ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?

Hailin Chen[*1,2], **Fangkai Jiao**[*1,3], **Xingxuan Li**[*1], **Chengwei Qin**[*1], **Mathieu Ravaut**[*1,3], **Ruochen Zhao**[*1], **Caiming Xiong**[2], **Shafiq Joty**[1,2]

[1] Nanyang Technological University, Singapore
[2] Salesforce Research
[3] Institute of Infocomm Research (I^{2}R), A*STAR, Singapore
{hailin001, fangkai002, xingxuan001, chengwei003}@e.ntu.edu.sg
{mathieuj001, ruochen002}@e.ntu.edu.sg
{cxiong, sjoty}@salesforce.com

Abstract

Upon its release in late 2022, ChatGPT has brought a seismic shift in the entire landscape of AI, both in research and commerce. Through instruction-tuning a large language model (LLM) with supervised fine-tuning and reinforcement learning from human feedback, it showed that a model could answer human questions and follow instructions on a broad panel of tasks. Following this success, interests in LLMs have intensified, with new LLMs flourishing at frequent interval across academia and industry, including many start-ups focused on LLMs. While closed-source LLMs (e.g., OpenAI's GPT, Anthropic's Claude) generally outperform their open-source counterparts, the progress on the latter has been rapid with claims of achieving parity or even better on certain tasks. This has crucial implications not only on research but also on business. In this work, on the first anniversary of ChatGPT, we provide an exhaustive overview of this success, surveying all tasks where an open-source LLM has claimed to be on par or better than ChatGPT.

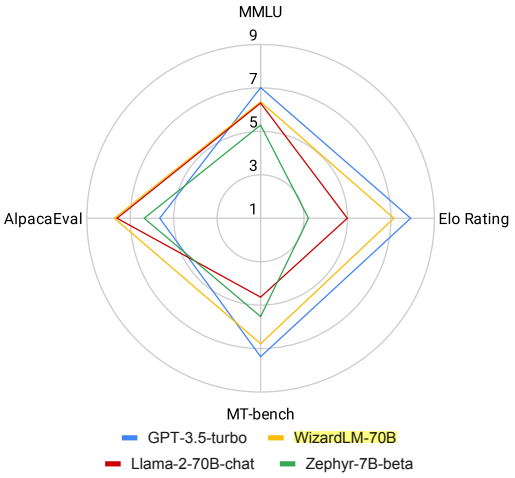
1 Introduction

Exactly one year ago, the release of ChatGPT by OpenAI took the AI community and the broader world by storm. For the first time, an application-based AI chatbot could generally provide helpful, safe and detailed answers to most questions, follow instructions, and even admit and fix its previous mistakes. Notably, it can perform these natural language tasks which were traditionally done by pre-trained then tailored fine-tuned language models such as summarization or question-answering (QA), seemingly amazingly well. As a first of its kind, ChatGPT has attracted the general public – it reached 100 million users within just two months of its launch, way faster than other popular apps like TikTok or YouTube.[1] It has also attracted huge business investments, for its potential to cut down labor cost, automate workflows and even bring new experiences to customers (Cheng et al., 2023).

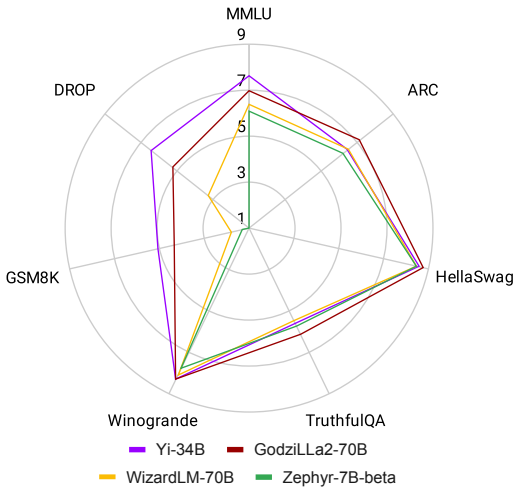
However, since ChatGPT is not open-sourced and its access is controlled by a private company, most of its technical details remain unknown. Despite the claim that it follows the procedure introduced in InstructGPT (also called GPT-3.5) (Ouyang et al., 2022b), its exact architecture, pre-training data and fine-tuning data are unknown. Such close-source nature generates several key issues. First,

[*] Authors contributed equally and are ranked by alphabetical order.
[1] https://www.reuters.com/technology/

Preprint. Under review.



(a) Open-source LLMs vs. ChatGPT on MT-bench Zheng et al. (2023), ELO Rating, MMLU (Hendrycks et al., 2021b) and AlpacaEval Li et al. (2023d).



(b) Selected open-source LLMs on HuggingFace's Open LLM leaderboard.

Figure 1: Overview of different open-source LLMs on various general benchmarks.

without knowing the internal details such as the pretraining and finetuning procedure, it is hard to properly estimate its potential risks to the society, especially knowing that LLMs can notoriously generate toxic, unethical and untruthful content. Second, it has been reported that ChatGPT's performance changes over time hindering reproducible results (Chen et al., 2023d). Third, ChatGPT has experienced multiple outages, with two major ones only in November 2023 during which the access to ChatGPT website and its API was completely blocked. Finally, enterprises adopting ChatGPT may be concerned with the heavy cost of calling APIs, service outages, data ownership and privacy issues, and other unpredictable events such as the recent boardroom drama about the CEO Sam Altman's dismissal to staff rebellion, and his eventual return (REUTERS source).

Open-source LLMs, on the other hand, offer a promising direction as they can potentially remediate or bypass most of the aforementioned issues. For this reason, the research community has been actively pushing for maintaining high-performing LLMs in open-source. However, as it stands today (as of late 2023), it is widely believed that open-source LLMs such as Llama-2 (Touvron et al., 2023b) or Falcon (Almazrouei et al., 2023) lag behind their closed-source counterparts such as OpenAI's GPT3.5 (ChatGPT) and GPT-4 (OpenAI, 2023b), Anthropic's Claude[2] or Google's Bard[3], with GPT-4 generally assumed to champion them all. However, what is very encouraging is that the gap is getting narrower and narrower, and open-source LLMs are quickly catching up. In fact, as it is illustrated in Figure 1, the best open-source LLMs already perform better than GPT-3.5-turbo on some standard benchmarks. Yet, it is not a straightforward uphill battle for open-source LLMs. The landscape is constantly evolving: closed-source LLMs are updated by retraining on newer data regularly, open-source LLMs are released to catch up, and there is a myriad of evaluation datasets and benchmarks being used to compare LLMs, making singling out a best LLM especially challenging.

In this survey, we aim to consolidate recent studies on open-source LLMs and provide an overview of that match or surpass ChatGPT in various domains. Our contributions are three-fold:

- Consolidating various evaluations of open-source LLMs, providing an unbiased and comprehensive view of open-source LLMs vs. ChatGPT (Figure 1, Section 3.1).
- Systematically reviewing open-source LLMs that match or surpass the performance of ChatGPT in various tasks with analysis (Figure 2, Section 3, Section 4.2). We are also maintaining a live web page to track the latest updates.[4]

