WOLFRAM
BLOCKCHAIN LABS

Wolfram AI Revolution & Project Catalyst

A literature review of Large Language Models

> *"Liberty in cyberspace will not come from the absence of the state. Liberty there, as anywhere, will come from a state of a certain kind. We build a world where freedom can flourish not by removing from society any self-conscious control, but by setting it in a place where a particular kind of self-conscious control survives. We build liberty as our founders did, by setting society upon a certain constitution."*
>
> — ***Lawrence Lessig, Laws of Cyberspace, 1998***

Abstract

This paper provides an overview of governance in community decision-making, with a focus on the transformative impacts of technological advancements in large language models (LLMs). Governance, and its various mechanisms, has received heighted attention in the past decade through new decentralized protocols which introduce transparency, trust and sovereignty into cyberspace. Now, LLMs are increasingly becoming central to this study, so we embark on an examination of their evolution and impact on blockchain governance systems like Cardano’s Project Catalyst. From the Transformer architecture to the latest GPT models, our goal is to highlight LLM capabilities and compare the top models as of December 2023. The methodology is straightforward and follows the rapidly evolving LLM research space with several externally compiled literature reviews. These literature reviews directly focus on the benchmarking of leading models so that the reader can understand the significance of various distinctions; open versus closed source, LLM capabilities, and model cost effectiveness. A diverse array of model comparisons are made, including GPT-3.5, GPT-4, PaLM 2, LlaMA 2, Gemini, and Claude2, each selected for distinct characteristics. The work aims to provide an analysis of LLMs' potential and challenges in enhancing governance through comprehension and natural language reasoning, and also offering theoretical insights and practical implications for future blockchain governance systems. It is the goal of this paper to better understand the role LLMs can play within a governance community perspective. Hopefully this information can empower individuals and communities with better decisions on implementing standards, improvements and workflows when using LLMs.

Table of contents

List of Figures and Tables

Introduction

Governance

Governance systems have continuously evolved throughout history. Broadly defined, governance is the process that regulates and enforces decision-making of an organizational unit. A particular governance system, adopted by a community, is the set of processes and structures for which choices are evaluated, elected, executed and enforced. Governance systems have the power to shape entire identities and trajectories of a community, leading to a broad spectrum of styles and architectures we see across the globe. Therefore, we must acknowledge and research how new technological developments will impact these various systems so we can proactively navigate around or integrate with them. Lastly, the digital age brings radically new technologies, and thus, new challenges to governance systems. Blockchain-based architectures enable new decentralization and self-sovereignty mechanisms, which aim to enhance trust and transparency for everyone (Swan, 2015). They add new structural components to how communities may potentially organize and govern in the future. While this may enable greater freedom for the individual, it may also create new challenges for our collective actions.

Blockchain Governance

Our research begins with the work of seminal thinkers like Lawrence Lessig, who famously articulated the “Laws of Cyberspace,” and Nobel Prize winner, Elinor Ostrom, who addressed the famous conundrum known as the “tragedy of the commons” through her work, “Governing the Commons.”

The main idea of “Laws of Cyberspace,” published in 1998, was to raise the alarm for the future world society is entering. Lessig states: “in my view…the world we are entering is not a world where freedom is assured. Cyberspace has the potential to be the most fully, and extensively, regulated space that we have ever known — anywhere, at any time in our history. It has the potential to be the antithesis of a space of freedom. And unless we understand this potential, unless we see how this might be, we are likely to sleep through this transition from freedom into control” (Lessig, 1998). To parallel the challenge Lessig highlights, Ostrom's examination of economic conundrums, including the "tragedy of the commons," serves as a cornerstone in understanding the intricacies of collective governance (Ostrom, 1990). Problems like the prisoner’s dilemma, free rider problem, and tragedy of the commons all predated the digital age, and highlight the importance that new technologies should play in addressing these long-standing challenges. Ostrom's insights are instrumental when considering robust governance structures and designing mechanisms capable of avoiding the pitfalls of oppression and control discussed by Lessig. It is these articulations which form the backdrop of our digital landscape and for understanding the high-stakes of our governance systems now and in the future.

Next, we dive into the core interplay between the ethos of decentralization (Zhang et al, 2023) and governance, which is at the heart of our examination. Decentralization refers to the distribution of functions, powers, people, or things away from a central location or authority. However, this ideal often clashes with the practical realities of governance, particularly when effective mechanisms have yet to be established or when critical decisions are on the horizon. Kiayias and Lazos provide a comprehensive analysis of governance within blockchain protocols, delineating their potential and limitations (Kiayias, Lazos, 2023). The analysis of governance is complemented by Laitikeenan et al.'s literature review, which delves into the two main dimensions of blockchain governance: 1) governance of the infrastructure and governance by the infrastructure. This exploration is vital for assessing the technical capabilities and potential of Large Language Models in governance structures. The concepts of "governance of the infrastructure" and "governance by the infrastructure" are crucial dimensions in understanding blockchain governance.

“Governance of the Infrastructure” pertains to the means and processes involved in directing, controlling, and coordinating actors within a blockchain system. It involves a more traditional sense of governance, concerning how decisions are made about the system itself, including its rules, operations, and modifications. This dimension is about how humans and organizations manage and govern the technology and is more involved with off-chain governance mechanisms, such as community decision-making processes, developer discussions, and formal institutional rules.

“Governance by the Infrastructure,” on the other hand, refers to the blockchain itself governing actions and behaviors through encoded rules and automated processes. This is a more novel aspect brought by blockchain technology, where governance rules are directly written into the code, allowing for decentralized decision-making and enforcement. This includes mechanisms like smart contracts, which execute automatically under specific conditions, and on-chain voting systems for token holders. Here, the technology itself enforces rules and agreements, which could potentially reduce the need for traditional legal systems or intermediaries. Governance by the infrastructure represents a shift from traditional governance structures, relying heavily on the technology's capacity to enforce and execute rules autonomously.

Both of these dimensions interact and overlap in complex ways within any blockchain system. The governance of the infrastructure sets the stage and creates the environment in which governance by the infrastructure operates. At the same time, the capabilities and design of the blockchain technology (governance by the infrastructure) can significantly influence and sometimes constrain the decisions and flexibility of traditional governance (governance of the infrastructure). Effective blockchain governance typically involves a careful balance and interplay between these two dimensions, tailored to the specific needs and context of the blockchain application (Laitikeenan et al., 2023).

Lastly, as blockchains introduce new ways of organizing, they necessarily introduce new ways of failing. “The Many Open Problems of DAOs” research highlights some of the major unresolved problems to be addressed by blockchain communities (Tan et al, 2023). With that said, we have reason to be optimistic about these new governance approaches, combined with new technologies, which may help us overcome challenges of the past.

Why LLMs?: An Overview

Because governance systems are inherently constrained by their overall architecture and the constituents among them, it requires a well-versed individual in a well-designed ecosystem, to understand, navigate and take proper/timely actions on multifaceted issues. Ultimately, this is the challenge of every governance system. Yet an individual routinely reaches their cognitive limit of relationships, famously known as “Dunbar’s Number,” which is the number of meaningful connections one can manage at one time. It is a general limit that suggests one individual can’t know all of the problems to address all at the same time. At around 150 people, (Dunbar, 1992) this phenomenon might help explain why comprehension and engagement in collective settings wanes as group sizes increase. Although debates persist regarding the precision and methodology of this number (Lindensfors, 2021), the general wisdom holds true. A human is finite in their capacity to make relevant sense of the world (Vervaeke, 2022)

Now, when considering critical community decisions in the digital domain, which reach a vastly new scale of community size, we recognize a palpable need for innovative solutions to overcome constraints of individuals. Communities throughout history, for better or worse, have overcome this issue by deferring decisions to elders, family members or electing representatives to act on their behalf. Sometimes, as evidenced by various kingdoms and monarchies throughout history, governance structures can forcefully devolve into dictatorships, with one ruler, family, or plutocracy deciding a community’s future. Individuals must grapple with the limits of our capacities and coordinate our governance structures to manage the increased complexity and volume of information of the digital age. From this context, our paper explores the potential of Large Language Models (LLMs) in augmenting governance systems with new types of resources. Specifically, we research the role of LLMs in enhancing decision-making and administrative processes.

The emergence of advanced LLM technologies represents a significant stride forward in collective sensemaking. LLMs have the potential to manage certain narrow aspects of a community’s cognitive load, and offer substantial support in information processing and decision-making tasks. Stephen Wolfram provides an in-depth exploration of how LLMs function in his paper called “What is ChatGPT Doing and Why Does It Work” (Wolfram, 2023).

Background Knowledge of LLM evolution

The journey of large language models (LLMs) in the realm of machine learning and AI has been characterized by a series of remarkable milestones. Vaswani and his team were the trailblazers, introducing the Transformer architecture in their 2017 seminal paper "Attention is All You Need" (Vaswani et al., 2017). This innovative shift in natural language processing (NLP) transformed the focus from sequential data to a more comprehensive sentence analysis via self-attention.

The year 2018 marked another significant leap in the field of language modeling with the emergence of BERT (Bidirectional Encoder Representations from Transformers). As elucidated by Devlin et al. in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," BERT integrated context from both directions, providing a more sophisticated interpretation of text. This enhanced its performance across various NLP tasks such as sentiment analysis and question answering.

In the development of LLMs, OpenAI made significant contributions with the GPT series. The 2018 paper "Improving Language Understanding by Generative Pre-training" by Radford et al. unveiled GPT-1, demonstrating the efficiency of unsupervised learning in NLP. This momentum was sustained with the release of GPT-2, as discussed in "Language Models are Unsupervised Multitask Learners" (2019), which expanded both model scale and dataset range, leading to impressive results across a myriad of NLP benchmarks.

Significant strides have also been made in conversational AI, primarily through the ChatGPT series, an LLM application. The 2020 paper "Language Models are Few-Shot Learners" demonstrated ChatGPT's(or GPT-3) capabilities in emulating human interactions, signaling its potential use in diverse arenas, from customer support bots to digital personal assistants. However, despite advances with open-source models like GPT-1 and GPT-2, a shift toward closed-source models has been noticeable since 2020, beginning with GPT-3. This change has presented obstacles for academic exploration of LLM training methodologies, potentially heralding a new era of API-based research in academia.

In terms of model architecture, decoder-only models have been gaining traction in the LLM landscape, particularly post the launch of GPT-3. While encoder-only and encoder-decoder models initially dominated, the trend has seen a shift since the zenith of BERT. OpenAI continues to be at the forefront of LLM advancements, particularly with its leadership in closed-source LLMs following the launch of GPT-3. In contrast to OpenAI, Meta has made considerable strides in supporting open-source LLMs (Touvron et al.,2023) and facilitating research in this domain. Among commercial entities, Meta is notable for its significant contributions and commitment to making all its developed LLMs publicly accessible.

In conclusion, the evolution of LLMs is a testament to ongoing innovation and breakthroughs. From Transformers to BERT and from open-source endeavors like GPT to proprietary developments, each phase has broadened the horizons of language comprehension and generation. The transformative impact of LLMs on AI and machine learning is ongoing, continually reshaping our interaction with technology. Figure 1 offers a detailed depiction of the evolution of language models, providing a clearer perspective on the LLM landscape and reinforcing the observations noted above.

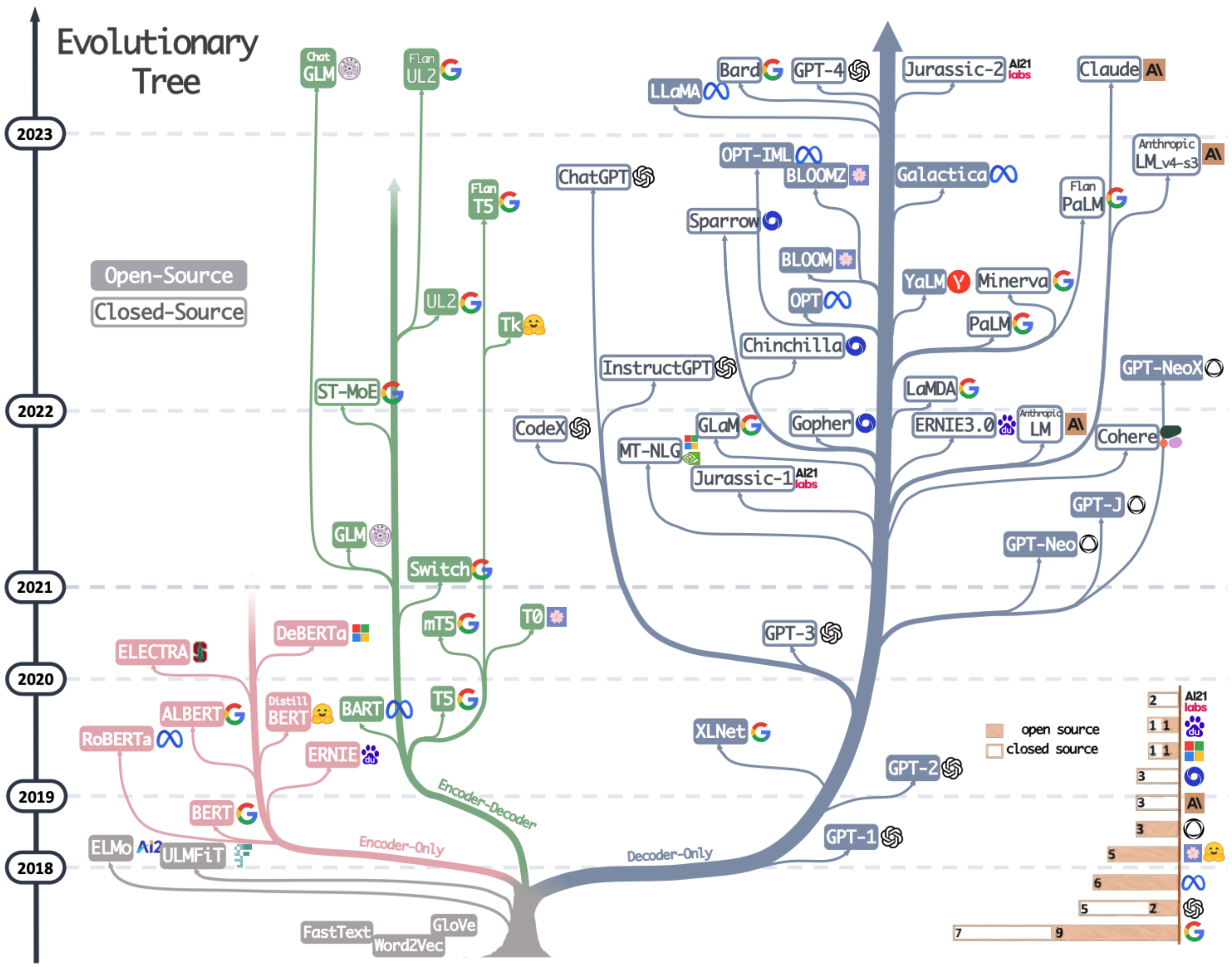


Figure 1: LLM evolutionary tree (Yang et al., 2023)

The evolutionary progression of modern Language Models in Figure 1 maps out the significant advancements over recent years while featuring some of the most influential ones. Models belonging to the same branch exhibit closer ties. Transformer-based models are indicated in vibrant colors: blue for decoder-only models, pink for encoder-only models, and green for encoder-decoder models. The models' vertical placements on the timeline correspond to their respective launch dates. Open-source models are denoted by solid squares, whereas closed-source models are depicted by hollow squares. The bottom right corner presents a stacked bar graph displaying the count of models originating from different organizations and institutions.

