Mini-Giants: "Small" Language Models and Open Source Win-Win

Zhengping Zhou
zpzhou@cs.stanford.edu

Lezhi Li
lli2@gsd.harvard.edu

Xinxi Chen
xc336@cornell.edu

Andy Li
andy@RLAI.institute

Abstract

ChatGPT is phenomenal. However, it is prohibitively expensive to train and refine such giant models. Fortunately, small language models are flourishing and becoming more and more competent. We call them "mini-giants". We argue that open source community like Kaggle and mini-giants will win-win in many ways, technically, ethically and socially. In this article, we present a brief yet rich background, discuss how to attain small language models, present a comparative study of small language models and a brief discussion of evaluation methods, discuss the application scenarios where small language models are most needed in the real world, and conclude with discussion and outlook.

1 Introduction

Large language models (LMs), like ChatGPT and GPT-4, have been taken us by storm. People compare it to the moment of the computer, the moment of the operating system, the moment of the Internet, or the moment of the iPhone. It is considered by many a paradigm shift in NLP and deep learning.

Large language models are large: OpenAI GPT-3 175B parameters, Google PALM 560B, and rumor has it that GPT-4 is as large as 8×220B. For most small/medium companies and independent researchers, it is prohibitively expensive to train or update such giant models. In addition, huge consumption of energy for language model training poses a serious concern to the environmental sustainability (Verdecchia et al., 2023).

Recent studies show that network size is not the sole determinant of model performance (Hoffmann et al., 2022). And thanks to the efforts from the ML open source community as well as private AI companies, we've recently seen more and more "small" LMs created out of these larger models. With their network parameter sizes of around or below 10B, and performance comparable or better than ChatGPT / GPT-4, these "small" LMs are indeed "mini-giants".

In this article, we survey the state-of-the-art for these small language models. We show that compared to their large counterparts, small language/foundation models offer particularly promising opportunities for various industries (including open source ML research and Kaggle competitions) to not only utilize but also to actively participate in the creation/adaptation of modern language models and AI in general. We center our arguments around the 3 key advantages of small models: adaptability, controllability, and affordability.

First of all, smaller models offer better adaptability by being more manageable to modify and fine-tune. In Section 3, we present various strategies of creating these small models through optimized fine-tuning techniques. This is important because in most industries (or even in a Kaggle competition), innovation typically arises from the ability to incorporate domain-specific data into the language model or to adjust the model's structure to accommodate their unique requirements. Relying solely on prompt engineering often falls short. Therefore, smaller language models bring forward great benefits to these industries, offering the much-needed flexibility for adaptation, allowing them to full leverage the power of AI and thus catalyzing innovation within them.

Second, smaller models can run on local infrastructure without resorting to GPU-rich third parties, improving the model's controllability by ensuring model users' autonomous data governance and result monitoring. In Section 5, we discuss real world scenarios where small language models fill in the gaps when their large counterparts are unacceptable due to privacy concerns. In Section 4, we also look into strategies for customized instruction following and other pioneer research directions for small models, underpinning the relevance of smaller language

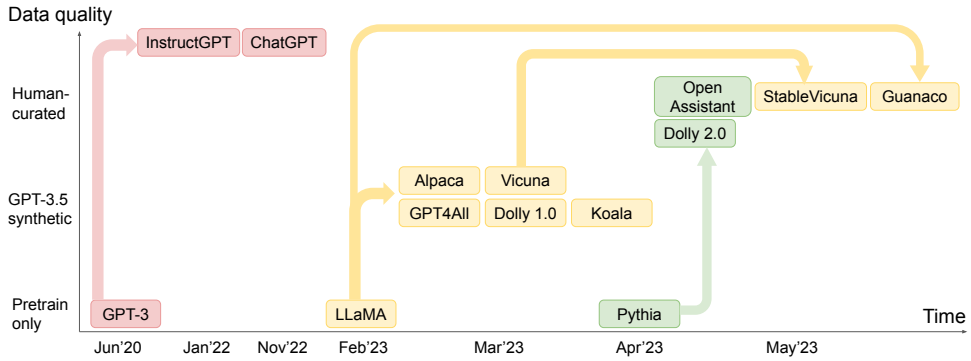

Figure 1: An evolution tree of recently released instruction-following small LMs. The color of the text boxes indicates the openness of the license under which the models are released: red stands for proprietary licenses, yellow stands for non-commercial licenses, and green stands for licenses permissive for commercial use.

models in ensuring compliance and mitigating the risk of misinformation. Understanding and managing the way a model operates, the data it accesses, and the outputs it produces, form the cornerstone of responsible AI usage.

Another crucial aspect of the superiority of small language models, is affordability. Taking an average Kaggle competitor as an example. The demanding nature of a Kaggle competition requires the competitor to iterate on the modeling solutions, often times by integrating a variety of data sources and trying different architectures. This necessitates transparent model components and fast iteration pace, which is at odds with the resource requirements that super large language models impose. Having access to fast and inexpensive training / inferencing options, means that he/she will not have to face the trade-off between being constrained in their innovation space, and moving away from language model solutions entirely. As another example which is elaborated in Section 5, privacy-sensitive sectors such as finance and healthcare face a more pressing challenge of choosing between regulation risks and the prohibitive cost of training massive models in-house. Small language models provide the opportunity for them to conform with regulations while not missing out on the power of latest AI technologies.

Outline

In the following sections, we first present a brief yet rich background. Next, we discuss how to attain small foundation models, including parameter reduction and efficient training/fine-tuning techniques. Then we present a comparative study of "small" foundation models a brief discussion of evaluation methods. After that, we discuss the application scenarios where small foundation models are most needed in the real world. We conclude with discussions and an outlook.

2 A brief yet rich background

The Giants are fast ChatGPT set a record for fastest-growing user base: one million users in five days, and 100 million monthly active users in January 2023, two months after launching.

