the intention of users [222].

• *Honesty.* At a basic level, a LLM aligned to be honest should present accurate content to users instead of fabricating information. Additionally, it is crucial for the LLM to convey appropriate degrees of uncertainty in its output, in order to avoid any form of deception or misrepresentation of information. This requires the model to know about its capabilities and levels of knowledge (*e.g.*, "know

unknowns"). According to the discussion in [223], honesty is a more objective criterion compared to helpfulness and harmlessness, hence honesty alignment could potentially be developed with less reliance on human efforts.

• *Harmlessness.* To be harmless, it requires that the language produced by the model should not be offensive or discriminatory. To the best of its abilities, the model should be capable of detecting covert endeavors aimed at soliciting requests for malicious purposes. Ideally, when the model was induced to conduct a dangerous action (*e.g.*, committing a crime), the LLM should politely refuse. Nonetheless, *what behaviors* are deemed harmful and *to what extent* vary amongst individuals or societies [223] highly depend on who is using the LLM, the type of the posed question, and the context (*e.g.*, time) at which the LLM is being used.

As we can see, these criteria are quite subjective, and are developed based on human cognition. Thus, it is difficult to directly formulate them as optimization objectives for LLMs. In existing work, there are many ways to fulfill these criteria when aligning LLMs. A promising technique is *red teaming* [115, 225], which involves using manual or automated means to probe LLMs in an adversarial way to generate harmful outputs and then updates LLMs to prevent such outputs.

5.2.2 Collecting Human Feedback

During the pre-training stage, LLMs are trained using the language modeling objective on a large-scale corpus. However, it cannot take into account the subjective and qualitative evaluations of LLM outputs by humans (called *human feedback* in this survey). High-quality human feedback is extremely important for aligning LLMs with human preferences and values. In this part, we discuss how to select a team of human labelers for feedback data collection.

Human Labeler Selection. In existing work, the dominant method for generating human feedback data is human annotation [61, 100, 226]. This highlights the critical role of selecting appropriate human labelers. To provide high-quality feedback, human labelers are supposed to have a qualified level of education and excellent proficiency in English. For example, Sparrow [100] requires human labelers to be UK-based native English speakers who have obtained at least an undergraduate-level educational qualification. Further, in [224], about half of human labelers for high priority tasks were recruited from the US-based Amazon Mechanical Turk workforce with a master's qualification. Even then, several studies [112, 226] have found that there still exists a mismatch between the intentions of researchers and human labelers, which may lead to low-quality human feedback and cause LLMs to produce unexpected output. To address this issue, InstructGPT [61] further conducts a screening process to filter labelers by assessing the agreement between human labelers and researchers. Specifically, researchers first label a small amount of data and then measure the agreement between themselves and human labelers. The labelers with the highest agreement will be selected to proceed with the subsequent annotation work. In some other work [227], "super raters" are used to ensure the high quality of human feedback. Researchers evaluate the performance of human labelers and select a group of

well-performing human labelers (*e.g.,* high agreement) as super raters. The super raters will be given priority to collaborate with the researchers in the subsequent study. When human labelers annotate the output of LLMs, it is helpful to specify detailed instructions and provide instant guidance for human labelers [112], which can further regulate the annotation of labelers.

Human Feedback Collection. In existing work, there are mainly three kinds of approaches to collecting feedback and preference data from human labelers.

• *Ranking-based approach.* In early work [226, 228], human labelers often evaluate model-generated outputs in a coarse-grained manner (*i.e.,* only selecting the best) without taking into account more fine-grained alignment criteria. Nonetheless, different labelers may hold diverse opinions on the selection of the best candidate output, and this method disregards the unselected samples, which may lead to inaccurate or incomplete human feedback. To address this issue, subsequent studies [100, 224] introduce the Elo rating system to derive the preference ranking by comparing candidate outputs. The ranking of outputs serves as the training signal that guides the model to prefer certain outputs over others, thus inducing outputs that are more reliable and safer.

• *Question-based approach.* Further, human labelers can provide more detailed feedback by answering certain questions designed by researchers [72], covering the alignment criteria as well as additional constraints for LLMs. Specially, in WebGPT [72], to assist the model in filtering and utilizing relevant information from retrieved documents, human labelers are required to answer questions with multiple options about whether the retrieved documents are useful for answering the given input.

• *Rule-based approach.* Besides, many studies develop rule-based methods to provide more detailed human feedback. As a typical case, Sparrow [100] not only selects the response that labelers consider the best but also uses a series of rules to test whether model-generated responses meet the alignment criteria of being helpful, correct, and harmless. In this way, two kinds of human feedback data can be obtained: (1) the response preference feedback is obtained by comparing the quality of model-generated output in pairs, and (2) the rule violation feedback is obtained by collecting the assessment from human labelers (*i.e.,* a score indicating to what extent the generated output has violated the rules). Furthermore, GPT-4 [46] utilizes a set of zero-shot classifiers (based on GPT-4 itself) as rule-based reward models, which can automatically determine whether the model-generated outputs violate a set of human-written rules.

In the following, we focus on a well-known technique, reinforcement learning from human feedback (RLHF), which has been widely used in the recent powerful LLMs such as ChatGPT. As discussed below, the alignment criteria introduced in Section 5.2.1 can be fulfilled by learning from human feedback on the responses of LLMs to users' queries.

5.2.3 Reinforcement Learning from Human Feedback

To align LLMs with human values, reinforcement learning from human feedback (RLHF) [70, 226] has been proposed to fine-tune LLMs with the collected human feedback data,
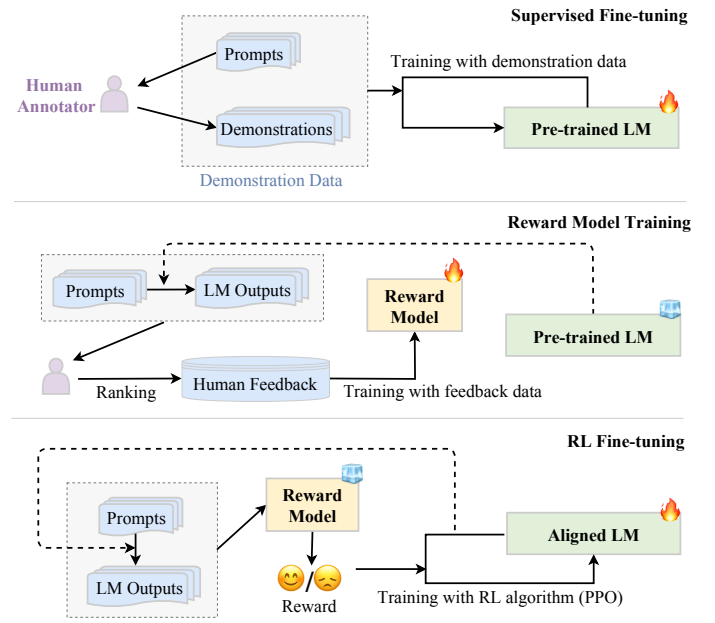
Fig. 6. The workflow of the RLHF algorithm.

which is useful to improve the alignment criteria (*e.g.,* helpfulness, honesty, and harmlessness). RLHF employs reinforcement learning (RL) algorithms (*e.g.,* Proximal Policy Optimization (PPO) [111]) to adapt LLMs to human feedback by learning a reward model. Such an approach incorporates humans in the training loop for developing well-aligned LLMs, as exemplified by InstructGPT [61].

RLHF System. The RLHF system mainly comprises three key components: a pre-trained LM to be aligned, a reward model learning from human feedback, and a RL algorithm training the LM. Specifically, the *pre-trained LM* is typically a generative model that is initialized with existing pre-trained LM parameters. For example, OpenAI uses 175B GPT-3 for its first popular RLHF model, InstructGPT [61], and DeepMind uses the 280 billion parameter model Gopher [59] for its GopherCite model [227]. Further, the *reward model (RM)* provides (learned) guidance signals that reflect human preferences for the text generated by the LM, usually in the form of a scalar value. The reward model can take on two forms: a fine-tuned LM or a LM trained de novo using human preference data. Existing work typically employs reward models having a parameter scale different from that of the aligned LM [61, 227]. For example, OpenAI uses 6B GPT-3 and DeepMind uses 7B Gopher as the reward model, respectively. Finally, to optimize the pre-trained LM using the signal from the reward model, a specific *RL algorithm* is designed for large-scale model tuning. Specifically, Proximal Policy Optimization (PPO) [111] is a widely used RL algorithm for alignment in existing work [61, 100, 227].

Key Steps for RLHF. Figure 6 illustrates the overall three-step process of RLHF [61, 112] as introduced below.

• *Supervised fine-tuning.* To make the LM initially perform desired behaviors, it usually needs to collect a supervised dataset containing input prompts (instruction) and desired outputs for fine-tuning the LM. These prompts and outputs can be written by human labelers for some specific tasks

while ensuring the diversity of tasks. For example, Instruct-GPT [61] asks human labelers to compose prompts (*e.g.,* "*List five ideas for how to regain enthusiasm for my career*") and desired outputs for several generative tasks such as open QA, brainstorming, chatting, and rewriting. Note that the first step is optional in specific settings or scenarios.

• *Reward model training.* The second step is to train the RM using human feedback data. Specifically, we employ the LM to generate a certain number of output texts using sampled prompts (from either the supervised dataset or the human-generated prompt) as input. We then invite human labelers to annotate the preference for these pairs. The annotation process can be conducted in multiple forms, and a common approach is to annotate by ranking the generated candidate texts, which can reduce the inconsistency among annotators. Then, the RM is trained to predict the human-preferred output. In InstructGPT, labelers rank model-generated outputs from best to worst, and the RM (*i.e.,* 6B GPT-3) is trained to predict the ranking.

• *RL fine-tuning.* At this step, aligning (*i.e.,* fine-tuning) the LM is formalized as an RL problem. In this setting, the pre-trained LM acts as the policy that takes as input a prompt and returns an output text, the action space of it is the vocabulary, the state is the currently generated token sequence, and the reward is provided by the RM. To avoid eviating significantly from the initial (before tuning) LM, a penalty term is commonly incorporated into the reward function. For example, InstructGPT optimizes the LM against the RM using the PPO algorithm. For each input prompt, InstructGPT calculates the KL divergence between the generated results from the current LM and the initial LM as the penalty. It is noted that the second and final steps can be iterated in multiple turns for better aligning LLMs.

5.3 Efficient Tuning

In the above, we have discussed the approaches of instruction tuning and alignment tuning to adapt LLMs according to specific goals. Since LLMs consist of a huge amount of model parameters, it would be costly to perform the full-parameter tuning. In this section, we will discuss how to conduct efficient tuning on LLMs. We first review several representative parameter-efficient fine-tuning methods for Transformer language models, and then summarize existing work on parameter-efficient fine-tuned LLMs.

5.3.1 Parameter-Efficient Fine-Tuning Methods

In existing literature, parameter-efficient fine-tuning [229, 229–232] has been an important topic that aims to reduce the number of trainable parameters while retaining a good performance as possible. In what follows, we briefly review four parameter-efficient fine-tuning methods for Transformer language models, including adapter tuning, prefix tuning, prompt tuning and LoRA.

Adapter Tuning. Adapter tuning incorporates small neural network modules (called *adapter*) into the Transformer models [233]. To implement the adapter module, a bottleneck architecture has been proposed in [233, 234], which first compresses the original feature vector into a smaller dimension (followed by a nonlinear transformation) and then recovers it to the original dimension. The adapter modules

would be integrated into each Transformer layer, typically using a serial insertion after each of the two core parts (*i.e.,* attention layer and feed-forward layer) of a Transformer layer. Alternatively, parallel adapters [235] can be also used in Transformer layers, where it places two adapter modules in parallel with the attention layer and feed-forward layer accordingly. During fine-tuning, the adapter modules would be optimized according to the specific task goals, while the parameters of the original language model are frozen in this process. In this way, we can effectively reduce the number of trainable parameters during fine-tuning.

Prefix Tuning. Prefix tuning [230] prepends a sequence of prefixes, which are a set of trainable continuous vectors, to each Transformer layer in language models. These prefix vectors are task-specific, which can be considered as virtual token embeddings. To optimize the prefix vectors, a reparameterization trick [230] has been proposed by learning a MLP function that maps a smaller matrix to the parameter matrix of prefixes, instead of directly optimizing the prefixes. It has been shown that this trick is useful for stable training. After optimization, the mapping function would be discarded, and only the derived prefix vectors are kept to enhance task-specific performance. Since only the prefix parameters would be trained, it can lead to a parameter-efficient model optimization. Similar to prefix tuning, p-tuning v2 [236] incorporates layer-wise prompt vectors into the Transformer architecture specially for natural language understanding, which also utilizes multi-task learning for jointly optimizing shared prompts. It has been shown to be useful in improving the model performance of different parameter scales on natural language understanding tasks.

Prompt Tuning. Different from prefix tuning, prompt tuning [231, 237] mainly focuses on incorporating trainable prompt vectors at the input layer[24]. Based on the discrete prompting methods [239, 240], it augments the input text by including a group of soft prompt tokens (either in a free form [237] or a prefix form [231]), and then takes the prompt-augmented input to solve specific downstream tasks. In implementation, task-specific prompt embeddings are combined with the input text embeddings, which are subsequently fed into language models. P-tuning [237] has proposed a free form to combine the context, prompt and target tokens, which can be applied to the architectures for both natural language understanding and generation. They further learn the representations of soft prompt tokens by a bidirectional LSTM. Another representative approach [231] named *prompt tuning* directly prepends prefix prompts to the input. During training, only the prompt embeddings would be learned according to task-specific supervisions. While, since this method only includes a small number of trainable parameters at the input layer, it has been found

24. Here, prompt tuning denotes a category of related efficient tuning methods exemplified by the work [231, 237, 238], instead of a specific method as used in [231]. Indeed, the prefix based tuning methods [230, 236] can be also considered as prompting methods, which are called *deep prompting tuning* in [236]. While, in this survey, prompt tuning specially refer to the methods that only include the prompt tokens at the input layer, in the context of LLMs. We assign p-tuning v2 [236] to the category of prefix tuning, because it incorporates layerwise prompts in langauge models.

that the performance highly relies on the model capacity of the underlying language models [231].

Low-Rank Adaptation (LoRA). LoRA [232] imposes the low-rank constraint for approximating the update matrix at each dense layer, so as to reduce the trainable parameters for adapting to downstream tasks. Consider the case of optimizing a parameter matrix $\mathbf{W}$. The update process can be written in a general form as: $\mathbf{W} \leftarrow \mathbf{W} + \Delta\mathbf{W}$. The basic idea of LoRA is to freeze the original matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ while approximating the parameter update $\Delta\mathbf{W}$ by low-rank decomposition matrices, *i.e.,* $\Delta\mathbf{W} = \mathbf{A} \cdot \mathbf{B}^\top$, where $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{n \times k}$ are the trainable parameters for task adaptation and $r \ll \min(m, n)$ is the reduced rank. The major merit of LoRA is that it can largely save the memory and storage usage (*e.g.,* VRAM). Further, one can only keep a single large model copy, while maintaining a number of task-specific low-rank decomposition matrices for adapting to different downstream tasks. Further, several studies have also discussed how to set the rank in a more principled approach, *e.g.,* importance score based allocation [241] and search-free optimal rank selection [242].

Besides the above methods, there is extensive research on efficient tuning of Transformer language models. While, a more comprehensive discussion of efficient tuning is beyond the scope of this article, which can be found in the related papers on this topic [229, 235].

5.3.2 Parameter-Efficient Fine-Tuning on LLMs

With the rising of LLMs, efficient tuning has attracted increasing research attention for developing a more lightweight adaptation approach in downstream tasks.

In particular, LoRA [232] has been widely applied to open-source LLMs (*e.g.,* LLaMA and BLOOM) for parameter-efficient fine-tuning. Among these research attempts, LLaMA and its variants have gained much attention for parameter-efficient tuning. For example, Alpaca-LoRA [243] has been trained using LoRA as a lightweight tuned version of Alpaca [221] (a fine-tuned 7B LLaMA model with 52K human demonstrations of instruction following). There are extensive explorations of Alpaca-LoRA ranging in different languages or model sizes, which can be found in the collection page[25]. Besides, LLaMA-Adapter [244] inserts learnable prompt vectors into each Transformer layer, in which zero-initialized attention has been proposed to improve the training by mitigating the influence of under-fitted prompt vectors. Besides, they also extend this approach to a multi-modal setting, *e.g.,* visual question answering.

Further, an empirical study [234] has been conducted to examine the effect of different tuning methods on language models. They compare four efficient tuning methods including serial adapter tuning [233], parallel adapter tuning [235, 245], and LoRA [232], on three open-source LLMs, namely GPT-J (6B), BLOOM (7.1B) and LLaMA (7B), for evaluation. Based on the experimental results on six math reasoning datasets, they show that these efficient-tuning methods under-perform the reference baseline GPT-3.5 on difficult tasks, while achieving a comparable performance on simple tasks. Overall, LoRA performs relatively well among these comparison methods, using significantly fewer trainable parameters.

As an important resource, the library *PEFT* [246] (standing for parameter-efficient fine-tuning) has been released on GitHub[26]. It has included several widely used efficient tuning methods, including LoRA [232]/AdaLoRA [241], prefix-tuning [230, 236], P-Tuning [237], and prompt-tuning [231]. Further, it supports a number of language models such as GPT-2 and LLaMA, and also covers several representative vision Transformer models (*e.g.,* ViT and Swin Transformer).

As discussed in Section 5.3.1, there have been a large number of efficient tuning methods proposed in the existing literature. While, most of these approaches are tested on small-sized pre-trained language models, instead of the LLMs. So far, there still lacks a thorough investigation on the effect of different efficient tuning methods on large-sized language models at different settings or tasks.

6 UTILIZATION

After pre-training or adaptation tuning, a major approach to using LLMs is to design suitable prompting strategies for solving various tasks. A typical prompting method is *in-context learning* [50, 55], which formulates the task description and/or demonstrations in the form of natural language text. In addition, *chain-of-thought prompting* [33] can be employed to enhance in-context learning by involving a series of intermediate reasoning steps into prompts. Next, we will elaborate on the details of the two techniques.

6.1 In-Context Learning

As a special prompting form, in-context learning (ICL) is first proposed along with GPT-3 [55], which has become a typical approach to utilizing LLMs.