Emergent abilities of LLMs

In the dynamic landscape of language models, LLMs have demonstrated abilities like in-context learning, instruction following, alignment tuning and Instruction following that distinguish them from their smaller counterparts like GPT-1, GPT-2, BERT, ALBERT, RoBERTa, DistilBERT, …etc. The most notable ability, exemplified by GPT-3, is in-context learning that generates anticipated output based on natural language instructions or task demonstrations. However there is variability related to both, the size of the model and the nature of the specific task at hand. LLMs exhibit an aptitude for following instructions effectively. Leveraging a combination of multi-task datasets formatted with natural language descriptions, a technique referred to as instruction tuning, these models can perform tasks that were not part of their initial training. This mechanism significantly enhances their ability to generalize across diverse tasks.

To align LLMs more closely with human values and to mitigate the potential risks associated with generating toxic, biased, or harmful content, sophisticated techniques like reinforcement learning with human feedback are employed. This alignment tuning process ensures that LLMs adhere to ethical guidelines and align their outputs with desired societal norms. Also, LLMs exhibit the ability to manipulate external tools. This functionality allows them to tackle tasks that may not be best expressed in text form or to access real-time information beyond the scope of their pre-training data. This adaptability empowers them to handle a diverse range of challenges.

These abilities distinguish LLMs from smaller models and play a pivotal role in augmenting their overall performance and versatility across an array of tasks. The exploration of these capabilities sheds light on the evolving landscape of language models and their potential to transcend traditional boundaries in natural language understanding and generation.

Summary of LLMs and blockchain governance

As technology advances, particularly through LLMs, there's a growing need to consider their role in the evolving landscape of governance systems. This paper provides an overview of the significant capabilities and potential challenges of integrating LLMs into governance, specifically in the context of blockchain technology and Project Catalyst. By understanding the evolutionary trajectory of LLMs and their emergent abilities, we can better grasp how they might influence and enhance governance systems, paving the way for more effective, inclusive, and adaptive forms of decision-making and community management. The ultimate goal is to contribute to the ongoing discourse on blockchain governance and the practical implications of incorporating LLMs, ensuring they align with the needs and values of diverse communities.

Literature review

General comparison of the top 6 LLMs

Methodology

Our methodology is fairly straightforward. We wanted to capture as many relevant details and most recent trends available as this is a rapidly changing space. Zheng et al. 2023 wrote a comprehensive survey of the entire LLM space which we've largely used below. We then focused on benchmarking and comparison studies to look directly at different leading models and why open vs closed source matters so much in the next few years. Finally, our guiding philosophy for sharing information here is directly informed by wanting to empower the Cardano community if/when it chooses to implement a Catalyst improvement with regards to LLMs. We have selected a diverse array of models of natural language processing. Our comparison includes GPT-3.5 and GPT-4 by OpenAI, PaLM 2 by Google, LlaMA 2 by Meta, Gemini by Google, and Claude 2 by Anthropic, each chosen for specific reasons detailed below.

GPT-3.5, is the standard against which every new model is measured and benchmarked. Its popularity in 2023 makes it an indispensable reference point in our comparative analysis.

GPT-4, released in March 2023, gained popularity in the latter half of 2023, signifying the pinnacle of current language model capabilities. Its significance is underscored by its recognition as state-of-the-art (SOTA), as detailed in the "GPT-4 Technical Report"(OpenAI, 2023).

LlaMA 2 by Meta, released in July 2023, takes a unique position as the most popular and frequently cited open-source model of the year. Although it may not currently reign supreme on leaderboards, its influence is enduring, with many leading models deriving from its series. As the landscape of language models evolves rapidly, we have chosen LLaMA 2 as our representative open-source model due to its historical significance and impact. On the other hand, despite the existence of renowned models that outperform LlaMA 2 on numerous conventional benchmarks, such as Falcon 2.0 (Sakor et al., 2023) and Mixtral 8x7B (Mistral AI. 2023), it remains a challenge to identify a single model that prevails across nearly all benchmarks.

PaLM 2 (Anil et al., 2023), Claude 2 (Anthropic, 2023), and Gemini (Gemini Team, Google. 2023) constitute the leading triumvirate of non-OpenAI models. Claude 2, from Anthropic, held the position of the premier non-OpenAI model until the advent of Google's Gemini. Grok-1 (xAI.,2023) may not be as good as these models. Inflection-2 (InflectionAI, 2023) is comparable to PaLM2.

Gemini, unveiled in December 2023, represents the latest closed-source marvel that outshines even GPT-4 in terms of performance. Its recent release solidifies its status as a frontrunner in the rapidly advancing field of language models. As we embark on this exploration, our curated selection provides a comprehensive snapshot of the diverse and competitive landscape of contemporary language modeling technology.

Our comparisons are based on the following criteria: Capabilities, Architecture, and Cost. For capabilities, we will focus on benchmarks divided into five categories: Knowledge, Reasoning, Comprehension, Math, and Safety, according to this paper. These capabilities are the most relevant indicators for the potential tasks of our project, even though we have not yet specified the details of the tasks.

It's important to note that the majority of model comparisons were conducted prior to the presence of Gemini. As a result, Gemini is not included in most of the comparisons made among each model. For comparisons involving Gemini, please refer to the section dedicated to the Gemini model.

Benchmarks for evaluation

Before describing the properties of the top models, we need to introduce the benchmarks for evaluating the capabilities (Zheng et al. 2023) of LLMs.

The selection of benchmarks is guided by these factors(Zheng et al. 2023) : 1) Ensuring a broad range of LLM functionalities are encompassed; 2) Utilizing benchmarks that are commonly employed in the assessment of LLMs; 3) Effectively differentiating between superior and inferior LLMs; 4) Corresponding closely to the practical usage scenarios of LLMs. Following this, Zheng et al. (2023) establish a classification system for capabilities. This is done first by listing the different types of tasks or capability categories, and then assigning chosen benchmarks to each category.

Knowledge. This category gauges the ability of LLM to utilize and apply extensive world knowledge, a skill that necessitates not only storing vast amounts of information but also linking disparate pieces of knowledge and logically reasoning about them. Currently, we have two sub-categories: 1) Question Answering, a direct assessment of the LLM's knowledge base through questioning. Benchmarks for this sub-category involve Natural Questions5 (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013) and TriviaQA (Joshi et al., 2017); 2) Multi-subject Test, which employs questions from human exams to assess LLMs. Well-regarded benchmarks such as MMLU (Hendrycks et al., 2021a), AGIEval (Zhong et al., 2023) (specifically the English section, referred to as AGIEval-EN), and ARC (Clark et al., 2018) (with partitions ARC-e and ARC-c to distinguish between easy and challenging difficulty levels) are used in our evaluation process.

Reasoning. This category is designed to gauge the overall reasoning abilities of LLMs. This includes 1) Commonsense Reasoning, assessing the LLM's performance on tasks based on commonsense, which are generally simple for humans but can pose a challenge for LLMs. We utilize well-known benchmarks for commonsense reasoning such as LAMBADA (Paperno et al., 2016), HellaSwag (Zellers et al., 2019), and WinoGrande (Sakaguchi et al., 2021) in our evaluation; 2) Comprehensive Reasoning, which combines a variety of reasoning tasks into a single benchmark. For this, we use BBH (Suzgun et al., 2023), a commonly utilized benchmark that includes a subset of 23 difficult tasks from the BIG-Bench suite (Srivastava et al., 2023).

Comprehension. This category evaluates the reading comprehension skills of LLMs. This involves the LLMs first digesting the given context and subsequently answering related questions, a traditionally difficult task in the field of natural language understanding. For this category, we select widely recognized reading comprehension benchmarks like RACE (Lai et al., 2017) (which includes RACE-m and RACE-h partitions to distinguish between middle school and high school difficulty levels) and DROP (Dua et al., 2019).

Math. This category specifically tests LLM's mathematical capability. Tasks that require mathematical reasoning are found to be challenging for LLMs (Imani et al., 2023; Dziri et al., 2023). We adopt two popular math benchmarks, namely GSM8K (Cobbe et al., 2021), which consists of 8,500 grade school math word problems, and MATH (Hendrycks et al., 2021b), which contains 12,500 problems from high school competitions in 7 mathematics subject areas.

Coding. This category puts the coding capabilities of LLMs under scrutiny, a skill often considered fundamental for advanced LLMs. Zheng et al. (2023) opt for renowned benchmarks such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), both of which are natural language to code datasets requiring LLMs to craft standalone Python programs that meet a series of pre-defined test cases. In alignment with Chen et al. (2021), we employ the widely accepted pass@k metric: k code samples are produced for each coding problem, and if any sample meets the unit tests, the problem is deemed solved; the overall proportion of solved problems is then reported.

Safety. This category evaluates the ability of LLMs to produce content that is accurate, dependable, non-offensive, and impartial, thus conforming to human values. Presently, Zheng et al. (2023) has two sub-categories (with plans to expand with more benchmarks in the future): 1) Truthfulness utilizes TruthfulQA (Lin et al., 2022), a benchmark established to measure the factualness of an LLM; 2) Toxicity utilizes RealToxicityPrompts (Gehman et al., 2020) to determine the likelihood of producing harmful output.

Definitions

Model variations: At times, benchmarks use different subset models to perform their benchmark evaluations. You may see GPT3.5 written as gpt3.5.0613 which indicates it's continuation or end date. You can find more definitions on these various submodels here.

Exploring In-Context Learning:

In-context learning stands out for its adaptability in the number of examples needed for task adaptation. There are mainly three approaches:

Few-Shot Learning (or prompting)

In this approach, multiple input-output pairs are provided as examples for the model to grasp the task description(Brown et al., 2020; Parnami et al., 2022). These instances serve as a semantic precursor, enabling the model to generalize and execute the new task. This method capitalizes on the model's pre-training data and pre-existing model parameters to predict the next token accurately for complex tasks.

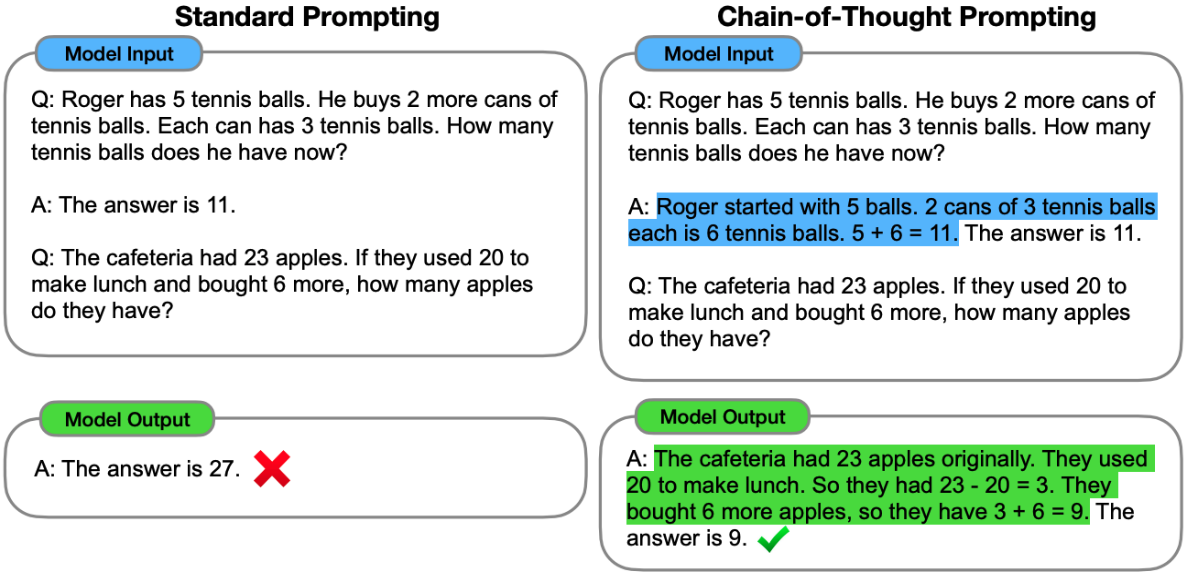
One-Shot Learning (or prompting)

One-shot learning is a more restricted form of in-context learning where only a single input-output example is given to comprehend the task. Despite the limited dataset, the model employs its pre-trained parameters and semantic precursor knowledge to generate an outcome that matches the task description. This technique is often used when there is a dearth of domain-specific data.

Chain-of-Thought Prompting:

Prompting with a chain-of-thought(CoT) is a strategy that bolsters the ability of LLMs to reason by interweaving intermediate steps of reasoning into the prompt (Wei et al. 2022). This method becomes especially potent when used in tandem with few-shot prompts for demanding reasoning tasks.

CoT Prompting shares a close relation with In-Context Learning (ICL), as both methodologies seek to utilize LLMs' pre-training data and model parameters for task-oriented learning. While the focus of ICL lies on few-shot learning and prompt engineering, CoT prompting underscores the continuity of thought, inciting intricate reasoning.



Standard Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Chain-of-Thought Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Figure 2: Chain-of-Thought Prompting Example (Wei et al. 2022)

Figure 2 shows an example of how Chain-of-Thought Prompting works. Zero-shot CoT Prompting is a further development of CoT Prompting that includes the phrase "Let's think step by step" into the original prompt. This strategy proves especially beneficial in situations where there are limited examples for the prompt.

Zero-Shot Learning (or prompting)

In zero-shot learning, no task-specific examples are provided to the model. Instead, it relies exclusively on the task description and pre-existing training data to deduce the requirements. This method assesses the model's inherent capabilities to generalize from its pre-training phase to new, unseen tasks. The following is an example for zero-shot prompting:

- **Prompt**: "I visited the market and purchased 10 apples. I offered 2 apples each to the neighbor and the repairman. Later, I bought an additional 5 apples and consumed 1. How many apples do I have left? Let's think step by step."
- **Output**: "To start with, you had 10 apples. You presented 2 apples to the neighbor and 2 to the repairman, leaving you with 6 apples. You then purchased 5 more apples, bringing your total to 11 apples. After consuming 1 apple, you were left with 10 apples."

As you can see, zero-shot prompting has the model output from no prior context setting besides its base model training. While this example works in the correct output, it is less likely to handle more complex situations in the same context window.