Radford et al. (2018) introduce generative pre-training for LMs, which could be regarded as "GPT-1". Radford et al. (2019) introduce GPT-2, an unsupervised multitask learning LM. Brown et al. (2020) introduce GPT-3, a few-shot learning LM, popularizing the concept of in-context learning. OpenAI (2022) introduces ChatGPT and OpenAI (2023) introduces GPT-4.

There are many LMs released in recent years: Google BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), Google T5, Text-To-Text Transfer Transformer (Raffel et al., 2020), Google LaMDA, Language Model for Dialogue Applications (Thoppilan et al., 2022), Google PaLM, Pathways Language Model (Chowdhery et al., 2022), Deepmind Sparrow (Glaese et al., 2022), Anthropic Claude (Bai et al., 2022a), Deepmind Chinchilla (Hoffmann et al., 2022) Nivedia Megatron-Turing NLG (Smith et al., 2022), Deepmind Gopher (Rae et al., 2022), HuggingFace BLOOM (BigScience Workshop et al., 2023), and Meta LLaMA, Large Language Model Meta AI (Touvron et al., 2023).

Language models as experts Besides general purpose LMs as above, there are many specialized models for various application, e.g., Table 1 shows a sample of them.

Model	Application	Reference
AlphaFold	Protein folding	Tunyasuvunakool et al. (2021)
Codex	Coding	Chen et al. (2023)
AlphaCode	Coding	Li et al. (2022)
RT-1	Robotics	Brohan et al. (2022)
BiomedGPT	Biomedical	Zhang et al. (2023a)
Clinical Camel	Clinical	Toma et al. (2023)
BloombergGPT	Finance	Wu et al. (2023b)
FinGPT	Finance	Yang et al. (2023)
Med-PaLM 2	Medical	Singhal et al. (2023)
MusicLM	Music	Agostinelli et al. (2023)
AudioGPT	Audio	Huang et al. (2023)

Table 1: A (small) sample of specialized LMs

Language and functional competence Mahowald et al. (2023) study language competence vs thought competence of LMs and show impressive but imperfect formal linguistic competence, i.e., "knowledge of rules and patterns of a given language", yet failures on many tests requiring functional linguistic competence, i.e.,"a host of cognitive abilities required for language understanding and use in the real world".

Then we can leverage LMs' competence as a good model of language, e.g., by prompt engineering. We can also manage to improve the functional competence, e.g., factuality, safety, and planning. With the capacity of in-context learning (Brown et al., 2020), prompting is a natural and popular way to utilize LMs. Prompting is the user interface for LMs, and can be formed with advanced methods like search and coding, e.g., Tree of Thoughts (ToT) (Yao et al., 2023), AdaPlanner (Sun et al., 2023), Code as Policies (Liang et al., 2023a). Fine-tuning can improve LMs further. A parameter efficient approach makes fine-tuning large LMs feasible considering the cost (Hu et al., 2021; Ding et al., 2023; Ruder et al., 2022). Augmenting LMs with tools can achieve various functionalities.

To approach artificial general intelligence (AGI) from language models, Mahowald et al. (2023) suggest that, "instead of or in addition to scaling up the size of the models, more promising solutions will come in the form of modular architectures . . . , like the human brain, integrate language processing with additional systems that carry out perception, reasoning, and planning". The authors believe that "a model that succeeds at real-world language use would include – in addition to the core language component – a successful problem solver, a grounded experiencer, a situation modeler, a pragmatic reasoner, and a goal setter".

Augmented LMs with tools A natural way to harnesses the language competence of LMs is by utilizing tools like a search engine, a vector database, a code interpreter, or a solver to handle tasks, e.g., LangChain[1], HuggingGPT (Shen et al., 2023), Visual ChatGPT (Wu et al., 2023a), TaskMatrix.AI (Liang et al., 2023b), RCI (Kim et al., 2023), LLM+P (Liu et al., 2023a), ChemCrow (Bran et al., 2023), etc. See Mialon et al. (2023) for a survey about augmented LMs.

Domain expertise is still required, e.g., the ChemCrow Bran et al. (2023) authors mention that "However, it is important to emphasize that potential risks may arise for non-experts who lack the chemical reasoning to evaluate results or the proper lab training, as conducting experiments still necessitates thorough laboratory experience." and the director of the movie trailer mentions that "For those who believe that AI will do everything for you: No!" and "I'll always prefer to put my own heart & soul in." [2]

Mini-Giants are coming Following the leakage of LLaMA (Touvron et al., 2023), many "small" LMs appear in the open source community, with neural network parameter sizes of around 10B or smaller, e.g., Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), Koala (Geng et al., 2023), Vicuna (Chiang et al., 2023), StableLM (Stability AI, 2023a), ChatGLM (Du et al., 2020; Zeng et al., 2023), Guanaco (Dettmers et al., 2023), Pythia (Biderman et al., 2023), GPT4All[3], OpenAssistant[4], ColossalChat (You, 2023).

See Kim (2023) for a list of open sourced fine-tuned LMs. In Section 4, we will discuss and com-

[1] https://langchain.com

[2] https://twitter.com/ChristianF369/status/1651607149804498946

[3] https://github.com/nomic-ai/gpt4all

[4] https://github.com/LAION-AI/Open-Assistant

pare these mini-giants in details.

Discussions & debates abound There are all sorts of discussions & debates, e.g. discussions about AI alignment with human value from Russell (2019); Mitchell (2020); Christian (2021). Table 2 lists a few representative examples.

3 How to make large foundation models "small"

Since the advent of ultra-capable large foundation models like ChatGPT and StableDiffusion, numerous efforts have been devoted to address the primary challenges for their wide-spread utilization: their humongous parameter sizes and the sheer time and compute resources needed to fine-tune them. Within 2 years, the research and open source community have arrived at several strategies to cope with this issue, which we will discuss in this section.