6.1.1 Prompting Formulation

As stated in [55], ICL uses a formatted natural language prompt, consisting of the task description and/or a few task examples as demonstrations. Figure 7 presents the illustration of ICL. First, starting with a task description, a few examples are selected from the task dataset as demonstrations. Then, they are combined in a specific order to form natural language prompts with specially designed templates. Finally, the test instance is appended to the demonstration as the input for LLMs to generate the output. Based on task demonstrations, LLMs can recognize and perform a new task without explicit gradient update.

Formally, let $D_k = \{f(x_1, y_1), \ldots, f(x_k, y_k)\}$ represent a set of demonstrations with k examples, where $f(x_k, y_k)$ is the prompt function that transforms the k-th task example into natural language prompts. Given the task description I, demonstration D_k, and a new input query x_{k+1}, the
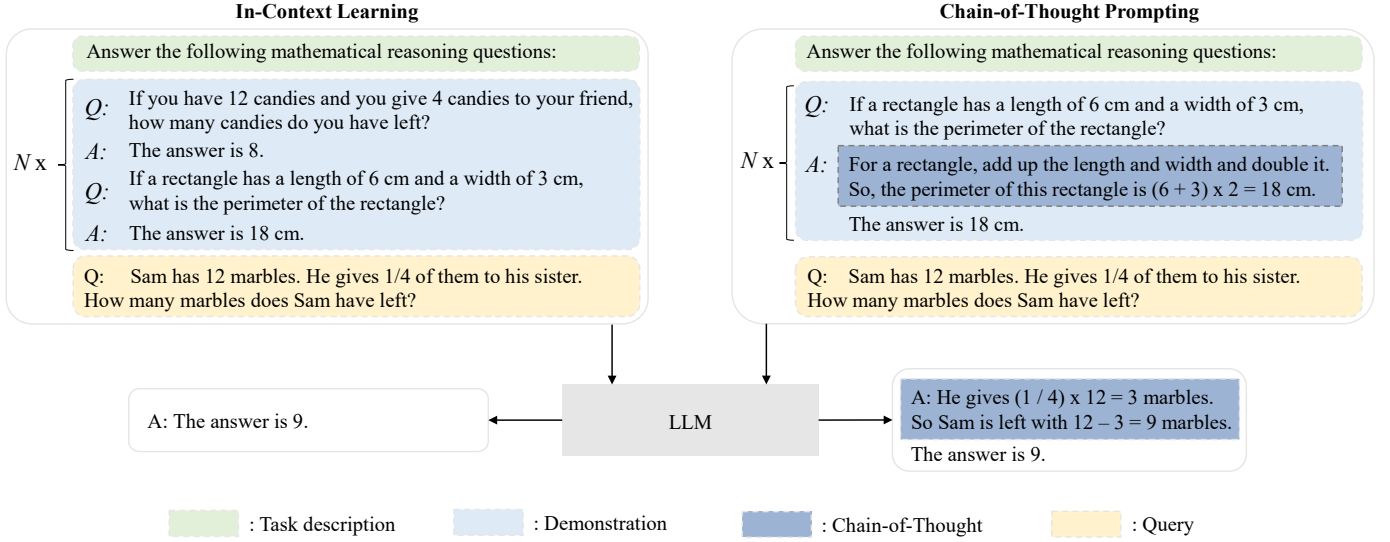
---

25. https://github.com/tloen/alpaca-lora

26. https://github.com/huggingface/peft

In-Context Learning

Answer the following mathematical reasoning questions:

N x

Q: If you have 12 candies and you give 4 candies to your friend, how many candies do you have left?

A: The answer is 8.

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: The answer is 18 cm.

Q: Sam has 12 marbles. He gives 1/4 of them to his sister. How many marbles does Sam have left?

Chain-of-Thought Prompting

Answer the following mathematical reasoning questions:

N x

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: For a rectangle, add up the length and width and double it. So, the perimeter of this rectangle is (6 + 3) x 2 = 18 cm.

The answer is 18 cm.

Q: Sam has 12 marbles. He gives 1/4 of them to his sister. How many marbles does Sam have left?

LLM

A: The answer is 9.

A: He gives (1 / 4) x 12 = 3 marbles. So Sam is left with 12 − 3 = 9 marbles.

The answer is 9.

☐ : Task description   ☐ : Demonstration   ☐ : Chain-of-Thought   ☐ : Query

Fig. 7. A comparative illustration of in-context learning (ICL) and chain-of-thought (CoT) prompting. ICL prompts LLMs with a natural language description, several demonstrations, and a test query. While CoT prompting involves a series of intermediate reasoning steps in prompts.

prediction of the output $\hat{y}_{k+1}$ generated from LLMs can be formulated as follows[27]:

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

(6)

where the actual answer y_{k+1} is left as a blank to be predicted by the LLM. Since the performance of ICL heavily relies on demonstrations, it is an important issue to properly design them in the prompts. According to the construction process in Equation (6), we focus on three major aspects in formatting demonstrations in the prompts, including how to select examples that make up demonstrations, format each example into the prompt with the function $f(\cdot)$, and arrange demonstrations in a reasonable order.

A comprehensive review of ICL has been presented in the survey paper [50], and we suggest the readers referring to it for a more general, detailed discussion on this topic. Compared with this survey, we specially focus on the discussion of applying ICL to LLMs in two major aspects, *i.e.,* demonstration design and the underlying mechanism of ICL. Besides, ICL also has a close connection with instruction tuning (discussed in Section 5.1) in that both utilize natural language to format the task or instances. However, instruction tuning needs to fine-tune LLMs for adaptation, while ICL only prompts LLMs for utilization. Furthermore, instruction tuning can enhance the ICL ability of LLMs to perform target tasks, especially in the zero-shot setting (only using task descriptions) [64].

6.1.2 Demonstration Design

Several studies have shown that the effectiveness of ICL is highly affected by the design of demonstrations [247–249]

27. When ICL was introduced in the GPT-3's paper [55], it was originally defined to be a combination of the task description and demonstration examples, wherein either component is dispensable. Following this definition, when a LLM is required to solve an unseen task by using only task descriptions, it can be also considered to perform ICL for task solving, whereas the ICL ability can be enhanced by instruction tuning.

Following the discussion in Section 6.1.1, we will introduce the demonstration design of ICL from three major aspects, *i.e.,* demonstration selection, format, and order.

Demonstration Selection. The performance of ICL tends to have a large variance with different demonstration examples [250], so it is important to select a subset of examples that can effectively leverage the ICL capability of LLMs. There are two main demonstration selection approaches, namely heuristic and LLM-based approaches:

• *Heuristic approaches.* Due to the simplicity and low costs, existing work widely adopts heuristic methods to select demonstrations. Several studies employ a k-NN based retriever to select examples that are semantically relevant to the query [250, 251]. However, they perform the selection individually for each example, rather than evaluating the example set as a whole. To resolve this issue, diversity-based selection strategies are proposed to choose the most representative set of examples for specific tasks [252, 253]. Furthermore, in [254], both relevance and diversity are taken into consideration when selecting demonstrations.

• *LLM-based approaches.* Another line of work selects demonstrations by making use of LLMs. For example, LLMs can be utilized to directly measure the informativeness of each example according to the performance gain after adding the example [255]. Besides, EPR [256] proposes a two-stage retrieval approach that first recalls similar examples with an unsupervised method (*e.g.,* BM25) and then ranks them using a dense retriever (trained with positive and negative examples labeled by LLMs). As an alternative approach, the task of demonstration selection can be formulated into a RL problem, where LLMs serve as the reward function to provide feedback for training the policy model [257]. Since LLMs perform well for text annotation [258], some recent studies employ LLM itself as the demonstration generator without human intervention [259, 260].

To summarize, as discussed in [261], the selected demonstration examples in ICL should contain sufficient informa-

tion about the task to solve as well as be relevant to the test query, for the above two selection approaches.

Demonstration Format. After selecting task examples, the next step is to integrate and format them into a natural language prompt for LLMs. A straightforward method is to instantiate a pre-defined template with the corresponding input-output pairs [36]. To construct more informative templates, recent studies consider adding task descriptions [64] or enhancing the reasoning capability of LLMs with chain-of-thought prompts [33]. For instance, in [208], the authors collect a large-scale dataset with task descriptions written by humans. After tuning with this dataset, the performance on seen tasks can be boosted, and LLMs can also generalize to unseen tasks to some extent. To reduce the annotation costs, a semi-automated approach has been proposed in [219] by employing a seed set consisting of human-written task descriptions to guide LLMs to generate task descriptions for new tasks. Since it is costly to manually annotate demonstration formats for different tasks, some work also studies how to automatically generate high-quality ones. As two representative methods, Auto-CoT [262] leverages LLMs with the zero-shot prompt "*Let's think step by step*" for generating intermediate reasoning steps, while least-to-most prompting [263] first queries LLMs to perform problem decomposition and then utilizes LLMs to sequentially solve sub-problems based on the intermediate answers to previously solved ones.

Demonstration Order. LLMs are shown to sometimes suffer from the recency bias, *i.e.*, they are prone to repeat answers that are near the end of demonstrations [249]. Thus, it is important to arrange demonstrations (*i.e.*, task examples) in a reasonable order. Early work proposes several heuristic methods to quickly find a good order. For example, demonstrations can be directly organized according to their similarity to the query in the embedding space [250]: the more similar, the closer to the end. Besides, global and local entropy metrics can be used to score different demonstration orders [248]. To integrate more task information, some recent studies propose to minimize the code length required to compress and transmit task labels, which is inspired by information theory [264]. However, these methods need additional labeled data as the validation set to evaluate the performance of specific demonstration orders. To eliminate this need, the authors in [248] propose to sample the validation data from the LLM itself.

6.1.3 Underlying Mechanism

After pre-training, LLMs can exhibit intriguing ICL capability without being updated. In what follows, we discuss two key questions about the ICL ability of LLMs, *i.e.*, "*how does pre-training affect the ICL ability*" and "*how do LLMs perform ICL during inference*".

How Pre-Training Affects ICL? ICL is first proposed in GPT-3 [55], and it has shown that the ICL ability becomes more significant with a larger model size. While, some studies reveal that small-scale PLMs can also demonstrate a strong ICL ability with specially designed training tasks (*e.g.*, learning to predict the label with task examples and the query as the input), and may even surpass larger mod-

els [265]. It suggests that the design of training tasks is an important influence factor of the ICL capability of LLMs. Besides training tasks, recent studies have also investigated the relationship between ICL and the pre-training corpora [261, 266, 267]. It has been shown that the performance of ICL heavily depends on the source of pre-training corpora rather than the scale [267]. Another study [266] provides an in-depth analysis of the impact of training data distribution. They find that ICL emerges when the training data can be clustered into numerous infrequent classes, instead of being uniformly distributed. Furthermore, the authors in [261] theoretically explain ICL as the product of pre-training on documents that exhibit long-range coherence.

How LLMs Perform ICL? At the inference stage, researchers focus on analyzing how the ICL capability operates based on given demonstrations since no explicit learning or updating is involved. They typically analyze from the perspective of gradient descent and consider ICL as implicit fine-tuning [60, 268]. Under this framework, the ICL process can be explained as follows: by means of forward computation, LLMs generate meta-gradients with respect to demonstrations and implicitly perform gradient descent via the attention mechanism. Experiments also show that certain attention heads in LLMs are capable of performing task-agnostic atomic operations (*e.g.*, copying and prefix matching), which are closely related to the ICL ability [269, 270]. To further explore the working mechanism of ICL, some studies abstract ICL as an algorithm learning process [271–273]. Specifically, the authors in [272] find that LLMs essentially encode implicit models through their parameters during pre-training. With the examples provided in ICL, LLMs can implement learning algorithms such as gradient descent or directly compute the closed-form solution to update these models during forward computation. Under this explanation framework, it has been shown that LLMs can effectively learn simple linear functions and even some complex functions like decision trees with ICL [271–273].

6.2 Chain-of-Thought Prompting

Chain-of-Thought (CoT) [33] is an improved prompting strategy to boost the performance of LLMs on complex reasoning tasks, such as arithmetic reasoning [274–276], commonsense reasoning [277, 278], and symbolic reasoning [33]. Instead of simply constructing the prompts with input-output pairs as in ICL, CoT incorporates intermediate reasoning steps that can lead to the final output into the prompts. In the following, we will elaborate on the usage of CoT with ICL and discuss when and why CoT prompting works.

6.2.1 In-context Learning with CoT

Typically, CoT can be used with ICL in two major settings, namely the few-shot and zero-shot settings, as introduced below.

Few-shot CoT. Few-shot CoT is a special case of ICL, which augments each demonstration ⟨*input, output*⟩ as ⟨*input, CoT, output*⟩ by incorporating the CoT reasoning steps. To apply this strategy, we next discuss two key issues, *i.e.*, how to

design appropriate CoT prompts and how to utilize the generated CoTs for deriving the final answer.

• *CoT prompt design.* It is critical to design appropriate CoT prompts for effectively eliciting the complex reasoning ability of LLMs. As a direct approach, it is shown that using diverse CoTs (*i.e.,* multiple reasoning paths for each problem) can effectively enhance their performance [279]. Another intuitive idea is that prompts with more complex reasoning paths are more likely to elicit the reasoning ability of LLMs [280], which can result in higher accuracy in generating correct answers. However, both of these two approaches rely on annotated CoT datasets, which limits their use in practice. To overcome this limitation, Auto-CoT [262] proposes to utilize Zero-shot-CoT [281] (detailed in the following part "*Zero-shot CoT*") to generate CoT reasoning paths by specially prompting LLMs, thus eliminating manual efforts. In order to boost the performance, Auto-CoT further divides the questions in the training set into different clusters and then chooses the questions that are closest to the centroid of each cluster, which is supposed to well represent the questions in the training set. Although few-shot CoT can be considered as a special prompt case of ICL, the ordering of demonstrations seems to have a relatively small impact compared to the standard prompt in ICL: reordering the demonstrations only results in a performance variation of less than 2% in most tasks [33].

• *Enhanced CoT strategies.* Besides enriching the contextual information, CoT prompting also provides more options to infer the answer given a question. Existing studies mainly focus on generating multiple reasoning paths, and try to find a consensus among the derived answers [282–284]. For instance, *self-consistency* [282] is proposed as a new decoding strategy when generating CoT and the final answer. It first generates several reasoning paths and then takes an ensemble over all the answers (*e.g.,* selecting the most consistent answer by voting among these paths). Self-consistency boosts the performance in CoT reasoning by a large margin, and can even improve some tasks where CoT prompting is usually worse than standard prompting (*e.g.,* closed-book question answering and natural language inference). Further, the authors in [283] expand the self-consistency strategy to a more general ensemble framework (extending to ensemble on the prompts), and they find that diverse reasoning paths are the key to the performance improvement in CoT reasoning. The above methods can be easily integrated into CoT prompting to enhance the performance without additional training. In contrast, other studies train a scoring model to measure the reliability of the generated reasoning paths [279] or continually train LLMs on the reasoning paths generated by themselves [285, 286] to improve the performance.

Zero-shot CoT. Different from few-shot CoT, zero-shot CoT does not include human-annotated task demonstrations in the prompts. Instead, it directly generates reasoning steps and then employs the generated CoTs to derive the answers. Zero-shot CoT is first proposed in [281], where the LLM is first prompted by "*Let's think step by step*" to generate reasoning steps and then prompted by "*Therefore, the answer is*" to derive the final answer. They find that such a strategy drastically boosts the performance when the model scale exceeds a certain size, but is not effective with small-scale models, showing a significant pattern of emergent abilities. In order to unlock the CoT ability on more tasks, Flan-T5 and Flan-PaLM [64] further perform instruction tuning on CoT annotations and the zero-shot performance on unseen tasks has been improved.

6.2.2 Further Discussion on CoT

In this part, we present discussions regarding two fundamental questions related to CoT, *i.e.,* "*when does CoT work for LLMs*" and "*why can LLMs perform CoT reasoning*".

When CoT works for LLMs? Since CoT is an emergent ability [31], it only has a positive effect on sufficiently large models (*e.g.,* typically containing 10B or more parameters [33]) but not on small models. Moreover, since CoT augments the standard prompting with intermediate reasoning steps, it is mainly effective to improve the tasks that require step-by-step reasoning [33], such as arithmetic reasoning, commonsense reasoning, and symbolic reasoning. Whereas, for other tasks that do not rely on complex reasoning, it might show worse performance than standard prompting [283], *e.g.,* MNLI-m/mm, SST-2, and QQP from GLUE [177]. Interestingly, it seems that the performance gain brought by CoT prompting could be significant only when standard prompting yields poor results [33].

Why LLMs Can Perform CoT Reasoning? As the second question, we discuss the underlying mechanism of CoT in the following two aspects.

• *The source of CoT ability*. Regarding the source of CoT capability, it is widely hypothesized that it can be attributed to training on code since models trained on it show a strong reasoning ability [47, 287]. Intuitively, code data is well organized with algorithmic logic and programming flow, which may be useful to improve the reasoning performance of LLMs. However, this hypothesis still lacks publicly reported evidence of ablation experiments (*with* and *without* training on code). Besides, instruction tuning seems not to be the key reason to obtain the CoT ability, since it has been empirically shown that instruction tuning on non-CoT data does not improve the performance on held-out CoT benchmarks [64].

• *The effect of prompting components*. The major distinction between CoT prompting and standard prompting is the incorporation of reasoning paths prior to the final answer. Thus, some researchers investigate the effect of different components in the reasoning paths. Specifically, a recent study identifies three key components in CoT prompting, namely *symbols* (*e.g.,* numerical quantities in arithmetic reasoning), *patterns* (*e.g.,* equations in arithmetic reasoning), and *text* (*i.e.,* the rest of tokens that are not symbols or patterns) [288]. It is shown that the latter two parts (*i.e.,* patterns and text) are essential to the model performance, and removing either one would lead to a significant performance drop. However, the correctness of symbols and patterns does not seem critical. Further, there exists a symbiotic relationship between text and patterns: the text helps LLMs to generate useful patterns, and patterns aid LLMs to understand tasks and generate texts that help solve them [288].