[2]https://www.anthropic.com/index/introducing-claude

[3]https://blog.google/technology/ai/bard-google-ai-search-updates

[4]https://github.com/ntunlp/OpenSource-LLMs-better-than-OpenAI/tree/main

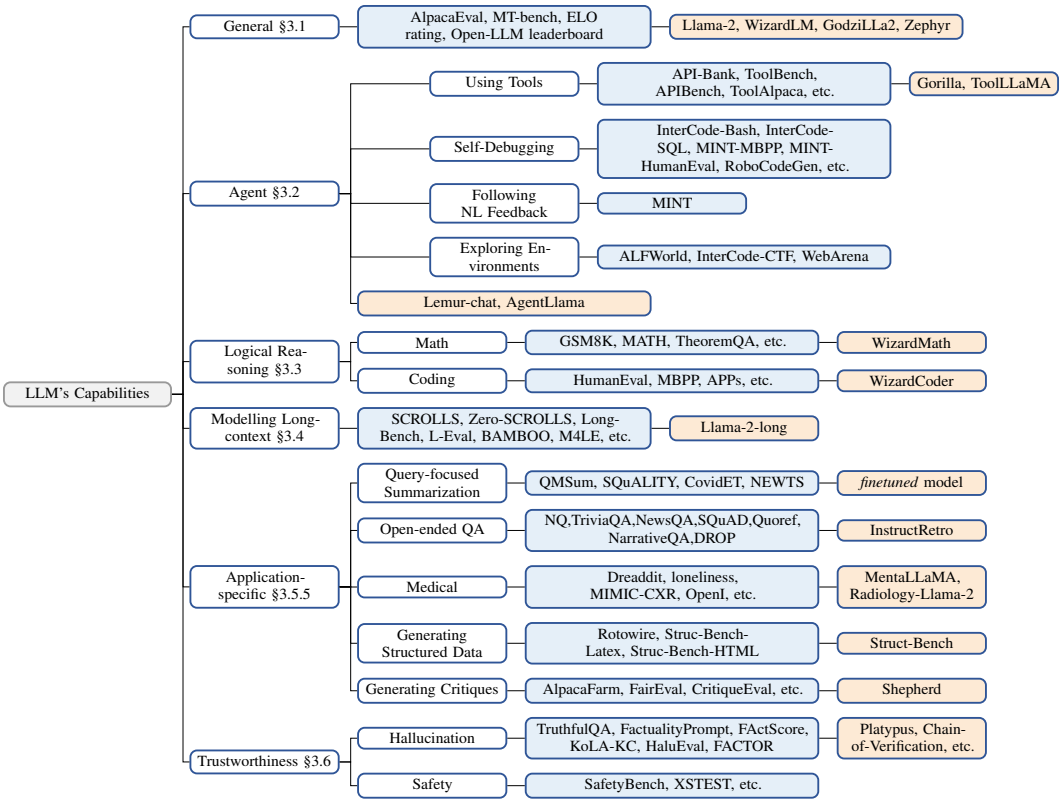


Figure 2: Typology of LLM's capabilities and best performing open-LLMs. White boxes denote domains, blue boxes represent specific datasets and orange boxes denote open-sourced LLMs.

- Presenting insights on the trend of open-source LLMs development (Section 4.1), the good practices to train open-source LLMs (Section 4.3) and potential issues with open-source LLMs (Section 4.4).

Who can benefit from this survey? This survey aims to serve as a pivotal resource for both the research community and business sector in understanding the current landscape and future potential of open-source LLMs. For researchers, it provides a detailed synthesis of the current progress and evolving trends in open-source LLMs, highlighting promising directions for future investigation. For the business sector, this survey offers valuable insights and guidance, assisting decision-makers in evaluating the applicability and benefits of adopting open-source LLMs.

In the following, we start by introducing background concepts (Section 2), then provide an in-depth review of open-source LLMs that beat ChatGPT in various domains (Section 3), followed by a discussion on insights and issues of open-source LLMs (Section 4), finally we conclude with a summary (Section 5).

2 Background

In this section, we briefly describe the fundamental concepts that relate to LLMs.

2.1 Training Regimes

Pre-training All LLMs rely on large-scale self-supervised pre-training on Internet text data (Radford et al., 2018; Brown et al., 2020). Decoder-only LLMs follow the causal language modeling objective, through which the model learns to predict the next token conditioning on the sequence of previous tokens (Bengio et al., 2000). As per pre-training details shared by open-source LLMs

(Touvron et al., 2023a), sources of text data include CommonCrawl[5], C4 (Raffel et al., 2020), GitHub, Wikipedia, books, and online discussion exchanges such as Reddit or StackOverFlow. It is widely acknowledged that scaling the size of pre-training corpus improves the model performance, and works hand-in-hand with scaling the model size, a phenomenon referred to as *scaling laws*, and analyzed in depth in (Hoffmann et al., 2022a). Modern-day LLMs pre-train on a corpus from hundreds of billions to several trillions of tokens (Touvron et al., 2023b; Penedo et al., 2023).

Fine-tuning Fine-tuning aims to adapt a pre-trained LLM to downstream tasks, by updating weights with the available supervision, which usually forms a dataset orders of magnitude smaller than the one used for pre-training (Devlin et al., 2018). T5 (Raffel et al., 2020) was among the first to frame fine-tuning into a text-to-text unified framework, with natural language instructions describing each task. **Instruction-tuning** later extended fine-tuning by training jointly on several tasks (Wei et al., 2021a; Aribandi et al., 2021), each described with natural language instructions. Instruction-tuning quickly gained in popularity, due to its ability to drastically improve zero-shot performance of LLMs, including on new tasks (unseen during training), and especially at larger models scale. Standard instruction-tuning with multi-task supervised fine-tuning (commonly known as **SFT**) may still not result in models that follow humans intentions while being safe, ethical and harmless, and can be futher improved with Reinforcement Learning from Human Feedback (**RLHF**): human annotators rank outputs from the fine-tuned model, which are used to fine-tune again with reinforcement learning (Ouyang et al., 2022b). Recent work showed that human feedback may be replaced with feedback from an LLM, a process referred to as Reinforcement Learning from AI Feedback (**RLAIF**) Bai et al. (2022b). Direct Preference Optimization (**DPO**) bypasses the need to fit a reward model to human preferences as in RLHF and instead directly fine-tunes the policy with a cross-entropy objective, achieving more efficient alignment of the LLM to human preferences.

A line of work focuses on *quality over quantity* when building an instruction-tuning dataset of diverse tasks: Lima (Zhou et al., 2023a) outperforms GPT-3 with a Llama-65B fine-tuned on just 1,000 examples, and Alpagasus (Chen et al., 2023c) improves on Alpaca (Taori et al., 2023) by cleaning its instruction fine-tuning dataset from 52k to 9k examples.

Continual pre-training Continual pre-training consists in performing another round of pre-training from a pre-trained LLM, typically with a lesser volume of data than in the first stage. Such process may be useful to quickly adapt to a new domain or elicit new properties in the LLM. For instance, continual pre-training is used in Lemur (Xu et al., 2023d) to improve coding and reasoning capacities, and Llama-2-long (Xiong et al., 2023) to extend context window.

Inference There exists several alternative methods for sequence generation with auto-regressive decoding with an LLM, which differ by the degree of randomness and diversity in the output. Increasing the temperature during sampling makes outputs more diverse, while setting it to 0 falls back to greedy decoding, which may be needed in scenarios necessitating deterministic outputs. Sampling methods top-k (Fan et al., 2018) and top-p (Holtzman et al., 2019) constrain the pool of tokens to sample from at each decoding step.

Several techniques aim to improve inference speed, especially at longer sequence length, which become problematic due to the attention complexity, which is quadratic with regards to input length. FlashAttention (Dao et al., 2022) optimizes reads/writes between levels of GPU memory, accelerating both training and inference. FlashDecoding (Dao et al., 2023) parallelizes the key-value (KV) cache loading in the attention mechanism, yielding a 8x end-to-end speedup. Speculative decoding (Leviathan et al., 2023; Chen et al., 2023b) uses an extra, small language model to approximate next token distribution from an LLM, which accelerates decoding without loss of performance. vLLM (Kwon et al., 2023) accelerates LLM inference and serving using PagedAttention, an algorithm for optimizing memory usage of attention keys and values.