We classify these strategies into 2 groups: ones that directly reduce the parameter sizes, and ones that makes fine-tuning large models more efficient.

3.1 Foundation models with reduced parameters

Chinchilla (Hoffmann et al., 2022) is the first influential study on computational efficiency of modern large language models. It put forward the argument that given a compute budget, the best model is attained not by larger parameter size, but by more training data tokens. Based on this principle, the authors produced the Chinchilla 70B model which out-performs prior large models 4 times as large, with the same amount of compute.

LLaMa (Touvron et al., 2023) further reduces the parameters and released a series of models ranging from 7 to 65B parameters, following the Chinchilla computation rule. Notably, the paper used only publicly available datasets as training corpus and proved comparable performance as closed source counterparts. This, as commented by (Harris et al.), started a revolution of open source LLM models. Along with parameter reduction, another contribution by the authors is efficient implementation of multi-headed attention layers through the open source *xformers* library, which optimizes the memory consumption in training.

3.2 Efficient fine-tuning strategies for foundation models

Compared with building even more compact models, the majority of research work by the ML community in the direction of "smaller" foundation models, is around making them easier to fine-tune. Here we list several key strategies to achieve this.

Adapter (Houlsby et al., 2019) is a strategy to add NN layers after existing layers (usually transformer blocks) in pretrained foundation models, so that they can be adapted to custom tasks without changing the weights of existing layers. This paper proposes an adapter module with two linear layers plus a non-linear activation in between. The first layer projects the hidden state to a lower-dimensional space, and the second layer projects it back to the original dimension. A newer paper (Lin et al., 2020) recommended only one linear layer plus an additional LayerNorm, as an Adapter module. Adapter achieves near state-of-the-art performance, while adding only a small amount of parameters per task - on GLUE, the added parameters accounted for 3.6% of the original model.

Prefix fine-tuning (Li and Liang, 2021) Unlike the Adapter architecture that focuses on modifying model behavior via model params, Prefix fine-tuning seeks to train a few params that are used as input prefixes, for each custom sub task. The authors commented that the method is inspired by prompting: similar to prepending a few sentences before a generation task, Prefix-tuning prepends a sequence of trained vectors to the input - just that the prefix vectors do not have to correspond to any real tokens. Compared to full fine-tuning, prefix-fine tuning achieves comparable or better performance with just 0.1% added parameters.

LoRA (Hu et al., 2021) Marks a substantial progress in parameter efficient fine-tuning. Performance-wise, it is more efficient than previous methods like Adapter and Prefix-finetuning. LoRA proposes that we add a low rank, trainable matrix in parallel to the frozen, pretrained model weights. The activation will be the sum of these two matrices. Formally:

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

where B and A are much "thinner" (i.e. low rank), trainable matrices compared to W_0 (the frozen pretrained matrix). The use of low rank matrices reduces trainable parameters to as much as

Issue to discuss	Reference
The dangers of stochastic parrots	Bender et al. (2021)
Limitation of neural networks	Delétang et al. (2023)
Limitation of autoregressive models	Lin et al. (2021)
Lack of causality	Jin et al. (2023)
Lack of compositionality	Dziri et al. (2023)
Lack of recursion	Zhang et al. (2023b)
Limitation of scaling laws	Deshpande et al. (2023)
Limitation of scaling laws	McKenzie et al. (2023)
Model collapse	Shumailov et al. (2023)
Artificial general intelligence (AGI)	Marcus (2023)
Evaluation of AI	Burnell et al. (2023)
Distortion of human beliefs	Kidd and Birhane (2023)
Social norms	Browning and LeCun (2023)
Risks and benefits	Goldman (2023)
Existential risk	Bengio (2023)
Court hearing due to hallucination	Novak (2023)
Risk of further concentration of wealth	Chiang (2023)
Eight things to know	Bowman (2023)

Table 2: Discussions and debates of LMs

by 10,000 times of the original model, compared to a full fine-tune of GPT-3 175B. The article suggests that LoRA can be used next to any model weights, not just transformer layers. The authors claim that LoRA is superior compared to Adapters in that it doesn't introduce additional inference latency; and it's better than Prefix fine-tuning in that it doesn't reduce the available sequence length like the latter does. Further more, since this architectural modification is orthogonal to the ideas of Adapter and Prefix fine-tuning, LoRA can be used in conjunction with them for even better results.

QLoRA (Dettmers et al., 2023) As an improvement of LoRA, QLoRA proposes optimization methods via quantized low rank fine tuning. Innovations of QLoRA include a 4-bit data type: NormalFloat4, which optimizes information efficiency for normally distributed data (e.g. weights) based on information theory. Apart from that, the paper uses Paged Optimizers (partial optimizer state stored on CPU rather than GPU) to manage memory spikes, like when processing mini batches with long sequence lengths. Experiment results show that fine-tuning using QLoRA reaches 99.3% of the performance of ChatGPT, and only requires training for 24 hours on one GPU.

ControlNet (Zhang and Agrawala, 2023) is proposed as a method to efficiently fine-tune image generation models (diffusion model) on user-defined tasks. Because image generation models in general have a larger design space in terms of user interaction than language models, we list this method here to inspire the readers to consider more complex scenarios of controlling / customizing large foundation model's outputs.

ControlNet copies weights of the original model to a frozen copy (ike all methods mentioned above). The trainable branch consists of an exact same copy as the frozen copy, as well as two convolution layers called "zero convolutions", both before and after the trainable copy. In the fine-tuning forward path, the activation from the trainable copy will be combined with that of the frozen copy by Zero Convolution. The so-called Zero Convolution is just a 1x1 convolution layer that are initiated with both weights and biases being zeros. The result of using ControlNet shows that in some tasks, ControlNets on a personal computer achieve comparable results as commercial models trained on terabytes of GPU memory and thousands of GPU hours.