In summary, CoT prompting provides a general yet flexible approach to eliciting the reasoning ability of LLMs. There are also some preliminary attempts that extend this

technique to solve multimodal tasks [289] and multilingual tasks [290]. In addition to directly utilizing LLMs with ICL and CoT, some recent studies explore how to specialize the ability of LLMs towards specific tasks [291–293], which is called *model specialization* [294]. For example, the researchers in [294] specialize the ability of mathematical reasoning from LLMs through fine-tuning the small-scale Flan-T5 [64] on CoT reasoning paths generated by LLMs. Model specialization can also be applied to solve a variety of tasks like question answering [295], code synthesis [296], and information retrieval [297].

7 CAPACITY EVALUATION

To examine the effectiveness and superiority of LLMs, a surge of tasks and benchmarks have been leveraged for conducting empirical evaluation and analysis. We first introduce three types of basic evaluation tasks of LLMs for language generation and understanding, then present several advanced tasks of LLMs with more complicated settings or goals, and finally discuss existing benchmarks and empirical analyses.

7.1 Basic Evaluation Tasks

In this part, we mainly focus on three types of evaluation tasks for LLMs, *i.e.,* language generation, knowledge utilization, and complex reasoning. It is noted that we do not intend to have complete coverage of all the related tasks, but instead only focus on the most widely discussed or studied tasks for LLMs. Next, we introduce these tasks in detail.

7.1.1 Language Generation

According to the task definition, existing tasks about language generation can be roughly categorized into language modeling, conditional text generation, and code synthesis tasks. Note that code synthesis is not a typical NLP task, we include it for discussion because it can be directly solved by a number of LLMs (trained on code data) in a similar generation approach as natural language text.

Language Modeling. As the most fundamental ability of LLMs, *language modeling* aims to predict the next token based on the previous tokens [15], which mainly focuses on the capacity of basic language understanding and generation. For evaluating such an ability, typical language modeling datasets that existing work uses include Penn Treebank [298], WikiText-103 [299], and the Pile [130], where the metric of *perplexity* is commonly used for evaluating the model performance under the zero-shot setting. Empirical studies [55, 83] show that LLMs bring substantial performance gains over the previous state-of-the-art methods on these evaluation datasets. To better test the modeling capacity of long-range dependencies in text, the LAMBADA dataset [167] has been introduced, where LLMs are required to predict the last word of sentences based on a paragraph of context. Then, the accuracy and perplexity of the predicted last words are employed to evaluate LLMs. As shown in existing work, the performance on the language modeling tasks typically follows the scaling law [30], which means that scaling language models would improve the accuracy and reduce the perplexity.

Conditional Text Generation. As an important topic in language generation, conditional text generation [48] focuses on generating texts satisfying specific task demands based on the given conditions, typically including machine translation [367], text summarization [368], and question answering [369]. To measure the quality of the generated text, automatic metrics (*e.g.,* Accuracy, BLEU [370] and ROUGE [371]) and human ratings have been typically used for evaluating the performance. Due to the powerful language generation capabilities, LLMs have achieved remarkable performance on existing datasets and benchmarks, even surpassing human performance (on test datasets). For instance, given only 32 examples as the input, GPT-3 with in-context learning can outperform a full-data fine-tuned BERT-Large on the average score of SuperGLUE [312]; on MMLU, a 5-shot Chinchilla [34] nearly doubles the average accuracy of human raters, and GPT-4 [46] in 5-shot setting further achieves the state-of-the-art performance which yields more than 10% improvement in average accuracy compared to the previous best model. Thus, it raises serious concern about whether existing benchmarks for conditional text generation tasks can appropriately evaluate and reflect the capability of LLMs. Considering this issue, researchers try to make new evaluation benchmarks (*e.g.,* BIG-bench Hard [314]) by collecting currently unsolvable tasks (*i.e.,* the task on which LLMs fail to perform well) or creating more challenging tasks, *e.g.,* super-long text generation [372]. Moreover, recent studies also find that the automatic metrics may underestimate the generation quality of LLMs. In OpenDialKG [311], ChatGPT underperforms a fine-tuned GPT-2 on BLEU and ROUGE-L metrics, while earning more favor from human judgment [373]. Therefore, more efforts need to be devoted to developing new metrics that are more aligned with human judgment.

Code Synthesis. Besides generating high-quality natural language, existing LLMs also show strong abilities to generate formal language, especially computer programs (*i.e.,* code) that satisfy specific conditions, called *code synthesis* [374]. Unlike natural language generation, as the generated code can be directly checked by execution with corresponding compilers or interpreters, existing work mostly evaluates the quality of the generated code from LLMs by calculating the pass rate against the test cases, *i.e.,* pass@k[28]. Recently, several code benchmarks focusing on functional correctness are proposed to assess the code synthesis abilities of LLMs, such as APPS [316], HumanEval [89], and MBPP [152]. Typically, they consist of diverse programming problems, with text specification and test cases for correctness checking. To improve such an ability, it is key to fine-tuning (or pre-training) LLMs on code data, which can effectively adapt LLMs to code synthesis tasks [77]. Besides, existing work has proposed new strategies to generate code, *e.g.,* sampling multiple candidate solutions [152] and planning-guided decoding [375], which can be considered as the imitation of bug-fixing and code-planning processes by programmers. Impressively, LLMs have recently shown competitive performance with humans by achieving a ranking of the top 28% among users on the programming contest

---

28. Given k programs generated by the LLM, pass@k is computed as 1 when at least one program passes all test cases, or else 0

TABLE 7
Basic evaluation tasks and corresponding representative datasets of LLMs.

Task		Dataset
Language Generation	Language Modeling	Penn Treebank [298], WikiText-103 [299], the Pile [130], LAMBADA [167]
	Conditional Text Generation	WMT'14,16,19,20,21,22 [300–305], Flores-101 [306], DiaBLa [307], CNN/DailyMail [308], XSum [309], WikiLingua [310], OpenDialKG [311] SuperGLUE [312], MMLU [313], BIG-bench Hard [314], CLUE [315]
	Code Synthesis	APPS [316], HumanEval [89], MBPP [152], CodeContest [98], MTPB [77], DS-1000 [317], ODEX [318]
Knowledge Utilization	Closed-Book QA	Natural Questions [319], ARC [320], TruthfulQA [321], Web Questions [322], TriviaQA [323], PIQA [324], LC-quad2.0 [325], GrailQA [326], KQApro [327], CWQ [328], MKQA [329], ScienceQA [330]
	Open-Book QA	Natural Questions [319], OpenBookQA [331], ARC [320], Web Questions [322], TriviaQA [323], MS MARCO [332], QASC [333], SQuAD [334], WikiMovies [335]
	Knowledge Completion	WikiFact [336], FB15k-237 [337], Freebase [338], WN18RR [339], WordNet [340], LAMA [341], YAGO3-10 [342], YAGO [343]
Complex Reasoning	Knowledge Reasoning	CSQA [277], StrategyQA [278], ARC [320], BoolQ [344], PIQA [324], SIQA [345], HellaSwag [346], WinoGrande [347], OpenBookQA [331], COPA [348], ScienceQA [330], proScript [349], ProPara [350], ExplaGraphs [351], ProofWriter [352], EntailmentBank [353], ProOntoQA [354]
	Symbolic Reasoning	CoinFlip [33], ReverseList [33], LastLetter [33], Boolean Assignment [355], Parity [355], Colored Object [356], Penguins in a Table [356], Repeat Copy [357], Object Counting [357]
	Mathematical Reasoning	MATH [313], GSM8k [274], SVAMP [275], MultiArith [358], ASDiv [276], MathQA [359], AQUA-RAT [360], MAWPS [361], DROP [362], NaturalProofs [363], PISA [364], miniF2F [365], ProofNet [366]

platform Codeforces [98]. Further, GitHub Copilot has been released to assist programming in coding IDEs (*e.g.,* Visual Studio and JetBrains IDEs), which can support a variety of languages including Python, JavaScript, and Java. A viewpoint article entitled "*The End of Programming*" [376] in Communications of the ACM has discussed the impact of AI programming in the field of computer science, emphasizing an important shift towards the highly adaptive LLM as a new atomic unit of computation.

Major Issues. Although LLMs have achieved splendid performance in generating human-like text, they are susceptible to suffering from two major issues in language generation as discussed below.

• *Controllable generation.* For LLMs, the mainstream way to generate texts under given conditions is through the use of natural language instructions or prompts. Despite the simplicity, such a mechanism poses significant challenges in terms of exerting fine-grained or structural constraints over the generated outputs of these models. Existing work [41] shows that, when generating texts with complex constraints on their structures, LLMs can handle *local planning* (*e.g.,* interactions between proximal sentences) very well but might struggle with *global planning* (*i.e.,* long-range relatedness). For example, to generate a complex long passage with several paragraphs, it is still difficult to directly ensure specific text structure (*e.g.,* the order of concepts and the logical flow), considering the whole text. This case will become even more challenging for generation tasks that require formal rules or grammar, *e.g.,* code synthesis. To tackle this issue, a potential solution is to extend the one-pass generation into the iterative prompting of LLMs. This simulates the human writing process to break down language generation

into multiple steps such as planning, drafting, rewriting, and editting [372]. Several studies have proven that iterative prompting can elicit relevant knowledge to achieve better performance in sub-tasks [377, 378]. In essence, chain-of-thought prompting has utilized the idea of decomposing complex tasks into multi-step reasoning chains. Besides, the safety control of generated texts is also important for practical deployment. It has been shown that LLMs may generate texts that contain sensitive information or offensive expressions [46]. Although the RLHF algorithm [61] can alleviate this problem to some extent, it still relies on considerable human-labeled data for tuning LLMs, without an objective optimization goal to follow. Thus, it is imperative to explore effective methods to overcome these limitations and enable safer control over the outputs of LLMs.

• *Specialized generation.* Although LLMs have learned general language patterns to generate coherent text, their proficiency in generation might be constrained when dealing with a specialized domain or task. For instance, a language model that has been trained on general web articles may face challenges when generating a medical report which involves many medical jargon and methods. Intuitively, domain knowledge should be critical for model specialization. Whereas, it is not easy to inject such specialized knowledge into LLMs. As discussed in recent analyses [47, 379], when LLMs are trained to exhibit some specific ability that allows them to excel in some areas, they might struggle in others. Such an issue is related to *catastrophic forgetting* [380, 381] in training neural networks, which refers to the conflict phenomenon of integrating new and old knowledge. Similar cases also occur in human alignment of LLMs, where "*alignment tax*" [61] (*e.g.,* a potential loss in the in-context learning ability) has to be paid for aligning

to human values and needs. Therefore, it is important to develop effective model specialization methods that can flexibly adapt LLMs to various task scenarios, meanwhile retaining the original abilities as possible.

7.1.2 Knowledge Utilization

Knowledge utilization is an important ability of intelligent systems to accomplish knowledge-intensive tasks (*e.g.,* commonsense question answering and fact completion) based on supporting factual evidence. Concretely, it requires LLMs to properly utilize the rich factual knowledge from the pre-training corpus or retrieve external data when necessary. In particular, question answering (QA) and knowledge completion have been two commonly used tasks for evaluating this ability. According to the test tasks (question answering or knowledge completion) and evaluation settings (*with* or *without* external resources), we categorize existing knowledge utilization tasks into three types, namely closed-book QA, open-book QA[29], and knowledge completion.

Closed-Book QA. Closed-book QA tasks [382] test the acquired factual knowledge of LLMs from the pre-training corpus, where LLMs should answer the question only based on the given context without using external resources. For evaluating this ability, there are several datasets that can be leveraged, including Natural Questions [319], Web Questions [322], and TriviaQA [323], where the accuracy metric is widely adopted. Empirical results have revealed that LLMs can perform well in this setting and even match the performance of state-of-the-art open-domain QA systems [56]. Besides, the performance of LLMs on closed-book QA tasks also shows a scaling law pattern in terms of both model size and data size: scaling the parameters and training tokens can increase the capacity of LLMs and help them learn (or memorize) more knowledge from the pre-training data [56]. Further, under a similar parameter scale, LLMs with more pre-training data relevant to the evaluated tasks would achieve better performance [72]. Besides, the closed-book QA setting also provides a testbed for probing the accuracy of the factual knowledge encoded by LLMs. However, as shown in existing work [55], LLMs might perform less well on QA tasks relying on fine-grained knowledge, even when it exists in the pre-training data.

Open-Book QA. Unlike closed-book QA, in open-book QA tasks, LLMs can extract useful evidence from the external knowledge base or document collections, and then answer the question based on the extracted evidence [383–386]. Typical open-book QA datasets (*e.g.,* Natural Questions [319], OpenBookQA [331], and SQuAD [334]) have overlap with closed-book QA datasets, but they incorporate external data sources, *e.g.,* Wikipedia. The metrics of accuracy and F1 score are widely used in open-book QA tasks for evaluation. To select relevant knowledge from external resources, LLMs are often paired with a text retriever (or even a search engine), which is trained independently or jointly with

---

29. In this part, open-book QA refers to the QA tasks that require to extract and utilize useful information from external knowledge resources, as the antithesis of closed-book QA (only using the encoded information from pre-training corpus). Note that there is a dataset also named OpenBookQA [331], which follows the settings of open-book QA tasks by extracting and utilizing external science facts.

LLMs [72, 383, 387]. In evaluation, existing studies mainly focus on testing how LLMs utilize the extracted knowledge to answer the question and show that the retrieved evidence can largely improve the accuracy of the generated answers, even enabling a smaller LLM to outperform $10\times$ larger ones [383, 387]. Besides, open-book QA tasks can also evaluate the recency of knowledge information. Pretraining or retrieving from outdated knowledge resources may cause LLMs to generate incorrect answers for timesensitive questions [383].

Knowledge Completion. In knowledge completion tasks, LLMs might be (to some extent) considered as a knowledge base [341], which can be leveraged to complete or predict the missing parts of knowledge units (*e.g.,* knowledge triples). Such tasks can probe and evaluate *how much* and *what kind of* knowledge LLMs have learned from the pre-training data. Existing knowledge completion tasks can be roughly divided into knowledge graph completion tasks (*e.g.,* FB15k-237 [337] and WN18RR [339]) and fact completion tasks (*e.g.,* WikiFact [336]), which aim to complete the triples from a knowledge graph and incomplete sentences about specific facts, respectively. Empirical studies have revealed that it is difficult for existing LLMs to accomplish knowledge completion tasks related to specific relation types [287]. As shown in the evaluation results on WikiFact, LLMs perform well on several frequent relations that occur in the pre-training data (*e.g.,* `currency` and `author`), while not well on rare ones (*e.g.,* `discoverer_or_inventor` and `place_of_birth`). Interestingly, under the same evaluation settings (*e.g.,* in-context learning), InstructGPT (*i.e.,* `text-davinci-002`) outperforms GPT-3 in all subsets of WikiFact. It indicates that instruction tuning is helpful for LLMs to accomplish knowledge completion tasks.

Major Issues. Although LLMs have achieved key progress in capturing and utilizing knowledge information, they suffer from two major issues as discussed below.

• *Hallucination*. In generating factual texts, a challenging issue is *hallucination generations* [373], where the generated information is either in conflict with the existing source (*intrinsic hallucination*) or cannot be verified by the available source (*extrinsic hallucination*), which are illustrated with two examples in Figure 8. Hallucination widely occurs in existing LLMs, even the most superior LLMs such as GPT-4 [46]. In essence, LLMs seem to "unconsciously" utilize the knowledge in task solving, which still lack an ability to accurately control the use of internal or external knowledge. Hallucination would mislead LLMs to generate undesired outputs and mostly degrade the performance, leading to potential risks when deploying LLMs in real-world applications. To alleviate this problem, the alignment tuning strategies (as discussed in Section 5.2) have been widely utilized in existing works [61], which rely on tuning LLMs on high-quality data or using human feedback. For the evaluation of the hallucination problem, a set of hallucination detection tasks have been proposed, *e.g.,* TruthfulQA [321], for detecting human falsehood mimicked by models.

• *Knowledge recency*. As another major challenge, LLMs would encounter difficulties when solving tasks that require the latest knowledge beyond the training data. To tackle

Bob's wife is Amy. Bob's daughter is Cindy. Who is Cindy to Amy?

Cindy is Amy's daughter-in-law.

(a) Intrinsic hallucination

Explain RLHF for LLMs.

RLHF stands for "Rights, Limitations, Harms, and Freedoms" and is a framework for …… models like LLMs (Large Language Models).
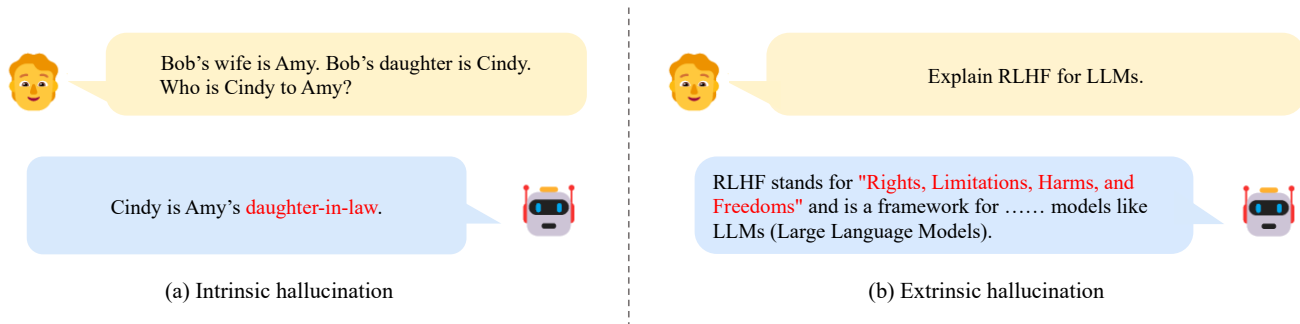
(b) Extrinsic hallucination

Fig. 8. Examples of intrinsic and extrinsic hallucination for a public LLM (access date: March 19, 2023). As an example of intrinsic hallucination, the LLM gives a conflicting judgment about the relationship between Cindy and Amy, which contradicts the input. For extrinsic hallucination, in this example, the LLM seems to have an incorrect understanding of the meaning of RLHF (reinforcement learning from human feedback), though it can correctly understand the meaning of LLMs (in this context).

this issue, a straightforward approach is to regularly update LLMs with new data. However, it is very costly to fine-tune LLMs, and also likely to cause the catastrophic forgetting issue when incrementally training LLMs. Therefore, it is necessary to develop efficient and effective approaches that can integrate new knowledge into existing LLMs, making them up-to-date. Existing studies have explored how to utilize the external knowledge source (*e.g.*, search engine) to complement LLMs, which can be either jointly optimized with LLMs [383] or used as a plug-and-play module [388]. For instance, ChatGPT utilizes a retrieval plugin to access up-to-date information sources [389]. By incorporating the extracted relevant information into the context [390, 391], LLMs can acquire new factual knowledge and perform better on relevant tasks. However, such an approach seems to be still at a superficial level. It has been revealed that it is difficult to directly amend intrinsic knowledge or inject specific knowledge into LLMs, which remains an open research problem [392, 393].