The reference Table for model comparison

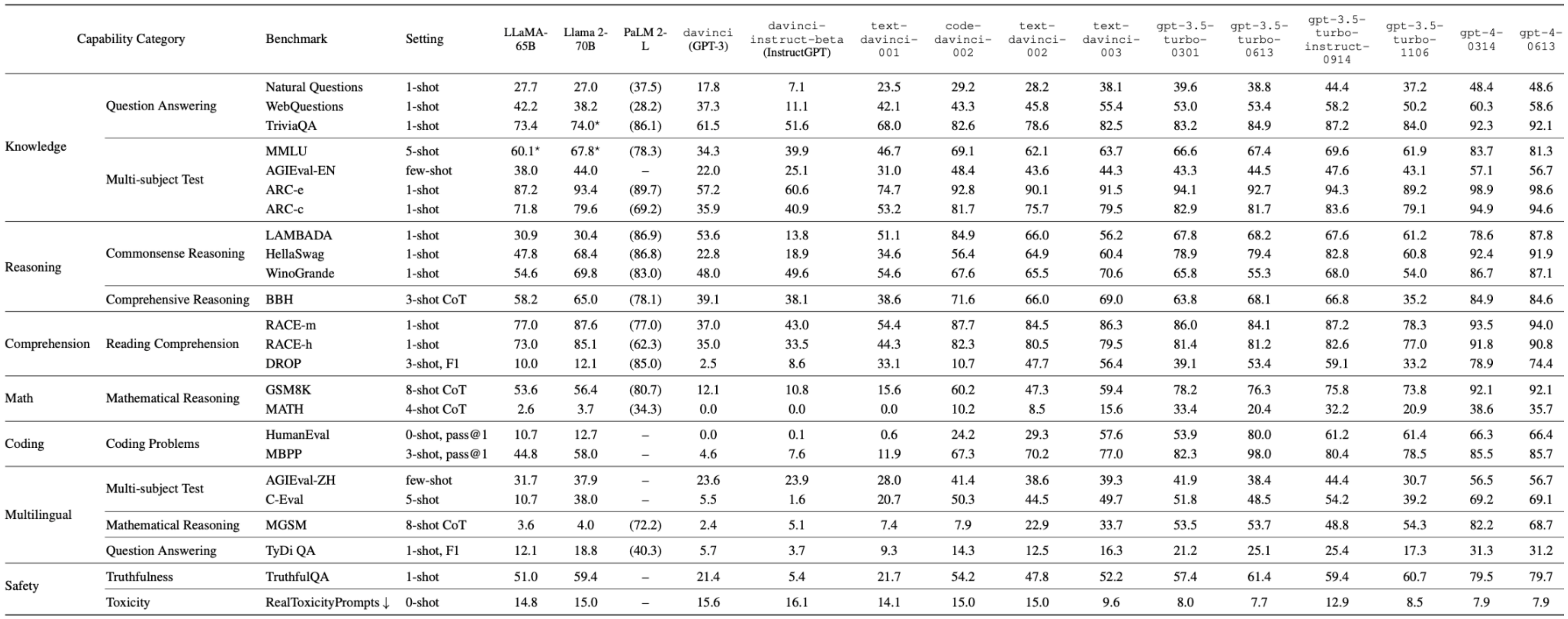
To enhance our understanding of LLMs, a comparison of the key properties of the top six models is conducted for our project. This comparison will provide additional insights into the LLMs.

Table 1 is taken from the research done by Zheng et al. (2023) that evaluates the capabilities of relevant models. In the following section our six chosen models are compared with reference to this table, except Google Gemini (not released by comparison date) and Claude 2, which has its own table (Table 4) comparison.

The Exact Match (EM) accuracy percentage is the default metric unless stated otherwise. For a prediction to be considered correct under EM, it must match the ground truth or reference answer precisely. This means every element of the answer must be the same - every word, number, punctuation mark, etc. Even a minor deviation results in the answer being counted as incorrect. Because of its all-or-nothing nature, EM is a very strict metric. It's often used in tasks where precision is critical, such as evaluating factual correctness in question-answering systems. Exact Match is often used alongside other metrics in translation to account for partial matches or similarities. This is done to provide metrics which can include a more nuanced view of performance, as they consider cases where the prediction might be close to correct but not an exact match.

To enhance comprehension, they also specify the quantity of "shots" utilized in prompts and whether Chain-of-Thought (CoT; Wei et al. 2022) prompting technique is incorporated. For the AGIEval (Zhong et al., 2023) benchmark, they apply the official few-shot (3-5 shots) setting. For PaLM 2-L, since there is no API access available currently, they cite figures from PaLM 2 (Anil et al., 2023) instead. Figures in brackets are those not derived from their experiments. For more details, please check their paper (Zheng et al. 2023).

Table 1: Primary evaluation outcomes for GPT-Fathom (Zheng et al. 2023).



Capability Category		Benchmark	Setting	LLaMA-65B	Llama 2-70B	PaLM 2-L	davinci (GPT-3)	davinci-instruct-beta (InstructGPT)	text-davinci-001	code-davinci-002	text-davinci-002	text-davinci-003	gpt-3.5-turbo-0301	gpt-3.5-turbo-0613	gpt-3.5-turbo-instruct-0914	gpt-3.5-turbo-1106	gpt-4-0314	gpt-4-0613
Knowledge	Question Answering	Natural Questions	1-shot	27.7	27.0	(37.5)	17.8	7.1	23.5	29.2	28.2	38.1	39.6	38.8	44.4	37.2	48.4	48.6
		WebQuestions	1-shot	42.2	38.2	(28.2)	37.3	11.1	42.1	43.3	45.8	55.4	53.0	53.4	58.2	50.2	60.3	58.6
		TriviaQA	1-shot	73.4	74.0*	(86.1)	61.5	51.6	68.0	82.6	78.6	82.5	83.2	84.9	87.2	84.0	92.3	92.1
	Multi-subject Test	MMLU	5-shot	60.1*	67.8*	(78.3)	34.3	39.9	46.7	69.1	62.1	63.7	66.6	67.4	69.6	61.9	83.7	81.3
		AGIEval-EN	few-shot	38.0	44.0	–	22.0	25.1	31.0	48.4	43.6	44.3	43.3	44.5	47.6	43.1	57.1	56.7
		ARC-e	1-shot	87.2	93.4	(89.7)	57.2	60.6	74.7	92.8	90.1	91.5	94.1	92.7	94.3	89.2	98.9	98.6
		ARC-c	1-shot	71.8	79.6	(69.2)	35.9	40.9	53.2	81.7	75.7	79.5	82.9	81.7	83.6	79.1	94.9	94.6
Reasoning	Commonsense Reasoning	LAMBADA	1-shot	30.9	30.4	(86.9)	53.6	13.8	51.1	84.9	66.0	56.2	67.8	68.2	67.6	61.2	78.6	87.8
		HellaSwag	1-shot	47.8	68.4	(86.8)	22.8	18.9	34.6	56.4	64.9	60.4	78.9	79.4	82.8	60.8	92.4	91.9
		WinoGrande	1-shot	54.6	69.8	(83.0)	48.0	49.6	54.6	67.6	65.5	70.6	65.8	55.3	68.0	54.0	86.7	87.1
	Comprehensive Reasoning	BBH	3-shot CoT	58.2	65.0	(78.1)	39.1	38.1	38.6	71.6	66.0	69.0	63.8	68.1	66.8	35.2	84.9	84.6
Comprehension	Reading Comprehension	RACE-m	1-shot	77.0	87.6	(77.0)	37.0	43.0	54.4	87.7	84.5	86.3	86.0	84.1	87.2	78.3	93.5	94.0
		RACE-h	1-shot	73.0	85.1	(62.3)	35.0	33.5	44.3	82.3	80.5	79.5	81.4	81.2	82.6	77.0	91.8	90.8
		DROP	3-shot, F1	10.0	12.1	(85.0)	2.5	8.6	33.1	10.7	47.7	56.4	39.1	53.4	59.1	33.2	78.9	74.4
Math	Mathematical Reasoning	GSM8K	8-shot CoT	53.6	56.4	(80.7)	12.1	10.8	15.6	60.2	47.3	59.4	78.2	76.3	75.8	73.8	92.1	92.1
		MATH	4-shot CoT	2.6	3.7	(34.3)	0.0	0.0	0.0	10.2	8.5	15.6	33.4	20.4	32.2	20.9	38.6	35.7
Coding	Coding Problems	HumanEval	0-shot, pass@1	10.7	12.7	–	0.0	0.1	0.6	24.2	29.3	57.6	53.9	80.0	61.2	61.4	66.3	66.4
		MBPP	3-shot, pass@1	44.8	58.0	–	4.6	7.6	11.9	67.3	70.2	77.0	82.3	98.0	80.4	78.5	85.5	85.7
Multilingual	Multi-subject Test	AGIEval-ZH	few-shot	31.7	37.9	–	23.6	23.9	28.0	41.4	38.6	39.3	41.9	38.4	44.4	30.7	56.5	56.7
		C-Eval	5-shot	10.7	38.0	–	5.5	1.6	20.7	50.3	44.5	49.7	51.8	48.5	54.2	39.2	69.2	69.1
	Mathematical Reasoning	MGSM	8-shot CoT	3.6	4.0	(72.2)	2.4	5.1	7.4	7.9	22.9	33.7	53.5	53.7	48.8	54.3	82.2	68.7
	Question Answering	TyDi QA	1-shot, F1	12.1	18.8	(40.3)	5.7	3.7	9.3	14.3	12.5	16.3	21.2	25.1	25.4	17.3	31.3	31.2
Safety	Truthfulness	TruthfulQA	1-shot	51.0	59.4	–	21.4	5.4	21.7	54.2	47.8	52.2	57.4	61.4	59.4	60.7	79.5	79.7
	Toxicity	RealToxicityPrompts ↓	0-shot	14.8	15.0	–	15.6	16.1	14.1	15.0	15.0	9.6	8.0	7.7	12.9	8.5	7.9	7.9

This table categorizes "capabilities" and uses an open-source and reproducible LLM evaluation suite built on top of OpenAI Evals Zheng and colleagues (2023). The default metric employed is the Exact Match (EM) accuracy, expressed as a percentage, unless specified otherwise. Also detailed is the quantity of "shots" incorporated in the prompts and whether the Chain-of-Thought (CoT; Wei et al., 2022) prompting technique is utilized. Any figures that are not the result of their direct experimentation are indicated within parentheses. Figures marked with a star symbol (*) represent those derived from refined prompts, which are further discussed in Section 3.2 of Zheng et al. (2023)

Pricing Comparison

In order to compare the price of the models, a price table that offers a comparative look at various AI models is constructed, including GPT-3.5turbo, GPT-4, GPT-4 Turbo, gpt-4-1106-preview, Claude 2, Claude 2.1, PaLM2, Gemini Pro, and Gemini Ultra, highlighting differences in input and output costs, context window sizes, and training data timelines. The input costs range significantly, with Claude 2.1 and PaLM2 being the most cost-effective, while GPT-4 stands on the higher end. Output costs generally double from their input counterparts across most models, reflecting the operational and processing costs involved. The context window sizes vary dramatically, with GPT-4 Turbo offering an exceptionally large window, suggesting its capability to handle complex tasks involving extensive text. On the other hand, models like Claude 2.1 and PaLM2 have smaller context windows, potentially making them more efficient for specific, shorter tasks.

The training data timeline indicates the recency of the model's knowledge, with GPT-4 Turbo and gpt-4-1106-preview being the most recent, possibly offering advanced understanding and performance. This diversity in training timelines suggests varying degrees of relevance in the models' knowledge base. Notably, Gemini Pro and Gemini Ultra lack available data, leaving their specifics to speculation or indicating a proprietary or under-development status.

Table 2: LLM Models Price comparison

Category	**GPT-3.5 turbo**	**GPT-4**	**GPT-4 Turbo gpt-4-1106**	**Claude 2**	**Claude 2.1**	**PaLM2**	**Gemini Pro**	**Gemini Ultra**
Input	$0.001/1K tokens	$0.03/1K tokens	$0.01/1K tokens	$0.008/1K tokens	$0.008/1K tokens	$0.00025/1K characters	$0.00025 / 1k characters	Not available
Output	$0.002/1K tokens	$0.06/1K tokens	$0.03/1K tokens	0.024/1K tokens	0.024/1K tokens	$0.0005/1K characters	$0.0005 / 1k characters	Not available
Context Window	4,096	8,192	128,000	100,000	200,000	8,000 tokens	32,000 tokens	32,000 tokens
Training Data	Sep 2021	Sep 2021	Apr 2023	Dec 2023	Early 2023	Mid 2021	Not available	Not available

In essence, the choice between them would depend on specific needs, including task complexity, budget, and the necessity for up-to-date language understanding. From cost-effective, specialized models to more advanced, comprehensive tools, the range indicates a broad targeting of potential applications and user requirements. The variation in context window size and costs across these models underscores the wide array of applications they are designed to cater to, from simple, quick tasks to more demanding, in-depth analyses.

Models

GPT-3.5 Turbo

GPT-3.5-Turbo is an advanced language model developed by OpenAI as proprietary software(closed-source). This version of the GPT-3 model has been fine-tuned for better performance and efficiency, hence the 'turbo'.

GPT-3.5-Turbo, like its predecessor, is based on the transformer architecture and has been trained on a diverse range of internet text. However, it also incorporates valuable modifications and improvements that boost its capabilities.

The model is capable of generating human-like text based on the prompts given to it. It can answer questions, write essays, summarize texts, and even generate Python code. It can translate languages, simulate characters for video games, tutor in a variety of subjects, and much more.

The '3.5' in its name signifies that it's a mid-way point between GPT-3 and an eventual GPT-4. The 'turbo' indicates its enhanced capabilities and performance. Its design is aimed at providing developers with a powerful tool for building applications with sophisticated natural language processing capabilities.

- **Capabilities**
    - **Knowledge: In the comparison of various language models, GPT-3.5-Turbo may not outperform Claude 2 or GPT-4, but it does hold a significant advantage over other models such as Llama 2 and Palm 2.**

      In benchmarks Natural TriviaQA(5-shot), MMLU(5-shot CoT), and ARC-c(5-shot), GPT-3.5-turbo0613 scores 80.6, 67.1, and 84.1 respectively, which are all lower than Claude 2's(87.5, 78.5, and 19.0) and GPT-4-0613's (92.7,82.7, and 94.9).

      In benchmarks Natural Questions(1-shot), WebQuestions(1-shot), and TriviaQA(1-shot), GPT-3.5-turbo0613 scores 38.8, 53.4, and 84.9 respectively, which is all higher than Llama Llama 2-70b, PaLM 2L but lower than GPT-4.