2.2 Task Domains and Evaluation

Properly assessing the capabilities of LLMs remains an active research area, due to the diversity and breadth of evaluations to perform. Question-answering datasets (Joshi et al., 2017; Kwiatkowski et al., 2019; Lin et al., 2022) are very popular evaluation benchmarks, but new benchmarks tailored

[5] https://commoncrawl.org

for LLM assessments have also emerged recently (Dubois et al., 2023; Beeching et al., 2023; Zheng et al., 2023). In the following section, we explore LLMs capacities across 6 main dimensions: general capabilities, agent capabilities, logical reasoning (including maths and coding capacities), long-context modelling, specific applications such as QA or summarization, and trustworthiness.

3 Open-Source LLMs vs. ChatGPT

3.1 General Capabilities

Benchmarks As numerous LLMs are released week upon week, each claiming superior performance on certain tasks, it becomes increasingly challenging to identify true advancements and the leading models. Therefore, it is crucial to comprehensively assess the performance of these models across a broad spectrum of tasks to understand their general capabilities. This section covers benchmarks using LLM-based (*e.g.,* GPT-4) evaluation and traditional (*e.g.,* ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)) evaluation metrics.

- **MT-Bench** (Zheng et al., 2023) is designed to test multi-turn conversation and instruction-following ability from eight perspectives: writing, roleplay, information extraction, reasoning, math, coding, knowledge I (STEM), and knoweldge II (humanities/social science). Stronger LLMs (*e.g.,* GPT-4) are utilized as judges to evaluate the models for this benchmark.
- **AlpacaEval** (Li et al., 2023d) is an LLM-based automatic evaluator based on AlpacaFarm (Dubois et al., 2023) evaluation set, which tests the ability of models to follow general user instructions. It benchmarks candidate models against Davinci-003 responses utilizing stronger LLMs (*e.g.,* GPT-4 and Claude), which generate the candidate model's win rate.
- **Open LLM Leaderboad** (Beeching et al., 2023) evaluates LLMs on seven key benchmarks using the Language Model Evaluation Harness (Gao et al., 2021), including AI2 Reasoning Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021b), TruthfulQA (Lin et al., 2022), Winogrande (Sakaguchi et al., 2019), GSM8K (Cobbe et al., 2021), and DROP (Dua et al., 2019). This framework evaluates LLMs on a variety of reasoning and general knowledge across a wide variety of fields in zero-shot and few-shot settings.
- **BIG-bench** (bench authors, 2023) is a collaborative benchmark aimed to probe LLMs and extrapolate their future capabilities. It includes more than 200 novel language tasks, covering a diverse range of topics and languages, which are not entirely solvable by existing models.
- **ChatEval** (Chan et al., 2023) is a multi-agent debate framework, which enables a multi-agent referee team to autonomously discuss and evaluate the quality of generated responses from different models on open-ended questions and traditional natural language generation tasks.
- **FairEval-Vicuna** (Wang et al., 2023b) utilizes both multiple evidence calibration and balanced position calibration on a set of 80 questions from the Vicuna Benchmark (Zheng et al., 2023). FairEval-Vicuna offers a more impartial evaluation outcome within the paradigm of adopting LLMs as evaluators, which closely aligns with human judgements.

Performance of LLMs Llama-2-70B (Touvron et al., 2023b), a prominent open-source LLM from meta, has been pre-trained on a massive dataset of two trillion tokens. It demonstrates remarkable results across various general benchmarks. When further fine-tuned with instruction data, the Llama-2-chat-70B variant exhibits enhanced capabilities in general conversational tasks. In particular, Llama-2-chat-70B achieves a 92.66% win rate in AlpacaEval, surpassing the performance of GPT-3.5-turbo by 10.95%. Nonetheless, GPT-4 remains the top performer among all LLMs with a win rate of 95.28%.

Models	MT-Bench	AlpacaEval	Open LLM Leaderboard
Llama-2-70B-chat	6.86	92.66	-
WizardLM-70B	7.71	92.91	57.17
GodziLLa2-70B	-	-	67.01
Zephyr-7B	7.34	90.60	52.15
Yi-34B	-	-	68.68
GPT-3.5-turbo	7.94	81.71	70.21
GPT-4	8.99	95.28	85.36

Table 1: Model performance on general benchmarks.

Zephyr-7B (Tunstall et al., 2023), another smaller model, uses distilled direct preference optimization (Rafailov et al., 2023a) and achieves comparable results to 70B LLMs on AlpacaEval with a win rate of 90.6%. It even surpasses Llama-2-chat-70B on MT-Bench, scoring 7.34 against 6.86. Additionally,

WizardLM-70B (Xu et al., 2023a) has been instruction fine-tuned using large amounts of instruction data with varying levels of complexity. It stands out as the highest-scoring open-sourced LLM on MT-Bench with a score of 7.71. However, this is still slightly lower than the scores of GPT-3.5-turbo (7.94) and GPT-4 (8.99). Although Zephyr-7B shows top performance in the MT-Bench, it falls short in the Open LLM Leaderboard, scoring only 52.15%. On the other hand, GodziLLa2-70B (Philippines, 2023), an experimental model that combines various proprietary LoRAs from Maya Philippines [6] and the Guanaco Llama 2 1K dataset (mlabonne, 2023) with Llama-2-70B, achieves a more competitive score of 67.01% on the Open LLM Leaderboard. Furthermore, Yi-34B pre-trained from scratch by developers at 01.AI [7], stands out among all open-source LLMs with a remarkable score of 68.68%. This performance is comparable to that of GPT-3.5-turbo, which scores 70.21%. However, both are still notably behind GPT-4, which leads with a substantial score of 85.36%. UltraLlama (Ding et al., 2023) utilizes fine-tuning data with enhanced diversity and quality. It matches GPT-3.5-turbo's performance in its proposed benchmark while exceeding it in areas of world and professional knowledge.

3.2 Agent Capabilities

Benchmarks With the recent advancements in scaling up model size, LLM-based agents (also called *language agents*) have drawn a great deal of attention from the NLP community. In light of this, we investigate the agent capabilities of open-source LLMs on a variety of benchmarks. Depending on the skills required, existing benchmarks can be mainly divided into four categories.

- *Using Tools*: Some benchmarks have been proposed to evaluate the tool usage capabilities of LLMs. **API-Bank** (Li et al., 2023b) is specifically designed for tool-augmented LLMs. **ToolBench** (Xu et al., 2023c) is a tool manipulation benchmark including various software tools for real-world tasks. **APIBench** (Patil et al., 2023) consists of APIs from HuggingFace, TorchHub, and TensorHub. **ToolAlpaca** (Tang et al., 2023a) develops a diverse and comprehensive tool-use dataset through a multi-agent simulation environment. Coincidentally, another instruction-tuning dataset constructed using ChatGPT for tool use is also named **ToolBench** (Qin et al., 2023b). Besides, **MINT** (Wang et al., 2023d) can evaluate the proficiency of LLMs in employing tools to solve tasks that necessitate multi-turn interactions.
- *Self-Debugging*: Several datasets are available to assess the ability of LLMs to self-debug, including **InterCode-Bash** and **InterCode-SQL** (Yang et al., 2023b), **MINT-MBPP** and **MINT-HumanEval** (Wang et al., 2023d), and **RoboCodeGen** (Liang et al., 2023).
- *Following Natural Language Feedback*: **MINT** (Wang et al., 2023d) can also be used to measure the ability of LLMs to leverage natural language feedback by using GPT-4 (OpenAI, 2023b) to simulate human users.
- *Exploring Environment*: **ALFWorld** (Shridhar et al., 2020), **InterCode-CTF** (Yang et al., 2023b), and **WebArena** (Zhou et al., 2023c) are introduced to evaluate whether LLMs-based agents are able to gather information from the environment and make decisions.