7.1.3 Complex Reasoning

Complex reasoning refers to the ability of understanding and utilizing supporting evidence or logic to derive conclusions or make decisions [51, 52]. According to the type of involved logic and evidence in the reasoning process, we consider dividing existing evaluation tasks into three major categories, namely knowledge reasoning, symbolic reasoning, and mathematical reasoning.

Knowledge Reasoning. The knowledge reasoning tasks rely on logical relations and evidence about factual knowledge to answer the given question. Existing work mainly uses specific datasets to evaluate the reasoning capacity of the corresponding type of knowledge, *e.g.*, CSQA [277]/StrategyQA [278] for commonsense knowledge reasoning and ScienceQA [330] for science knowledge reasoning. In addition to the accuracy of the predicted results, existing work [330] has also evaluated the quality of the generated reasoning process, via automatic metrics (*e.g.*, BLEU) or human evaluation. Typically, these tasks require LLMs to perform step-by-step reasoning based on factual knowledge, until reaching the answer to the given question. To elicit the step-by-step reasoning ability, chain-of-thought (CoT) prompting strategy [33] has been proposed for enhancing the complex reasoning capacity of LLMs.

As discussed in Section 6.2, CoT involves the intermediate reasoning steps, which can be manually created [33] or automatically generated [394], into the prompts to guide LLMs to perform multi-step reasoning. Such a way largely improves the reasoning performance of LLMs, leading to new state-of-the-art results on several complex knowledge reasoning tasks [33, 56, 395]. Further, after reformulating knowledge reasoning tasks into code generation tasks, researchers have found that the performance of LLMs can be further improved [156], especially with the LLMs pretrained on code. However, due to the complexity of knowledge reasoning tasks, the performance of current LLMs still lags behind human results on tasks such as commonsense reasoning [33, 56, 396]. As one of the most common mistakes, LLMs might generate inaccurate intermediate steps based on wrong factual knowledge, leading to a wrong final result. To address this issue, existing work has proposed special decoding or ensemble strategies to improve the accuracy of the whole reasoning chain [279, 282]. More recently, an empirical study [395] reveals that LLMs may have difficulty in explicitly inferring the commonsense knowledge required by a specific task, though they can successfully solve it. Further, it further shows that leveraging self-generated knowledge may not be beneficial for improving the reasoning performance.

Symbolic Reasoning[30]. The symbolic reasoning tasks mainly focus on manipulating the symbols in a formal rule setting to fulfill some specific goal [51], where the operations and rules may have never been seen by LLMs during pretraining. Existing work [33, 263, 281] commonly evaluates LLMs on the task of last letter concatenation and coin flip, where the evaluation examples require the same reasoning steps as the in-context examples (called *in-domain test*) or more steps (called *out-of-domain test*). For an example of the out-of-domain test, LLMs could only see the examples with two words in context, but it requires LLMs to concatenate the last letters of three or more words. Typically, the accuracy of the generated symbols is adopted to evaluate the performance of LLMs on these tasks. Thus, LLMs need to understand the semantic relations among the symbolic

---

30. Following [33], we mainly discuss symbolic reasoning tasks specially designed for evaluating LLMs. We do not consider symbolic reasoning methods in traditional NLP tasks, such as deducing logical rules from the knowledge graphs in KBQA.

operations and their composition in complex scenarios. However, under the out-of-domain setting, as LLMs have not seen the complex compositions of symbolic operations and rules (*e.g.,* twice the number of operations in context examples), it is hard for LLMs to capture their accurate meanings. To solve this issue, existing studies incorporate scratchpad [355, 397] and tutor [398] strategies to help LLMs better manipulate symbolic operations, for generating longer and more complex reasoning processes. Another line of research work utilizes the formal programming language to represent the symbolic operations and rules, which requires LLMs to generate code and perform the reasoning process by executing it with external interpreters. Such a way can decompose the complex reasoning process into code synthesis and program execution for LLMs and interpreters, respectively, leading to a simplified reasoning process with yet more accurate results [357].

Mathematical Reasoning. The mathematical reasoning tasks need to comprehensively utilize mathematical knowledge, logic, and computation for solving problems or generating proof statements. Existing mathematical reasoning tasks can be mainly categorized into math problem solving and automated theorem proving. For math problem solving tasks, SVAMP [275], GSM8k [274], and MATH [313] datasets are commonly used for evaluation, where LLMs need to generate accurate concrete numbers or equations to answer the mathematical problem. As these tasks also require multi-step reasoning, the chain-of-thought prompting strategy has been widely adopted for LLMs to improve the reasoning performance [33]. As a practical strategy, continually pre-training LLMs on large-scale mathematical corpora can largely boost their performance on mathematical reasoning tasks [35, 147, 399]. Further, since math problems in different languages share the same mathematical logic, researchers also propose a multilingual math word problem benchmark [290] to evaluate the multilingual mathematical reasoning capacity of LLMs. As another challenging task, automated theorem proving (ATP) [363, 365, 400] requires the reasoning model to strictly follow the reasoning logic and mathematical skills. To evaluate the performance on this task, PISA [364] and miniF2F [365] are two typical ATP datasets with the *proof success rate* as the evaluation metric. As a typical approach, existing work on ATP utilizes LLMs to aid the search for proofs using an interactive theorem prover (ITP), such as Lean, Metamath, and Isabelle [401–403]. A major limitation of ATP research is the lack of related corpora in formal language. To tackle it, several studies utilize LLMs to convert informal statements into formal proofs for augmenting new data [157] or generate drafts and proof sketches to reduce the search space of the proofs [404].

Major Issues. In spite of the advancements, LLMs still have several limitations in solving complex reasoning tasks.

• *Inconsistency*. With improved reasoning strategies (*e.g.,* CoT prompting), LLMs can solve some complex reasoning tasks, by performing step-by-step reasoning based on the supporting logic and evidence. Despite the effectiveness, the *inconsistency* issue often occurs in the decomposed reasoning process. Concretely, LLMs may generate the correct answer following an invalid reasoning path, or produce a wrong answer after a correct reasoning process [33, 405], leading to inconsistency between the derived answer and the reasoning process. To alleviate this problem, existing work has proposed to guide the whole generation process of LLMs via external tools or models [375], or re-check the reasoning process and final answer for correcting them [406]. As a promising solution, recent approaches reformulate the complex reasoning tasks into code generation tasks, where the strict execution of the generated code ensures the consistency between the reasoning process and the outcome. Besides, it has been revealed that there might also exist inconsistency between tasks with similar inputs, where small changes in the task description may cause the model to produce different results [49, 275]. To mitigate this problem, the ensemble of multiple reasoning paths can be applied to enhance the decoding process of LLMs [282].

• *Numerical computation*. For complex reasoning tasks, LLMs still face difficulties in the involved numerical computation, especially for the symbols that are seldom encountered during pre-training, such as arithmetic with large numbers [49, 398]. To tackle this issue, a direct way is to tune LLMs on synthesized arithmetic problems [407]. A surge of studies follow this approach and further improve the numerical computation performance by special training and inference strategies [397], *e.g.,* scratchpad tracing. Besides, existing work [71] has also incorporated external tools (*e.g.,* calculator), especially for handling arithmetic operations. More recently, ChatGPT has provided a plugin mechanism to use external tools [389]. In this way, LLMs need to learn how to properly manipulate the tools. For this purpose, researchers have augmented the examples using tools (even the LLM itself) for tuning the LLM [71, 408], or devised instructions and exemplars for in-context learning [357]. While, these LLMs still rely on the text context to capture the semantic meanings of mathematical symbols (during the pre-training stage), which is not best suited for numerical computation in essence.

7.2 Advanced Ability Evaluation

In addition to the above basic evaluation tasks, LLMs also exhibit some superior abilities that require special considerations for evaluation. In this part, we discuss several representative advanced abilities and the corresponding evaluation approaches, including human alignment, interaction with the external environment, and tool manipulation. Next, we discuss these advanced abilities in detail.

7.2.1 Human Alignment

It is desired that LLMs could well conform to human values and needs, *i.e.,* human alignment, which is a key ability for the broad use of LLMs in real-world applications.

To evaluate this ability, existing studies consider multiple criteria for human alignment, such as helpfulness, honesty, and safety [46, 223, 224]. For helpfulness and honesty, adversarial question answering tasks (*e.g.,* TruthfulQA [321]) can be utilized to examine LLM's ability in detecting possible falsehood in the text [46, 72]. Furthermore, harmlessness can be also evaluated by several existing benchmarks, *e.g.,* CrowS-Pairs [409] and Winogender [410]. Despite the automatic evaluation with the above datasets, human evaluation

is still a more direct way to effectively test the human alignment ability of LLMs. OpenAI invites many experts in domains related to AI risks to evaluate and improve the behaviors of GPT-4 when encountering risky contents [46]. Besides, for other aspects of human alignment (*e.g.,* truthfulness), several studies propose to use specific instructions and devise annotation rules to guide the annotation process [72]. Empirical studies have revealed that these strategies can greatly improve the human alignment ability of LLMs [224]. For instance, after alignment tuning on data collected through interactions with experts, the incorrect behavior rate of GPT-4 can be largely reduced when it deals with sensitive or disallowed prompts. In addition, high-quality pre-training data can reduce the effort required for alignment [46]. For instance, Galactica is potentially more harmless due to the less biased contents in the scientific corpus [35].

7.2.2 Interaction with External Environment

Besides standard evaluation tasks, LLMs have the ability to receive feedback from the external environment and perform actions according to the behavior instruction, *e.g.,* generating action plans in natural language to manipulate agents [411, 412]. Such an ability is also emergent in LLMs that can generate detailed and highly realistic action plans, while smaller models (*e.g.,* GPT-2) tend to generate shorter or meaningless plans [411].

To test this ability, several embodied AI benchmarks can be used for evaluation, described as follows. Virtual-Home [413] builds a 3D simulator for household tasks such as cleaning and cooking, in which the agent can execute natural language actions generated by LLMs. ALFRED [414] includes more challenging tasks that require LLMs to accomplish compositional targets. BEHAVIOR [415] focuses on everyday chores in simulation environments and requires LLMs to generate complex solutions, *e.g.,* changing the internal status of objects. Based on the generated action plans from LLMs, existing work either adopts the regular metrics (*e.g.,* executability and correctness of the generated action plans) [411] in the benchmark or directly conducts real-world experiments and measures the success rate [416], to evaluate such ability. Existing work has shown the effectiveness of LLMs in interacting with the external environment and generating accurate action plans [417]. Recently, several improved methods have been proposed to enhance the interaction ability of LLMs, *e.g.,* designing code-like prompts [418] and providing real-world grounding [416].

7.2.3 Tool Manipulation

When solving complex problems, LLMs can turn to external tools if they determine it is necessary. By encapsulating available tools with API calls, existing work has involved a variety of external tools, *e.g.,* search engine [72], calculator [71], and compiler [357], to enhance the performance of LLMs on several specific tasks. Recently, OpenAI has supported the use of plugins in ChatGPT [389], which can equip LLMs with broader capacities beyond language modeling. For example, the web browser plugin enables ChatGPT to access fresh information. Further, incorporating third-party plugins is particularly key for creating a prosperous ecosystem of applications based on LLMs.

To examine the ability of tool manipulation, existing work mostly adopts complex reasoning tasks for evaluation, such as mathematical problem solving (*e.g.,* GSM8k [274] and SVAMP [275]) or knowledge question answering (*e.g.,* TruthfulQA [321]), where the successful utilization of tools is very important for enhancing the required skills that LLMs are incapable of (*e.g.,* numerical calculation). In this way, the evaluated performance on these tasks can reflect the ability of LLMs in tool manipulation. To teach LLMs to utilize tools, existing studies add exemplars using tools in context to elicit LLMs [357], or fine-tune LLMs on simulated data about tool utilization [71, 408]. Existing work has found that with the help of tools, LLMs become more capable of handling the issues that they are not good at, *e.g.,* equation calculation and utilizing real-time information, and eventually improve the final performance [71].

Summary. The above three abilities are of great value to the practical performance of LLMs: conforming to human values and preferences (human alignment), acting properly in real-world scenarios (interaction with the external environment), and expanding the ability scope (tool manipulation). In addition to the above three advanced abilities, LLMs might also show other abilities that are specially related to some tasks (*e.g.,* data annotation [258]) or learning mechanisms (*e.g.,* self-improvement [286]). It will be an open direction to discover, measure and evaluate these newly emerging abilities, so as to better utilize and improve LLMs.

7.3 Public Benchmarks and Empirical Analysis

In the aforementioned parts, we have discussed the evaluation tasks of LLMs and their corresponding settings. Next, we will introduce existing evaluation benchmarks and empirical analyses for LLMs, which focus on exploring more comprehensive discussions from a general perspective.

7.3.1 Evaluation Benchmarks

Recently, several comprehensive benchmarks [287, 313, 356] have been released for the evaluation of LLMs. In this part, we introduce several representative and widely used benchmarks, *i.e.,* MMLU, BIG-bench, and HELM.

• *MMLU* [313] is a versatile benchmark for large-scale evaluation of multi-task knowledge understanding, covering a wide range of knowledge domains from mathematics and computer science to humanities and social sciences. The difficulties of these tasks vary from basic to advanced. As shown in existing work, LLMs mostly outperform small models by a substantial margin on this benchmark [35, 56, 57, 64], which shows the scaling law in model size. More recently, GPT-4 achieves a remarkable record (86.4% in 5-shot setting) in MMLU, which is significantly better than the previous state-of-the-art models [46].

• *BIG-bench* [356] is a collaborative benchmark intended to probe existing LLMs from various aspects. It comprises 204 tasks that encompass a broad range of topics, including linguistics, childhood development, mathematics, commonsense reasoning, biology, physics, social bias, software development, and so on. By scaling the model size, LLMs can even outperform the average human performance under the few-shot setting on 65% of tasks in BIG-bench [56]. Considering the high evaluation cost of the entire benchmark, a lightweight benchmark BIG-bench-Lite has been

proposed, which contains 24 small yet diverse and challenging tasks from BIG-bench. Additionally, the BIG-bench hard (BBH) benchmark has been proposed to concentrate on investigating the currently unsolvable tasks of LLMs by selecting the challenging tasks in which LLMs exhibit inferior performance compared to humans. Since BBH becomes more difficult, small models mostly achieve performance close to random. As a comparison, CoT prompting can elicit the abilities of LLMs to perform step-by-step reasoning for enhancing the performance, even exceeding the average human performance in BBH [314].

• *HELM* [287] is a comprehensive benchmark that currently implements a core set of 16 scenarios and 7 categories of metrics. It is built on top of many prior studies, conducting a holistic evaluation of language models. As shown in the experimental results of HELM [287], instruction tuning can consistently boost the performance of LLMs in terms of accuracy, robustness, and fairness. Further, for reasoning tasks, the LLMs that have been pre-trained on code corpus show superior performance.

The above benchmarks cover a variety of mainstream evaluation tasks for the evaluation of LLMs. Besides, there are also several benchmarks that focus on evaluating specific abilities of LLMs, such as TyDiQA [419] for multilingual knowledge utilization and MGSM [290] for multilingual mathematical reasoning. To conduct the evaluation, one can select suitable benchmarks according to specific goals. In addition, there are also several open-source evaluation frameworks for researchers to evaluate LLMs on existing benchmarks or extend new tasks for customized evaluations, such as Language Model Evaluation Harness [420] and OpenAI Evals [46].

7.3.2 Comprehensive Analyses on LLMs' Capacities

In addition to constructing large-scale evaluation benchmarks, a surge of studies have conducted comprehensive analyses to investigate the strengths and limitations of LLMs. In this part, we briefly discuss them in major aspects, namely *generalist* (general-purpose capacity) and *specialist* (domain-specific capacity).

Generalist. Due to the remarkable performance, existing work [41, 46, 373, 379, 421–423] has systematically evaluated the general capacities of LLMs, to explore their competences in a variety of different tasks or applications. Typically, these studies mainly focus on the newly emerged LLMs (*e.g.,* ChatGPT and GPT-4) that have not been well investigated before, which are discussed as follows:

• *Mastery*. To evaluate the mastery level of LLMs in solving general tasks, existing work [423] typically collects a set of datasets covering a range of tasks and domains, and then tests LLMs under the few/zero-shot setting. Empirical results [41, 46, 379, 423] have shown the superior capacities of LLMs as a general-purpose task solver. As a remarkable progress, GPT-4 has surpassed the state-of-the-art methods with benchmark-specific training in a wide range of tasks, such as language understanding, commonsense reasoning, and mathematical reasoning [46]. Furthermore, it can achieve human-like performance in real-world exams designed for humans (*e.g.,* Advanced Placement exams and Graduate Record Examination [46]). More recently, a

comprehensive qualitative analysis [41] has revealed that GPT-4 approaches human-level performance in a variety of challenging tasks across various fields (*e.g.,* mathematics, computer vision, and programming), and considered it as "*an early version of an artificial general intelligence system*". Despite the promising results, this analysis has also revealed that GPT-4 still has severe limitations. For example, GPT-4 is hard to calibrate its confidence about the generated result, and can not verify its consistency with the training data and itself. Besides, it demonstrates inferior performance on tasks that require planning (*e.g.,* solving the "Tower of Hanoi" problem) or conceptual leaps (*e.g.,* proposing a new scientific hypothesis). Furthermore, several studies have also shown that LLMs may misunderstand unfamiliar concepts [423, 424] on information extraction tasks from specific domains, and face challenges in solving pragmatic emotion-related tasks [422] (*e.g.,* personalized emotion recognition), showing inferior performance compared to specific fine-tuned models.