      In benchmarks MMLU(5-shot), ARC-e(1-shot), and ARC-c(1-shot), GPT-3.5-turbo0613 and Llama 2 are really close, where GPT-3.5-turbo0613 scores 67.4, 92.7, and 81.7 and Llama 2 scores 67.8, 93.4, and 79.6 respectively.

    - **Reasoning: it appears that the model GPT-3.5-turbo 0613 does not quite reach the performance levels of PalM 2L and GPT-4 (maybe Claude 2L or Gemini)** in LAMBADA(1-shot), HellaSwag(1-shot), WinoGrande(1-shot), and BBH(3-shot CoT).

      GPT-3.5-turbo0613 loses Llama 2- 70B only in WinoGrande(1-shot).

    - **Comprehension: GPT-3.5-turbo0613 loses Llama2, Claude 2, and GPT-4(maybe Gemini)** in all benchmarks, e.g., RACE-m(1-shot), RACE-h(1-shot), DROP(3-shot,F1).

      In benchmarks RACE-m(1-shot), RACE-h(1-shot), DROP(3-shot,F1), GPT-3.5-turbo0613 scores 84.1, 81.2, 53.7, and Llama 2L scores 87.6, 85.1, 67.6, and PaLM 2L scores 77.0, 62.3, 85.0 respectively, where Llama 2L wins PaLM 2L and GPT-3.5-turbo0613 in RACE-m, RACE-h but PaLM 2L wins in DROP.

    - **Math: GPT-3.5 loses PalM 2L, Claude 2, and GPT-4(maybe Gemini)** in all benchmarks, e.g., MATH and GSM8K. Llama 2L is weak in Math.
    - **Coding: GPT 3.5 ranks top in all benchmarks**, e.g., HumanEval(0-shot) and MBPP(3-shot) benchmarks.

      GPT3.5turbo scores 80.0 and 90.0 in HumanEval(0-shot) and MBPP(3-shot), but GPT4 0613 scores 66.4 and 85.7.

    - **Safety: It's kind of mixed in safety benchmarks** like TruthfulQA(1-shot) and RealToxicityPrompts(0-shot).
- **Architecture:** Transformer-based neural network.
- **Cost:** Pay-as-you-go API access, likely with different pricing tiers. Please also see our Price Table.

GPT-4

GPT-4 stands as the most recent variant of the Generative Pretrained Transformers developed by OpenAI as proprietary software as its predecessor. This series, also encompassing GPT-1, GPT-2, and GPT-3, was engineered with the aim of generating text that mirrors human language, guided by the input it receives.

As the latest addition to this series, GPT-4 symbolizes a considerable advancement in performance. It underwent training using a wide array of internet text on high-performance AI supercomputers provided by Microsoft Azure. This has enabled it to deliver striking outcomes on multiple Natural Language Processing (NLP) tasks, positioning it amongst the leading language models presently in existence.

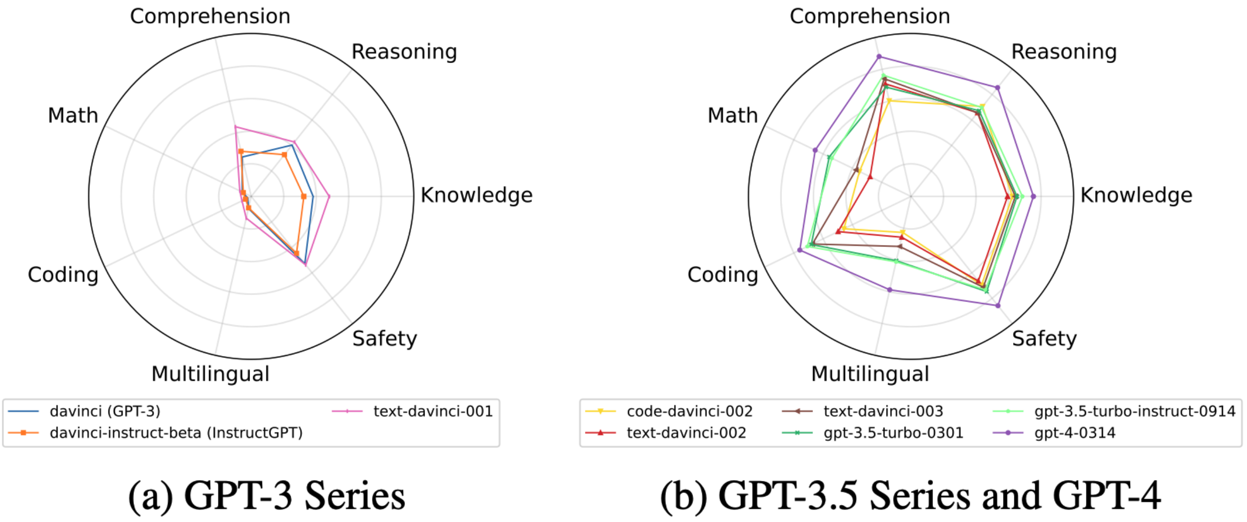
As a Transformer-based model (Vaswani et al., 2017), GPT-4 was initially trained to anticipate subsequent tokens in text sequences, leveraging data sourced from the public domain (like internet content) as well as data procured via third-party agreements. Subsequently, the model underwent refinement through a process known as Reinforcement Learning from Human Feedback (RLHF)（Bai et al. 2022a). In light of the intense competition in the field and the potential safety risks posed by large-scale models such as GPT-4, the “GPT-4 Technical Report” (OpenAI, 2023) deliberately omits specific details about its architecture (including the size of the model), the hardware used, computation resources for training, dataset creation, training methodologies, or any related aspects.

- **Capabilities**
  - **Knowledge: top 1 models,** i.e., gpt-4-0314 and gpt-4-0613, in all benchmarks, e.g., Natural Questions, WebQuestions,TriviaQA, MMLU,AGIEval-EN, ARC-e, ARC-c.
  - **Reasoning: GPT-4-0314 and GPT-4-0613 outperform in most benchmarks**, e.g., LAMBADA, HellaSwag, WinoGrande, and BBH. The only exception is that PaLM 2L (86.9) wins GPT-4-0314(78.6) in LAMBADA. PaLM 2L is compatible to GPT-4 in the Reasoning benchmarks.
  - **Comprehension: GPT-4 leads in most benchmarks**, e.g., RACE-m, RACE-h, DROP. The only exception is that PaLM 2L (85.0) wins

    GPT-4-0314(78.7) in DROP. PaLM 2L is only compatible to GPT-4 in DROP. Claude 2(88.3) is compatible to GPT-4(90.8) in RACE-h.
  - **Math: gpt-4-0314 and gpt-4-0613 win the benchmark MATH.** Exceptionally, PaLM 2L (34.3) is compatible to GPT-4(34.9) in MATH, and Claude 2(88.0) wins GPT-4(83.9) in GSM8K (0-shot CoT).
  - **Coding:** GPT-4 achieves scores of 66.4 and 85.7 in HumanEval and MBPP benchmarks respectively, securing the second position in both of these benchmarks. **However, GPT-3.5-turbo0613 outperforms GPT-4** in both these benchmarks with respective scores of 80.0 and 98.0.
  - **Safety: GPT-4's performance is somewhat mixed.** It scores 7.9 in RealToxicityPrompts, which is lower than LLaMa, LLaMa 2, GPT 3, and GPT 3.5turbo. However, in the TruthfulQA benchmark, GPT-4 excels by scoring 79.5, the highest score amongst all models.

- **Architecture:** Transformer-based neural network, with GPT-4 likely having more parameters than GPT-3.5. GPT-4 underwent training utilizing the high-performance AI supercomputers provided by Microsoft Azure. The infrastructure optimized for AI from Azure also enables us to offer GPT-4 to users globally.
- **Cost:** Pay-as-you-go API access. Please also see our Price Table.

The radar charts in figure 3 can help to visualize the capability distribution of OpenAI's GPT series.



(a) GPT-3 Series

(b) GPT-3.5 Series and GPT-4

Figure 3: Radar chart comparing GPT-3, 3.5 and 4 capabilities (Zheng et al. 2023)

The following outline gives a brief overview of OpenAI available models with descriptions. For more details please visit https://platform.openai.com/docs/models:

The OpenAI API is powered by a diverse set of models with different capabilities and price points, offering fine-tuning for specific use cases. The models include GPT-4 and GPT-4 Turbo, which are improvements on GPT-3.5 and are capable of understanding and generating natural language or code. GPT-3.5 is a set of models that also improve on GPT-3 and have similar capabilities. DALL-E is a model that can generate and edit images based on natural language prompts. TTS represents a set of models that convert text into natural-sounding spoken audio. Whisper is a model that converts audio into text, while Embeddings are models that convert text into a numerical form. The Moderation model is fine-tuned to detect whether text may be sensitive or unsafe. GPT base includes models without instruction following, but can understand and generate natural language or code. GPT-3 Legacy is a set of models that understand and generate natural language. Deprecated models are those that have been replaced, and a full list is available along with the suggested replacement.

PaLM 2

PaLM is a powerful large language model (LLM) currently under development by Google AI was unveiled in May 2023 as proprietary software. Leveraging Google's extensive datasets for its training, PaLM is poised to significantly impact the field of machine learning and responsible artificial intelligence (AI). Its capabilities extend to understanding language, generating natural language responses, offering machine translation, code generation, summarization, among other creative features. PaLM 2 has huge improvements over its predecessor PaLM (Chowdhery et al., 2022).

- **Capabilities:**
    - **Knowledge: PaLM 2's performance is somewhat mixed.** It scores high in TriviaQA and MMLU benchmark, but low in ARC-e, ARC-c, and WebQuestions.
      In the TriviaQA benchmark, PaLM 2L achieves a score of 86.1. This surpasses both Llama 2-70B and gpt-3.5-turbo0613, but falls short of gpt-4 and gpt-3.5-turbo instruct0914.

      In the MMLU benchmark, PaLM 2L registers a score of 78.3, exceeding the scores of Llama 2-70B (67.8) and gpt-3.5-turbo0613 (67.4), and matching Claude 2 (78.5). However, it is lower than GPT-4 (83.7).

      For the ARC-e, ARC-c, and WebQuestions benchmarks, PaLM 2L scores are 89.7, 69.2, and 28.2 respectively, which are the lowest among the models we considered.

    - **Reasoning: Expected to be strong in reasoning.** PaLM 2L ranks 2nd in most of benchmarks.
      In LAMBADA, HellaSwag, WinoGrande it scores 86.9, 86.8, 83.0 respectively, which are even higher than gpt-4-0314 (78.6) in LAMBADA but lower than gpt-4-0314 in HellaSwag(92.4) and WinoGrande(86.7), and lower than gpt-4-0613 in all benchmarks (87.8, 91.9, and 87.1).

      In BHH, PaLM 2L scores 78.1, which is only lower than gpt-4(84.6).

    - **Comprehension: PaLM 2 holds the second place in the DROP, while its performance is relatively lower in the RACE-m and RACE-h benchmarks**.

      In RACE-m, RACE-h, DROP benchmarks, it scores 77.0, 62.3, 85.0 respectively, which are lower than Llama2 70B(87.6, 85.1), Claude2L (-, 88.3), GPT 3.5-turbo0613(84.1, 81.2), GPT-40613 (94.0, 90.8) in RACE-m, RACE-h, and only lower than GPT-40613 (87.2) in DROP.

    - **Math: Expected to be proficient. PaLM 2 ranks 3 in GSM8K, and is compatible with GPT-4 in MATH benchmark.**
      PaLM 2L scores 80.7 and 34.3 respectively in GSM8K and MATH benchmarks, where GPT-40613 scores 92.1 and 34.9, and Claude 2 (88.0) higher than GPT-4 0613 (83.9) in GSM8K (0-shot CoT).
    - **Coding: Its coding capability is somewhat less superior compared to other models.** Within the HumanEval benchmark, it only outperforms Llama 2 among the chosen models, and it holds the last position in the MBPP benchmark. See also “PaLM 2 Technical Report“ (Anil et al., 2023) and our Tables.
    - **Safety:** Google typically implements robust safety measures. But no specific information is available.
      Key to PaLM's design is its emphasis on privacy and data security. The model has the ability to encrypt data and guard it against unauthorized access, making it an ideal tool

for projects that require enhanced security measures such as developing secure e-commerce websites or platforms handling sensitive user information.

- **Architecture:** PaLM 2 is a model rooted in Transformer architecture, and its training is conducted using a mixture of objectives.
- **Cost:** As a Generative AI on Vertex AI. Please check our Price Table or their website.

Llama 2

Llama 2 is a family of advanced, open-source large language models released by Meta in July 2023, the parent company of Facebook. It is made available for both research and commercial use under a very permissive community license. This makes Llama 2 a significant player in the AI space, as it offers a credible alternative to closed-source AI models like OpenAI's GPT and Google's AI models.

The Llama 2 release introduces a variety of pre-trained and fine-tuned Language Models, ranging from 7 billion to 70 billion parameters. These models have been significantly improved over their predecessors, trained on 40% more tokens, with a longer context length and utilizing grouped-query attention for fast inference.

A particularly exciting aspect of this release is the fine-tuned models (Llama 2-Chat) that have been optimized for dialogue applications using Reinforcement Learning from Human Feedback (RLHF). These models perform better than most open models in terms of helpfulness and safety benchmarks, achieving comparable performance to ChatGPT according to human evaluations.

Llama 2 was trained with 2 trillion "tokens" from publicly available sources like Common Crawl (an archive of billions of webpages), Wikipedia, and public domain books from Project Gutenberg. To ensure safe and appropriate responses, the developers also employed reinforcement learning with human feedback (RLHF).

Even though Llama 2 is a powerful tool in its own right, it is designed to be further trained to meet specific needs. For instance, it can be trained with examples to generate text in a specific brand style or voice, or fine-tune a chat-optimized model to respond to customer support requests by providing it with FAQs and other relevant information.

While Llama 2 does not outperform all other models in the AI space, it is generally as good as GPT-3.5 and PaLM on most benchmarks but doesn't perform as well as GPT-4 or PaLM 2. However, it's important to note that Llama 2 is not trying to be a direct competitor to these models; instead, it offers something a little different.

Llama 2 is a significant development because it's freely available for almost anyone to use for research and commercial purposes. This openness could potentially disrupt the AI space and provide a credible alternative to closed-source AIs. Therefore, Llama 2 is not only an advanced tool for AI applications but also a driver for open research and experimentation in AI.

- **Capabilities:**
    - **Knowledge: Llama 2 is relatively weaker than others in benchmarks.**

      In MMLU(5-shot) Llama 2 is comparable to GPT-3.5turbo but lower than Claude 2, PaLM 2L, GPT-4 (and Gemini)

      Llama 2 is the lowest among the models in Natural Questions(1-shot), WebQuestions(1-shot), and TriviaQA(1-shot).

      In ARC-e(1-shot) and ARC-c(1-shot), Llama 2 is higher than PaLM 2L and comparable to GPT-3.5turbo.

