GPT4All: An Ecosystem of Open Source Compressed Language Models

Yuvanesh Anand
Nomic AI
yuvanesh@nomic.ai

Zach Nussbaum
Nomic AI
zach@nomic.ai

Adam Treat
Nomic AI
adam@nomic.ai

Aaron Miller
Nomic AI
aaron@nomic.ai

Richard Guo
Nomic AI
richard@nomic.ai

Ben Schmidt
Nomic AI
ben@nomic.ai

GPT4All Community
Planet Earth

Brandon Duderstadt*
Nomic AI
brandon@nomic.ai

Andriy Mulyar*
Nomic AI
andriy@nomic.ai

Abstract

Large language models (LLMs) have recently achieved human-level performance on a range of professional and academic benchmarks. The accessibility of these models has lagged behind their performance. State-of-the-art LLMs require costly infrastructure, are only accessible via rate-limited, geo-locked, and censored web interfaces, and lack publicly available code and technical reports.

In this paper, we tell the story of GPT4All, a popular open source repository that aims to democratize access to LLMs. We outline the technical details of the original GPT4All model family, as well as the evolution of the GPT4All project from a single model into a fully fledged open source ecosystem. It is our hope that this paper acts as both a technical overview of the original GPT4All models as well as a case study on the subsequent growth of the GPT4All open source ecosystem.

1 Introduction

On March 14 2023, OpenAI released GPT-4, a large language model capable of achieving human level performance on a variety of professional and academic benchmarks. Despite the popularity of the release, the GPT-4 technical report (OpenAI, 2023) contained virtually no details regarding the architecture, hardware, training compute, dataset construction, or training method used to create the model. Moreover, users could only access the model through the internet interface at chat.openai.com, which was severely rate limited and unavailable in several locales (e.g. Italy) (BBC News, 2023). Additionally, GPT-4 refused to answer a wide variety of queries, responding only with the now infamous "As an AI Language Model, I cannot..." prefix (Vincent, 2023). These transparency and accessibility concerns spurred several developers to begin creating open source large language model (LLM) alternatives. Several grassroots efforts focused on fine tuning Meta's open code LLaMA model (Touvron et al., 2023; McMillan, 2023), whose weights were leaked on BitTorrent less than a week prior to the release of GPT-4 (Verge, 2023). GPT4All started as one of these variants.

In this paper, we tell the story of GPT4All. We comment on the technical details of the original GPT4All model (Anand et al., 2023), as well as the evolution of GPT4All from a single model to an ecosystem of several models. We remark on the impact that the project has had on the open source community, and discuss future directions. It is our hope that this paper acts as both a technical overview of the original GPT4All models as well as a case study on the subsequent growth of the GPT4All open source ecosystem.

2 The Original GPT4All Model

2.1 Data Collection and Curation

To train the original GPT4All model, we collected roughly one million prompt-response pairs using the GPT-3.5-Turbo OpenAI API between March 20, 2023 and March 26th, 2023. In particular, we gathered GPT-3.5-Turbo responses to prompts of three publicly available datasets: the unified chip2 subset of LAION OIG, a random sub-sample of Stackoverflow Questions, and a sub-sample of Bigscience/P3 (Sanh et al., 2021). Following the approach in Stanford Alpaca (Taori et al., 2023), an open source LLaMA variant that came just before GPT4All, we focused substantial effort on dataset curation.

The collected dataset was loaded into Atlas (Nomic, 2023)—a visual interface for exploring and tagging massive unstructured datasets —for data curation. Using At-

* Shared Senior Authorship

las, we identified and removed subsets of the data where GPT-3.5-Turbo refused to respond, had malformed output, or produced a very short response. This resulted in the removal of the entire Bigscience/P3 subset of our data, as many P3 prompts induced responses that were simply one word. After curation, we were left with a set of 437,605 prompt-response pairs, which we visualize in Figure 1a.

2.2 Model Training

The original GPT4All model was a fine tuned variant of LLaMA 7B. In order to train it more efficiently, we froze the base weights of LLaMA, and only trained a small set of LoRA (Hu et al., 2021) weights during the fine tuning process. Detailed model hyper-parameters and training code can be found in our associated code repository[1].

2.3 Model Access

We publicly released all data, training code, and model weights for the community to build upon. Further, we provided a 4-bit quantized version of the model, which enabled users to run it on their own commodity hardware without transferring data to a 3rd party service.