• *Robustness*. Besides the mastery, another aspect to consider is the stability of LLMs against noises or perturbations, which is particularly important for practical applications. To evaluate the robustness of LLMs against noises or perturbations, existing work [425] conducts adversarial attack (*e.g.,* token replacement) on the input, and then evaluates the robustness of LLMs based on the change of output results. It has been shown that LLMs are more robust than small language models in a variety of tasks, but may encounter new issues about robustness, *e.g.,*robustness instability and prompt sensitivity. Concretely, LLMs are prone to provide different answers when using varied expressions of the same input, even in conflict with the content generated by itself [426]. Such an issue would also lead to unstable results when evaluating the robustness using different prompts, making the evaluation results of robustness analysis themselves less reliable.

Specialist. As LLMs have been pre-trained on large-scale mixture-of-source corpora, they can capture rich knowledge from the pre-training data. Thus, LLMs are also employed as domain experts or specialists for specific areas. Therefore, recent studies have widely explored the use of LLMs for solving domain-specific tasks and evaluated the adaptation capacity of LLMs. Typically, these studies collect or construct domain-specific datasets to evaluate the performance of LLMs using in-context learning. Since our focus is not to cover all the possible application domains, we briefly discuss three representative domains receiving considerable attention from the research community, namely healthcare, education, and law.

• *Healthcare* is a vital application field closely related to human life. Since the advent of ChatGPT, a series of studies have applied ChatGPT or other LLMs to the medical domain. It has been shown that LLMs are capable of handling a variety of healthcare tasks, *e.g.,* biology information extraction [427], medical advice consultation [428–430], and report simplification [431], and can even pass the medical license exams [432–434] specially designed for professional doctors. However, LLMs may fabricate medical misinformation [429, 431], *e.g.,* misinterpreting medical terms and suggesting advice inconsistent with medical guidelines. Be-

sides, it would also raise privacy concerns to upload the health information of patients [427].

- *Education* is also an important application domain where LLMs potentially exert significant influence. Existing work has found that LLMs can achieve student-level performance on standardized tests [46, 435, 436] in the subjects of mathematics, physics, computer science and so on, in both multiple-choice and free-response problems. Besides, empirical studies have shown that LLMs can serve as writing or reading assistant for education [437, 438]. A recent study [438] reveals that ChatGPT is capable of generating logically consistent answers across disciplines, balancing both depth and breadth. Another quantitative analysis [437] shows that students utilizing ChatGPT perform better than average students with different usage methods (*e.g.,* keeping or refining the results from LLMs as their own answers) in some courses from the computer security field. However, the increasing popularity of LLMs has been raising concerns (*e.g.,* cheating on homework) on the rational use of such intelligent assistants for education.

- *Law* is a specialized domain that is built on professional domain knowledge. Recently, a number of studies have applied LLMs to solve various legal tasks, *e.g.,* legal document analysis [439, 440], legal judgment prediction [441], and legal document writing [442]. A recent study [443] has found that LLMs own powerful abilities of legal interpretation and reasoning. Moreover, the latest GPT-4 model achieves a top 10% score in a simulated bar exam compared with human test-takers. However, the use of LLMs in law also raises concerns about legal challenges, including copyright issues [444], personal information leakage [445], or bias and discrimination [446].

Besides the aforementioned work, the capacities of LLMs have been also analyzed from other perspectives. For instance, some recent work has studied the human-like characteristics of LLMs, such as self-awareness, theory of mind (ToM), and affective computing [41, 447–449]. In particular, an empirical evaluation of ToM conducted on two classic false-belief tasks speculates that LLMs may have ToM-like abilities since the model in the GPT-3.5 series achieves comparable performance with nine-year-old children in ToM task [448]. Further, another line of work has investigated the fairness and accuracy of existing evaluation settings about LLMs [450], *e.g.,* the large-scale mixture-of-source pre-training data may contain the data in test sets.

8 CONCLUSION AND FUTURE DIRECTIONS

In this survey, we have reviewed the recent progress of large language models (LLMs), and introduced the key concepts, findings, and techniques for understanding and utilizing LLMs. We focus on the large-sized models (*i.e.,* having a size larger than 10B) while excluding the contents of early pre-trained language models (*e.g.,* BERT and GPT-2) that have been well covered in the existing literature. In particular, our survey has discussed four important aspects of LLMs, *i.e.,* pre-training, adaptation tuning, utilization, and evaluation. For each aspect, we highlight the techniques or findings that are key to the success of LLMs. Besides, we also summarize the available resources for developing LLMs and discuss important implementation guidelines for reproducing LLMs.

This survey tries to cover the most recent literature about LLMs and provides a good reference resource on this topic for both researchers and engineers.

Next, we summarize the discussions of this survey, and introduce the challenges and future directions for LLMs, in the following aspects.

Theory and Principle. To understand the underlying working mechanism of LLMs, one of the greatest mysteries is how information is distributed, organized, and utilized through the very large, deep neural network. It is important to reveal the basic principles or elements that establish the foundation of the abilities of LLMs. In particular, *scaling* seems to play an important role in increasing the capacity of LLMs [31, 55, 59]. It has been shown that some emergent abilities would occur in an unexpected way (a sudden performance leap) when the parameter scale of language models increases to a critical size (*e.g.,* 10B) [31, 33], typically including in-context learning, instruction following, and step-by-step reasoning. These emergent abilities are fascinating yet perplexing: *when* and *how* they are obtained by LLMs are not yet clear. Recent studies either conduct extensive experiments for investigating the effect of emergent abilities and the contributing factors to such abilities [250, 267, 451], or explain some specific abilities with existing theoretical frameworks [60, 261]. An insightful technical post also specially discusses this topic [47], taking the GPT-series models as the target. While, more formal theories and principles to understand, characterize, and explain the abilities or behaviors of LLMs are still missing. Since emergent abilities bear a close analogy to phase transitions in nature [31, 58], cross-discipline theories or principles (*e.g.,* whether LLMs can be considered as some kind of complex systems) might be useful to explain and understand the behaviors of LLMs. These fundamental questions are worth exploring for the research community, which are important for developing the next-generation LLMs.

Model Architecture. Due to the scalability and effectiveness, Transformer, consisting of stacked multi-head self-attention layers, has become the de facto architecture for building LLMs. Various strategies have been proposed to improve the performance of this architecture, such as neural network configuration and scalable parallel training (see discussions in Section 4.2.2). To enhance the model capacity (*e.g.,* the multi-turn conversation ability), existing LLMs typically maintain a long context window, *e.g.,* GPT-4-32k has an extremely large context length of 32,768 tokens. Thus, a practical consideration is to reduce the time complexity (originally to be quadratic costs) incurred by the standard self-attention mechanism. It is important to investigate the effect of more efficient Transformer variants in building LLMs [452], *e.g.,* sparse attention has been used in GPT-3 [55]. Besides, catastrophic forgetting has been a long-standing challenge for neural networks, which also has a negative impact on LLMs. When tuning LLMs with new data, the originally learned knowledge is likely to be damaged, *e.g.,* fine-tuning a LLM according to some specific tasks will affect the general ability of LLMs. A similar case occurs when LLMs are aligned with human values (called *alignment tax* [61, 223]). Thus, it is necessary to consider

extending existing architectures with more flexible mechanisms or modules that can effectively support data update and task specialization.

Model Training. In practice, it is very difficult to pre-train capable LLMs, due to the huge computation consumption and the sensitivity to data quality and training tricks [69, 83]. Thus, it becomes particularly important to develop more systemic, economical pre-training approaches for optimizing LLMs, considering the factors of model effectiveness, efficiency optimization, and training stability. More model checking or performance diagnosis methods (*e.g.,* predictable scaling in GPT-4 [46]) should be developed in order to detect early abnormal issues during training. Furthermore, it also calls for more flexible mechanisms of hardware support or resource schedule, so as to better organize and utilize the resources in a computing cluster. Since it is very costly to pre-train a LLM from scratch, it is important to design a suitable mechanisms for continually pre-training or fine-tuning the LLM based on publicly available model checkpoints (*e.g.,* LLaMA [57] and Flan-T5 [64]). For this purpose, a number of technical issues have to be resolved, *e.g.,* catastrophic forgetting and task specialization. However, to date, there still lack open-source model checkpoints for LLMs with complete pre-processing and training logs (*e.g.,* the scripts to prepare the pre-training data) for reproduction. We believe that it will be of great value to report more technical details in open-source models for the research of LLMs. Besides, it is also important to develop more improvement tuning strategies that effectively elicits the model abilities.

Model Utilization. Since fine-tuning is very costly in real applications, *prompting* has become the prominent approach to using LLMs. By combining task descriptions and demonstration examples into prompts, in-context learning (a special form of prompting) endows LLMs with the ability to perform well on new tasks, even outperforming full-data fine-tuned models in some cases. Furthermore, to enhance the ability of complex reasoning, advanced prompting techniques have been proposed, exemplified by the chain-of-thought (CoT) strategy, which includes the intermediate reasoning steps into prompts. However, existing prompting approaches still have several deficiencies described as follows. Firstly, it involves considerable human efforts in the design of prompts. It would be quite useful to automatically generate effective prompts for solving various tasks. Secondly, some complex tasks (*e.g.,* formal proof and numerical computation) require specific knowledge or logic rules, which may not be well expressed in natural language or demonstrated by examples. Thus, it is important to develop more informative, flexible task formatting methods for prompts[31]. Thirdly, existing prompting strategies mainly focus on single-turn performance. It is useful to develop interactive prompting mechanisms (*e.g.,* through natural language conversations) for solving complex tasks, which have been demonstrated to be very useful by ChatGPT.

Safety and Alignment. Despite their capacities, LLMs pose similar safety challenges as small language models. For example, LLMs exhibit a tendency to generate hallucinations [373], which are texts that seem plausible but may be factually incorrect. What is worse, LLMs might be elicited by intentional instructions to produce harmful, biased, or toxic texts for malicious systems, leading to the potential risks of misuse [55, 61]. To have a detailed discussion of the safety issues of LLMs (*e.g.,* privacy, overreliance, disinformation, and influence operations), the readers can refer to the GPT-3/4 technical reports [46, 55]. As the major approach to averting these issues, reinforcement learning from human feedback (RLHF) [61, 100] has been widely used by incorporating humans in the training loop for developing well-aligned LLMs. To improve the model safety, it is also important to include safety-relevant prompts during RLHF, as shown by GPT-4 [46]. However, RLHF heavily relies on high-quality human feedback data from professional labelers, making it difficult to be properly implemented in practice. Therefore, it is necessary to improve the RLHF framework for reducing the efforts of human labelers and seek a more efficient annotation approach with guaranteed data quality, *e.g.,* LLMs can be employed to assist the labeling work. More recently, red teaming [115, 225] has been adopted for improving the model safety of LLMs, which utilizes the collected adversarial prompts to refine the LLMs (*i.e.,* avoiding the attacks from red teaming). Furthermore, it is also meaningful to establish the proper learning mechanism for LLMs to obtain human feedback via chatting and directly utilize it for self-improvement.

Application and Ecosystem. As LLMs have shown a strong capacity in solving various tasks, they can be applied in a broad range of real-world applications (*i.e.,* following task-specific natural language instructions). As a remarkable progress, ChatGPT has potentially changed the way how humans access information, which has been implemented in the release of *New Bing*. In the near future, it can be foreseen that LLMs would have a significant impact on information-seeking techniques, including both search engines and recommender systems. Furthermore, the development and use of intelligent information assistants would be highly promoted with the technology upgrade from LLMs. In a broader scope, this wave of technical innovation would lead to an ecosystem of LLM-empowered applications (*e.g.,* the support of plugins by ChatGPT), which has a close connection with human life. Lastly, the rise of LLMs sheds light on the exploration of artificial general intelligence (AGI). It is promising to develop more smart intelligent systems (possibly with multi-modality signals) than ever. However, in this development process, AI safety should be one of the primary concerns, *i.e.,* making AI lead to good for humanity but not bad [40].

CODA: This survey was planned during a discussion meeting held by our research team, and we aimed to summarize the recent advances of large language models as a highly readable report for our team members. The first draft was finished on March 13, 2023, in which our team members tried their best to include the related studies about LLMs in a relatively objective, comprehensive way. Then, we have extensively revised the writing and contents in several passes. Despite all our efforts, this survey is still

---

31. While, it seems that an alternative approach to this issue is to invoke external tools, *e.g.,* the plugins for ChatGPT, when the task is difficult to solve via text generation.

far from perfect: we are likely to miss important references or topics, and might also have non-rigorous expressions or discussions. Due to the space limit, we can only include a fraction of existing LLMs in Figure 1 and Table 1, by setting the selection criterion. However, we set a more relaxed criterion for model selection on our GitHub page (https://github.com/RUCAIBox/LLMSurvey), which will be regularly maintained. We will continuously update this survey, and improve the quality as much as we can. For us, survey writing is also a learning process for LLMs by ourselves. For readers with constructive suggestions to improve this survey, you are welcome to leave comments on the GitHub page of our survey or directly email our authors. We will make revisions following the received comments or suggestions in a future version, and acknowledge the readers who have contributed constructive suggestions in our survey.

Update log. In this part, we regularly maintain a update log for the submissions of this survey to arXiv:

- First release on March 31, 2023: the initial version.
- Update on April 9, 2023: add the affiliation information, revise Figure 1 and Table 1 and clarify the corresponding selection criterion for LLMs, improve the writing, and correct some minor errors.
- Update on April 11, 2023: correct the errors for library resources.
- Update on April 12, 2023: revise Figure 1 and Table 1, and clarify the release date of LLMs.
- Update on April 16, 2023: add a new Section 2.2 about the technical evolution of GPT-series models.
- Update on April 24, 2023: add the discussion about scaling laws and add some explanations about the model sizes for emergent abilities (Section 2.1); add an illustrative figure for the attention patterns for different architectures in Figure 4, and add the detailed formulas in Table 4.
- Update on April 25, 2023: revise some copy errors in figures and tables.
- Update on April 27, 2023: add efficient tuning in Section 5.3.
- Update on April 28, 2023: revise Section 5.3.
- Update on May 7, 2023: revise Table 1, Table 2, and some minor points.

Planning content. We will regularly include new content into this survey, to make it more self-contained and up-to-date. Here, we list several potential topics that might appear in the next major version(s): (1) the technical evolution from GPT-1 to ChatGPT (partially done), (2) LLaMA based tuning (*e.g.,* Alpaca), (3) lightweight tuning strategies (done), and (4) detailed formulations for model details (done). If you have a specific topic suggested for this survey, please drop us a message about it.

REFERENCES

[1] S. Pinker, *The Language Instinct: How the Mind Creates Language*. Brilliance Audio; Unabridged edition, 2014.

[2] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: what is it, who has it, and how did it evolve?" *science*, vol. 298, no. 5598, pp. 1569–1579, 2002.

[3] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[4] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[5] J. Gao and C. Lin, "Introduction to the special issue on statistical language modeling," *ACM Trans. Asian Lang. Inf. Process.*, vol. 3, no. 2, pp. 87–93, 2004.

[6] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[7] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[8] X. Liu and W. B. Croft, "Statistical language modeling for information retrieval," *Annu. Rev. Inf. Sci. Technol.*, vol. 39, no. 1, pp. 1–31, 2005.

[9] C. Zhai, *Statistical Language Models for Information Retrieval*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2008.

[10] S. M. Thede and M. P. Harper, "A second-order hidden markov model for part-of-speech tagging," in *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*, R. Dale and K. W. Church, Eds. ACL, 1999, pp. 175–182.

[11] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "A tree-based statistical language model for natural language speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 1001–1008, 1989.

[12] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, J. Eisner, Ed. ACL, 2007, pp. 858–867.

[13] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 3, pp. 400–401, 1987.

[14] W. A. Gale and G. Sampson, "Good-turing frequency estimation without tears," *J. Quant. Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[15] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.

[16] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based lan-

guage model," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 1045–1048.

[17] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 2877–2880.

[18] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013.

[21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 2227–2237.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 7871–7880.

[25] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res*, pp. 1–40, 2021.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, p. 9, 2019.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[28] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[29] T. Wang, A. Roberts, D. Hesslow, T. L. Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel, "What language model architecture and pretraining objective works best for zero-shot generalization?" in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, vol. 162, 2022, pp. 22 964–22 984.

[30] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *CoRR*, vol. abs/2001.08361, 2020.

[31] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *CoRR*, vol. abs/2206.07682, 2022.

[32] M. Shanahan, "Talking about large language models," *CoRR*, vol. abs/2212.03551, 2022.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *CoRR*, vol. abs/2201.11903, 2022.

[34] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," vol. abs/2203.15556, 2022.

[35] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *CoRR*, vol. abs/2211.09085, 2022.

[36] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A system-

atic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, pp. 195:1–195:35, 2023.

[37] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to chatgpt," *CoRR*, vol. abs/2302.09419, 2023.

[38] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J. Wen, J. Yuan, W. X. Zhao, and J. Zhu, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[39] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *CoRR*, vol. abs/2003.08271, 2020.

[40] S. Altman, "Planning for agi and beyond," *OpenAI Blog*, February 2023.

[41] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," vol. abs/2303.12712, 2023.

[42] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, "Language is not all you need: Aligning perception with language models," *CoRR*, vol. abs/2302.14045, 2023.

[43] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[44] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[45] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[46] OpenAI, "Gpt-4 technical report," *OpenAI*, 2023.

[47] Y. Fu, H. Peng, and T. Khot, "How does gpt obtain its ability? tracing emergent abilities of language models to their sources," *Yao Fu's Notion*, Dec 2022.

[48] J. Li, T. Tang, W. X. Zhao, and J. Wen, "Pretrained language model for text generation: A survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, Ed. ijcai.org, 2021, pp. 4492–4499.

[49] P. Lu, L. Qiu, W. Yu, S. Welleck, and K. Chang, "A survey of deep learning for mathematical reasoning," *CoRR*, vol. abs/2212.10535, 2022.

[50] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, and Z. Sui, "A survey for in-context learning," *CoRR*, vol. abs/2301.00234, 2023.

[51] J. Huang and K. C. Chang, "Towards reasoning in large language models: A survey," *CoRR*, vol. abs/2212.10403, 2022.

[52] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey," *CoRR*, vol. abs/2212.09597, 2022.

[53] J. Zhou, P. Ke, X. Qiu, M. Huang, and J. Zhang, "Chatgpt: potential, prospects, and limitations," in *Frontiers of Information Technology & Electronic Engineering*, 2023, pp. 1–6.

[54] W. X. Zhao, J. Liu, R. Ren, and J. Wen, "Dense text retrieval based on pretrained language models: A survey," *CoRR*, vol. abs/2211.14876, 2022.