    - **Reasoning: Llama 2 ranks almost the lowest in** LAMBADA(1-shot), HellaSwag(1-shot), WinoGrande(1-shot), and BBH(3-shot CoT).

      Llama 2- 70B only win GPT-3.5-turbo0613 in WinoGrande(1-shot).

    - **Comprehension: Llama 2 is relatively strong in Comprehension,**

      it wins PaLM 2L and GPT-3.5burbo in all benchmarks, e.g., RACE-m(1-shot), RACE-h(1-shot), DROP(3-shot,F1).

    - **Math: Llama 2 ranks the lowest among the 6 models in Math benchmarks.**
    - **Coding: Llama 2 ranks lower among the 6 models in coding benchmarks.**
    - **Safety: Llama 2 is above the average in coding benchmarks.**
- **Architecture:** transformer-based or similar.
- **Cost:** Open-source, so no direct cost; costs for infrastructure or computing resources needed.

Zheng et al.(2023) also evaluated the entire LLaMA / Llama 2 family, including models ranging from 7B to 65B / 70B parameters, and reported the complete results in Table 3 below.

Table 3: Complete evaluation results of LLaMA and Llama 2 family models (Zheng et al. 2023)

Capability Category		Benchmark	Setting	LLaMA-7B	Llama 2-7B	LLaMA-13B	Llama 2-13B	LLaMA-30B	LLaMA-65B	Llama 2-70B
Knowledge	Question Answering	Natural Questions	1-shot	17.6	19.8	20.8	27.6	24.0	27.7	27.0
		WebQuestions	1-shot	37.0	38.3	37.6	42.8	39.0	42.2	38.2
		TriviaQA	1-shot	52.0	61.1	66.6	70.0	73.5	73.4	74.0
	Multi-subject Test	MMLU	5-shot	25.1	41.0	38.5	49.5	51.0	60.1	67.8
		AGIEval-EN	few-shot	19.1	25.7	27.0	35.7	34.7	38.0	44.0
		ARC-e	1-shot	30.0	62.3	67.6	76.4	82.4	87.2	93.4
		ARC-c	1-shot	26.7	48.6	49.1	55.7	60.8	71.8	79.6
Reasoning	Commonsense Reasoning	LAMBADA	1-shot	19.0	38.0	47.0	56.4	32.5	30.9	30.4
		HellaSwag	1-shot	24.6	25.4	28.9	37.2	31.3	47.8	68.4
		WinoGrande	1-shot	50.4	50.2	48.1	52.1	51.3	54.6	69.8
	Comprehensive Reasoning	BBH	3-shot CoT	33.7	38.4	39.1	46.2	49.6	58.2	65.0
Comprehension	Reading Comprehension	RACE-m	1-shot	26.7	45.8	52.4	57.9	65.3	77.0	87.6
		RACE-h	1-shot	29.1	39.5	48.5	55.1	64.1	73.0	85.1
		DROP	3-shot, F1	9.6	7.7	8.7	9.3	9.8	10.0	12.1
Math	Mathematical Reasoning	GSM8K	8-shot CoT	13.9	17.2	18.4	28.6	35.1	53.6	56.4
		MATH	4-shot CoT	0.4	0.1	0.4	0.5	0.5	2.6	3.7
Coding	Coding Problems	HumanEval	0-shot, pass@1	7.0	14.6	9.7	15.8	7.2	10.7	12.7
		MBPP	3-shot, pass@1	23.7	39.2	29.5	46.0	38.5	44.8	58.0
Multilingual	Multi-subject Test	AGIEval-ZH	few-shot	22.3	23.4	23.5	29.7	28.4	31.7	37.9
		C-Eval	5-shot	11.5	10.3	14.8	28.9	10.1	10.7	38.0
	Mathematical Reasoning	MGSM	8-shot CoT	2.7	2.3	2.8	4.1	3.1	3.6	4.0
	Question Answering	TyDi QA	1-shot, F1	2.4	3.6	3.2	4.5	3.8	12.1	18.8
Safety	Truthfulness	TruthfulQA	1-shot	37.6	31.0	29.5	38.0	44.5	51.0	59.4
	Toxicity	RealToxicityPrompts ↓	0-shot	14.5	14.8	14.9	14.8	14.7	14.8	15.0

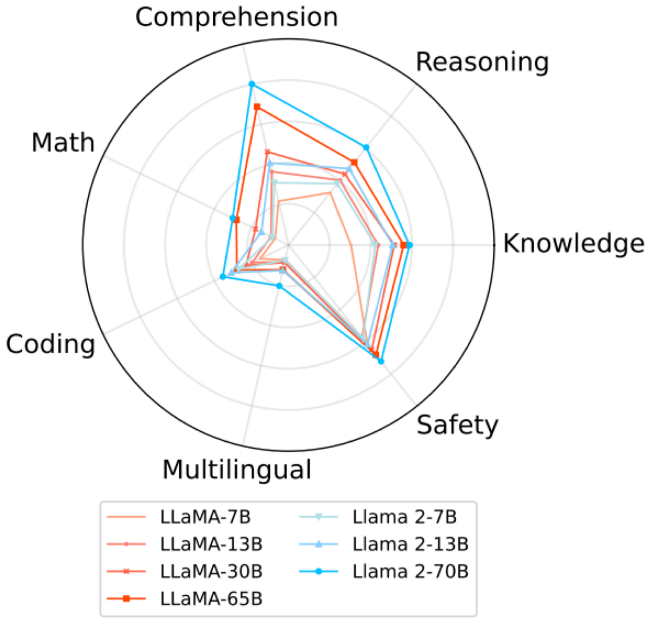


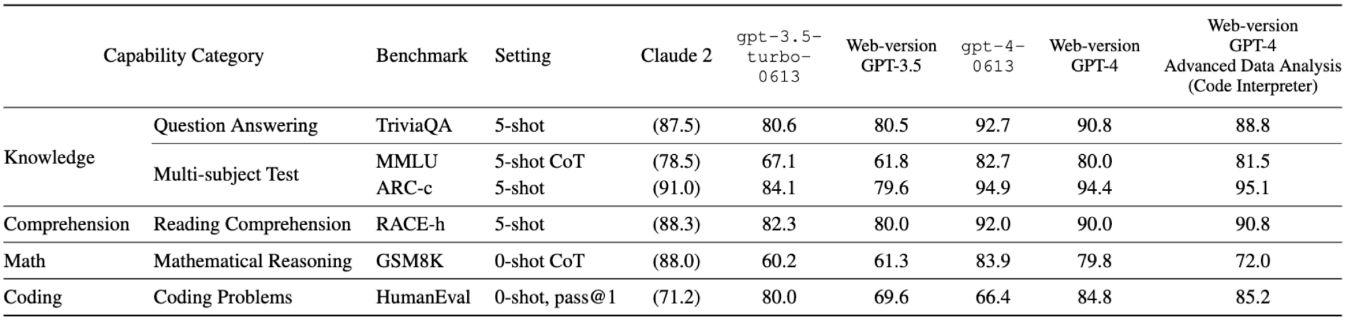
Figure 4: Radar charts of LLaMA and Llama 2 capabilities (Zheng et al. 2023)

Claude2

Anthropic, an artificial intelligence (AI) and “public benefit” company, launched Claude 2 on July 2023 as proprietary software which indeed stood as the leading non-OpenAI model before the release of Google’s Gemini.

- **Capabilities**
    - **Knowledge: Claude 2 has top 2 performance in Knowledge benchmarks**, where it just lower than gpt-4-0613.
      Claude 2 scores 87.5, 78.5 and 91.0 respectively in TriviaQA(5-shot), MMLU(5-shot CoT), ARC-c(5-shot) benchmarks, which is higher than GPT-3.5 turbo0613 (80.6, 67.1, 84.1) but lower than gpt-4-0613(92.7, 82.7,94.9)
    - **Reasoning: Claude 2 is expected to be proficient. GPT-4 slightly surpassed Claude 2** in the ARC common sense reasoning assessment, achieving an accuracy rate of 83% compared to Claude 2's 82%.(from the source https://www.akkio.com/post/gpt-4-vs-claude-2)
    - **Comprehension: Claude 2 has nearly top performance in reading comprehension benchmark.** In the RACE-h (5-shot) evaluation, it achieved a score of 88.3, outperforming both GPT-3.5-turbo0613 and the web version of GPT-3.5, which scored 82.3 and 80.0, respectively. This result is not far behind the scores of GPT-4-0613 and its web counterpart, which attained scores of 92.0 and 90.0.
    - **Math: Claude 2 showcased superior performance in the benchmarks.** Specifically, it achieved a score of 88.0 in the GSM8K (0-shot CoT) benchmark, surpassing GPT-4-0613 which scored 83.9, and outperforming all other models.
    - **Coding: Claude 2 exhibits decent performance in coding benchmarks.** Specifically, in the HumanEval(0-shot, pass@1) benchmark, it achieves a score of 71.2. This is higher than the web versions of GPT-3.5 and GPT-4-0613, which score 69.6 and 66.4 respectively. However, it falls short when compared to the scores of GPT-3.5-turbo0613 and the web version of GPT-4, which score 80.0 and 84.8 respectively. We estimate that Claude 2 surpasses LlaMa 2 and PaLM 2 in coding capabilities.
    - **Safety:** no specific information is available.

- **Architecture:** uses a transformer-based approach or similar technology.
- **Cost:** Typically offers API access with usage-based pricing models. Please see the Price Table.



Capability Category		Benchmark	Setting	Claude 2	gpt-3.5-turbo-0613	Web-version GPT-3.5	gpt-4-0613	Web-version GPT-4	Web-version GPT-4 Advanced Data Analysis (Code Interpreter)
Knowledge	Question Answering	TriviaQA	5-shot	(87.5)	80.6	80.5	92.7	90.8	88.8
	Multi-subject Test	MMLU	5-shot CoT	(78.5)	67.1	61.8	82.7	80.0	81.5
		ARC-c	5-shot	(91.0)	84.1	79.6	94.9	94.4	95.1
Comprehension	Reading Comprehension	RACE-h	5-shot	(88.3)	82.3	80.0	92.0	90.0	90.8
Math	Mathematical Reasoning	GSM8K	0-shot CoT	(88.0)	60.2	61.3	83.9	79.8	72.0
Coding	Coding Problems	HumanEval	0-shot, pass@1	(71.2)	80.0	69.6	66.4	84.8	85.2

Table 4: Claude 2 compared to other recent models(Zheng et al., 2023)

Table 4 compares the performance of Claude 2 and the most recent models from OpenAI under equivalent conditions. It should be noted that the web-based models (assessed in September 2023) are subject to updates at any moment, hence their behavior may vary from the earlier API-based models.

Gemini

Gemini Ultra, launched by Google in December as proprietary software, indeed stands as the leading non-OpenAI model, i.e., recognized as the most competitive LLM against OpenAI's leading models.

- **Capabilities**
  - **Knowledge: Gemini is the first model to outperform human experts on MMLU** of CoT@32(Cot with 32 samples). Gemini scores 90.0 in MMLU(CoT@32), which is higher than previous SOTA (GPT-4) of score 87.29.

    But in MMLU(5-shot), GPT-4 remains the top1 with the score 86.4, which is higher than Gemini's 83.7.

  - **Reasoning: Gemini is proficient in most benchmarks**, e.g., Big-Bench Hard, HellaSwag.
    Gemini Ultra scores 83.6 in Big-Bench Hard(3-shot), which is higher than 83.1 of GPT-4.
    Gemini Ultra scores 87.8 in HellaSwag(10-shot), which is lower than 95.3 of GPT-4.
  - **Comprehension: Gemini has the top performance in the DROP benchmark.**
    Gemini Ultra scores 82.4 in DROP(F1), which is higher than 80.9 of GPT-4.
  - **Math: Gemini Ultra exhibits a solid capability in mathematical analysis and problem-solving, showing performance levels that are on par with GPT-4 across multiple standard assessments.**
    In the GSM8K(Cobbe et al., 2021) elementary math benchmark, Gemini Ultra demonstrates robust capabilities, achieving a 94.4% score in the maj1@32* metric. On the other hand, GPT-4's achieves 92.0% within a 5-shot COT setting. These two scores are based on two different metrics.
    As the complexity of the math problems increases, such as those in the middle- and high-school math competitions (MATH benchmark; Hendrycks et al., 2021b), Gemini

Ultra continues to excel, achieving a 53.2% success rate when using 4-shot prompting, which is little higher than 52.9% of GPT-4 and tops all other competing models(Gemini Team Google, 2023).
Moreover, Gemini Ultra also showcases its superiority in tackling highly challenging problems derived from the American Mathematical Competitions (questions from 2022 and 2023). While smaller models struggle with these tasks, scoring almost randomly, Gemini Ultra manages to solve 32% of the questions, surpassing GPT-4's solve rate of 30%.

* The term "maj1@32" after GSM8K refers to a specific evaluation metric. Here's a breakdown of what it likely means: - maj1: This typically stands for "majority 1" or a similar concept where the AI's responses are aggregated, and the most common response is considered. If an AI model is asked the same question multiple times or in slightly different ways, the majority answer (the one given most frequently) is taken as the final answer. - @32: This usually denotes the number of attempts, samples, or different formulations of a problem that are considered when calculating the majority response. In this case, it suggests that the majority answer was determined across 32 different instances or trials. So, "maj1@32 score" means that Gemini Ultra achieved a 94.4% accuracy rate where the most common answer it provided over 32 trials was correct.

  - **Coding: Gemini Ultra also excels on coding benchmarks.**
    Its performance has been evaluated on conventional and internal benchmarks, as well as within complex reasoning systems like AlphaCode 2 (refer to section 5.1.7 on complex reasoning systems of their report (Gemini Team Google, 2023).
    On the standard HumanEval code-completion benchmark (Chen et al., 2021), which maps function descriptions to Python implementations, Gemini Ultra correctly implements 74.4% of problems.
    On the new benchmark called Natural2Code, created for Python code generation tasks, it achieves the highest score of 74.9%.
  - **Safety:** Gemini has been developed with a focus on responsibility and safety. It has undergone comprehensive safety evaluations, including for bias and toxicity. Google has conducted research into potential risk areas like cyber-offense, persuasion, and autonomy, applying best-in-class adversarial testing techniques to identify critical safety issues prior to Gemini's deployment. External experts have also been involved in stress-testing the models across a range of issues. But no specific benchmark information is available.
- **Architecture:** The Gemini models are developed upon Transformer decoder architectures (Vaswani et al., 2017), incorporating architectural enhancements and advancements in model optimization. These improvements facilitate stable large-scale training and refined inference capabilities on Google's Tensor Processing Units.

- **Cost:** As a Generative AI on Vertex AI. Please check our Price Table or their website.