4 A brief survey of "small" instruction-following LMs

Over the past few months, we have seen small LMs flourish. See Figure 1 for an evolution tree. This is a very fast progressing field, and it is challenging to even keep ahead with the latest progress. Quoting (Tunguz, 2023), "Trying to get ahead in AI these days feels like wrestling a rabid 5,000 lbs hippo covered in baby oil".

4.1 Closed-source milestones

GPT-3 (Brown et al., 2020) gained public attention when it was released in 2020. As reported by New York Times, it "generates tweets, pens poetry,

Basic info			Scale			Openness			
Time MM/YY	Model	Institute	# parameters	Training hardware cost	Training data size	L	I	TC	TD
06/20	GPT-3	OpenAI	175B	3.64k PT-days	300B tokens	P	P	P	✓
02/23	LLaMA-7B	Meta	7B	82k GPU-hours	1.4T tokens	NC	✓	✗	✓
02/23	LLaMA-13B	Meta	13B	135k GPU-hours	1.4T tokens	NC	✓	✗	✓
04/23	Pythia-7B	Eleuther AI	7B	33.5k GPU-hours	300B tokens	C	✓	✓	✓
04/23	Pythia-12B	Eleuther AI	12B	72k GPU-hours	300B tokens	C	✓	✓	✓

Table 3: Comparison of recent base LMs. In the Openness section, L stands for License, I stands for Inference, TC stands for Training Codes, and TD stands for Training Data. In the License column, P stands for Proprietary, NC stands for Non-Commercial, and C stands for permissive for Commercial use.

Basic info				Scale		Openness			
Time MM/YY	Model	Institute	Backbone	# parameters	Training hardware cost	L	I	TC	TD
01/22	InstructGPT	OpenAI	GPT-3	1.3B	N/A	P	P	P	P
11/22	ChatGPT	OpenAI	GPT-3	N/A	N/A	P	P	P	P
03/23	Alpaca-7B	Stanford	LLaMA-7B	7B	< $100	NC	✓	✓	✓
03/23	GPT4All-Lora	Nomic AI	LLaMA-7B	7B	$100	NC	✓	✓	✓
03/23	ChatGLM-6B	Tsinghua	GLM	6B	N/A	NC	✓	✓	✗
03/23	Vicuna-7B/13B	LMSYS	LLaMA-7B/13B	7B/13B	$140/$300	NC	✓	✓	✓
03/23	Dolly-6B	Databricks	GPT-J-6B	6B	< $30	NC	✓	✓	✓
04/23	OASST-12B	LAION AI	Pythia-12B	12B	N/A	C	✓	✓	✓
04/23	Koala-13B	Berkeley	LLaMA-13B	13B	< $100	NC	✓	✓	✓
04/23	Dolly-v2-12B	Databricks	Pythia-12B	12B	N/A	C	✓	✓	✓
04/23	StableVicuna-13B	Stability AI	Vicuna-13B	13B	N/A	NC	✓	✓	✓
05/23	Guanaco-7B/13B	UW	LLaMA-7B/13B	7B/13B	< 12 GPU-hours	NC	✓	✓	✓

Table 4: Comparison of recent instruction-following small LMs. The abbreviations of the column names follow Table 3.

summarizes emails, answers trivia questions, translates languages and even writes its own computer programs" (Markoff, 2020). It shows that decent few-shot performance can be achieved without gradient update, and the unprecedented model scale (175B parameters) is a key ingredient for success.

InstructGPT Although GPT-3 is already powerful, Ouyang et al. (2022) points out that the model output may not align well with human intent and may contain harmful content. For example, when prompted to generate a story, the LM should generate a story instead of rambling around the prompt itself. This necessitated an extra step called *model alignment*, and the desired model behavior is called *instruction-following*. In InstructGPT, this is achieved by applying the reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) technique on top of a GPT-3 backbone. Despite having 100x less parameters, InstructGPT outperforms the unaligned GPT-3 model in human evaluation, giving rise to the phenomenal success of ChatGPT ten months later.

ChatGPT (OpenAI, 2022) brings AIGC to the attention of the general public. It uses the same technique as InstructGPT, but extends InstructGPT by incorporating dialogue data into the supervised fine-tuning and the RLHF stage. It acquired 1 million users in just 5 days and revolutionizes the way people interact with modern AIs. As a proprietary product, although the web UI is free, the underlying model can only be accessed via a paid API.

4.2 Open-source backbone LMs

LLaMA Despite the recent success of GPT-3 and ChatGPT, training and deploying LLMs remain a major challenge to the open source community due to the high training infra cost. For instance, the GPT-3 training is estimated to cost millions of dollars. (Touvron et al., 2023) propose LLaMA, an open source LLM pretrained with public data available at several sizes. Remarkably, the 13B LLaMA model benefited from large scale pretraining data (1.4T tokens), and outperforms the 175B GPT-3 on most benchmarks. It soon becomes a highly influential milestone in the open source world, serving as a powerful yet lightweight backbone for a wide range of subsequent instruction-following small LMs. The non-commercial bespoke license, under which it is released, limits the usage to research purpose only.

Pythia (Biderman et al., 2023) Published two months later than LLaMA, Pythia releases a suite of 16 LLMs ranging from 70M to 12B parameters. Trained with 300B tokens from the Pile (Gao et al., 2020), it consumed a similar amount of data as GPT-3 but around four times less than LLaMA (see Table 3 for comparison). Released under the Apache 2.0 license, Pythia is free for commercial use, making it an appealing backbone for many subsequent instruction-following small LMs (e.g. Open Assistant (Köpf et al., 2023), Dolly 2.0 (Conover et al., 2023)).