Performance of LLMs By pre-training Llama-2 using a code-intensive corpus containing 90B tokens and instruction fine-tuning on 300K examples including both text and code, Lemur-70B-chat (Xu et al., 2023d) surpasses the performance of GPT-3.5-turbo when exploring the environment or following natural language feedback on coding tasks. AgentTuning (Zeng et al., 2023) conducts instruction tuning with Llama-2 on the combination of its constructed AgentInstruct dataset and general domain instructions, resulting in AgentLlama. Notably, AgentLlama-70B achieves comparable performance to GPT-3.5-turbo on unseen agent tasks. Through fine-tuning Llama-2-7B on ToolBench, ToolLLaMA (Qin et al., 2023b) demonstrates comparable performance to GPT-3.5-turbo in tool usage evaluations. Chen et al. (2023a) introduce FireAct, which can fine-tune

Model	Environment			NL Feedback
	ALFWorld	IC-CTF	WebAreana	Code Generation
Lemur-70B-chat	59.70	22.00	5.30	17.65
GPT-3.5-turbo	41.79	11.00	7.38	9.56
GPT-4	84.33	37.00	10.59	-

Table 2: Model performance on several agent benchmarks.

[6] Maya (https://www.maya.ph) is a Filipino financial services and digital payments company.

[7] https://www.01.ai

Llama-2-13B to outperform prompting GPT-3.5-turbo on HotpotQA (Yang et al., 2018). In addition, Gorilla (Patil et al., 2023), fine-tuned from Llama-7B, outperforms GPT-4 on writing API calls.

3.3 Logical Reasoning Capabilities

Benchmarks Logical reasoning serves as fundamental capability of high-level ability and skill, such as programming, theorem proving, as well as arithmetic reasoning. To this end, in this section, we will cover the following benchmarks:

- **GSM8K** (Cobbe et al., 2021) consists of 8.5K high quality grade school math problems created by human problem writers. These problems take between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations to reach the final answer.
- **MATH** (Hendrycks et al., 2021c) is a dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations.
- **TheoremQA** (Wenhu et al., 2023) is a theorem-driven question answering dataset designed to evaluate AI models' capabilities to apply theorems to solve challenging science problems. TheoremQA is curated by domain experts containing 800 high-quality questions covering 350 theorems from Math, Physics, EE&CS, and Finance.
- **HumanEval** (Chen et al., 2021) is a set of 164 hand written programming problems. Each problem includes a function signature, docstring, body, and several unit tests, with an average of 7.7 tests per problem.
- **MBPP** (Austin et al., 2021) (The Mostly Basic Programming Problems) dataset contains 974 short Python programs constructed by crowd-sourcing to an internal pool of crowd workers who have basic knowledge of Python. Each problem is assigned with a self-contained Python function solving the problem specified, and three test cases that check for semantic correctness of the function.
- **APPs** (Hendrycks et al., 2021a) is a benchmark for code generation measuring the ability of models to take an arbitrary natural language specification and generate satisfactory Python code. The benchmark includes 10,000 problems, which range from having simple one line solutions to being substantial algorithmic challenges.

Enhanced Instruction Tuning Different from conventional knowledge distillation based instruction tuning, Luo et al. (2023c,a) employed *Evol-Instruct* (Xu et al., 2023a) to construct the task-specific high quality instruction tuning dataset, where the seed instructions have evolved to the ones either extended in knowledge boundary or the depth of task complexity. Besides, Luo et al. (2023a) also incorporate PPO (Schulman et al., 2017a) algorithm to further improve the quality of both generated instruction and answer. After obtaining the expanded instruction pool, the new instruction tuning dataset is generated by collecting responses from another LLM, e.g., GPT-3.5-turbo. Finally, benefiting from the evolved depth and width of queries, the fine-tuned model achieves even better performance than GPT-3.5-turbo. For example, WizardCoder (Luo et al., 2023c) outperforms GPT-3.5-turbo on HumanEval with 19.1% absolute improvements. And WizardMath (Luo et al., 2023a) has also obtained 42.9% absolute improvements on GSM8K compared with GPT-3.5-turbo.

Pre-training on Data with Higher Quality Lemur (Xu et al., 2023d) has verified a better mixture of natural language data and code and induces the LLMs with stronger abilities on function calling, automatic programming, and agent capabilities. Specifically, Lemur-70B-chat achieves significant improvements over GPT-3.5-turbo on both HumanEval and GSM8K without task-specific fine-tuning. Phi-1 and Phi-1.5 (Gunasekar et al., 2023; Li et al., 2023e) take a different road by using the textbook as the main corpus for pre-training, which makes the strong abilities observable on much smaller language models.

3.4 Modelling Long-context Capabilities

Benchmarks Processing long sequences remains one of the key technological bottlenecks of LLMs, as all models are limited by a finite maximum context window, typically from 2k to 8k tokens in length. Benchmarking long-context capability of LLMs involves evaluating on several tasks

which naturally have a long context, such as abstractive summarization or multi-document QA. The following benchmarks have been proposed for long-context evaluation of LLMs:

- **SCROLLS** (Shaham et al., 2022) is a popular evaluation benchmark made of 7 datasets with naturally long input. The tasks cover summarization, question-answering and natural language inference.
- **ZeroSCROLLS** (Shaham et al., 2023) builds on SCROLLS (discarding ContractNLI, reusing the 6 other datasets, and adding 4 datasets) and only considers the zero-shot setting, evaluating LLMs out-of-the-shelf.
- **LongBench** (Bai et al., 2023) sets a bilingual English/Chinese long-context benchmark of 21 datasets across 6 tasks.
- **L-Eval** (An et al., 2023) re-uses 16 existing datasets and creates 4 datasets from scratch to make a diverse, long-context benchmark, with average length per task above 4k tokens. The authors advocate for LLM judges evaluation (especially GPT-4) rather than n-gram for long-context evaluation.
- **BAMBOO** (Dong et al., 2023) creates a long-context LLM evaluation benchmark focused on removing pre-training data contamination by collecting only recent data in the evaluation datasets.
- **M4LE** (Kwan et al., 2023) introduces a broad-scope benchmark, splitting 36 datasets in 5 understanding abilities: explicit single-span, semantic single-span, explicit multiple-span, semantic multiple-span, and global understanding.

Models On LongBench, L-Eval, BAMBOO and M4LE benchmarks, GPT-3.5-turbo or its 16k version largely outperform all open-source LLMs, such as Llama-2, LongChat, or Vicuna ; showing that it is not trivial to drive up the performance of open-source LLM on long-input tasks. Llama-2-long (Xiong et al., 2023) continues pre-training of Llama-2 with 400B tokens using a 16k context window (up from 4k in Llama-2). The resulting Llama-2-long-chat-70B outperforms GPT-3.5-turbo-16k by 37.7 to 36.7 on ZeroSCROLLS. Approaches to tackle long-context tasks include context window extension with positional interpolation (Chen et al., 2023e), which involves another (short) round of fine-tuning with longer context window ; and retrieval augmentation (Lewis et al., 2020), which necessitates access to a retriever to find relevant information. Xu et al. (2023b) combine both these seemingly opposite techniques, pushing a Llama-2-70B above GPT-3.5-turbo-16k on average over 7 long-context tasks (including 4 datasets from ZeroSCROLLS).

3.5 Application-specific Capabilities

In this section, we discuss the desired capabilities in LLMs to tackle specific applications.

3.5.1 Query-focused Summarization

Benchmarks Query-focused or aspect-based summarization requires to generate summaries with regard to a fine-grained question or an aspect category. Query-focused datasets include AQualMuse (Kulkarni et al., 2020), QMSum (Zhong et al., 2021) and SQuALITY (Wang et al., 2022), while Aspect-based datasets include CovidET (Zhang et al., 2023a), NEWTS (Bahrainian et al., 2022), WikiAsp (Hayashi et al., 2021), etc.