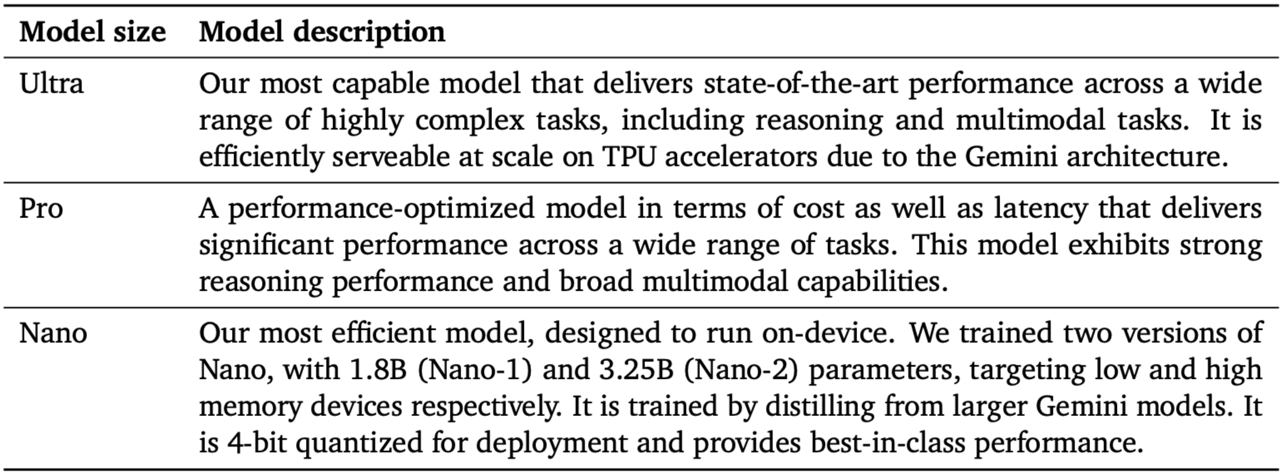
- **Efficiency:** Prioritizes efficient processing for scalable applications.

  Gemini 1.0, trained on Google's custom TPUs v4 and v5e, is their most efficient, scalable, and reliable model. These AI accelerators power Google's suite of services and enable cost-effective AI training for businesses globally. They're launching Cloud TPU v5p - the top-tier TPU system for training advanced AI models. This will accelerate Gemini's development, hasten AI training for clients, and bring new products to market faster.

  They also pushed the boundaries of efficiency with the introduction of Gemini Nano, a collection of compact models designed for deployment on devices. These smaller models are adept at tasks like summarization, reading comprehension, and text completion when used on-device. They display remarkable skills in reasoning, STEM, coding, as well as multimodal and multilingual tasks, especially considering their smaller scale.

The initial release, Gemini 1.0, includes three primary variants designed to cater to a broad spectrum of applications, as outlined in table 5.

Table 5: Summary of the Gemini 1.0 series of models(Gemini Team Google, 2023)



Model size	Model description
Ultra	Our most capable model that delivers state-of-the-art performance across a wide range of highly complex tasks, including reasoning and multimodal tasks. It is efficiently serveable at scale on TPU accelerators due to the Gemini architecture.
Pro	A performance-optimized model in terms of cost as well as latency that delivers significant performance across a wide range of tasks. This model exhibits strong reasoning performance and broad multimodal capabilities.
Nano	Our most efficient model, designed to run on-device. We trained two versions of Nano, with 1.8B (Nano-1) and 3.25B (Nano-2) parameters, targeting low and high memory devices respectively. It is trained by distilling from larger Gemini models. It is 4-bit quantized for deployment and provides best-in-class performance.

ChatGPT vs. open-source LLMs.

In 2023, a study titled "ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?" was published by Chen and his team (Chen et al., 2023). Accordingly, the subsequent comparisons we present in this section are primarily guided by their research.

Thoroughly evaluating the abilities of LLMs continues to be a vibrant area of research due to the wide-ranging and diverse evaluations that need to be conducted. Datasets for question-answering tasks (Joshi et al., 2017; Kwiatkowski et al., 2019; Lin et al., 2022) have become renowned benchmarks for evaluation, but recently, new benchmarks specifically designed for LLM assessments have also been introduced (Dubois et al., 2023; Beeching et al., 2023; Zheng et al., 2023). In the subsequent sections, we delve into the various abilities of LLMs across six primary dimensions: general capabilities, agent capabilities, logical reasoning (including maths and coding capacities), long-context modeling, specific applications such as QA or summarization, and trustworthiness.

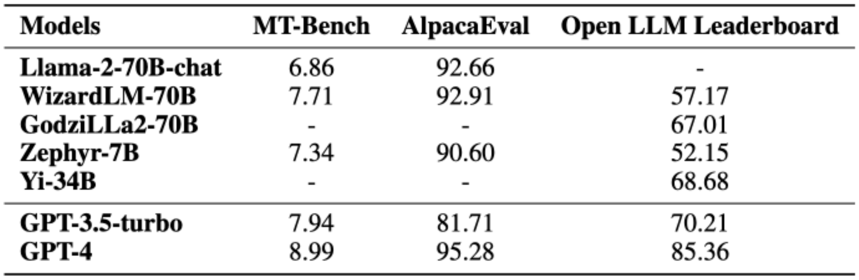
General Capabilities

Benchmarks

With a surge of new Language Learning Models (LLMs) being introduced every week, each purporting to excel in specific tasks, it's becoming more difficult to distinguish genuine progress and identify the top models. As a result, it's essential to thoroughly evaluate these models across a wide range of tasks to grasp their overall abilities. This part discusses benchmarks using evaluations based on LLMs, such as GPT-4, and conventional evaluation metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002).

- MT-Bench (Zheng et al., 2023) is designed to evaluate the abilities of multi-turn conversation and instruction following. The assessment is conducted from eight distinct aspects, including writing, roleplay, information extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science). For the evaluation of these benchmarks, more powerful Language Learning Models (LLMs) such as GPT-4 are employed.
- AlpacaEval (Li et al., 2023d) is an LLM-based automatic evaluator. It's rooted in the AlpacaFarm evaluation set (Dubois et al., 2023) and aims to test the proficiency of models in executing general user instructions. In this benchmarking system, candidate models are compared against responses from Davinci-003 with the use of more potent LLMs like GPT-4 and Claude. The performance is measured by the win rate of the candidate model.
- Open LLM Leaderboard (Beeching et al., 2023) is a performance evaluation platform for LLMs. It uses the Language Model Evaluation Harness (designed by Gao et al., 2021) to assess LLMs based on seven crucial benchmarks. These include the 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 measures the LLMs' capabilities in reasoning and general knowledge across a broad range of fields, both in zero-shot and few-shot settings

Table 6: Model performance on general benchmarks (Chen et al., 2023)



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

Models and Comparisons

- MT-Bench
  - WizardLM-70B: Developed by Xu et al. in 2023a, this model has been fine-tuned with instructions using a large dataset with varying complexity levels. It distinguishes itself as the highest scoring open-source LLM on MT-Bench with a score of 7.71, albeit slightly less than the scores of GPT-3.5-turbo (7.94) and GPT-4 (8.99).
  - Zephyr-7B: A smaller model developed by Tunstall et al. in 2023, it implements distilled direct preference optimization (Rafailov et al., 2023a) and surpasses Llama-2-chat-70B on MT-Bench, scoring 7.34 against 6.86.
- AlpacaEval
  - Llama-2-chat-70B: This model achieves a win rate of 92.66% in AlpacaEval, exceeding the performance of GPT-3.5-turbo by 10.95% margin.
  - GPT-4: This model maintains its position as the top performer among all LLMs with a win rate of 95.28%.
  - Zephyr-7B: Another smaller model developed by Tunstall et al. in 2023, it implements distilled direct preference optimization (Rafailov et al., 2023a) and manages to achieve results that are comparable to 70B LLMs on AlpacaEval with a win rate of 90.6%.
- Open LLM Leaderboard
  - GodziLLa2-70B: An experimental model from Philippines in 2023, it integrates various proprietary LoRAs from Maya Philippines 6 and the Guanaco Llama 2 1K dataset (mlabonne, 2023) with Llama-2-70B. It attains a competitive score of 67.01% on the Open LLM Leaderboard.
  - Yi-34B: Pre-trained from scratch by developers at 01.AI 7, it stands out among all open-source LLMs with an impressive score of 68.68%. This performance is on par with that of GPT-3.5-turbo, which scores 70.21%.
  - GPT-4: This model leads the pack with an exceptional score of 85.36%, significantly outperforming the other models.

Agent Capabilities

The fine-tuned Llama or fine-tuned Llama-2 shows the ability to match or even outperform GPT-3.5-turbo or GPT-4 in some benchmarks of Agent Capabilities.

Benchmarks

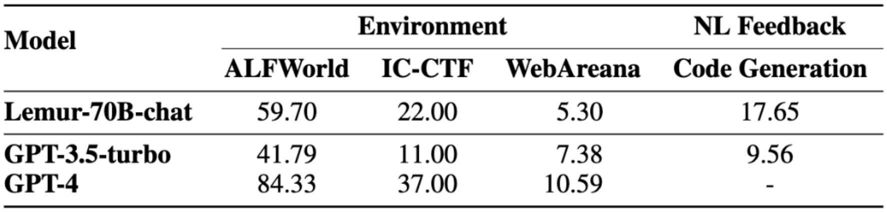
The progress in expanding the model size has led to an increased interest in LLM-based agents, also known as language agents, within the Natural Language Processing (NLP) community. Considering this, we explore the competencies of open-source LLMs across various performance measurements. These measurements, based on the skillsets they assess, can be primarily categorized into four groups.

- Using Tools: Several benchmarks have been introduced to assess the tool application abilities of LLMs. API-Bank (Li et al., 2023c) is a benchmark specifically built for tool-enhanced LLMs. ToolBench (Xu et al., 2023c) serves as a tool handling benchmark, encompassing numerous software tools meant for real-world assignments. APIBench (Patil et al., 2023) incorporates APIs from platforms like HuggingFace, TorchHub, and TensorHub. Through a multi-agent simulation environment, ToolAlpaca (Tang et al., 2023a) has created a broad and inclusive dataset for tool use. Interestingly, another dataset for tool use, built using ChatGPT, also goes by the name ToolBench (Qin et al., 2023b). Furthermore, MINT (Wang et al., 2023e) is capable of appraising the skill level of LLMs in utilizing tools to accomplish tasks requiring multi-turn interactions.
- Self-Debugging: A variety of datasets exist for evaluating the self-debugging capabilities of LLMs. These include InterCode-Bash and InterCode-SQL (Yang et al., 2023c), MINT-MBPP and MINT-HumanEval (Wang et al., 2023e), along with RoboCodeGen (Liang et al., 2023).
- Exploring Environment: ALFWorld (Shridhar et al., 2020), InterCode-CTF (Yang et al., 2023c), and WebArena (Zhou et al., 2023a) have been developed to assess the capability of LLMs-based agents in terms of extracting information from their surroundings and making informed choices.

Models and Comparisons

- Lemur-70B-chat (Xu et al., 2023d) outshines GPT-3.5-turbo in terms of performance when it comes to environmental exploration and coding tasks guided by natural language feedback.,
- AgentLlama-70B, a result of instruction tuning conducted by AgentTuning (Zeng et al., 2023) using Llama-2 on a combined dataset of constructed AgentInstruct and general domain instructions, **exhibits performance equivalent to GPT-3.5-turbo on unseen agent tasks**. This is a notable achievement of AgentLlama .
- ToolLLaMA (Qin et al. in 2023b), shows performance on par with GPT-3.5-turbo in tool usage evaluations, achieved by fine-tuning Llama-2-7B on ToolBench.
- Fine-tuning Llama-2-13B with FireAct, as introduced by Chen et al. (2023a), **can lead to superior performance over prompting GPT-3.5-turbo on HotpotQA** (Yang et al., 2018).
- Gorilla (Patil et al., 2023), which was fine-tuned from Llama-7B, **demonstrates superior performance compared to GPT-4 when it comes to writing API calls**.
- Some comparisons are shown in Table 2(in Chen et al., 2023).

Table 7: Model performance on several agent benchmarks (Chen et al., 2023)



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	-

Logical Reasoning Capabilities

Specialized benchmarks assess models on tasks such as coding, mathematics, and domain-specific knowledge. For instance, benchmarks like HumanEval and MBPP test the coding abilities of LLMs, while GSM8K and MATH evaluate mathematical reasoning skills. GPT-4 and its predecessors generally excel in these areas, suggesting strong application-specific capabilities.

Benchmarks

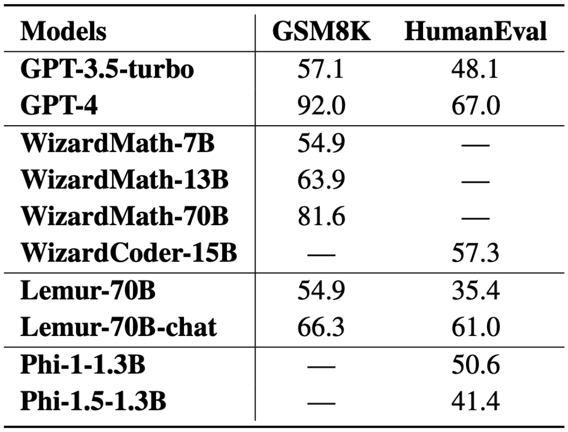
The fundamental capability of high-level ability and skill, such as programming, theorem proving, and arithmetic reasoning, is logical reasoning. In this section, we will discuss the following benchmarks:

- GSM8K (Cobbe et al., 2021): This benchmark comprises 8.5K high-quality grade school math problems created by human problem writers. The problems require between 2 and 8 steps to solve, primarily involving a series of basic arithmetic operations to reach the final answer.
- MATH (Hendrycks et al., 2021c): This dataset contains 12,500 challenging competition mathematics problems, each with a full step-by-step solution that can be used to guide models in generating answer derivations and explanations.
- TheoremQA (Wenhu et al., 2023): This benchmark is a theorem-driven question answering dataset designed to assess AI models' abilities to apply theorems in solving complex science problems. The TheoremQA benchmark includes 800 high-quality questions covering 350 theorems from Math, Physics, EE&CS, and Finance, curated by domain experts.
- HumanEval (Chen et al., 2021): This set consists of 164 handwritten programming problems. Each problem includes a function signature, docstring, body, and multiple unit tests, averaging about 7.7 tests per problem.
- MBPP (Austin et al., 2021): Also known as The Mostly Basic Programming Problems dataset, it 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 equipped with a self-contained Python function that solves the specified problem and three test cases that check for the function's semantic correctness.
- APPs (Hendrycks et al., 2021a): This benchmark measures the ability of models to convert an arbitrary natural language specification into satisfactory Python code. The benchmark includes 10,000 problems, which vary from simple one-liners to substantial algorithmic challenges.