4.3 Small LMs trained with GPT synthetic data

Since the release of LLaMA, open-source instruction fine-tuned small LMs emerge at a rapid speed. Viewing LLaMA as an open-source counterpart of GPT-3, these small LMs can be seen as the open-source counterparts of InstructGPT or ChatGPT. Most of them can be fine-tuned under a feasible budget (the training hardware cost can be capped under several hundred dollars).

A major challenge is to obtain high-quality instruction-following data, a key ingredient in the model alignment stage. At an early stage, the open-source community tackles this challenge by using GPT-3.5 (OpenAI, 2022) to synthesize the response of a given prompt. This imposes a non-commercial license on the fine-tuned model.

Alpaca (Taori et al., 2023) is the first newborn in this family. It fine-tunes LLaMA-7B with 52k instruction-following data generated using the self-instruct method, which leverages GPT-3.5 to synthesize prompt-response pairs from a manually created seed set. According to human evaluation, it achieves similar performance to GPT-3.5 on a small sample data.

GPT4All (Anand et al., 2023) fine-tunes LLaMA-7B with 437k prompt-response pairs. The instructions are collected from the `unified_chip2` and Stackoverflow Questions, while the responses are generated by GPT-3.5. The model is fine-tuned using the LoRA (Hu et al., 2021) algorithm. Evaluated using the ground truth perplexity on the Self-Instruct (Wang et al., 2023) human evaluation data, GPT4All stochastically outperforms Alpaca.

Vicuna (Chiang et al., 2023) fine-tunes LLaMA-13B with 70k user-shared conversations with ChatGPT (from ShareGPT.com). Compared to Alpaca, it accounts for multi-turn conversation in training, and made several optimizations to cut the training cost. Vicuna uses GPT-4 as an automatic chatbot judge, based on which it outperforms LLaMA and Alpaca, while achieving more than 90% quality of ChatGPT. A more rigorous analysis validating this evaluation approach is later presented in the Guanaco work (Dettmers et al., 2023).

Koala (Geng et al., 2023) is another instruction fine-tuned LLaMA model, with 13B parameters. It is a concurrent effort with Vicuna, released at a similar time. Like Vicuna, it is fine-tuned on ChatGPT-distilled data, with a focus on the dialogue scenario. In human evaluation, Koala achieves comparable or superior results compared to Alpaca.

4.4 Small LMs trained with human-curated data

Dolly 1.0 (Conover et al., 2023) trains a two-year-old GPT-J-6B backbone using the same data as Alpaca, showcasing that the instruction-following capability does not necessarily require state-of-the-art backbone model as long as the data quality is decent. **Dolly 2.0**, released one month later, upgrades to the newly released Pythia-12B (Biderman et al., 2023) backbone and is instruction fine-tuned using a newly crowd-sourced dataset, `databricks-dolly-15k` which contains 15k human-generated prompt-response pairs. Notably, it is the first open-source instruction-following small LM that permits commercial use.

Open Assistant (Köpf et al., 2023) uses LLaMA-13B and Pythia-12B as the backbones, allowing it to release chatbots under either non-commercial and commercial licenses. It also releases the OpenAssistant Conversations (`oasst1`) dataset, which contains 66k conversations generated by human, accompanied with quality ratings. It also includes human preferences for the model responses, which enables RLHF training. After fine-tuning on this dataset, Open Assistant achieves a 48.3% v.s. 51.7% as compared to ChatGPT. As a high quality human-generated dataset free of GPT-synthesized content, `oasst1` is widely used in follow-up works.

StableVicuna After the release of the `oasst1` dataset, (Stability AI, 2023b) proposes StableVicuna, "the AI world's first open-source RLHF LLM chatbot". It is fine-tuned on the Vicuna-13B model using a mix of the prompt-response datasets from Open Assistant, GPT4All, and Alpaca. The model

is further optimized using RLHF with human preference data from Open Assistant, HH-RLHF (Bai et al., 2022b), and SHP (Ethayarajh et al., 2022). By the time StableVicuna is released, it outperforms other similarly sized open-source chatbots on a number of question-answering benchmarks.

Guanaco (Dettmers et al., 2023) introduces an efficient fine-tuning approach called QLoRA. As a by product, the chatbot Guanaco-65B fine-tuned on top of LLaMA achieves state-of-the-art results in human evaluation. It also releases the 7B/13B versions which are of a similar scale as previously mentioned small LMs. The fine-tuning dataset is a mix of `oasst1` (Köpf et al., 2023) and some other public datasets.

4.5 Community trends and research directions

In addition to trained models shown above, we would like to point out a few research trends around the topic of making small language models more efficient and performant. We discuss studies on accelerated training for large language models, performance improvement strategies, the scaling rules of large models, as well as the evaluation frameworks.

Acceleration and optimization Hewitt et al. (2023) propose Backpack, a new network architecture that takes all of performance, interpretability and control into consideration. In Backpack, each word in a vocabulary is associated with multiple learned non-contextual sense vectors, and a word in a sequence is represented as a context-dependent, non-negative linear combination of its associated sense vectors. The authors show that a 170M-parameter Backpack LM on OpenWebText has a comparable loss of a 124M parameter GPT-2 small, and, Backpack sense vectors outperform word embeddings of a 6B-parameter Transformer LM on lexical similarity evaluations.

Liu et al. (2023b) propose Sophia, Second-order Clipped Stochastic Optimization, an optimizer with light-weight estimate of the diagonal Hessian as the pre-conditioner to improve the popular, state-of-the-art optimizer Adam. Sophia attains half the number of steps, total compute, and wall-clock time compared with Adam with GPT-2 of sizes from 125M to 770M. The authors also prove theoretical properties of Sophia.

Lin et al. (2023) propose Activation-aware Weight Quantization (AWQ), "a hardware-friendly approach for LLM low-bit weight-only quantization", exploiting the observation that "protecting only 1% of salient weights can greatly reduce quantization error".