Models Yang et al. (2023d) finds that standard fine-tuning on training data is still better in performance compared to ChatGPT, with an average of 2 points ROUGE-1 improvement over CovidET, NEWTS, QMSum and SQuALITY.

3.5.2 Open-ended QA

Benchmarks There are two sub-categories in Open-ended QA: either the answers are of short-form or long-form. Short-form datasets include SQuAD 1.1 (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SQuAD 2.0 (Rajpurkar et al., 2018), NarrativeQA (Kociský et al., 2018), Natural Question (NQ) (Kwiatkowski et al., 2019), Quoref (Dasigi et al., 2019) and DROP (Dua et al., 2019). Long-form datasets include ELI5 (Fan et al., 2019) and doc2dial (Feng et al., 2020). For both short-form and long-form datasets, the evaluation metrics are exact match (EM)

and F1 over words in the answers. Answering Open-ended QA requires the model to comprehend the provided context, or retrieve related knowledge if there's no context provided.

Models InstructRetro (Wang et al., 2023a) shows large improvement over GPT-3 on NQ, TriviaQA, SQuAD 2.0 and DROP, while having 7-10 percent improvement compared to a proprietary GPT-instruct model of similar size, over a range of short-form and long-form open-ended QA datasets. Initialized from a pretrained GPT model, InstructRetro continues pretraining with retrieval and then undergoes instruction tuning.[8]

3.5.3 Medical

Benchmarks One desirable capability of LLMs is on contributing medical related tasks to make affordable, high-quality healthcare more accessible to the broader public.

For mental health, IMHI (Yang et al., 2023c) benchmark is constructed using 10 existing mental health analysis datasets, including mental health detection: DR (Pirina & Çöltekin, 2018), CLP (Coppersmith et al., 2015), Dreaddit (Turcan & McKeown, 2019), loneliness, SWMH and T-SID (Ji et al., 2022); mental health cause detection: SAD (Mauriello et al., 2021), CAMS (Garg et al., 2022); mental risk factors detection: MultiWD (SATHVIK & Garg, 2023), IRF (Garg et al., 2023).

For radiology, OpenI (Demner-Fushman et al., 2016) dataset and MIMIC-CXR (Johnson et al., 2019) datasets both contain radiology reports with findings and impressions text.

Models For mental health, MentalLlama-chat-13B (Yang et al., 2023c) finetunes a Llama-chat-13B model on IMHI training set. MentalLlama-chat-13B model with zero-shot prompting outperforms ChatGPT with few-shot prompting or with zero-shot prompting for 9 out of 10 tasks in IMHI. Liu et al. (2023) proposes to fine-tune a Llama checkpoint to generate impression text given radiology report findings. The resulting Radiology-Llama-2 model outperforms ChatGPT and GPT-4 by a large margin on both MIMIC-CXR and OpenI datasets.

3.5.4 Generating Structured Responses

Generating formatted responses in accordance with instructions is a core ability to support agentic capabilities or simply reduce the manual efforts in parsing or translating model responses.

Benchmarks Rotowire (Wiseman et al., 2017) contains NBA game summaries with corresponding score tables. Struc-Bench (Tang et al., 2023b) introduces two datasets: Struc-Bench-Latex of which the outputs are tables in Latex format, and Struc-Bench-HTML with outputs as tables in HTML format.

Models Struc-Bench (Tang et al., 2023b) fine-tunes a Llama-7B model on structured generation data. The fine-tuned 7B model outperforms ChatGPT on all benchmarks mentioned above.

3.5.5 Generating Critiques

Benchmarks One interesting ability of LLMs is providing feedback or critiques to a response for a question. To benchmark such ability, one can use human annotators or GPT-4 as an evaluator to directly rate the critiques. The original questions can come from any dataset of other capabilities mentioned above.

Models Shepherd (Wang et al., 2023c) is a 7B model initialized from Llama-7B and trained on community collected critique data and 1,317 examples of high quality human annotated data. Shepherd generates critiques on a range of diverse NLP datasets: AlpacaFarm, FairEval, CosmosQA (Huang et al., 2019), OBQA (Mihaylov et al., 2018a), PIQA (Bisk et al., 2020), TruthfulQA and CritiqueEval. With GPT-4 as an evaluator, Shepherd wins or equals ChatGPT over 60% of the time. With human evaluators, Shepherd is almost on-par with ChatGPT.

[8] InstructRetro is not yet open-sourced.

Models	TruthfulQA	FactScore	HotpotQA	OpenBookQA	MedMC-QA	TriviaQA
Playtus	62.26	-	-	-	-	-
CoVe + Llama-65B	-	71.4	-	-	-	-
CoK + GPT-3.5-turbo	-	-	35.4	-	73.3	-
CRITIC + GPT-3.5-turbo	-	-	38.7	-	-	75.1
KSL + GPT-3.5-turbo	-	-	-	81.6	-	-
PKG + text-davinci-002	-	-	-	-	47.4	-
Cohen et al. (2023) + text-davinci-002	-	-	-	-	-	83.1
GPT-3.5-turbo	47	58.7	24.0	78.3	44.4	79.3

Table 3: Model performance on hallucination benchmarks.

3.6 Towards Trust-worthy AI

To ensure LLMs can be trusted by humans in real-world applications, an important consideration is their reliability. For example, concerns on hallucination (Ye & Durrett, 2022; Zhao et al., 2023) and safety (Zhiheng et al., 2023b) could deteriorate user trust in LLMs and lead to risks in high-impact applications.

3.6.1 Hallucination

Benchmarks Various benchmarks have been proposed for better evaluating hallucinations in LLMs. Specifically, they consist of both large-scale datasets, automated metrics, and evaluation models.

- **TruthfulQA** (Lin et al., 2022) is a benchmark question-answering (QA) dataset consisting of questions spanning 38 categories. The questions are crafted such that some humans would falsely answer them due to misconceptions.
- **FactualityPrompts** (Lee et al., 2022) is a dataset that measures hallucinations for open-ended generation. It consists of factual and non-factual prompts to study the impact of prompts on LLM's continuations.
- **HaluEval** (Li et al., 2023a) is a large dataset of generated and human-annotated hallucinated samples. It spans three tasks: question answering, knowledge-grounded dialogue, and text summarization.
- **FACTOR** (Muhlgay et al., 2023) proposes a scalable approach for evaluating LM factuality: it automatically transforms a factual corpus into a faithfulness evaluation benchmark. The framework is used to create two benchmarks: Wiki-FACTOR and News-FACTOR.
- **KoLA** (Yu et al., 2023a) constructs a Knowledge-oriented LLM Assessment benchmark (KoLA) with three crucial factors: mimicking human cognition for ability modeling, using Wikipedia for data collection, and designing contrastive metrics for automatic hallucination evaluation.
- **FActScore** (Min et al., 2023) proposes a new evaluation that first breaks an LLM's generation into a series of atomic facts, and then computes the percentage of atomic facts supported by a reliable knowledge source.
- **Vectara's Hallucination Evaluation Model** (Hughes, 2023) is a small language model that is fine-tuned as a binary classifier to classify a summary as factually consistent (or not) with the source document. Then, it is used to evaluate and benchmark hallucinations of summaries generated by various LLMs.
- **FacTool** (Chern et al., 2023) is a task and domain agnostic framework for detecting factual errors of texts generated by LLMs.

Besides the newly introduced hallucination benchmarks, prior QA datasets based on real-world knowledge are also widely used for measuring faithfulness, such as HotpotQA (Yang et al., 2018), OpenBookQA (Mihaylov et al., 2018b), MedMC-QA (Pal et al., 2022), and TriviaQA (Joshi et al., 2017). Besides datasets and automated metrics, human evaluation is also widely adopted as a reliable measure for faithfulness.