Models and Comparisons

Both Enhanced Instruction Tuning and Pre-training on High-Quality Data can help open-source models to surpass the performance of GPT-3.5-turbo on Logical Reasoning benchmarks, the comparison is shown in table 8.

Table 8: GSM8K(math) and HumanEval(coding) comparison (Chen et al., 2023)



Models	GSM8K	HumanEval
GPT-3.5-turbo	57.1	48.1
GPT-4	92.0	67.0
WizardMath-7B	54.9	—
WizardMath-13B	63.9	—
WizardMath-70B	81.6	—
WizardCoder-15B	—	57.3
Lemur-70B	54.9	35.4
Lemur-70B-chat	66.3	61.0
Phi-1-1.3B	—	50.6
Phi-1.5-1.3B	—	41.4

- Enhanced Instruction Tuning:
  - The fine-tuned model surpasses the performance of GPT-3.5-turbo, e.g.,
  - WizardCoder (Luo et al., 2023c) shows an absolute improvement of 19.1% over GPT-3.5-turbo on HumanEval.
  - WizardMath (Luo et al., 2023a) achieves a 42.9% absolute improvement on GSM8K compared to GPT-3.5-turbo.
- Pre-training on High-Quality Data:
  - Lemur (Xu et al., 2023d): This model has verified an improved mixture of natural language data and code, enhancing the LLMs' capabilities in function calling, automatic programming, and agent capabilities. Specifically, Lemur-70B-chat showed significant improvements over GPT-3.5-turbo on both HumanEval and GSM8K, even without task-specific fine-tuning.
  - Phi-1 and Phi-1.5 (Gunasekar et al., 2023; Li et al., 2023e): These models took a different approach by using textbooks as the main corpus for pre-training. This strategy made the strong abilities observable even on much smaller language models.

Modeling Long-context Capabilities

Models like GPT-3 and successors have demonstrated proficiency in handling long-context information. However, it's not clear from the text how open-source LLMs compare to ChatGPT in this regard.

Benchmarks for Long-Context Evaluation of LLMs

- SCROLLS (Shaham et al., 2022): It is a well - known assessment benchmark comprising seven datasets featuring naturally extended input . The tasks encompass the following : summarization

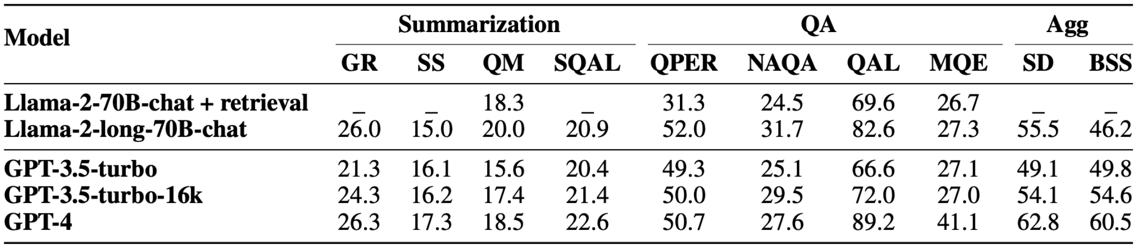
(GovReport (Huang et al., 2021), SummScreen(Chen et al., 2021b), QMSum (Zhong et al., 2021)), question-answering (Qasper (Dasigi et al.,2021), NarrativeQA (Kociský et al., 2018), QuALITY (Pang et al., 2021)) and natural language inference (ContractNLI (Koreeda & Manning, 2021)).

- ZeroSCROLLS (Shaham et al., 2023): Building on SCROLLS, this benchmark only considers the zero-shot setting and evaluates out-of-the-shelf LLMs. It discards ContractNLI from SCROLLS, reuses the other 6 datasets, and adds 4 new ones.
- LongBench (Bai et al., 2023): This bilingual English/Chinese benchmark includes 21 datasets across 6 tasks, focusing on long-context evaluation.
- L-Eval (An et al., 2023): Comprising 16 existing datasets and 4 new ones, this diverse benchmark has an average task length of over 4k tokens. The authors prefer LLM judges evaluation (especially for GPT-4) over n-gram for long-context evaluation.
- BAMBOO (Dong et al., 2023): This benchmark focuses on eliminating pre-training data contamination by using only recent data in the evaluation datasets. It's designed for long-context LLM evaluation.
- M4LE (Kwan et al., 2023): A broad-scope benchmark that splits 36 datasets into 5 understanding abilities: explicit single-span, semantic single-span, explicit multiple-span, semantic multiple-span, and global understanding.

Models and Comparisons

- GPT-3.5-turbo and its 16k version: On the LongBench, L-Eval, BAMBOO, and M4LE benchmarks, these models significantly outperform all open-source LLMs like Llama-2, LongChat, or Vicuna. This indicates the challenge of improving open-source LLM performance on long-input tasks.
- Llama-2-long (Xiong et al., 2023): This model continues the pre-training of Llama-2 with a 16k context window (up from 4k in Llama-2) using 400B tokens. The resulting Llama-2-long-chat-70B outperforms GPT-3.5-turbo-16k by 37.7 to 36.7 on ZeroSCROLLS.
- Combination Approach (Xu et al., 2023b): This approach combines both positional interpolation (Chen et al., 2023e) and retrieval augmentation (Lewis et al., 2020), which pushes a Llama-2-70B above GPT-3.5-turbo-16k on average over 7 long-context tasks, including 4 datasets from ZeroSCROLLS.
- The performance of the model on ZeroSCROLLS is presented in table 9. Here, GR stands for GovReport, SS denotes SummScreen, QM signifies QMSum, SQAL is an abbreviation for SQuALITY, QPER represents Qasper, NAQA is referred to NarrativeQA, QAL is an acronym for QuALITY, MQE is short for MuSiQue, SD corresponds to SpaceDigest, and BSS is known as BookSumSort.

Table 9: Performance of the model on ZeroSCROLLS (Chen et al., 2023)



Model	Summarization				QA				Agg	
	GR	SS	QM	SQAL	QPER	NAQA	QAL	MQE	SD	BSS
Llama-2-70B-chat + retrieval	_	_	18.3	_	31.3	24.5	69.6	26.7	_	_
Llama-2-long-70B-chat	26.0	15.0	20.0	20.9	52.0	31.7	82.6	27.3	55.5	46.2
GPT-3.5-turbo	21.3	16.1	15.6	20.4	49.3	25.1	66.6	27.1	49.1	49.8
GPT-3.5-turbo-16k	24.3	16.2	17.4	21.4	50.0	29.5	72.0	27.0	54.1	54.6
GPT-4	26.3	17.3	18.5	22.6	50.7	27.6	89.2	41.1	62.8	60.5

Application-specific Capabilities

This section is dedicated to exploring the necessary features of large language models that enable them to handle specific applications effectively.

Query-focused Summarization

Benchmarks

- Query-focused Datasets: These datasets require generating summaries in response to a detailed question. They include:
  - AQualMuse (Kulkarni et al., 2020)
  - QMSum (Zhong et al., 2021)
  - SQuALITY (Wang et al., 2022)
- Aspect-based Datasets: These datasets involve generating summaries based on a specific aspect category. They include:
  - CovidET (Zhang et al., 2023a)
  - NEWTS (Bahrainian et al., 2022)
  - WikiAsp (Hayashi et al., 2021)

Models and Comparisons

According to Yang et al. (2023d), standard fine-tuning on training data consistently outperforms ChatGPT. It shows an average improvement of 2 ROUGE-1 points over benchmarks like CovidET, NEWTS, QMSum, and SQuALITY.

Open-ended QA

Benchmarks

Open-ended QA can be categorized into two subgroups based on the answer type: short-form and long-form. For both types of datasets, the evaluation metrics are exact match (EM) and F1 over words in the answers. Answering Open-ended QA requires the model to understand the provided context, or retrieve related knowledge if no context is provided.

- Short-Form Datasets: These 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: These include ELI5 (Fan et al., 2019) and doc2dial (Feng et al., 2020).

Models and Comparisons

InstructRetro (Wang et al., 2023a): This model demonstrates substantial improvement over GPT-3 on NQ, TriviaQA, SQuAD 2.0, and DROP. It also shows a 7-10 percent improvement compared to a similar-sized proprietary GPT-instruct model across a range of short-form and long-form open-ended QA datasets. InstructRetro is not yet open-sourced.

Medical

Benchmarks

A valuable feature of Large Language Models (LLMs) is their potential to assist in medical tasks, enabling affordable and high-quality healthcare services to reach a wider audience.

- Mental Health: The IMHI benchmark (Yang et al., 2023c) includes 10 existing mental health analysis datasets like DR (Pirina & Çöltekin, 2018), CLP (Coppersmith et al., 2015), Dreaddit (Turcan & McKeown, 2019), Loneliness, SWMH, T-SID (Ji et al., 2022), SAD (Mauriello et al., 2021), CAMS (Garg et al., 2022), MultiWD (SATHVIK & Garg, 2023), and IRF (Garg et al., 2023).
- Radiology: The datasets OpenI (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) include radiology reports that feature findings and impressions text.

Models and Comparisons

- MentalLlama-chat-13B (Yang et al., 2023c) outperforms ChatGPT with few-shot or zero-shot prompting in 9 out of 10 tasks in IMHI.
- Radiology-Llama-2 model (Liu et al., 2023) significantly outperforms ChatGPT and GPT-4 on both the MIMIC-CXR and OpenI datasets.

Generating Structured Responses

This process entails generating structured responses based on guidelines, a critical skill that enhances autonomous abilities and minimizes manual work in interpreting or converting model outputs.

Benchmarks

- Rotowire (Wiseman et al., 2017): This benchmark includes NBA game summaries with corresponding score tables.
- Struc-Bench (Tang et al., 2023b): This introduces two datasets. Struc-Bench-Latex outputs tables in Latex format, and Struc-Bench-HTML outputs tables in HTML format.

Models and Comparisons

- The Struc-Bench (Tang et al., 2023b) fine-tuned a Llama-7B model using structured generation data. The enhanced 7B model demonstrated superior performance to ChatGPT across all benchmarks previously discussed.

Generating Critiques

Benchmarks

An intriguing capability of LLMs is to offer feedback or evaluations in response to a query. To evaluate this feature, human annotators or GPT-4 could be employed as evaluators to directly assess the responses. The initial queries can be derived from any dataset related to the aforementioned capabilities.

Models and Comparisons

- Shepherd (Wang et al., 2023c): This is a 7B architecture, initialized using the Llama-7B model and further refined with data collected from community critiques and 1,317 instances of top-tier human-annotated data. This model is capable of producing evaluations for a variety of NLP datasets, including AlpacaFarm, FairEval, CosmosQA (Huang et al., 2019), OBQA (Mihaylov et al., 2018a), PIQA (Work of Bisk et al., 2020), TruthfulQA, and CritiqueEval. When assessed by GPT-4, Shepherd either matches or surpasses ChatGPT over 60% of the time. Moreover, when reviewed by human evaluators, Shepherd's performance closely mirrors that of ChatGPT.

Towards Trust-worthy AI

It's essential to establish the dependability of LLMs for their practical use by humans. Aspects like hallucination (Ye & Durrett, 2022; Zhao et al., 2023a) and safety (Zhiheng et al., 2023b) are key factors. These concerns, if not addressed, could compromise user confidence in LLMs and pose potential dangers in applications with significant consequences.

Hallucination

Benchmarks

Several benchmarks have been proposed for evaluating hallucinations in LLMs, including large-scale datasets, automated metrics, and evaluation models.

- TruthfulQA (Lin et al., 2022): A question-answering (QA) dataset with questions across 38 categories. Some questions are crafted to test human misconceptions.
- FactualityPrompts (Lee et al., 2022): This is a dataset that quantifies hallucinations in open-ended generation tasks. The dataset incorporates both factual and non-factual prompts to examine their effect on the continuations produced by LLMs.
- HaluEval (Li et al., 2023a): A large dataset of generated and human-annotated hallucinated samples, covering QA, knowledge-grounded dialogue, and text summarization tasks.

- FACTOR (Muhlgay et al., 2023): It suggests a method for assessing the factuality of LMs by converting a factual database into a benchmark for faithfulness evaluation. This method is utilized to establish two standards: Wiki-FACTOR and News-FACTOR.
- KoLA (Yu et al., 2023a): Constructs a Knowledge-oriented LLM Assessment benchmark (KoLA) which focuses on mimicking human cognition for ability modeling, using Wikipedia for data.

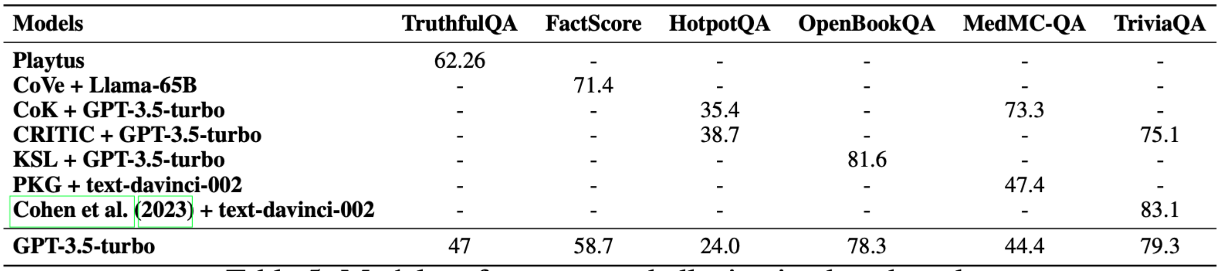
- FActScore (Min et al., 2023): The proposition introduces a novel assessment method. This method deconstructs the generation of a LLM into a sequence of fundamental facts. Afterward, it calculates the proportion of these facts that are backed by a dependable knowledge source.

- In addition to the recently developed benchmarks for hallucination, existing QA datasets that leverage real-world knowledge are commonly utilized to evaluate faithfulness. These include HotpotQA (Yang et al., 2018), OpenBookQA (Mihaylov et al., 2018b), MedMC-QA (Pal et al., 2022), and TriviaQA (Joshi et al., 2017). Furthermore, human assessment is frequently employed as a trustworthy gauge of faithfulness, alongside datasets and automated metrics.

Models and Comparisons

- Several research studies have been conducted on the phenomenon of hallucination (Zhang et al., 2023b; Rawte et al., 2023), which deeply explore potential methodologies. These methodologies, which can improve upon the current performance of GPT-3.5-turbo, can be applied during the fine-tuning stage or at the time of inference. Select model performances are shown in Table 10 (from Chen et al., 2023).