Performance improvement Liu and Low (2023) propose a fine-tuned LLaMA-based model Goat to outperform GPT-4 on arithmetic tasks, due to consistent tokenization of numbers by LLaMA. The authors decompose challenging tasks like multi-digit multiplication and division into learnable tasks and leverage basic arithmetic principles. The authors show that Goat-7B can be trained with LoRA on a 24GB VRAM GPU.

Patil et al. (2023) propose a finetuned LLaMA-based model Gorilla to surpass GPT-4 on writing API calls. With a document retriever, Gorilla adapts to document changes like user updates and version changes and mitigates hallucination. The author also introduce APIBench, a dataset including HuggingFace, TorchHub, and TensorHub APIs.

Study of the scaling law Eldan and Li (2023) show that LMs with <10M parameters and one Transformer block can generate fluent and consistent stories of several paragraphs with close to perfect grammar.

Gunasekar et al. (2023) introduce phi-1 and show good coding performance with 1.3B parameters and 7B training tokens, with a selection of "textbook quality" data.

Deshpande et al. (2023) study downscaling effects with the shrunk language, showing the benefits of pre-training for models of 1.25M parameters and that compute-optimal models break the power law. McKenzie et al. (2023) provide 11 datasets for empirical analysis of inverse scaling laws and discuss the important of data and objectives for training LMs. Zhang et al. (2023c) propose NeQA, a dataset containing questions with negation and exhibit inverse scaling, U-shaped scaling, or positive scaling. Before this, the popular view follows scaling laws that the overall cross-entropy loss of an LM improves with the increased scale of model, dataset and compute for training (Kaplan et al., 2020), and that the model and data should be scaled equally for compute-optimal training (Hoffmann et al., 2022).

Evaluation for instruction-following LMs Fairly assessing the performance of instruction-following LMs poses a challenging task, given the extensive variety of tasks it must handle, including

question answering, mathematics problem solving, coding and debugging, translation, and more. Furthermore, assessing the quality of chatbot responses is highly subjective in nature.

Most works in Section 4.3 and 4.4 are evaluated by a few human evaluators on a small sample data. For instance, Alpaca (Taori et al., 2023) is evaluated by five students on around two hundred comparisons against `text-davinci-003`. Koala (Geng et al., 2023) is evaluated by 100+ people on 180 test queries. Open Assistant (Köpf et al., 2023) is evaluated using 7,042 manual comparisons on a sample of 22 prompts.

On the other side, Vicuna (Chiang et al., 2023) employs GPT-4 as a proxy evaluator across 80 questions. This approach gains further support from Guanaco (Dettmers et al., 2023), wherein both GPT-4 and humans are used to evaluate 953 user queries. The comparison demonstrate that GPT-4 evaluations serve as a "cheap and reasonable" substitute for human evaluation.

Evaluation of LMs in general, not just the instruction-following ones, continues to be a significant challenge and an active area of research. We delve deeper into this topic in Section 4.6.

4.6 Evaluation

Evaluation feedback is valuable for researchers and engineers to improve learning algorithms. Evaluation and benchmarks for natural language processing, in particular, language models and interactive applications, have been enjoying steady progress. However, it is still challenging for research and development.

Burnell et al. (2023) present guidelines for robust evaluation practices with more granular reporting, in particular, in-depth performance breakdowns beyond aggregate metrics and instance-by-instance evaluation results.

Gehrmann et al. (2022) survey obstacles in evaluation of test generation and propose to evaluate a model with multiple datasets via multiple metrics and document human evaluation well. The authors propose the following best best practice & implementation: make informed evaluation choices and document them, measure specific generation effects, analyze and address issues in the used dataset(s), evaluate in a comparable setting, run a well-documented human evaluation, produce robust human evaluation results, document results in model cards, and release model outputs and annotations.

Srivastava et al. (2022) propose the Beyond the Imitation Game benchmark (BIG-bench) with more than 200 tasks.

Liang et al. (2022) propose Holistic Evaluation of Language Models (HELM) to improve transparency of LMs, with 1) a taxonomy of LM evaluation design space w.r.t. scenarios and metrics, 2) a broad coverage of 16 core scenarios with 7 metrics, i.e., accuracy, calibration, robustness, fairness, bias, toxicity, efficiency, together with 7 targeted evaluations of skills and risks and 21 new scenarios, and 3) evaluation of 30 existing models.

Lee et al. (2022) propose Human-AI Language-based Interaction Evaluation (HALIE) beyond non-interactive evaluation by considering targets (full process and final output), perspectives (first-person and third-party), and criteria (preference and quality).

Pythia (Biderman et al., 2023) is a suite of 16 LMs with sizes from 70M to 12B parameters and public access to checkpoints for each models to analyze the developments and evolutions of LMs over the course of training.

Shumailov et al. (2023) discuss the issue of model collapse due to training with generated data from LMs and show the importance of genuine human data for LMs.

5 Applying "Mini-Giants" to real-world

"Mini-giants" are uniquely positioned to solve two important issues unaddressed by larger language models: privacy protection and local computation. We examine the application of these smaller models in real-world scenarios, using the therapeutic chatbot Woebot as an example. Cognitive Based Therapy (CBT) took several years from being popular in Woebot, to become closer to clinical ready.

Before delving into the discussion, let's clarify the definition of small language models. Recall that by today's standard, small LMs are the models with parameter sizes of around 10B or lower and with performance comparable or better than ChatGPT / GPT-4. However, this is a definition based on today's technology capabilities. With the development of hardware and other optimization softwares, there will definitely be "mini-giants" with much more network parameters in the future. Therefore, to future-proof our discussion on applications, we use a more extensible definition for a "mini-giant": a language model which can be

trained/modified/used with affordable resources, like with a single GPU and an open source developer today.