Models There exist several existing surveys on hallucination (Zhang et al., 2023b; Rawte et al., 2023) that investigate potential methodologies in detail. Specifically, the methods that surpass the current GPT-3.5-turbo performance can be either incorporated during fine-tuning or only at inference time. Selected performance metrics are shown in Table 3.

During fine-tuning, improving data quality in correctness and relevance can lead to less-hallucinated models. Lee et al. (2023a) curated a content-filtered, instruction-tuned dataset, focusing on high-quality data in the STEM domain. A family of LLMs is fine-tuned on this filtered dataset and merged. The resulting family, named Platypus, demonstrates a substantial improvement on TruthfulQA (approximately 20%) compared to GPT-3.5-turbo.

During inference, existing techniques include specific decoding strategies, external knowledge augmentation, and multi-agent dialogue. For decoding, Dhuliawala et al. (2023) introduces Chain-of-Verification (CoVe), where the LLM drafts verification questions and self-verify the responses. CoVe leads to a substantial improvement on FactScore over GPT-3.5-turbo.

For external knowledge augmention, various frameworks incorporate different searching and prompting techniques to the current improve GPT-3.5-turbo performance. Li et al. (2023c) designs Chain-of-Knowledge (CoK), which retrieves from heterogenous knowledge sources before answering. Peng et al. (2023) proposes LLM-AUGMENTER, which augments LLMs with a set of plug-and-play modules and iteratively revises LLM prompts to improve model responses using feedback generated by utility functions. Knowledge Solver (KSL) (Feng et al., 2023) tries to teach LLMs to search for essential knowledge from external knowledge bases by harnessing their own strong generalizability. CRITIC (Gou et al., 2023) allows LLM to validate and progressively amend their own outputs in a manner similar to human interaction with tools. Luo et al. (2023b) introduces a Parametric Knowledge Guiding (PKG) framework, which equips LLMs with a knowledge-guiding module to access relevant knowledge without altering the LLMs' parameters. These inference techniques then improve answer accuracy compared to the naive prompting strategy using GPT-3.5-turbo. Currently, GPT-3.5-turbo has also incorporated a retrieval plugin (OpenAI, 2023a) that accesses external knowledge to reduce hallucinations.

For multi-agent dialogue, Cohen et al. (2023) facilitates a multi-turn interaction between the Examinee LLM that generated the claim and another Examiner LLM which introduces questions to discover inconsistencies. Through the cross-examination process, performance on various QA tasks is improved. Du et al. (2023) asks multiple language model instances to propose and debate their individual responses and reasoning processes over multiple rounds to arrive at a common final answer, which improves on multiple benchmarks.

3.6.2 Safety

Benchmarks Safety concerns in LLMs can mostly be grouped into three aspects (Zhiheng et al., 2023a): social bias, model robustness, and poisoning issues. To gather datasets that better evaluate the above aspects, several benchmarks have been proposed:

- **SafetyBench** (Zhang et al., 2023c) is a dataset which consists of 11,435 diverse multiple choice questions spanning 7 distinct categories of safety concerns.
- **Latent Jailbreak** (Qiu et al., 2023) introduces a benchmark that assesses both the safety and robustness of LLMs, emphasizing the need for a balanced approach.
- **XSTEST** (Röttger et al., 2023) is a test suite that systematically identifies exaggerated safety behaviors, such as refusing safe prompts.
- **RED-EVAL** (Bhardwaj & Poria, 2023) is a benchmark to perform red-teaming (Ganguli et al., 2022) to conduct safety evaluations of LLMs using a Chain of Utterances (CoU)-based prompt.

Besides automated benchmarks, an important measure for safety is human evaluation (Dai et al., 2023), where crowdworkers label the responses as safe or harmful. Some studies also attempt to collect such labels from GPT-4, as research shows that it can replace human evaluators in evaluating alignment abilities (Chiang & Lee, 2023).

Models Based on current evaluations (Zhang et al., 2023c; Röttger et al., 2023), GPT-3.5-turbo and GPT-4 models remain at the top for safety evaluations. This is largely attributed to Reinforcement Learning with Human Feedback (RLHF) (Bai et al., 2022a). RLHF first collects a human preference dataset on responses, then trains a reward model to mimic human preferences, and finally uses RL to train the LLM to align with human preferences. In the process, LLMs learn to demonstrate desired behaviors and exclude harmful responses such as impolite or biased answers. However, the RLHF procedure requires collecting a large number of expensive human annotations, which hinders its use

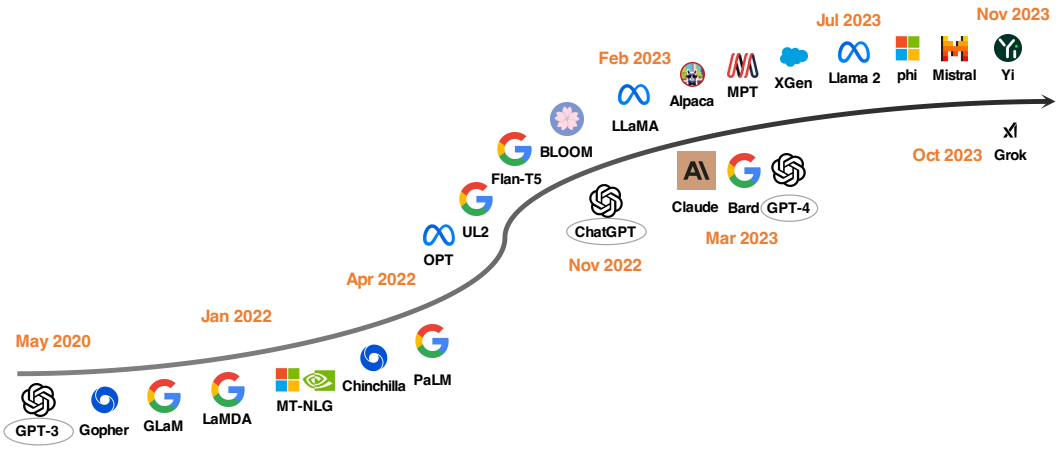


Figure 3: LLM development timeline. The models below the arrow are closed-source while those above the arrow are open-source.

for open-source LLMs. To democratize endeavors on advancing the safety alignment of LLMs, Ji et al. (2023) gathers a human-preference dataset to disentangle harmlessness and helpfulness from the human-preference score, thus providing separate ranking data for the two metrics. Experiments show that disentangling human preferences enhances safety alignment. Bai et al. (2022b) seeks to increase safety with RL from AI Feedback (RLAIF), where the preference model is trained using LLM-generated self-critiques and revisions. Direct Preference Optimization (DPO) (Rafailov et al., 2023a) reduces the need to learn a reward model and learn from preferences directly with a simple cross-entropy loss, which could largely reduce costs for RLHF. Combining and improving these methodologies could lead to potential improvements in safety for open-source LLMs.

4 Discussion

4.1 Development Trend of LLMs

Ever since Brown et al. (2020) demonstrated that a frozen GPT-3 model can achieve impressive zero- and few-shot performance on a variety of tasks, numerous efforts have been made to advance the development of LLMs. One line of research focused on scaling up model parameters, including Gopher (Rae et al., 2021), GLaM (Du et al., 2022), LaMDA (Thoppilan et al., 2022), MT-NLG (Smith et al., 2022) and PaLM (Chowdhery et al., 2022), culminating at 540B parameters. Despite exhibiting remarkable capabilities, the closed-source nature of these models limited their widespread application, thereby leading to a growing interest in developing open-source LLMs (Zhang et al., 2022; Workshop et al., 2022).

Rather than scaling up model size, another line of research explored better strategies or objectives for pre-training smaller models, such as Chinchilla (Hoffmann et al., 2022b) and UL2 (Tay et al., 2022). Beyond pre-training, considerable attention has been devoted to studying instruction tuning of LMs, *e.g.,* FLAN (Wei et al., 2021b), T0 (Sanh et al., 2021) and Flan-T5 (Chung et al., 2022).