Our research and development costs were dominated by ∼$800 in GPU spend (rented from Lambda Labs and Paperspace) and ∼$500 in OpenAI API spend. Our final GPT4All model could be trained in about eight hours on a Lambda Labs DGX A100 8x 80GB for a total cost of ∼$100.

2.4 Model Evaluation

We performed a preliminary evaluation of our model using the human evaluation data from the Self Instruct paper (Wang et al., 2023). We reported the ground truth perplexity of our model against what was, to our knowledge, the best openly available alpaca-lora model at the time, provided by user *chainyo* on HuggingFace. Both models had very large perplexities on a small number of tasks, so we reported perplexities clipped to a maximum of 100. We found that GPT4All produces stochastically lower ground truth perplexities than alpaca-lora (Anand et al., 2023).

3 From a Model to an Ecosystem

3.1 GPT4All-J: Repository Growth and the implications of the LLaMA License

The GPT4All repository grew rapidly after its release, gaining over 20000 GitHub stars in just one week, as shown in Figure 2. This growth was supported by an in-person hackathon hosted in New York City three days after the model release, which attracted several hundred participants. As the Nomic discord, the home of online discussion about GPT4All, ballooned to over 10000 people, one thing became very clear - there was massive demand for a model that could be used commercially.

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

The LLaMA model that GPT4All was based on was licensed for research only, which severely limited the set of domains that GPT4All could be applied in. As a response to this, the Nomic team repeated the model training procedure of the original GPT4All model, but based on the already open source and commercially licensed GPT-J model (Wang and Komatsuzaki, 2021). GPT4All-J also had an augmented training set, which contained multi-turn QA examples and creative writing such as poetry, rap, and short stories. The creative writing prompts were generated by filling in schemas such as "Write a [CREATIVE STORY TYPE] about [NOUN] in the style of [PERSON]." We again employed Atlas to curate the prompt-response pairs in this data set.

Our evaluation methodology also evolved as the project grew. In particular, we began evaluating GPT4All models using a suite of seven reasoning tasks that were used for evaluation of the Databricks Dolly (Conover et al., 2023b) model, which was released on April 12, 2023. Unfortunately, GPT4All-J did not outperform other prominent open source models on this evaluation. As a result, we endeavoured to create a model that did.

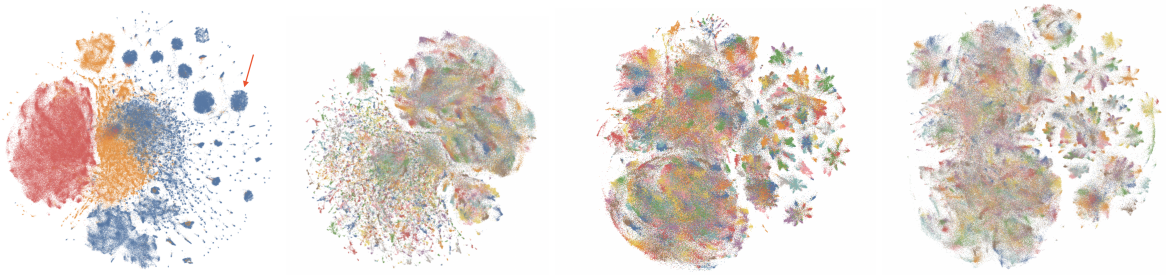
3.2 GPT4All-Snoozy: the Emergence of the GPT4All Ecosystem

GPT4All-Snoozy was developed using roughly the same procedure as the previous GPT4All models, but with a few key modifications. First, GPT4All-Snoozy used the LLaMA-13B base model due to its superior base metrics when compared to GPT-J. Next, GPT4All-Snoozy incorporated the Dolly's training data into its train mix. After data curation and deduplication with Atlas, this yielded a training set of 739,259 total prompt-response pairs. We dubbed the model that resulted from training on this improved dataset GPT4All-Snoozy. As shown in Figure 1, GPT4All-Snoozy had the best average score on our evaluation benchmark of any model in the ecosystem at the time of its release.