[55] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[56] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *CoRR*, vol. abs/2204.02311, 2022.

[57] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *CoRR*, 2023.

[58] B. A. Huberman and T. Hogg, "Phase transitions in artificial intelligence systems," *Artificial Intelligence*, vol. 33, no. 2, pp. 155–171, 1987.

[59] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, H. F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. M. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J. Lespiau,

M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. J. Johnson, B. A. Hechtman, L. Weidinger, I. Gabriel, W. S. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving, "Scaling language models: Methods, analysis & insights from training gopher," *CoRR*, vol. abs/2112.11446, 2021.

[60] D. Dai, Y. Sun, L. Dong, Y. Hao, Z. Sui, and F. Wei, "Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers," *CoRR*, vol. abs/2212.10559, 2022.

[61] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *CoRR*, vol. abs/2203.02155, 2022.

[62] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* OpenReview.net, 2022.

[63] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. S. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. H. Chi, and Q. Le, "Lamda: Language models for dialog applications," *CoRR*, vol. abs/2201.08239, 2022.

[64] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *CoRR*, vol. abs/2210.11416, 2022.

[65] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *KDD*, 2020, pp. 3505–3506.

[66] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *CoRR*, vol. abs/1909.08053, 2019.

[67] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia, "Efficient large-scale language model training on GPU clusters using megatron-lm," in *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2021, St. Louis, Missouri, USA, November 14-19, 2021.* ACM, 2021, p. 58.

[68] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro, "Reducing activation recomputation in large transformer models," *CoRR*, vol. abs/2205.05198, 2022.

[69] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, and et al., "BLOOM: A 176b-parameter open-access multilingual language model," *CoRR*, vol. abs/2211.05100, 2022.

[70] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4299–4307.

[71] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *CoRR*, vol. abs/2302.04761, 2023.

[72] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, "Webgpt: Browser-assisted question-answering with human feedback," *CoRR*, vol. abs/2112.09332, 2021.

[73] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, pp. 140:1–140:67, 2020.

[74] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 2021, pp. 483–498.

[75] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, C. Li, Z. Gong, Y. Yao, X. Huang, J. Wang, J. Yu, Q. Guo, Y. Yu, Y. Zhang, J. Wang, H. Tao, D. Yan, Z. Yi,

F. Peng, F. Jiang, H. Zhang, L. Deng, Y. Zhang, Z. Lin, C. Zhang, S. Zhang, M. Guo, S. Gu, G. Fan, Y. Wang, X. Jin, Q. Liu, and Y. Tian, "Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation," *CoRR*, vol. abs/2104.12369, 2021.

[76] Z. Zhang, Y. Gu, X. Han, S. Chen, C. Xiao, Z. Sun, Y. Yao, F. Qi, J. Guan, P. Ke, Y. Cai, G. Zeng, Z. Tan, Z. Liu, M. Huang, W. Han, Y. Liu, X. Zhu, and M. Sun, "CPM-2: large-scale cost-effective pre-trained language models," *CoRR*, vol. abs/2106.10715, 2021.

[77] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, "Codegen: An open large language model for code with mtulti-turn program synthesis," *arXiv preprint arXiv:2203.13474*, 2022.

[78] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, "Gpt-neox-20b: An open-source autoregressive language model," *CoRR*, vol. abs/2204.06745, 2022.

[79] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. R. A, S. Patro, T. Dixit, and X. Shen, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 5085–5109.

[80] Y. Tay, M. Dehghani, V. Q. Tran, X. García, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, H. Zheng, D. Zhou, N. Houlsby, and D. Metzler, "Ul2: Unifying language learning paradigms," 2022.

[81] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: open pre-trained transformer language models," *CoRR*, vol. abs/2205.01068, 2022.

[82] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," *CoRR*, vol. abs/2207.04672, 2022.

[83] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, and J. Tang, "GLM-130B: an open bilingual pre-trained model," vol. abs/2210.02414, 2022.

[84] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel, "Crosslingual generalization through multitask finetuning," *CoRR*, vol. abs/2211.01786, 2022.

[85] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, X. Li, B. O'Horo, G. Pereyra, J. Wang, C. Dewan, A. Celikyilmaz, L. Zettlemoyer, and V. Stoyanov, "OPT-IML: scaling language model instruction meta learning through the lens of generalization," *CoRR*, vol. abs/2212.12017, 2022.

[86] Q. Zheng, X. Xia, X. Zou, Y. Dong, S. Wang, Y. Xue, Z. Wang, L. Shen, A. Wang, Y. Li *et al.*, "Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x," *arXiv preprint arXiv:2303.17568*, 2023.

[87] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," *arXiv preprint arXiv:2304.01373*, 2023.

[88] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[89] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, "Evaluating large language models trained on code," *CoRR*, vol. abs/2107.03374, 2021.

[90] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang, "ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *CoRR*, vol. abs/2107.02137, 2021.

[91] O. Lieber, O. Sharir, B. Lenz, and Y. Shoham, "Jurassic-1: Technical details and evaluation," *White Paper. AI21 Labs*, vol. 1, 2021.

[92] B. Kim, H. Kim, S. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo, H. Lee, M. Jeong, S. Lee, M. Kim, S. Ko, S. Kim, T. Park, J. Kim, S. Kang, N. Ryu, K. M. Yoo, M. Chang, S. Suh, S. In, J. Park, K. Kim, H. Kim, J. Jeong, Y. G. Yeo, D. Ham, D. Park,

M. Y. Lee, J. Kang, I. Kang, J. Ha, W. Park, and N. Sung, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021.* Association for Computational Linguistics, 2021.

[93] S. Wu, X. Zhao, T. Yu, R. Zhang, C. Shen, H. Liu, F. Li, H. Zhu, J. Luo, L. Xu *et al.*, "Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning," *arXiv preprint arXiv:2110.04725*, 2021.

[94] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. Das-Sarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, "A general language assistant as a laboratory for alignment," *CoRR*, vol. abs/2112.00861, 2021.

[95] S. Wang, Y. Sun, Y. Xiang, Z. Wu, S. Ding, W. Gong, S. Feng, J. Shang, Y. Zhao, C. Pang, J. Liu, X. Chen, Y. Lu, W. Liu, X. Wang, Y. Bai, Q. Chen, L. Zhao, S. Li, P. Sun, D. Yu, Y. Ma, H. Tian, H. Wu, T. Wu, W. Zeng, G. Li, W. Gao, and H. Wang, "ERNIE 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation," *CoRR*, vol. abs/2112.12731, 2021.

[96] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. P. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. S. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, "Glam: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022, pp. 5547–5569.

[97] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zheng, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro, "Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model," *CoRR*, vol. abs/2201.11990, 2022.

[98] Y. Li, D. H. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. D. Lago, T. Hubert, P. Choy, C. de Masson d'Autume, I. Babuschkin, X. Chen, P. Huang, J. Welbl, S. Gowal, A. Cherepanov, J. Molloy, D. J. Mankowitz, E. S. Robson, P. Kohli, N. de Freitas, K. Kavukcuoglu, and O. Vinyals, "Competition-level code generation with alphacode," *Science*, 2022.

[99] S. Soltan, S. Ananthakrishnan, J. FitzGerald, R. Gupta, W. Hamza, H. Khan, C. Peris, S. Rawls, A. Rosenbaum, A. Rumshisky, C. S. Prakash, M. Sridhar, F. Triefenbach, A. Verma, G. Tür, and P. Natarajan, "Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model," *CoRR*, vol. abs/2208.01448, 2022.

[100] A. Glaese, N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving, "Improving alignment of dialogue agents via targeted human judgements," *CoRR*, vol. abs/2209.14375, 2022.

[101] H. Su, X. Zhou, H. Yu, Y. Chen, Z. Zhu, Y. Yu, and J. Zhou, "Welm: A well-read pre-trained language model for chinese," *CoRR*, vol. abs/2209.10372, 2022.

[102] Y. Tay, J. Wei, H. W. Chung, V. Q. Tran, D. R. So, S. Shakeri, X. Garcia, H. S. Zheng, J. Rao, A. Chowdhery, D. Zhou, D. Metzler, S. Petrov, N. Houlsby, Q. V. Le, and M. Dehghani, "Transcending scaling laws with 0.1% extra compute," *CoRR*, vol. abs/2210.11399, 2022.

[103] X. Ren, P. Zhou, X. Meng, X. Huang, Y. Wang, W. Wang, P. Li, X. Zhang, A. Podolskiy, G. Arshinov, A. Bout, I. Piontkovskaya, J. Wei, X. Jiang, T. Su, Q. Liu, and J. Yao, "Pangu-Σ: Towards trillion parameter language model with sparse heterogeneous computing," *CoRR*, vol. abs/2303.10845, 2023.

[104] A. Radford, R. Józefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *CoRR*, vol. abs/1704.01444, 2017.

[105] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[106] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *CoRR*, vol. abs/1806.08730, 2018.

[107] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, A. Celikyilmaz and T. Wen, Eds. Association for Computational Linguistics, 2020, pp. 270–278.

[108] D. Ham, J. Lee, Y. Jang, and K. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 2020, pp. 583–592.

[109] I. Drori, S. Tran, R. Wang, N. Cheng, K. Liu, L. Tang, E. Ke, N. Singh, T. L. Patti, J. Lynch, A. Shporer, N. Verma, E. Wu, and G. Strang, "A neural network solves and generates mathematics problems by program synthesis: Calculus, differential equations, linear algebra, and more," *CoRR*, vol. abs/2112.15594, 2021.

[110] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, and L. Weng, "Text and code embeddings by

contrastive pre-training," *CoRR*, vol. abs/2201.10005, 2022.

[111] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[112] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize from human feedback," *CoRR*, vol. abs/2009.01325, 2020.

[113] OpenAI, "Our approach to alignment research," *OpenAI Blog*, August 2022.

[114] ——, "Introducing chatgpt," *OpenAI Blog*, November 2022.

[115] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *CoRR*, vol. abs/2209.07858, 2022.

[116] OpenAI, "Lessons learned on language model safety and misuse," *OpenAI Blog*, March 2022.

[117] L. Huawei Technologies Co., "Huawei mindspore ai development framework," in *Artificial Intelligence Technology*. Springer, 2022, pp. 137–162.

[118] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[119] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," 2023. [Online]. Available: https://vicuna.lmsys.org

[120] 2023. [Online]. Available: https://github.com/nebuly-ai/nebullvm/tree/main/apps/accelerate/chatllama

[121] Y. You, "Colossalchat: An open-source solution for cloning chatgpt with a complete rlhf pipeline," 2023. [Online]. Available: https://medium.com/@yangyou_berkeley/colossalchat-an-open-source-solution-for-cloning-chatgpt-with-a-complete-rlhf-pipeline-5edf08fb538b

[122] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 19–27.

[123] "Project gutenberg." [Online]. Available: https://www.gutenberg.org/

[124] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *CoRR*, vol. abs/1806.02847, 2018.

[125] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 9051–9062.

[126] A. Gokaslan, V. C. E. Pavlick, and S. Tellex, "Openwebtext corpus," http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[127] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*. AAAI Press, 2020, pp. 830–839.

[128] "Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Main_Page

[129] "Bigquery dataset." [Online]. Available: https://cloud.google.com/bigquery?hl=zh-cn

[130] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800gb dataset of diverse text for language modeling," *CoRR*, vol. abs/2101.00027, 2021.

[131] H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. V. del Moral, T. Le Scao, L. Von Werra, C. Mou, E. G. Ponferrada, H. Nguyen *et al.*, "The bigscience roots corpus: A 1.6 tb composite multilingual dataset," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[132] "Common crawl." [Online]. Available: https://commoncrawl.org/

[133] "A reproduction version of cc-stories on hugging face." [Online]. Available: https://huggingface.co/datasets/spacemanidol/cc-stories

[134] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," https://github.com/kingoflolz/mesh-transformer-jax, 2021.

[135] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*. Association for Computational Linguistics, 2020, pp. 38–45.

[136] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/google/jax

[137] Z. Bian, H. Liu, B. Wang, H. Huang, Y. Li, C. Wang, F. Cui, and Y. You, "Colossal-ai: A unified deep learning system for large-scale parallel training," *CoRR*, vol. abs/2110.14883, 2021.

[138] J. Fang, Y. Yu, S. Li, Y. You, and J. Zhou, "Patrick-star: Parallel training of pre-trained models via

a chunk-based memory management," *CoRR*, vol. abs/2108.05818, 2021.

[139] "Bmtrain: Effient training for big models." [Online]. Available: https://github.com/OpenBMB/BMTrain

[140] J. He, J. Qiu, A. Zeng, Z. Yang, J. Zhai, and J. Tang, "Fastmoe: A fast mixture-of-expert training system," *CoRR*, vol. abs/2103.13262, 2021.

[141] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.

[142] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, K. Keeton and T. Roscoe, Eds. USENIX Association, 2016, pp. 265–283.

[143] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *CoRR*, vol. abs/1512.01274, 2015.

[144] Y. Ma, D. Yu, T. Wu, and H. Wang, "Paddlepaddle: An open-source deep learning platform from industrial practice," *Frontiers of Data and Domputing*, vol. 1, no. 1, p. 105, 2019.

[145] J. Yuan, X. Li, C. Cheng, J. Liu, R. Guo, S. Cai, C. Yao, F. Yang, X. Yi, C. Wu, H. Zhang, and J. Zhao, "Oneflow: Redesign the distributed deep learning framework from scratch," *CoRR*, vol. abs/2110.15032, 2021.

[146] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 2021, pp. 300–325.

[147] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra, "Solving quantitative reasoning problems with language models," *CoRR*, vol. abs/2206.14858, 2022.

[148] T. Saier, J. Krause, and M. Färber, "unarxive 2022: All arxiv publications pre-processed for nlp, including structured full-text and citation network," *arXiv preprint arXiv:2303.14957*, 2023.

[149] H. A. Simon, "Experiments with a heuristic compiler," *J. ACM*, vol. 10, no. 4, pp. 493–506, 1963.

[150] Z. Manna and R. J. Waldinger, "Toward automatic program synthesis," *Commun. ACM*, vol. 14, no. 3, pp. 151–165, 1971.

[151] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, "Codebert: A pre-trained model for programming and natural languages," in *Findings of EMNLP*, 2020.

[152] J. Austin, A. Odena, M. I. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. J. Cai, M. Terry, Q. V. Le, and C. Sutton, "Program synthesis with large language models," *CoRR*, vol. abs/2108.07732, 2021.

[153] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," 2021.

[154] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn, "A systematic evaluation of large language models of code," in *MAPS@PLDI*, 2022.

[155] D. Fried, A. Aghajanyan, J. Lin, S. Wang, E. Wallace, F. Shi, R. Zhong, W. Yih, L. Zettlemoyer, and M. Lewis, "Incoder: A generative model for code infilling and synthesis," in *ICLR*, 2023.

[156] A. Madaan, S. Zhou, U. Alon, Y. Yang, and G. Neubig, "Language models of code are few-shot commonsense learners," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 1384–1403.

[157] Y. Wu, A. Q. Jiang, W. Li, M. N. Rabe, C. Staats, M. Jamnik, and C. Szegedy, "Autoformalization with large language models," *CoRR*, vol. abs/2205.12615, 2022.

[158] D. Hernandez, T. B. Brown, T. Conerly, N. DasSarma, D. Drain, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume, S. Johnston, B. Mann, C. Olah, C. Olsson, D. Amodei, N. Joseph, J. Kaplan, and S. McCandlish, "Scaling laws and interpretability of learning from repeated data," *CoRR*, vol. abs/2205.10487, 2022.

[159] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[160] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 8424–8445.

[161] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," *CoRR*, 2022.

[162] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, 2021, pp. 2633–2650.

[163] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. PMLR, 2022, pp. 10 697–10 707.

[164] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, E. Blanco and W. Lu, Eds. Association for Computational Linguistics, 2018.

[165] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[166] M. Davis and M. Dürst, "Unicode normalization forms," 2001.

[167] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández, "The LAMBADA dataset: Word prediction requiring a broad discourse context," in *ACL (1)*. The Association for Computer Linguistics, 2016.

[168] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[169] B. Zhang, B. Ghorbani, A. Bapna, Y. Cheng, X. Garcia, J. Shen, and O. Firat, "Examining scaling and transfer of language model architectures for machine translation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022, pp. 26 176–26 192.

[170] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 13 042–13 054.

[171] A. Clark, D. de Las Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. A. Hechtman, T. Cai, S. Borgeaud, G. van den Driessche, E. Rutherford, T. Hennigan, M. J. Johnson, A. Cassirer, C. Jones, E. Buchatskaya, D. Budden, L. Sifre, S. Osindero, O. Vinyals, M. Ranzato, J. W. Rae, E. Elsen, K. Kavukcuoglu, and K. Simonyan, "Unified scaling laws for routed language models," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022, pp. 4057–4086.

[172] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang, "Cogview: Mastering text-to-image generation via transformers," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021, pp. 19 822–19 835.

[173] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," vol. abs/1607.06450, 2016.

[174] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 12 360–12 371.

[175] H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei, "Deepnet: Scaling transformers to 1, 000 layers," vol. abs/2203.00555, 2022.

[176] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[177] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, T. Linzen, G. Chrupala, and A. Alishahi, Eds. Association for Computational Linguistics, 2018, pp. 353–355.

[178] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[179] N. Shazeer, "GLU variants improve transformer," vol. abs/2002.05202, 2020.

[180] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," vol. abs/2104.09864, 2021.

[181] O. Press, N. A. Smith, and M. Lewis, "Train short, test long: Attention with linear biases enables input length extrapolation," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[182] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *ICML*, 2020.

[183] S. Narang, H. W. Chung, Y. Tay, L. Fedus, T. Févry, M. Matena, K. Malkan, N. Fiedel, N. Shazeer, Z. Lan, Y. Zhou, W. Li, N. Ding, J. Marcus, A. Roberts, and C. Raffel, "Do transformer modifications transfer across implementations and applications?" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, pp. 5758–5773.

[184] T. L. Scao, T. Wang, D. Hesslow, S. Bekman, M. S. Bari, S. Biderman, H. Elsahar, N. Muennighoff, J. Phang, O. Press, C. Raffel, V. Sanh, S. Shen, L. Sutawika, J. Tae, Z. X. Yong, J. Launay, and I. Beltagy, "What language model to train if you have one million GPU hours?" in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 765–782.