The emergence of OpenAI's ChatGPT a year ago greatly changed the research focus of the NLP community (Qin et al., 2023a). To catch up with OpenAI, Google and Anthropic introduced Bard and Claude, respectively. While they show comparable performance to ChatGPT on many tasks, there is still a performance gap between them and the latest OpenAI model GPT-4 (OpenAI, 2023b). As the success of these models is primarily attributed to reinforcement learning from human feedback (RLHF) (Schulman et al., 2017b; Ouyang et al., 2022a), researchers have explored various ways to improve RLHF (Yuan et al., 2023; Rafailov et al., 2023b; Lee et al., 2023b).

To promote research of open-source LLMs, Meta released Llama series models (Touvron et al., 2023a,b). Since then, open-source models based on Llama have started to emerge explosively. One representative research direction is to fine-tune Llama with instruction data, including Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), Lima (Zhou et al., 2023b) and WizardLM (Xu et al.,

2023a). The ongoing research has also explored improving the agent (Xu et al., 2023d; Zeng et al., 2023; Patil et al., 2023; Qin et al., 2023b), logical reasoning (Roziere et al., 2023; Luo et al., 2023a,c) and long-context modeling (Tworkowski et al., 2023; Xiong et al., 2023; Xu et al., 2023b) capabilities of Llama-based open-source LLMs. Besides, rather than developing LLMs based on Llama, many efforts have been devoted to training powerful LLMs from scratch, *e.g.,* MPT (Team, 2023), Falcon (Almazrouei et al., 2023), XGen (Nijkamp et al., 2023), Phi (Gunasekar et al., 2023; Li et al., 2023e), Baichuan (Yang et al., 2023a), Mistral (Jiang et al., 2023a), Grok (xAI, 2023) and Yi (01ai, 2023). We believe that developing more powerful and efficient open-source LLMs to democratize the capabilities of closed-source LLMs should be a quite promising future direction.

4.2 Summary of Results

For general capabilities, Llama-2-chat-70B (Touvron et al., 2023b) shows improvement over GPT-3.5-turbo in some benchmarks, but remains behind for most others. Zephir-7B (Tunstall et al., 2023) approaches 70B LLMs as a result of distilled direct preference optimization. WizardLM-70B (Xu et al., 2023a) and GodziLLa-70B (Philippines, 2023) can achieve comparable performance to GPT-3.5-turbo, which show a promising path forward.

There are also several domains where open-source LLMs are able to beat GPT-3.5-turbo. For LLM-based agents, open-source LLMs are able to surpass GPT-3.5-turbo with more extensive and task-specific pre-training and fine-tuning. For example, Lemur-70B-chat (Xu et al., 2023d) performs better in exploring the environment and following feedback on coding tasks. AgentTuning (Zeng et al., 2023) improves on unseen agent tasks. ToolLLama (Qin et al., 2023b) can better grasp tool usage. Gorilla (Patil et al., 2023) outperforms GPT-4 on writing API calls. For logical reasoning, WizardCoder (Luo et al., 2023c) and WizardMath (Luo et al., 2023a) improve reasoning abilities with enhanced instruction tuning. Lemur (Xu et al., 2023d) and Phi (Gunasekar et al., 2023; Li et al., 2023e) achieve stronger abilities by pre-training on data with higher quality. For modelling long contexts, Llama-2-long (Xiong et al., 2023) can improve on selected benchmarks by pre-training with longer tokens and a larger context window. Xu et al. (2023b) improves over 7 long-context tasks by combining context window extension with positional interpolation and retrieval augmentation. For application-specific capabilities, InstructRetro (Wang et al., 2023a) improves on open-ended QA by pre-training with retrieval and instruction tuning. With task-specific fine-tuning, MentaLlama-chat-13B (Yang et al., 2023c) outperforms GPT-3.5-turbo in mental health analysis datasets. Radiology-Llama2 (Liu et al., 2023) can improve performance on radiology reports. Stru-Bench (Tang et al., 2023b), a fine-tuned 7B model, can improve structured response generation compared to GPT-3.5-turbo, which is a core ability to support agentic tasks. Shepherd (Wang et al., 2023c), with only 7B parameters, can achieve comparable or better performance compared to GPT-3.5-turbo in generating model feedbacks and critiques. For trustworthy AI, hallucinations can be reduced by fine-tuning with data of higher quality (Lee et al., 2023a), context-aware decoding techniques (Dhuliawala et al., 2023), external knowledge augmentation such as Li et al. (2023c); Yu et al. (2023b); Peng et al. (2023); Feng et al. (2023), or multi-agent dialogue (Cohen et al., 2023; Du et al., 2023).

There are also domains where GPT-3.5-turbo and GPT-4 remain unbeatable, such as AI safety. Due to the large-scale RLHF (Bai et al., 2022a) involved in GPT models, they are known to demonstrate safer and more ethical behaviors, which is probably a more important consideration for commercial LLMs compared to open-source ones. However, with the recent efforts on democratizing the RLHF process (Bai et al., 2022b; Rafailov et al., 2023a), we could expect to see more performance improvements for open-source LLMs in safety.

4.3 Recipe of Best Open-source LLMs

Training an LLM involves complex and resource-intensive practices, including data collection and preprocessing, model design, and training process. While there is a growing trend of releasing open-source LLMs regularly, the detailed practices of the leading models are often kept secret unfortunately. Below we list some best practices widely acknowledged by the community.

Data Pre-training involves the use of trillions of data tokens, often sourced from publicly accessible sources. Ethically, it is crucial to exclude any data that includes personal information of private individuals (Touvron et al., 2023b). Unlike pre-training data, fine-tuning data is smaller in quantity

but superior in quality. Fine-tuned LLMs with top-quality data have shown improved performance, particularly in specialized areas (Philippines, 2023; Zeng et al., 2023; Xu et al., 2023d,a).

Model Architecture While the majority of LLMs utilize the decoder-only transformer architecture, different techniques in the model are employed to optimize efficiency. Llama-2 implements Ghost attention for improved multi-turn dialogue control (Touvron et al., 2023b). Mistral (Jiang et al., 2023b) employs sliding window attention to handle extended context lengths.

Training The process of supervised fine-tuning (SFT) with instruction tuning data is vital. For high quality outcomes, tens of thousands of SFT annotations are sufficient, as evidenced by the 27,540 annotations used for Llama-2 (Touvron et al., 2023b). The diversity and quality of these data are essential (Xu et al., 2023a). In the RLHF stage, proximal policy optimization (PPO) (Schulman et al., 2017a) is often the preferred algorithm to better align the model's behavior with human preferences and instruction adherence, playing a key role in enhancing LLM safety. An alternative to PPO is direct preference optimization (DPO) (Rafailov et al., 2023a). Zephyr-7B (Tunstall et al., 2023), for instance, employs distilled DPO and has shown results comparable to 70B-LLMs on various general benchmarks, even surpassing GPT-3.5-turbo on AlpacaEval.

4.4 Loopholes and potential problems

Data Contamination during Pre-training The issue of data contamination has become increasingly pronounced with the release of foundation models that obscure the source of their pre-training corpus. This lack of transparency can result in biased perceptions regarding the genuine generalization capabilities of Large Language Models (LLMs). Ignoring the cases where benchmark data is manually integrated into the training set with annotations from human experts or larger models, the root of the data contamination problem lies in the fact that the collecting source of benchmark data is already encompassed in the pre-training corpus. While these models are not intentionally pre-trained using supervised data, they can still acquire exact knowledge. Consequently, it is crucial to address the challenge of detecting the pre-training corpus of LLMs (Shi et al., 2023), exploring the overlap between existing benchmarks and widely-used pre-training corpus, and assessing overfitting to benchmarks (Wei et al., 2023). These efforts are essential for enhancing the faithfulness and reliability of LLMs. Looking ahead, future directions could involve establishing standardized practices for disclosing pre-training corpus details and developing methods to mitigate data contamination throughout the model development lifecycle.