Concurrently with the development of GPT4All, several organizations such as LMSys, Stability AI, BAIR, and Databricks built and deployed open source language models. We heard increasingly from the community that they wanted quantized versions of these models for local use. As we realized that organizations with ever more resources were developing source language models, we decided to pivot our effort away from training increasingly capable models and towards providing easy access to the plethora of models being produced by the open source community. Practically, this meant spending our time compressing open source models for use on commodity hardware, providing stable and simple high level model APIs, and supporting a GUI for no code model experimentation.

3.3 The Current State of GPT4All

Today, GPT4All is focused on improving the accessibility of open source language models. The repository



(a) (b) (c) (d)

Figure 1: TSNE visualizations showing the progression of the GPT4All train set. Panel (a) shows the original uncurated data. The red arrow denotes a region of highly homogeneous prompt-response pairs. The coloring denotes which open dataset contributed the prompt. Panel (b) shows the original GPT4All data after curation. This panel, as well as panels (c) and (d) are 10 colored by topic, which Atlas automatically extracts. Notice that the large homogeneous prompt-response blobs no longer appearl. Panel (c) shows the GPT4All-J dataset. The "starburst" clusters introduced on the right side of the panel correspond to the newly added creative data. Panel (d) shows the final GPT4All-snoozy dataset. All datasets have been released to the public, and can be interactively explored online. In the web version of this article, you can click on a panel to be taken to its interactive visualization.

Model	BoolQ	PIQA	HellaSwag	WinoG.	ARC-e	ARC-c	OBQA	Avg.
GPT4All-J 6B v1.0*	73.4	74.8	63.4	64.7	54.9	36	40.2	58.2
GPT4All-J v1.1-breezy*	74	75.1	63.2	63.6	55.4	34.9	38.4	57.8
GPT4All-J v1.2-jazzy*	74.8	74.9	63.6	63.8	56.6	35.3	41	58.6
GPT4All-J v1.3-groovy*	73.6	74.3	63.8	63.5	57.7	35	38.8	58.1
GPT4All-J Lora 6B*	68.6	75.8	66.2	63.5	56.4	35.7	40.2	58.1
GPT4All LLaMa Lora 7B*	73.1	77.6	72.1	67.8	51.1	40.4	40.2	60.3
GPT4All 13B snoozy*	83.3	79.2	75	71.3	60.9	44.2	43.4	65.3
GPT4All Falcon	77.6	79.8	74.9	70.1	67.9	43.4	42.6	65.2
Nous-Hermes (Nous-Research, 2023b)	79.5	78.9	80	71.9	74.2	50.9	**46.4**	68.8
Nous-Hermes2 (Nous-Research, 2023c)	**83.9**	**80.7**	80.1	71.3	75.7	**52.1**	46.2	**70.0**
Nous-Puffin (Nous-Research, 2023d)	81.5	**80.7**	**80.4**	**72.5**	**77.6**	50.7	45.6	69.9
Dolly 6B* (Conover et al., 2023a)	68.8	77.3	67.6	63.9	62.9	38.7	41.2	60.1
Dolly 12B* (Conover et al., 2023b)	56.7	75.4	71	62.2	64.6	38.5	40.4	58.4
Alpaca 7B* (Taori et al., 2023)	73.9	77.2	73.9	66.1	59.8	43.3	43.4	62.5
Alpaca Lora 7B* (Wang, 2023)	74.3	79.3	74	68.8	56.6	43.9	42.6	62.8
GPT-J* 6.7B (Wang and Komatsuzaki, 2021)	65.4	76.2	66.2	64.1	62.2	36.6	38.2	58.4
LLama 7B* (Touvron et al., 2023)	73.1	77.4	73	66.9	52.5	41.4	42.4	61.0
LLama 13B* (Touvron et al., 2023)	68.5	79.1	76.2	70.1	60	44.6	42.2	63.0
Pythia 6.7B* (Biderman et al., 2023)	63.5	76.3	64	61.1	61.3	35.2	37.2	56.9
Pythia 12B* (Biderman et al., 2023)	67.7	76.6	67.3	63.8	63.9	34.8	38	58.9
Fastchat T5* (Zheng et al., 2023)	81.5	64.6	46.3	61.8	49.3	33.3	39.4	53.7
Fastchat Vicuña* 7B (Zheng et al., 2023)	76.6	77.2	70.7	67.3	53.5	41.2	40.8	61.0
Fastchat Vicuña 13B* (Zheng et al., 2023)	81.5	76.8	73.3	66.7	57.4	42.7	43.6	63.1
StableVicuña RLHF* (Stability-AI, 2023)	82.3	78.6	74.1	70.9	61	43.5	44.4	65.0
StableLM Tuned* (Stability-AI, 2023)	62.5	71.2	53.6	54.8	52.4	31.1	33.4	51.3
StableLM Base* (Stability-AI, 2023)	60.1	67.4	41.2	50.1	44.9	27	32	46.1
Koala 13B* (Geng et al., 2023)	76.5	77.9	72.6	68.8	54.3	41	42.8	62.0
Open Assistant Pythia 12B*	67.9	78	68.1	65	64.2	40.4	43.2	61.0
Mosaic MPT7B (MosaicML-Team, 2023)	74.8	79.3	76.3	68.6	70	42.2	42.6	64.8
Mosaic mpt-instruct (MosaicML-Team, 2023)	74.3	80.4	77.2	67.8	72.2	44.6	43	65.6
Mosaic mpt-chat (MosaicML-Team, 2023)	77.1	78.2	74.5	67.5	69.4	43.3	44.2	64.9
Wizard 7B (Xu et al., 2023)	78.4	77.2	69.9	66.5	56.8	40.5	42.6	61.7
Wizard 7B Uncensored (Xu et al., 2023)	77.7	74.2	68	65.2	53.5	38.7	41.6	59.8
Wizard 13B Uncensored (Xu et al., 2023)	78.4	75.5	72.1	69.5	57.5	40.4	44	62.5
GPT4-x-Vicuna-13b (Nous-Research, 2023a)	81.3	75	75.2	65	58.7	43.9	43.6	63.2
Falcon 7b (Almazrouei et al., 2023)	73.6	**80.7**	76.3	67.3	71	43.3	44.4	65.2
Falcon 7b instruct (Almazrouei et al., 2023)	70.9	78.6	69.8	66.7	67.9	42.7	41.2	62.5
text-davinci-003	88.1	83.8	83.4	75.8	83.9	63.9	51.0	75.7