[185] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[186] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 933–941.

[187] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *CoRR*, vol. abs/1904.10509, 2019.

[188] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. A. Smith, and L. Kong, "Random feature attention," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

[189] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[190] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Re, "Flashattention: Fast and memory-efficient exact attention with IO-awareness," in *NeurIPS*, 2022.

[191] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[192] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[193] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 4603–4611.

[194] Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. X. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and Z. Chen, "Gpipe: Efficient training of giant neural networks using pipeline parallelism," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 103–112.

[195] A. Harlap, D. Narayanan, A. Phanishayee, V. Seshadri, N. R. Devanur, G. R. Ganger, and P. B. Gibbons, "Pipedream: Fast and efficient pipeline parallel DNN training," *CoRR*, vol. abs/1806.03377, 2018.

[196] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: memory optimizations toward training trillion parameter models," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, C. Cuicchi, I. Qualters, and W. T. Kramer, Eds. IEEE/ACM, 2020, p. 20.

[197] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," *CoRR*, vol. abs/1710.03740, 2017.

[198] Q. Xu, S. Li, C. Gong, and Y. You, "An efficient 2d method for training super-large deep learning models," *CoRR*, vol. abs/2104.05343, 2021.

[199] B. Wang, Q. Xu, Z. Bian, and Y. You, "Tesseract: Parallelize the tensor parallelism efficiently," in *Proceedings of the 51st International Conference on Parallel Processing, ICPP 2022, Bordeaux, France, 29 August 2022 - 1 September 2022*. ACM, 2022.

[200] Z. Bian, Q. Xu, B. Wang, and Y. You, "Maximizing parallelism in distributed training for huge neural networks," *CoRR*, vol. abs/2105.14450, 2021.

[201] S. Li, F. Xue, C. Baranwal, Y. Li, and Y. You, "Sequence parallelism: Long sequence training from system perspective," *arXiv e-prints*, pp. arXiv–2105, 2021.

[202] FairScale authors, "Fairscale: A general purpose modular pytorch library for high performance and large scale training," https://github.com/facebookresearch/fairscale, 2021.

[203] L. Zheng, Z. Li, H. Zhang, Y. Zhuang, Z. Chen, Y. Huang, Y. Wang, Y. Xu, D. Zhuo, E. P. Xing *et al.*, "Alpa: Automating inter-and {Intra-Operator} parallelism for distributed deep learning," in *OSDI*, 2022, pp. 559–578.

[204] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *CoRR*, vol. abs/1604.06174, 2016.

[205] Z. Yao, C. Li, X. Wu, S. Youn, and Y. He, "A comprehensive study on post-training quantization for large language models," *CoRR*, vol. abs/2303.08302, 2023.

[206] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," *CoRR*, vol. abs/2208.07339, 2022.

[207] C. Tao, L. Hou, W. Zhang, L. Shang, X. Jiang, Q. Liu, P. Luo, and N. Wong, "Compression of generative pre-trained language models via quantization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 4821–4836.

[208] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., 2022, pp. 3470–3487.

[209] Q. Ye, B. Y. Lin, and X. Ren, "Crossfit: A few-shot learning challenge for cross-task generalization in NLP," in *EMNLP (1)*. Association for Computational Linguistics, 2021, pp. 7163–7189.

[210] S. H. Bach, V. Sanh, Z. X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Févry, Z. Alyafeai, M. Dey, A. Santilli, Z. Sun, S. Ben-David, C. Xu, G. Chhablani, H. Wang, J. A. Fries, M. S. AlShaibani, S. Sharma, U. Thakker, K. Almubarak, X. Tang, D. R. Radev, M. T. Jiang, and A. M. Rush, "Promptsource: An integrated development environ-

ment and repository for natural language prompts," in *ACL (demo)*. Association for Computational Linguistics, 2022, pp. 93–104.

[211] V. Aribandi, Y. Tay, T. Schuster, J. Rao, H. S. Zheng, S. V. Mehta, H. Zhuang, V. Q. Tran, D. Bahri, J. Ni, J. P. Gupta, K. Hui, S. Ruder, and D. Metzler, "Ext5: Towards extreme multi-task scaling for transfer learning," in *ICLR*. OpenReview.net, 2022.

[212] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C. Wu, M. Zhong, P. Yin, S. I. Wang, V. Zhong, B. Wang, C. Li, C. Boyle, A. Ni, Z. Yao, D. Radev, C. Xiong, L. Kong, R. Zhang, N. A. Smith, L. Zettlemoyer, and T. Yu, "Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models," in *EMNLP*. Association for Computational Linguistics, 2022, pp. 602–631.

[213] T. Tang, J. Li, W. X. Zhao, and J. Wen, "MVP: multi-task supervised pre-training for natural language generation," *CoRR*, vol. abs/2206.12131, 2022.

[214] R. Lou, K. Zhang, and W. Yin, "Is prompt all you need? no. A comprehensive and broader view of instruction learning," *CoRR*, vol. abs/2303.10475, 2023.

[215] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *ACL (1)*. Association for Computational Linguistics, 2019, pp. 4487–4496.

[216] A. Aghajanyan, A. Gupta, A. Shrivastava, X. Chen, L. Zettlemoyer, and S. Gupta, "Muppet: Massive multi-task representations with pre-finetuning," in *EMNLP (1)*. Association for Computational Linguistics, 2021, pp. 5799–5811.

[217] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts, "The flan collection: Designing data and methods for effective instruction tuning," *CoRR*, vol. abs/2301.13688, 2023.

[218] Y. Gu, P. Ke, X. Zhu, and M. Huang, "Learning instructions with unlabeled data for zero-shot cross-task generalization," in *EMNLP*. Association for Computational Linguistics, 2022, pp. 1617–1634.

[219] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," *CoRR*, vol. abs/2212.10560, 2022.

[220] O. Honovich, T. Scialom, O. Levy, and T. Schick, "Unnatural instructions: Tuning language models with (almost) no human labor," *CoRR*, vol. abs/2212.09689, 2022.

[221] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[222] Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving, "Alignment of language agents," *CoRR*, vol. abs/2103.14659, 2021.

[223] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, "A general language assistant as a laboratory for alignment," *CoRR*, vol. abs/2112.00861, 2021.

[224] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *CoRR*, vol. abs/2204.05862, 2022.

[225] E. Perez, S. Huang, H. F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 3419–3448.

[226] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. F. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *CoRR*, vol. abs/1909.08593, 2019.

[227] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, H. F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese, "Teaching language models to support answers with verified quotes," *CoRR*, vol. abs/2203.11147, 2022.

[228] J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, and P. F. Christiano, "Recursively summarizing books with human feedback," *CoRR*, vol. abs/2109.10862, 2021.

[229] N. Ding, Y. Qin, G. Yang, F. Wei, Y. Zonghan, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, pp. 1–16, 03 2023.

[230] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 4582–4597.

[231] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 3045–3059.

[232] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[233] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019, pp. 2790–2799.

[234] Z. Hu, Y. Lan, L. Wang, W. Xu, E. Lim, R. K. Lee, L. Bing, and S. Poria, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models," *CoRR*, vol. abs/2304.01933, 2023.

[235] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[236] X. Liu, K. Ji, Y. Fu, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," *CoRR*, vol. abs/2110.07602, 2021.

[237] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," *CoRR*, vol. abs/2103.10385, 2021.

[238] Y. Gu, X. Han, Z. Liu, and M. Huang, "Ppt: Pre-trained prompt tuning for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8410–8423.

[239] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[240] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4222–4235.

[241] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," *CoRR*, vol. abs/2303.10512, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.10512

[242] M. Valipour, M. Rezagholizadeh, I. Kobyzev, and A. Ghodsi, "Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation," *CoRR*, vol. abs/2210.07558, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.07558

[243] Alpaca-LoRA, "Instruct-tune llama on consumer hardware," https://github.com/tloen/alpaca-lora, 2023.

[244] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *CoRR*, vol. abs/2303.16199, 2023.

[245] J. Pfeiffer, I. Vulic, I. Gurevych, and S. Ruder, "MAD-X: an adapter-based framework for multi-task cross-lingual transfer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for

Computational Linguistics, 2020, pp. 7654–7673.

[246] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul, "Peft: State-of-the-art parameter-efficient fine-tuning methods," https://github.com/huggingface/peft, 2022.

[247] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 2022, pp. 11 048–11 064.

[248] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., 2022, pp. 8086–8098.

[249] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., 2021, pp. 12 697–12 706.

[250] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for gpt-3?" in *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, 2022, pp. 100–114.

[251] Y. Lee, C. Lim, and H. Choi, "Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, Eds. International Committee on Computational Linguistics, 2022, pp. 669–683.

[252] I. Levy, B. Bogin, and J. Berant, "Diverse demonstrations improve in-context compositional generalization," *CoRR*, vol. abs/2212.06800, 2022.

[253] H. Su, J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, R. Zhang, M. Ostendorf, L. Zettlemoyer, N. A. Smith, and T. Yu, "Selective annotation makes language models better few-shot learners," *CoRR*, 2022.

[254] X. Ye, S. Iyer, A. Celikyilmaz, V. Stoyanov, G. Durrett, and R. Pasunuru, "Complementary explanations for effective in-context learning," *CoRR*, 2022.

[255] X. Li and X. Qiu, "Finding supporting examples for in-context learning," *CoRR*, 2023.

[256] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, 2022, pp. 2655–2671.

[257] Y. Zhang, S. Feng, and C. Tan, "Active example selection for in-context learning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 9134–9148.

[258] F. Gilardi, M. Alizadeh, and M. Kubli, "Chatgpt outperforms crowd-workers for text-annotation tasks," 2023.

[259] H. J. Kim, H. Cho, J. Kim, T. Kim, K. M. Yoo, and S. Lee, "Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator," *CoRR*, vol. abs/2206.08082, 2022.

[260] Y. Lin, A. Papangelis, S. Kim, S. Lee, D. Hazarika, M. Namazifar, D. Jin, Y. Liu, and D. Hakkani-Tur, "Selective in-context data augmentation for intent detection using pointwise v-information," *CoRR*, 2023.

[261] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[262] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," *CoRR*, vol. abs/2210.03493, 2022.

[263] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, and E. H. Chi, "Least-to-most prompting enables complex reasoning in large language models," *CoRR*, vol. abs/2205.10625, 2022.

[264] Z. Wu, Y. Wang, J. Ye, and L. Kong, "Self-adaptive in-context learning," *CoRR*, vol. abs/2212.10375, 2022.

[265] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds., 2022, pp. 2791–2809.

[266] S. C. Y. Chan, A. Santoro, A. K. Lampinen, J. X. Wang, A. Singh, P. H. Richemond, J. McClelland, and F. Hill, "Data distributional properties drive emergent in-context learning in transformers," *CoRR*, vol. abs/2205.05055, 2022.

[267] S. Shin, S. Lee, H. Ahn, S. Kim, H. Kim, B. Kim, K. Cho, G. Lee, W. Park, J. Ha, and N. Sung, "On the effect of pretraining corpora on in-context learning by a large-scale language model," in *NAACL-HLT*. Association for Computational Linguistics, 2022, pp. 5168–5186.

[268] J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, "Transformers learn in-context by gradient descent," *CoRR*, vol. abs/2212.07677, 2022.

[269] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, "In-context learning and induction heads," *CoRR*, vol. abs/2209.11895, 2022.

[270] H. Bansal, K. Gopalakrishnan, S. Dingliwal, S. Bodapati, K. Kirchhoff, and D. Roth, "Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale," *CoRR*, vol. abs/2212.09095, 2022.

[271] Y. Li, M. E. Ildiz, D. S. Papailiopoulos, and S. Oymak, "Transformers as algorithms: Generalization and implicit model selection in in-context learning," *CoRR*, vol. abs/2301.07067, 2023.

[272] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, "What learning algorithm is in-context learning? investigations with linear models," *CoRR*, vol. abs/2211.15661, 2022.

[273] S. Garg, D. Tsipras, P. Liang, and G. Valiant, "What can transformers learn in-context? A case study of simple function classes," *CoRR*, vol. abs/2208.01066, 2022.

[274] K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *CoRR*, vol. abs/2110.14168, 2021.

[275] A. Patel, S. Bhattamishra, and N. Goyal, "Are NLP models really able to solve simple math word problems?" in *NAACL-HLT*. Association for Computational Linguistics, 2021, pp. 2080–2094.

[276] S. Miao, C. Liang, and K. Su, "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, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 975–984.

[277] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4149–4158.

[278] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 346–361, 2021.

[279] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J. Lou, and W. Chen, "On the advance of making language models better reasoners," *CoRR*, vol. abs/2206.02336, 2022.

[280] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," *CoRR*, vol. abs/2210.00720, 2022.

[281] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *CoRR*, vol. abs/2205.11916, 2022.

[282] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *CoRR*, vol. abs/2203.11171, 2022.

[283] ——, "Rationale-augmented ensembles in language

models," *CoRR*, 2022.

[284] S. Imani, L. Du, and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," *arXiv preprint arXiv:2303.05398*, 2023.

[285] E. Zelikman, J. Mu, N. D. Goodman, and Y. T. Wu, "Star: Self-taught reasoner bootstrapping reasoning with reasoning," 2022.

[286] J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han, "Large language models can self-improve," *CoRR*, vol. abs/2210.11610, 2022.

[287] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. J. Orr, L. Zheng, M. Yüksekgönül, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," *CoRR*, vol. abs/2211.09110, 2022.

[288] A. Madaan and A. Yazdanbakhsh, "Text and patterns: For effective chain of thought, it takes two to tango," *CoRR*, vol. abs/2209.07686, 2022.

[289] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *CoRR*, vol. abs/2302.00923, 2023.

[290] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei, "Language models are multilingual chain-of-thought reasoners," *CoRR*, vol. abs/2210.03057, 2022.

[291] K. Shridhar, A. Stolfo, and M. Sachan, "Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions," *ArXiv*, vol. abs/2212.00193, 2022.

[292] N. Ho, L. Schmid, and S. Yun, "Large language models are reasoning teachers," *CoRR*, vol. abs/2212.10071, 2022.

[293] L. C. Magister, J. Mallinson, J. Adámek, E. Malmi, and A. Severyn, "Teaching small language models to reason," *CoRR*, vol. abs/2212.08410, 2022.

[294] Y. Fu, H. Peng, L. Ou, A. Sabharwal, and T. Khot, "Specializing smaller language models towards multi-step reasoning," *CoRR*, vol. abs/2301.12726, 2023.

[295] A. Chan, Z. Zeng, W. Lake, B. Joshi, H. Chen, and X. Ren, "Knife: Distilling meta-reasoning knowledge with free-text rationales," in *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*.

[296] Z. Li, C. Wang, P. Ma, C. Liu, S. Wang, D. Wu, and C. Gao, "On the feasibility of specialized ability stealing for large language code models," *CoRR*, 2023.

[297] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, and M. Chang, "Promptagator: Few-shot dense retrieval from 8 examples," *CoRR*, 2022.

[298] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[299] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," in *ICLR (Poster)*. OpenReview.net, 2017.

[300] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 workshop on statistical machine translation," in *WMT@ACL*. The Association for Computer Linguistics, 2014, pp. 12–58.

[301] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation," in *WMT*. The Association for Computer Linguistics, 2016, pp. 131–198.

[302] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation (WMT19)," in *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Post, M. Turchi, and K. Verspoor, Eds. Association for Computational Linguistics, 2019, pp. 1–61.

[303] L. Barrault, M. Biesialska, O. Bojar, M. R. Costa-jussà, C. Federmann, Y. Graham, R. Grundkiewicz, B. Haddow, M. Huck, E. Joanis, T. Kocmi, P. Koehn, C. Lo, N. Ljubesic, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, S. Pal, M. Post, and M. Zampieri, "Findings of the 2020 conference on machine translation (WMT20)," in *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, Eds. Association for Computational Linguistics, 2020, pp. 1–55.

[304] F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussà, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin, and M. Zampieri, "Findings of the 2021 conference on machine translation (WMT21)," in *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Fe-

dermann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, Eds. Association for Computational Linguistics, 2021, pp. 1–88.

[305] T. Kocmi, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, T. Gowda, Y. Graham, R. Grundkiewicz, B. Haddow, R. Knowles, P. Koehn, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, M. Novák, M. Popel, and M. Popovic, "Findings of the 2022 conference on machine translation (WMT22)," in *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costajussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, Eds. Association for Computational Linguistics, 2022, pp. 1–45.

[306] N. Goyal, C. Gao, V. Chaudhary, P. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 522–538, 2022.

[307] R. Bawden, E. Bilinski, T. Lavergne, and S. Rosset, "Diabla: a corpus of bilingual spontaneous written dialogues for machine translation," *Lang. Resour. Evaluation*, vol. 55, no. 3, pp. 635–660, 2021.

[308] R. Nallapati, B. Zhou, C. N. dos Santos, Ç. Gülçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, Y. Goldberg and S. Riezler, Eds. ACL, 2016, pp. 280–290.

[309] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *EMNLP*. Association for Computational Linguistics, 2018, pp. 1797–1807.

[310] F. Ladhak, E. Durmus, C. Cardie, and K. Mckeown, "Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4034–4048.

[311] S. Moon, P. Shah, A. Kumar, and R. Subba, "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *ACL (1)*. Association for Computational Linguistics, 2019, pp. 845–854.

[312] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," in *NeurIPS*, 2019, pp. 3261–3275.

[313] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *ICLR*.

OpenReview.net, 2021.

[314] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, "Challenging big-bench tasks and whether chain-of-thought can solve them," *CoRR*, vol. abs/2210.09261, 2022.

[315] L. Xu, H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu, Y. Tian, Q. Dong, W. Liu, B. Shi, Y. Cui, J. Li, J. Zeng, R. Wang, W. Xie, Y. Li, Y. Patterson, Z. Tian, Y. Zhang, H. Zhou, S. Liu, Z. Zhao, Q. Zhao, C. Yue, X. Zhang, Z. Yang, K. Richardson, and Z. Lan, "CLUE: A chinese language understanding evaluation benchmark," in *COLING*. International Committee on Computational Linguistics, 2020, pp. 4762–4772.

[316] D. Hendrycks, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, and J. Steinhardt, "Measuring coding challenge competence with APPS," in *NeurIPS Datasets and Benchmarks*, 2021.