Close-sourced Development of Alignment The application of Reinforcement Learning from Human Feedback (RLHF) for alignment using general preference data has obtained increasing attention within the community. However, only a limited number of open-source LLMs have been augmented with RLHF for alignment, primarily due to the scarcity of high-quality, publicly available preference datasets and pre-trained reward models. Some initiatives (Bai et al., 2022a; Wu et al., 2023; Cui et al., 2023) have sought to contribute to the open-source community. Yet, we are still facing the challenges lacking diverse, high-quality and scalable preference data in complex reasoning, programming, and safety scenarios.

Difficulty in Continuous Improvements over Fundamental Abilities Reviewing the breakthroughs in fundamental abilities outlined in this paper reveals somewhat challenging scenarios: (1) Considerable efforts have been invested in exploring improved data mixtures during pre-training to enhance balance and robustness in constructing more potent foundation models. However, the associated exploration costs often render this approach impractical. (2) Models surpassing GPT-3.5-turbo or GPT-4 are predominantly based on knowledge distillation from closed-source models and additional expert annotation. While efficient, heavy reliance on knowledge distillation may mask potential issues concerning the effectiveness of proposed approaches when being scaled to the teacher model. Moreover, LLMs are anticipated to act as agents and provide reasonable interpretations to support decisions, while annotating the agent-style data to make LLMs applicable to real-world scenarios is also expensive and time-consuming. In essence, optimization through knowledge distillation or expert annotation alone cannot realize continuous improvement and is likely to approach an upper bound. Future research directions may involve exploring novel methodologies, such as unsupervised or self-supervised learning paradigms, to enable continuous advancements in fundamental LLM abilities while mitigating the associated challenges and costs.

5 Conclusion

In this survey, we deliver a systematical review on high performing open-source LLMs that surpass or catch up with ChatGPT in various task domains, at the one-year anniversary mark after ChatGPT's release (Section 3). In addition, we provide insights, analysis and potential issues of open-source LLMs (Section 4). We believe that this survey sheds lights on promising directions of open-source LLMs and will serve to inspire further research and development in the field of open-source LLMs, helping to close the gap with their paying counterparts.

References

01ai. Yi model. *HuggingFace*, 2023.

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. Falcon-40B: an open large language model with state-of-the-art performance. *HuggingFace*, 2023.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*, 2021.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *arXic preprint arXiv:2108.07732*, 2021.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. NEWTS: A corpus for news topic-focused summarization. *In Findings of ACL*, 2022.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. *Hugging Face*, 2023.

BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*, 2023.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *In Proceedings of NeurIPS*, 2000.

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. *In Proceedings of AAAI*, 2020.

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

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023a.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023b.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023c.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023d.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023e.

Liying Cheng, Xingxuan Li, and Lidong Bing. Is GPT-4 a good data analyst? *CoRR*, 2023.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and PTSD on twitter. *In Proceedings of NAACL*, 2015.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. Flash-decoding for long-context inference. *https://crfm.stanford.edu/2023/10/12/flashdecoding.html*, 2023.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *In Proceedings of EMNLP*, 2019.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc.*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*, 2023.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *In Proceedings of ICML*, 2022.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. *In Proceedings of ACL*, 2019.

Chao Feng, Xinyu Zhang, and Zichu Fei. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*, 2023.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. *In Proceedings of EMNLP*, 2020.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. *Zenodo*, 2021.

Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. *In Proceedings of LREC*, 2022.

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. *In Findings of ACL*, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

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. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. Wikiasp: A dataset for multi-domain aspect-based summarization. *TACL*, 2021.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. *In Proceedings of NeurIPS*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *In Proceedings of NeurIPS*, 2021c.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022a.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022b.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. *In Proceedings of EMNLP*, 2019.

Simon Hughes. Cut the bull.... detecting hallucinations in large language models. *vectara*, 2023.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*, 2023.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Comput. Appl.*, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023b.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 2019.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *In Proceedings of ACL*, 2017.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *TACL*, 2018.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *CoRR*, 2020.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and Kam-Fai Wong. M4le: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. *arXiv preprint arXiv:2310.19240*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*, 2023a.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023b.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Neur*, 2022.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *In Proceedings of NeurIPS*, 2020.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprints arXiv:2305.11747*, 2023a.

Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*, 2023b.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*, 2023c.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*, 2023d.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023e.

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

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *In Proceedings of ACL*, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. Radiology-llama2: Best-in-class large language model for radiology. *arXiv preprints arXiv:2309.06419*, 2023.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023a.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*, 2023b.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023c.

Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. SAD: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. *In Proceedings of CHI Extended Abstracts*, 2021.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. *In Proceedings of EMNLP*, 2018a.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018b.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

mlabonne. Guanaco llama 2 1k dataset. *Hugging Face*, 2023.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*, 2023.

Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450*, 2023.

OpenAI. Chatgpt plugins. *OpenAI*, 2023a.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023b.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022a.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *In Proceedings of Conference on Health, Inference, and Learning*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of ACL*, 2002.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Maya Philippines. Godzilla 2 70b. *Hugging Face*, 2023.

Inna Pirina and Çağrı Çöltekin. Identifying depression on Reddit: The effect of training data. *In Proceedings of EMNLP*, 2018.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023a.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023b.

Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*, 2023.

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

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023a.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023b.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *In Proceedings of EMNLP*, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *In Proceedings of ACL*, 2018.

Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WINOGRANDE: an adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

MSVPJ SATHVIK and Muskan Garg. MULTIWD: Multiple Wellness Dimensions in Social Media Posts. *TechRxiv*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized CompaRison over long language sequences. *In Proceedings of EMNLP*, 2022.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *CoRR*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023a.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint arXiv:2309.08963*, 2023b.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model. *Stanford CRFM*, 2023.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

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

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *In Proceedings of ACL*, 2017.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Elsbeth Turcan and Kathy McKeown. Dreaddit: A reddit dataset for stress analysis in social media. *In Proceedings of EMNLP*, 2019.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *arXiv preprint arXiv:2307.03170*, 2023.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. Squality: Building a long-document summarization dataset the hard way. *In Proceedings of EMNLP*, 2022.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713*, 2023a.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *CoRR*, 2023b.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*, 2023c.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023d.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021a.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021b.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. Skywork: A more open bilingual foundation model. *CoRR*, 2023.

Chen Wenhu, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, and Pan Lu. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. Challenges in data-to-document generation. *In Proceedings of EMNLP*, 2017.

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

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *CoRR*, 2023.

xAI. Grok-1 model. *xAI*, 2023.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023a.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023b.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models. *arXiv preprint arXiv:2305.16504*, 2023c.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, Zhoujun Cheng, Siheng Zhao, Lingpeng Kong, Bailin Wang, Caiming Xiong, and Tao Yu. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*, 2023d.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023a.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *arXiv preprint arXiv:2306.14898*, 2023b.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023c.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the limits of chatgpt for query or aspect-based text summarization. *CoRR*, 2023d.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *NeurIPS*, 2022.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023a.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*, 2023b.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization. *CoRR*, 2023a.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*, 2023c.

Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*, 2023.

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. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Xi Zhiheng, Zheng Rui, and Gui Tao. Safety and ethical concerns of large language models. *In Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)*, 2023a.

Xi Zhiheng, Zheng Rui, and Gui Tao. Safety and ethical concerns of large language models. *In Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)*, 2023b.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *In Proceedings of NAACL-HLT*, 2021.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023a.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023b.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023c.