Table 1: Evaluations of all language models in the GPT4All ecosystem as of August 1, 2023. Code models are not included. OpenAI's text-davinci-003 is included as a point of comparison. The best overall performing model in the GPT4All ecosystem, Nous-Hermes2, achieves over 92% of the average performance of text-davinci-003. Models marked with an asterisk were available in the ecosystem as of the release of GPT4All-Snoozy. Note that at release, GPT4All-Snoozy had the best average performance of any model in the ecosystem. Bolded numbers indicate the best performing model as of August 1, 2023.

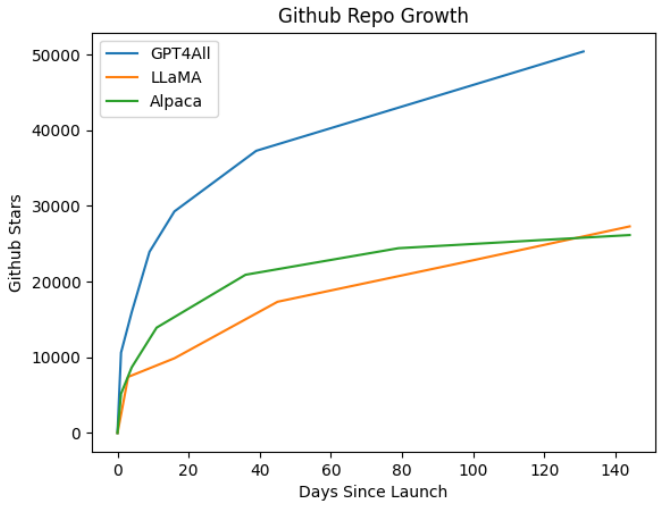


Figure 2: Comparison of the github start growth of GPT4All, Meta's LLaMA, and Stanford's Alpaca. We conjecture that GPT4All achieved and maintains faster ecosystem growth due to the focus on access, which allows more users to meaningfully participate.

provides compressed versions of open source models for use on commodity hardware, stable and simple high level model APIs, and a GUI for no code model experimentation. The project continues to increase in popularity, and as of August 1 2023, has garnered over 50000 GitHub stars and over 5000 forks.