Compared to their larger counterparts, “Mini-giants” offer two advantages: privacy protection and computation efficiency. Users wishing to utilize language models have two primary choices. They can either utilize APIs provided by organizations like OpenAI, or build their own "mini-giants". If they choose the former, it is expected that their proprietary data will go through third party’s servers and be logged, which would be unacceptable to sensitive industries such as financial or health care institutes. On the other hand, "Mini-giants" permit centralized user data storage, potentially on a single GPU. For example, “Alpaca-Lora” can run locally on affordable hardware like a Raspberry Pi. In terms of computation efficiency, in industries like autonomous driving, high network latency may occur when connecting to remote data centers. Hence, it’s crucial that the language model can function independently.

To demonstrate "mini-giants" advantages, we examine Cognitive Based Therapy (CBT), an effective technique for treating clinical depression. Moving CBT from casual to clinical use is a demanding process, involving extensive clinical trials. Woebot, an AI chatbot, incorporates CBT into daily use, providing around-the-clock mental health support and anxiety reduction. The company Woebot was founded by Alison Darcy, a psychology student who worked as a software engineer, and then joined Stanford as a postdoctoral researcher in clinical psychology in 2017. Since its establishment, it received endorsement from AI pioneers such as Andrew Ng, who became one of the board of directors in 2017. The chatbot is a popular App with 4.7 rating out of 5, and more than 5,900 reviews in July 2023, and exchanges millions of messages with users every week in 2021 (Steven Loeb, 2021).

However, despite great user reviews, it took more than two years for the company to go through the clinical trials process and get closer to being endorsed by mental health doctors. Woebot first posted their clinical trials recruitment notice on ClinicalTrials.gov in 2019, and designed a process to recruit 101 participants to evaluate whether this chatbot can help in alcohol use disorders etc. It took around 5 months to complete the study in 2020, and the results were first posted in Aug 2022. (Woebot Health, 2022). In 2023, Woebot announced the enrollment of the first patient in a pivotal clinical trial to evaluate if it can help women with postpartum depression (Woebot Health, 2023). Their paper published in Expert Review of Medical Devices (Darcy et al., 2022) documented the clinical trial process.

The reader might ask why it takes such a complicated experimentation process to adopt a new technology in clinical trials and go through the U.S. Food & Drug Administration (FDA) process. The answer is simple. If your families and friends are going to go to a doctor and look for some mental health help, what evidence would you need to decice a chatbot is as trust-worthy as a doctor?

In short, “mini-giants” a.k.a. “small” language models had some unique advantages in privacy protection and computation efficiency. However, their successful integration into specific domains like healthcare requires adherence to industry standards, a frequently long process involving more than just technological considerations.

6 Discussion and outlook

As the capability of large foundation models and AI becomes increasingly well-known to the general public, the demand for AI democracy becomes an issue of societal fairness and equity. In our opinion, the open source community and "small" language models mark one step towards facilitating AI democracy, making it easier for everyone to control, adapt, interpret and afford the power of AI.

- **Adaptability**: For the open source communities including Kaggle, the ability to innovate comes from the capability to use the model in ways that are best suited to domain specific scenarios. Prompt engineering alone is not enough. Thanks to methods mentioned in Section 3, fine-tuning even complex model architectures can mostly be achieved on a single or a few GPUs. Without this, the role of ML researchers without an unimaginable amount of resources risks being diminished to prompt engineers.

- **Controllability**: Being able to choose where to run the model, what data is seen by the model, and what model outputs are used relies heavily on the model being easy enough to run on local infrastructure, and model components are transparent and interpretable. Section 4 listed a wide range of options to select from

for research and/or business use, which leverage the power of large foundation models and at the same time keeps data local. Moreover, with smaller models, users will have a better chance tuning it with instruction following strategies, to further reduce mis-information and ensure the compliance requirements for model outputs. This increases the chance of successful AI application in compliance-demanding domains.

- **Affordability**: Having access to smaller models and cheaper training / fine-tuning options is the only way that privacy-sensitive industries and applications can avoid the trade-off between giving up the right of autonomous data governance, and squandering unreasonable amounts of funds on training gigantic models in-house. As mentioned in Section 5, the affordable option to build domain-specific "small" language models enables industries like finance and healthcare to leverage AI without risking leaking sensitive data to unwarranted third parties. In this sense, lowered costs brought about by these "small" models can prevent the privilege of using AI from falling into the hands of a few exclusive entities.

To sum up, being users of new achievements like GPT-4 is great. Being builders and/or owners of innovations is even better. As technology optimists, the authors believe that it is only through the ability to understand and leverage AI that the society as a whole can mitigate the potential AI risks. With a well designed paradigm, the open source community ans small language models can increase the chance for all to benefit from, and to contribute to, the power of AI.

Acknowledgments

The authors would like to thank Yiyao Liu and Qibin Chen for offering constructive feedback and valuable insights. The authors used ChatGPT (OpenAI, 2022) to edit several sentences in the essay with the following prompt: *Revise to more concise, formal, and fluent, following the style of an academic research paper: [Insert sentence].*

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. MusicLM: Generating music from text. *arXiv*.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022b. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *ACM conference on fairness, accountability, and transparency*.

Yoshua Bengio. 2023. How rogue AIs may arise. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv*.

BigScience Workshop et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv*.

Samuel R. Bowman. 2023. Eight things to know about large language models. *arXiv*.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv*.

Anthony Brohan et al. 2022. RT-1: Robotics transformer for real-world control at scale. *arXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Jacob Browning and Yann LeCun. 2023. AI chatbots don't care about your social norms. https://www.noemamag.com/ai-chatbots-dont-care-about-your-social-norms/.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138.

Mark Chen et al. 2023. Evaluating large language models trained on code. *arXiv*.