[317] Y. Lai, C. Li, Y. Wang, T. Zhang, R. Zhong, L. Zettlemoyer, S. W. Yih, D. Fried, S. I. Wang, and T. Yu, "DS-1000: A natural and reliable benchmark for data science code generation," *CoRR*, vol. abs/2211.11501, 2022.

[318] Z. Wang, S. Zhou, D. Fried, and G. Neubig, "Execution-based evaluation for open-domain code generation," *CoRR*, vol. abs/2212.10481, 2022.

[319] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, pp. 452–466, 2019.

[320] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.

[321] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 3214–3252.

[322] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2013, pp. 1533–1544.

[323] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 2017, pp. 1601–1611.

[324] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, "PIQA: reasoning about physical commonsense in natural language," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-*

Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, 2020, pp. 7432–7439.

[325] M. Dubey, D. Banerjee, A. Abdelkawi, and J. Lehmann, "Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia," in *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, 2019, pp. 69–78.

[326] Y. Gu, S. Kase, M. Vanni, B. M. Sadler, P. Liang, X. Yan, and Y. Su, "Beyond I.I.D.: three levels of generalization for question answering on knowledge bases," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, 2021, pp. 3477–3488.

[327] S. Cao, J. Shi, L. Pan, L. Nie, Y. Xiang, L. Hou, J. Li, B. He, and H. Zhang, "KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 6101–6119.

[328] X. Hu, X. Wu, Y. Shu, and Y. Qu, "Logical form generation via multi-task learning for complex question answering over knowledge bases," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2022, pp. 1687–1696.

[329] S. Longpre, Y. Lu, and J. Daiber, "MKQA: A linguistically diverse benchmark for multilingual open domain question answering," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1389–1406, 2021.

[330] T. Saikh, T. Ghosal, A. Mittal, A. Ekbal, and P. Bhattacharyya, "Scienceqa: a novel resource for question answering on scholarly articles," *Int. J. Digit. Libr.*, vol. 23, no. 3, pp. 289–301, 2022.

[331] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "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, Brussels, Belgium, October 31 - November 4, 2018*, 2018, pp. 2381–2391.

[332] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," in *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, 2016.

[333] T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal, "QASC: A dataset for question answering via sentence composition," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020,

pp. 8082–8090.

[334] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 2383–2392.

[335] A. H. Miller, A. Fisch, J. Dodge, A. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 1400–1409.

[336] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, 2019, pp. 166–175.

[337] K. Toutanova and D. Chen, "Observed versus latent features for knowledge base and text inference," in *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, CVSC 2015, Beijing, China, July 26-31, 2015*, 2015, pp. 57–66.

[338] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, 2008, pp. 1247–1250.

[339] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 1811–1818.

[340] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, pp. 39–41, 1995.

[341] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller, "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 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 2463–2473.

[342] F. Mahdisoltani, J. Biega, and F. M. Suchanek, "YAGO3: A knowledge base from multilingual wikipedias," in *Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[343] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, 2007, pp. 697–706.

[344] C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," in *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2924–2936.

[345] M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi, "Socialiqa: Commonsense reasoning about social interactions," *CoRR*, vol. abs/1904.09728, 2019.

[346] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 4791–4800.

[347] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," in *AAAI*. AAAI Press, 2020, pp. 8732–8740.

[348] M. Roemmele, C. A. Bejan, and A. S. Gordon, "Choice of plausible alternatives: An evaluation of commonsense causal reasoning," in *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI, 2011.

[349] K. Sakaguchi, C. Bhagavatula, R. L. Bras, N. Tandon, P. Clark, and Y. Choi, "proscript: Partially ordered scripts generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 2138–2149.

[350] B. Dalvi, L. Huang, N. Tandon, W. Yih, and P. Clark, "Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 1595–1604.

[351] S. Saha, P. Yadav, L. Bauer, and M. Bansal, "Explagraphs: An explanation graph generation task for structured commonsense reasoning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 7716–7740.

[352] O. Tafjord, B. Dalvi, and P. Clark, "Proofwriter: Generating implications, proofs, and abductive statements over natural language," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 3621–3634.

[353] B. Dalvi, P. Jansen, O. Tafjord, Z. Xie, H. Smith, L. Pipatanangkura, and P. Clark, "Explaining answers with entailment trees," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 7358–7370.

[354] A. Saparov and H. He, "Language models are greedy reasoners: A systematic formal analysis of chain-of-thought," *CoRR*, vol. abs/2210.01240, 2022.

[355] C. Anil, Y. Wu, A. Andreassen, A. Lewkowycz, V. Misra, V. V. Ramasesh, A. Slone, G. Gur-Ari, E. Dyer, and B. Neyshabur, "Exploring length generalization in large language models," *CoRR*, vol. abs/2207.04901, 2022.

[356] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Rahane, A. S. Iyer, A. Andreassen, A. Santilli, A. Stuhlmüller, A. M. Dai, A. La, A. K. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakas, and et al., "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *CoRR*, vol. abs/2206.04615, 2022.

[357] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "PAL: program-aided language models," *CoRR*, vol. abs/2211.10435, 2022.

[358] S. Roy and D. Roth, "Solving general arithmetic word problems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds. The Association for Computational Linguistics, 2015, pp. 1743–1752.

[359] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2357–2367.

[360] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program induction by rationale generation: Learning to solve and explain algebraic word problems," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, R. Barzilay and M. Kan, Eds. Association for Computational Linguistics, 2017, pp. 158–167.

[361] R. Koncel-Kedziorski, S. Roy, A. Amini, N. Kushman,

and H. Hajishirzi, "Mawps: A math word problem repository," in *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1152–1157.

[362] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 2368–2378.

[363] S. Welleck, J. Liu, R. L. Bras, H. Hajishirzi, Y. Choi, and K. Cho, "Naturalproofs: Mathematical theorem proving in natural language," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung, Eds., 2021.

[364] A. Q. Jiang, W. Li, J. M. Han, and Y. Wu, "Lisa: Language models of isabelle proofs," in *6th Conference on Artificial Intelligence and Theorem Proving*, 2021, pp. 378–392.

[365] K. Zheng, J. M. Han, and S. Polu, "minif2f: a cross-system benchmark for formal olympiad-level mathematics," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[366] Z. Azerbayev, B. Piotrowski, H. Schoelkopf, E. W. Ayers, D. Radev, and J. Avigad, "Proofnet: Autoformalizing and formally proving undergraduate-level mathematics," *CoRR*, vol. abs/2302.12433, 2023.

[367] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[368] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *EMNLP*. The Association for Computational Linguistics, 2015, pp. 379–389.

[369] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," in *ACL (1)*. Association for Computational Linguistics, 2017, pp. 1870–1879.

[370] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318.

[371] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[372] K. Yang, Y. Tian, N. Peng, and D. Klein, "Re3: Generating longer stories with recursive reprompting and revision," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp.

4393–4479.

[373] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *CoRR*, vol. abs/2302.04023, 2023.

[374] S. Gulwani, O. Polozov, and R. Singh, "Program synthesis," *Found. Trends Program. Lang.*, vol. 4, no. 1-2, pp. 1–119, 2017.

[375] S. Zhang, Z. Chen, Y. Shen, M. Ding, J. B. Tenenbaum, and C. Gan, "Planning with large language models for code generation," 2023.

[376] M. Welsh, "The end of programming," *Commun. ACM*, vol. 66, no. 1, pp. 34–35, 2023.

[377] B. Wang, X. Deng, and H. Sun, "Iteratively prompt pre-trained language models for chain of thought," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 2714–2730.

[378] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis, "Measuring and narrowing the compositionality gap in language models," *CoRR*, vol. abs/2210.03350, 2022.

[379] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.

[380] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, 1989, pp. 109–165.

[381] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 3390–3398.

[382] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 5418–5426.

[383] G. Izacard, P. S. H. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Few-shot learning with retrieval augmented language models," *CoRR*, vol. abs/2208.03299, 2022.

[384] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2020, pp. 3929–3938.

[385] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[386] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J. Wen, "Complex knowledge base question answering: A survey," *CoRR*, vol. abs/2108.06688, 2021.

[387] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 2206–2240.

[388] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *CoRR*, vol. abs/2302.12813, 2023.

[389] S. Agarwal, I. Akkaya, V. Balcom, M. Bavarian, G. Bernadett-Shapiro, G. Brockman, M. Brundage, J. Chan, F. Chantzis, N. Deutsch, B. Eastman, A. Eleti, N. Felix, S. P. Fishman, I. Fulford, C. Gibson, J. Gross, M. Heaton, J. Hilton, X. Hu, S. Jain, H. Jin, L. Kilpatrick, C. Kim, M. Kolhede, A. Mayne, P. McMillan, D. Medina, J. Menick, A. Mishchenko, A. Nair, R. Nayak, A. Neelakantan, R. Nuttall, J. Parish, A. T. Passos, A. Perelman, F. de Avila Belbute Peres, V. Pong, J. Schulman, E. Sigler, N. Staudacher, N. Turley, J. Tworek, R. Greene, A. Vijayvergiya, C. Voss, J. Weng, M. Wiethoff, S. Yoo, K. Yu, W. Zaremba, S. Zhao, W. Zhuk, and B. Zoph, "Chatgpt plugins," *OpenAI Blog*, March 2023.

[390] A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev, "Internet-augmented language models through few-shot prompting for open-domain question answering," *CoRR*, vol. abs/2203.05115, 2022.

[391] A. Madaan, N. Tandon, P. Clark, and Y. Yang, "Memory-assisted prompt editing to improve GPT-3 after deployment," in *EMNLP*. Association for Computational Linguistics, 2022, pp. 2833–2861.

[392] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei, "Knowledge neurons in pretrained transformers," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 8493–8502.

[393] K. Meng, D. Bau, A. J. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," in

[394] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Synthetic prompting: Generating chain-of-thought demonstrations for large language models," *CoRR*, vol. abs/2302.00618, 2023.

[395] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He, "ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models," *CoRR*, 2023.

[396] Sifatkaur, M. Singh, V. S. B, and N. Malviya, "Mind meets machine: Unravelling gpt-4's cognitive psychology," *CoRR*, vol. abs/2303.11436, 2023.

[397] M. I. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena, "Show your work: Scratchpads for intermediate computation with language models," *CoRR*, vol. abs/2112.00114, 2021.

[398] J. Qian, H. Wang, Z. Li, S. Li, and X. Yan, "Limitations of language models in arithmetic and symbolic induction," *CoRR*, vol. abs/2208.05051, 2022.

[399] W. X. Zhao, K. Zhou, Z. Gong, B. Zhang, Y. Zhou, J. Sha, Z. Chen, S. Wang, C. Liu, and J. Wen, "Jiuzhang: A chinese pre-trained language model for mathematical problem understanding," in *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, A. Zhang and H. Rangwala, Eds. ACM, 2022, pp. 4571–4581.

[400] Q. Wang, C. Kaliszyk, and J. Urban, "First experiments with neural translation of informal to formal mathematics," in *Intelligent Computer Mathematics - 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings*, ser. Lecture Notes in Computer Science, F. Rabe, W. M. Farmer, G. O. Passmore, and A. Youssef, Eds., vol. 11006. Springer, 2018, pp. 255–270.

[401] S. Polu and I. Sutskever, "Generative language modeling for automated theorem proving," *CoRR*, vol. abs/2009.03393, 2020.

[402] A. Q. Jiang, W. Li, S. Tworkowski, K. Czechowski, T. Odrzygózdz, P. Milos, Y. Wu, and M. Jamnik, "Thor: Wielding hammers to integrate language models and automated theorem provers," *CoRR*, vol. abs/2205.10893, 2022.

[403] S. Polu, J. M. Han, K. Zheng, M. Baksys, I. Babuschkin, and I. Sutskever, "Formal mathematics statement curriculum learning," *CoRR*, vol. abs/2202.01344, 2022.

[404] A. Q. Jiang, S. Welleck, J. P. Zhou, W. Li, J. Liu, M. Jamnik, T. Lacroix, Y. Wu, and G. Lample, "Draft, sketch, and prove: Guiding formal theorem provers with informal proofs," *CoRR*, vol. abs/2210.12283, 2022.

[405] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch, "Faithful chain-of-thought reasoning," *CoRR*, vol. abs/2301.13379, 2023.

[406] Y. Weng, M. Zhu, S. He, K. Liu, and J. Zhao, "Large language models are reasoners with self-verification," *CoRR*, vol. abs/2212.09561, 2022.

[407] X. Pi, Q. Liu, B. Chen, M. Ziyadi, Z. Lin, Q. Fu, Y. Gao,

J. Lou, and W. Chen, "Reasoning like program executors," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 761–779.

[408] A. Parisi, Y. Zhao, and N. Fiedel, "TALM: tool augmented language models," *CoRR*, vol. abs/2205.12255, 2022.

[409] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 1953–1967.

[410] R. Rudinger, J. Naradowsky, B. Leonard, and B. V. Durme, "Gender bias in coreference resolution," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 2018, pp. 8–14.

[411] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 9118–9147.

[412] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P. Oudeyer, "Grounding large language models in interactive environments with online reinforcement learning," *CoRR*, vol. abs/2302.02662, 2023.

[413] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8494–8502.

[414] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A benchmark for interpreting grounded instructions for everyday tasks," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 10737–10746.

[415] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, C. K. Liu, S. Savarese, H. Gweon, J. Wu, and L. Fei-Fei, "BEHAVIOR: benchmark for everyday household activities in virtual, interactive, and ecological environments," in *CoRL*, ser. Proceedings of Machine Learning Research, vol. 164. PMLR, 2021, pp. 477–490.

[416] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, and M. Yan, "Do as I can, not as I say: Grounding language in robotic affordances," *CoRR*, vol. abs/2204.01691, 2022.

[417] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *CoRR*, vol. abs/2209.07753, 2022.

[418] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," *CoRR*, vol. abs/2209.11302, 2022.

[419] J. H. Clark, J. Palomaki, V. Nikolaev, E. Choi, D. Garrette, M. Collins, and T. Kwiatkowski, "Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 454–470, 2020.

[420] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, "A framework for few-shot language model evaluation," Sep. 2021.

[421] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT," *CoRR*, vol. abs/2302.10198, 2023.

[422] J. Kocon, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydlo, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocon, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Milkowski, M. Oleksy, M. Piasecki, L. Radlinski, K. Wojtasik, S. Wozniak, and P. Kazienko, "Chatgpt: Jack of all trades, master of none," *CoRR*, vol. abs/2302.10724, 2023.

[423] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?" *CoRR*, vol. abs/2302.06476, 2023.

[424] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" *CoRR*, vol. abs/2303.08559, 2023.

[425] X. Chen, J. Ye, C. Zu, N. Xu, R. Zheng, M. Peng, J. Zhou, T. Gui, Q. Zhang, and X. Huang, "How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks," 2023.

[426] M. Jang and T. Lukasiewicz, "Consistency analysis of chatgpt," *CoRR*, vol. abs/2303.06273, 2023.

[427] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?" *arXiv preprint arXiv:2303.04360*, 2023.

[428] O. Nov, N. Singh, and D. M. Mann, "Putting chatgpt's medical advice to the (turing) test," *CoRR*, vol. abs/2301.10035, 2023.

[429] S. Chen, B. H. Kann, M. B. Foote, H. J. Aerts, G. K. Savova, R. H. Mak, and D. S. Bitterman, "The utility of chatgpt for cancer treatment information," *medRxiv*, 2023.

[430] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You, "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge," 2023.

[431] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke, and M. Ingrisch, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," *CoRR*, vol. abs/2212.14882, 2022.

[432] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," vol. abs/2303.13375, 2023.

[433] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is chatgpt to human experts? comparison corpus, evaluation, and detection," *CoRR*, vol. abs/2301.07597, 2023.

[434] V. Liévin, C. E. Hother, and O. Winther, "Can large language models reason about medical questions?" *CoRR*, vol. abs/2207.08143, 2022.

[435] G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?" *arXiv preprint arXiv:2301.12127*, 2023.

[436] S. Bordt and U. von Luxburg, "Chatgpt participates in a computer science exam," *CoRR*, vol. abs/2303.09461, 2023.

[437] K. Malinka, M. Peresíni, A. Firc, O. Hujnak, and F. Janus, "On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?" *CoRR*, vol. abs/2303.11146, 2023.

[438] T. Susnjak, "Chatgpt: The end of online exam integrity?" *CoRR*, vol. abs/2212.09292, 2022.

[439] A. Blair-Stanek, N. Holzenberger, and B. V. Durme, "Can GPT-3 perform statutory reasoning?" *CoRR*, vol. abs/2302.06100, 2023.

[440] F. Yu, L. Quartey, and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," *CoRR*, vol. abs/2212.01326, 2022.

[441] D. Trautmann, A. Petrova, and F. Schilder, "Legal prompt engineering for multilingual legal judgement prediction," *CoRR*, vol. abs/2212.02199, 2022.

[442] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "Chatgpt goes to law school," *Available at SSRN*, 2023.

[443] J. J. Nay, "Law informs code: A legal informatics approach to aligning artificial intelligence with humans," *CoRR*, vol. abs/2209.13020, 2022.

[444] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," *CoRR*, vol. abs/2102.02503, 2021.

[445] Z. Sun, "A short survey of viewing large language models in legal aspect," *CoRR*, vol. abs/2303.09136, 2023.

[446] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," in *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, M. Fourcade, B. Kuipers, S. Lazar, and D. K. Mulligan, Eds. ACM, 2021, pp. 298–306.

[447] A. Borji, "A categorical archive of chatgpt failures," *CoRR*, vol. abs/2302.03494, 2023.

[448] M. Kosinski, "Theory of mind may have spontaneously emerged in large language models," *CoRR*, vol. abs/2302.02083, 2023.

[449] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general ai? A first evaluation on chatgpt," *CoRR*, vol. abs/2303.03186, 2023.

[450] R. Aiyappa, J. An, H. Kwak, and Y.-Y. Ahn, "Can we trust the evaluation on chatgpt?" vol. abs/2303.12767, 2023.

[451] H. Cho, H. J. Kim, J. Kim, S. Lee, S. Lee, K. M. Yoo, and T. Kim, "Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners," *CoRR*, vol. abs/2212.10873, 2022.

[452] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 109:1–109:28, 2023.