GPT4All currently provides native support and benchmark data for over 35 models (see Figure 1), and includes several models co-developed with industry partners such as Replit and Hugging Face. GPT4All also provides high level model APIs in languages including Python, Typescript, Go, C#, and Java, among others. Furthermore, the GPT4All no code GUI currently supports the workflows of over 50000 monthly active users, with over 25% of users coming back to the tool every day of the week. (Note that all GPT4All user data is collected on an *opt in* basis.) GPT4All has become the top language model integration in the popular open source AI orchestration library LangChain (Chase, 2022), and powers many popular open source projects such as PrivateGPT (imartinez, 2023), Quiver (StanGirard, 2023), and MindsDB (MindsDB, 2023), among others. GPT4All is the 3rd fastest growing GitHub repository of all time (Leo, 2023), and is the 185th most popular repository on the platform, by star count.

4 The Future of GPT4All

In the future, we will continue to grow GPT4All, supporting it as the de facto solution for LLM accessibility. Concretely, this means continuing to compress and distribute important open-source language models developed by the community, as well as compressing and distributing increasingly multimodal AI models. Furthermore, we will expand the set of hardware devices that GPT4All models run on, so that GPT4All models "just work" on any machine, whether it comes equipped with Apple Metal silicon, NVIDIA, AMD, or other edge-accelerated hardware. Overall, we envision a world where anyone, anywhere, with any machine, can access and contribute to the cutting edge of AI.

Limitations

By enabling access to large language models, the GPT4All project also inherits many of the ethical concerns associated with generative models. Principal among these is the concern that unfiltered language models like GPT4All enable malicious users to generate content that could be harmful and dangerous (e.g., instructions on building bioweapons). While we recognize this risk, we also acknowledge the risk of concentrating this technology in the hands of a limited number of increasingly secretive research groups. We believe that the risk of concentrating the benefits of language model technology in the hands of a small number of people significantly outweighs the risk of misuse, and hence we prefer to make the technology as widely available as possible.

Finally, we realize the challenge in assigning credit for large-scale open source initiatives. We make a first attempt at fair credit assignment by explicitly including the GPT4All open source developers as authors on this work, but recognize that this is insufficient fully characterize everyone involved in the GPT4All effort. Furthermore, we acknowledge the difficulty in citing open source works that do not necessarily have standardized citations, and do our best in this paper to provide URLs to projects whenever possible. We encourage further research in the area of open source credit assignment, and hope to be able to support some of this research ourselves in the future.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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.

BBC News. 2023. Chatgpt banned in italy over privacy concerns. *BBC News*.

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.

Harrison Chase. 2022. langchain. https://github.com/langchain-ai/langchain.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. 2023a. Hello dolly: Democratizing the magic of chatgpt with open models.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023b. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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.

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.

imartinez. 2023. privategpt. https://github.com/imartinez/privateGPT.

Oscar Leo. 2023. GitHub: The Fastest Growing Repositories of All Time.

Robert McMillan. 2023. A meta platforms leak put powerful ai in the hands of everyone. *The Wall Street Journal*.

MindsDB. 2023. Mindsdb. https://github.com/mindsdb/mindsdb. GitHub repository.

MosaicML-Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-08-07.

Nomic. 2023. Atlas. https://atlas.nomic.ai/.

Nous-Research. 2023a. gpt4-x-vicuna-13b. https://huggingface.co/NousResearch/gpt4-x-vicuna-13b. Model on Hugging Face.

Nous-Research. 2023b. Nous-hermes-13b. https://huggingface.co/NousResearch/Nous-Hermes-13b. Model on Hugging Face.

Nous-Research. 2023c. Nous-hermes-llama-2-7b. https://huggingface.co/NousResearch/Nous-Hermes-llama-2-7b. Model on Hugging Face.

Nous-Research. 2023d. Redmond-puffin-13b. https://huggingface.co/NousResearch/Redmond-Puffin-13B. Model on Hugging Face.

OpenAI. 2023. Gpt-4 technical report.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Stability-AI. 2023. Stablelm. https://github.com/Stability-AI/StableLM. GitHub repository.

StanGirard. 2023. quivr. https://github.com/StanGirard/quivr. GitHub repository.

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.

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.

The Verge. 2023. Meta's powerful ai language model has leaked online — what happens now? *The Verge*.

James Vincent. 2023. As an ai generated language model: The phrase that shows how ai is polluting the web. *The Verge*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Eric J. Wang. 2023. alpaca-lora. https://github.com/tloen/alpaca-lora. GitHub repository.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.