Ted Chiang. 2023. Will A.I. become the new McKinsey? https://www.newyorker.com/science/annals-of-artificial-intelligence/will-ai-become-the-new-mckinsey.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv*.

Brian Christian. 2021. *The Alignment Problem: Machine Learning and Human Values*. WW Norton.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. https://tinyurl.com/3v9jss39.

Alison Darcy, Aaron Beaudette, Emil Chiauzzi, Jade Daniels, Kim Goodwin, Timothy Y. Mariano, Paul Wicks, and Athena Robinson. 2022. Anatomy of a woebot® (wb001): agent guided cbt for women with postpartum depression. *Expert Review of Medical Devices*, 19(4):287–301. PMID: 35748029.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. 2023. Neural networks and the chomsky hierarchy. In *ICLR*.

Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin, and Anna Rumshisky. 2023. Honey, i shrunk the language: Language model behavior at reduced scale. In *ACL*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2020. GLM: General language model pretraining with autoregressive blank infilling. In *ACL*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv*.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent english? *arXiv*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv*.

Sharon Goldman. 2023. Top AI researcher dismisses AI 'extinction' fears, challenges 'hero scientist' narrative. https://tinyurl.com/bdd772p5.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv*.

Derrick Harris, Matt Bornstein, and Guido Appenzeller. Ai canon. https://a16z.com/2023/05/25/ai-canon/. Accessed: 2023-07-02.

John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang. 2023. Backpack language models. In *ACL*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. AudioGPT: Understanding and generating speech, music, sound, and talking head. *arXiv*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv*.

Celeste Kidd and Abeba Birhane. 2023. How ai can distort human beliefs. *Science*, 380(6651):1222–1223.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv*.

Sung Kim. 2023. List of open sourced fine-tuned large language models (LLM). https://tinyurl.com/ykf57jd6.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating human-language model interaction. *arXiv*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023a. Code as policies: Language model programs for embodied control. *arXiv*.

Percy Liang et al. 2022. Holistic evaluation of language models. *arXiv*.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023b. TaskMatrix.AI: Completing tasks by connecting foundation models with millions of APIs. *arXiv*.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R. Gormley, and Jason Eisner. 2021. Limitations of autoregressive models and their alternatives. In *NAACL*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *arXiv*.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv*.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. 2023b. Sophia: A scalable stochastic second-order optimizer for language model pretraining. *arXiv*.

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned LLaMA outperforms GPT-4 on arithmetic tasks. *arXiv*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv*.

Gary Marcus. 2023. The sparks of AGI? or the end of science? https://cacm.acm.org/blogs/blog-cacm/271354-the-sparks-of-agi-or-the-end-of-science/fulltext.

John Markoff. 2020. The minds behind the ai 'arms race'. *The New York Times*.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn't better. *arXiv*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *arXiv*.

Melanie Mitchell. 2020. *Artificial Intelligence: A Guide for Thinking Humans*. Picador.

Matt Novak. 2023. Lawyer uses chatgpt in federal court and it goes horribly wrong. https://tinyurl.com/5n7uk84m.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *arXiv*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *arXiv*.

Jack W. Rae et al. 2022. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *JMLR*, 21(140):1–67.

Sebastian Ruder, Jonas Pfeiffer, and Ivan Vulic. 2022. Modular and parameter-efficient fine-tuning for nlp models. In *EMNLP: Tutorial Abstracts*.

Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace. *arXiv*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *arXiv*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *arXiv*.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv*.

Stability AI. 2023a. StableLM: Stability AI language models. https://github.com/stability-AI/stableLM/.

Stability AI. 2023b. StableVicuna: Open-Source RLHF Chatbot. https://stability.ai/blog/stablevicuna-open-source-rlhf-chatbot. Accessed on July 4, 2023.

Steven Loeb. 2021. Woebot CEO Michael Evers on AI in mental health, and how to get a chatbot to bond with a human. https://vator.tv/news/2021-07-30-woebot-ceo-michael-evers-on-ai-in-mental-health-and-how-to-get-a-chatbot-to-bond-with-a-hu Accessed on July 1, 2023.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. AdaPlanner: Adaptive planning from feedback with language models. *arXiv*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Romal Thoppilan et al. 2022. LaMDA: Language models for dialog applications. *arXiv*.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv*.

Bojan Tunguz. 2023. Tweet by bojan tunguz. https://twitter.com/tunguz/status/1673760614576189441?s=20. [Accessed July 1, 2023].

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. 2021. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596.

Roberto Verdecchia, June Sallou, and Luis Cruz. 2023. A systematic review of green AI. *Data Mining Knowledge Discovery*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language model with self generated instructions. In *ACL*.

Woebot Health. 2022. Woebot for Substance Use Disorders . https://classic.clinicaltrials.gov/ct2/show/study/NCT04096001. Accessed on July 1, 2023.

Woebot Health. 2023. Woebot Health Enrolls First Patient in Pivotal Clinical Trial of WB001 for Postpartum Depression . https://woebothealth.com/woebot-health-enrolls-first-patient-in-pivotal-clinical Accessed on July 1, 2023.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. BloombergGPT: A large language model for finance. *arXiv*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *arXiv*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv*.

Yang You. 2023. ColossalChat: An open-source solution for cloning ChatGPT with a complete RLHF pipeline. https://bit.ly/42ZTwW4.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An open bilingual pre-trained model. In *ICLR*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, Lifang He, Brian D. Davison, Quanzheng Li, Yong Chen, Hongfang Liu, and

Lichao Sun. 2023a. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv*.

Shizhuo Dylan Zhang, Curt Tigges, Stella Biderman, Maxim Raginsky, and Talia Ringer. 2023b. Can transformers learn to solve problems recursively? *arXiv*.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023c. Beyond positive scaling: How negation impacts scaling trends of language models. In *ACL*.