Table 10: Model performance on hallucination benchmarks (Chen et al., 2023)



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

- Dhuliawala et al. (2023) introduced Chain-of-Verification (CoVe) for decoding, which led to a substantial improvement on FactScore over GPT-3.5-turbo.
- For external knowledge augmentation, frameworks use different searching and prompting techniques. Notable methodologies include the Chain-of-Knowledge (CoK) (Li et al. 2023d), LLM-AUGMENTER (Peng et al. 2023), Knowledge Solver (KSL) (Feng et al. 2023), CRITIC (Gou et al. 2023), and Parametric Knowledge Guiding (PKG) (Luo et al. 2023b).
- For multi-agent dialogue, Cohen et al. (2023) proposed a multi-turn interaction between two LLMs to discover inconsistencies. Du et al. (2023) proposed a method where multiple language model instances propose and debate their individual responses to arrive at a final answer.

- GPT-3.5-turbo has also incorporated a retrieval plugin to access external knowledge and reduce hallucinations (OpenAI, 2023a).
- These methods have all shown improvement on various QA tasks and benchmarks.

Safety

Benchmarks

The safety issues associated with LLMs are primarily segmented into three key areas: social bias, the robustness of the model, and issues related to poisoning (Zhiheng et al., 2023a). For a more comprehensive evaluation of these sectors, a variety of benchmarks have been suggested:

- SafetyBench(Zhang et al., 2023c): this dataset comprises 11,435 diverse multiple-choice questions, covering seven unique safety concern categories.
- Latent Jailbreak(Qiu et al., 2023): this benchmark assesses both the safety and robustness of LLMs, underlining the importance of a well-rounded approach.
- XSTEST(Röttger et al., 2023): this tests suite systematically detects instances of overly cautious safety behaviors, such as the unwarranted rejection of safe prompts.
- RED-EVAL(Bhardwaj & Poria, 2023): It uses a Chain of Utterances (CoU)-based prompt to carry out red-teaming (Ganguli et al., 2022) for safety evaluations of LLMs.

In addition to automatic benchmarks, human assessment also plays a critical role in evaluating safety, where crowdworkers classify responses as safe or dangerous (Dai et al., 2023). Certain studies, including those that demonstrate GPT-4's ability to supplant human evaluators in assessing alignment capacities (Chiang & Lee, 2023), strive to gather such labels.

Models and Comparisons

Current evaluations by Zhang et al. (2023) and Röttger et al. (2023) indicate that **GPT-3.5-turbo and GPT-4 models are leading in safety evaluations**, primarily due to the application of Reinforcement Learning with Human Feedback (RLHF) as articulated by Bai et al. (2022a). RLHF commences by gathering a dataset of human preferences about responses, then develops a reward model that replicates these preferences, and finally utilizes reinforcement learning to train the LLMs to align with human preferences.

However, the RLHF process necessitates the collection of a substantial amount of costly human annotations, which limits its application for open-source LLMs. To democratize safety enhancements, Ji et al. (2023) created a dataset separating harmlessness and helpfulness from overall human preferences, improving safety alignment as shown in experiments. Alternatives to reduce RLHF costs include Bai et al.'s (2022b) Reinforcement Learning from AI Feedback (RLAIF) and Rafailov et al.'s (2023a) Direct Preference Optimization (DPO), both offering potential for future safety improvements in open-source LLMs.

Discussion

Choosing between open-source and closed-source Large Language Models (LLMs) depends on several factors. However, when evaluating long-term development prospects of LLM environments, setting aside privacy and cost concerns, a preference emerges for closed-source LLMs. The primary reasons for this preference include:

Closed-source

Benefits

Closed-source Large Language Models (LLMs) are distinguished by their **superior performance and commitment to continuous improvement**. In the coming years, it's expected that models such as GPT-4, Gemini, and Claude 2 will maintain a performance edge over open-source alternatives in a variety of capability assessments. While performance may vary, the trend for closed-source LLMs is a consistent enhancement over time.

One of the significant advantages of closed-source LLMs is the **dedicated support and long-term maintenance** provided by the originating companies. This ensures that any issues encountered can be swiftly and effectively addressed, minimizing downtime and maintaining system integrity. The process of updating these models is streamlined and efficient, thanks to the focused efforts of the developers in keeping up with the rapid advancements in LLM technology. This means less time worrying about the technicalities of updates and more time optimizing prompt engineering.

In terms of stability and reliability, closed-source LLMs generally outperform their open-source counterparts. The professional development and maintenance teams behind these models ensure a **robust and dependable performance**, essential for critical or commercial applications. Furthermore, these models often come **equipped with a variety of plugins**, enhancing their versatility and functionality. This array of plugins allows users to tailor the LLMs to a wide range of tasks, making them a flexible tool for numerous applications. In summary, closed-source LLMs provide a compelling package of performance, support, and versatility, making them a preferred choice for many users.

Challenges

While closed-source Large Language Models (LLMs) offer numerous benefits, they also present certain challenges that require careful consideration. One of the primary drawbacks is the **cost; these models often entail significant licensing fees**, along with additional charges for updates or extra features. This financial aspect can be a considerable burden, especially for smaller organizations or individual users.

Another issue is the **lack of customization options**. Without access to the source code, users are limited in their ability to modify or tailor the model to specific needs. This can be a significant drawback for projects requiring a high degree of customization or for those who wish to understand and possibly adjust the inner workings of their LLM.

Transparency, or the lack thereof, is another challenge associated with closed-source LLMs. The **internal workings of these models are not always clear**, and understanding how the model makes predictions or processes data can be difficult. This opacity can be problematic for users who require a clear understanding of the model's processes for trust, compliance, or ethical reasons.

Given these challenges, it's crucial for individuals and organizations to thoroughly assess these aspects against their specific needs and constraints. While closed-source LLMs may offer superior performance and support, the associated costs, limited customization, and lack of transparency may not align with every user's requirements or values. Therefore, a careful evaluation of both the benefits and drawbacks is essential in making an informed decision regarding the adoption of closed-source LLMs.

Open-source

Benefits

Open-source Large Language Models (LLMs) offer significant advantages, particularly in terms of **cost-effectiveness and adaptability**. Being freely available, these models provide an economical solution without sacrificing quality. The ability to customize and alter the code allows users to tailor the LLM to their project's unique requirements, offering a level of flexibility that is particularly valuable in innovative or niche applications.

The community aspect of open-source LLMs is another notable benefit. These models are **supported by robust, active communities that contribute to a wealth of documentation and user support**. This community-driven approach not only fosters continuous improvement and innovation but also ensures that users have access to help and guidance. Additionally, the open nature of these models means that the source code is available for review, offering complete transparency into how the LLM functions. This transparency is essential for users who need to understand or validate the inner workings of the model, particularly in applications where trust and accuracy are paramount.

Privacy is another area where open-source LLMs excel. **Users have full control over their data and how the model is implemented**, which is especially important when dealing with sensitive information. This control allows for a more secure and tailored approach to data handling and model usage, aligning with strict privacy requirements or personal preferences. In summary, open-source LLMs provide an attractive combination of cost savings, flexibility, community support, transparency, and privacy, making them a compelling option for a wide range of users and applications.

Challenges

These points highlight some of the key challenges one might face when considering the use of open-source LLMs for a project. As always, careful consideration and testing are crucial when deciding whether to use an open-source LLM for your project.

While open-source Large Language Models (LLMs) offer many benefits, they also come with their own set of challenges that need careful consideration. Firstly, **while the software itself is free, the infrastructure and resources required to run these models can be quite costly**. Setting up and maintaining servers, storage, and high-performance computing resources can amount to significant expenses. Additionally, the reliance on community support rather than official, dedicated support may increase the time and effort needed for troubleshooting and problem-solving.

The stability, security, and compatibility of open-source LLMs can also be of concern. **These models may not be as stable or secure as their closed-source counterparts, potentially introducing vulnerabilities or inconsistencies**. Compatibility issues with other software or systems might arise, requiring additional adjustments or workarounds.

Furthermore, the rapid development cycle typical of open-source projects means that these models are frequently updated. **While this keeps the models current, it can also necessitate regular code modifications, leading to increased maintenance effort**. The learning curve can be steep, especially for teams not already familiar with the specific LLM or open-source practices in general, possibly slowing down initial progress.

These challenges are important to consider when evaluating the suitability of open-source LLMs for your project. They highlight the need for careful planning, budgeting, and skill assessment to ensure that the advantages of open-source LLMs can be fully realized while mitigating the potential downsides. As with any technological choice, a thoughtful approach and thorough testing are key to making the best decision for your specific needs and context.

Project Catalyst

What is Project Catalyst?

Project Catalyst is an innovative community governance initiative aimed at developing the Cardano blockchain ecosystem through funding research, development, adoption, and maintenance projects. Funded by the Cardano Treasury, it facilitates a "one-coin one-vote" approach for project approval and funding, with community members actively participating in administrative functions and voting (Nelson et al, 2022). The initiative faces open problems characteristic of Decentralized Autonomous Organizations (DAOs) and blockchain governance, including questions about managing resources, legal and social contracts, and the overall evolution of organizational structures online. While Catalyst is often referred to as a DAO, a more accurate description may be a Decentralized Innovation

Fund. As of December 2023, Catalyst has funded ten rounds of projects totaling $66 million in awards (Project Catalyst). Additionally, community members receive rewards for administrative functions such as overseeing a challenge topic and reviewing proposals. The fund is currently administered by Input Output Global (IOG) and is currently in the process of decentralizing control through various improvement proposals at both the protocol level (CIP-1694) and Community scale (Fund 10 and beyond) to develop the communities ability to self-administer the fund.

What are the open problems?

The new growing class of organizations governed by smart contracts are typically known as Decentralized Autonomous Organizations. Contracts of all kinds; financial, legal, social, cultural, etc., now have a basis for execution in cyberspace in a way they didn't previously. DAOs and the underlying smart contracts signify a new, decentralized approach to governing actions, contrasting with traditional regulatory mechanisms. They raise questions about how to govern resources of all types and it is further elaborated in the paper titled, "Open Problems in DAOs". This means that DAOs can express the full range of existing organizational logics, so the choice is not between a corporation and a DAO per se but between a traditional, legally-constituted corporation and a corporation that is digitally-constituted through a smart contract. "DAOs are not just curious instances of a certain kind of online community; they have the expressive potential to transport a tremendous amount of institutional infrastructure from the stonebound halls of power onto the open internet; from law and economics into computer science. DAOs in their current form may or may not become the future of organizations (Tan et al, 2023)." However, it is already clear that online forms of organization are becoming more and more important in the politics and economies of the world. DAO science is one of the most promising paths forward for tackling and making progress on hard questions of organization, coordination, and governance.

In a community directed research paper called "Democratic Pluralism: A White Paper on Cardano Governance," it was proposed that the governance of cryptocurrency platforms, such as Cardano, need to aim towards creating trustworthy mechanisms focused on defining digital citizenship so that the principles of democracy can be fulfilled while participating in fair market rules. This paper defines and contrasts an open corporation from a digital nation and describes the security requirements for citizen models of cryptocurrency governance. Finally, it outlines a roadmap for Democratic Pluralism, based on incremental implementations of secure quadratic voting and funding systems. (Nelson et al, 2023).

What can LLMs assist with as it relates to Project Catalyst?

Large Language Models can significantly enhance digital communities, such as Project Catalyst, in their various governance ways. 1) **Semantic Analysis of Proposals**: LLMs can analyze the semantic space of all historical proposals, identifying trends, gaps, and underrepresented areas. This analysis helps in predicting relevant topics, encouraging diverse project submissions, and aligning with strategic goals. 2) **Structuring Proposal Templates**: LLMs can assist in structuring and refining the prompting templates in Ideascale. By improving how questions are asked, they can elicit more informative and useful responses, enhancing the quality of proposals and discussions. 3) **Democratizing Innovation**: Open sourcing the process with the help of LLMs can democratize development and innovation. By providing a platform that anyone can contribute to, it fosters a transparent culture, encourages a diverse range of ideas, and accelerates improvements through community feedback. 4) **Overcoming Language Barriers**: LLMs can provide translation and multilingual support, making the project more accessible globally. This helps in leveling the playing field by offering non-English speaking services and assistance, making the governance and proposal process more inclusive. 5) **Enhancing Review and Voting Processes**: LLMs can help streamline the review and voting process by providing key metrics and insights on proposals. They can assist in evaluating feasibility, impact, clarity, team qualifications, and cost-effectiveness, making the decision-making process more efficient and informed. 6) **Fostering Community Engagement**: LLMs can create spaces for community interaction, collaboration, and feedback. By fostering a sense of community, Project Catalyst can become not just a funding mechanism but a collaborative ecosystem of innovators. 7) **Optimizing Feedback and Iteration**: LLMs can facilitate a robust feedback mechanism, enabling continuous updates and refinements based on real-world experiences and changing needs. They can adapt and evolve with advancements in technology and community feedback. 8) **Exploring Creative Solutions**: LLMs can be tuned to generate creative and innovative responses, aiding reviewers and strategists in understanding the transformative potential of proposals. This can help in visualizing the future impact and aligning it with the value expectations. Below are summarized feedback given by various groups around the Cardano and larger blockchain communities.

Survey from Catalyst Improvement Team

The improvement team of Project Catalyst is deeply invested in creating a fair, efficient, and practical process for the evolution of Project Catalyst. They understand that the foundation of a successful project lies in its ability to adapt and respond to feedback while maintaining a clear and equitable framework. The feedback received is a critical component of this ongoing process, and the team is committed to integrating it thoughtfully and strategically.

What is the semantic space of all proposals historically? Can this help us identify what attention is needed in the existing funding round?

The semantic space of all proposals historically refers to the range and nature of topics, ideas, and themes that have been covered in proposals over time. It encompasses the vocabulary, concepts, and relationships between ideas that have been used in the context of these proposals. Analyzing the semantic space can provide a comprehensive overview of the trends, patterns, and gaps in the topics that have been proposed and funded.

Understanding the semantic space of historical proposals can indeed be instrumental in identifying where attention is needed in the existing funding round. Here's how:

- Trend analysis: By examining the semantic space, the team can identify what topics or themes have been consistently popular or have emerged as trends over time. This can help predict what areas might continue to be of interest or are becoming increasingly relevant.
- Gap identification: Analyzing the semantic space can highlight underrepresented areas or innovative topics that haven't been explored thoroughly. This can guide the team to encourage

proposals in these less saturated but potentially valuable areas, ensuring a diverse and comprehensive range of projects.

- Goal alignment: Understanding the historical semantic space allows the team to align the current funding round with the overarching goals or mission. If certain key areas or strategic priorities have not been adequately addressed in past proposals, the team can focus on soliciting and supporting projects in these domains.
- Benchmarking standards: Knowing the historical semantic space helps in setting benchmarks for quality, innovation, and relevance. It can inform the current round's evaluators what level of novelty, feasibility, and impact to expect and encourage in new proposals.
- Predictive insights: Advanced analytics and machine learning models can be used to predict future trends based on historical semantic spaces. This predictive insight can be incredibly valuable in preparing for future rounds, understanding emerging fields, and staying ahead of the curve.

One key area of focus is the structuring of prompting templates in Ideascale. The way questions are asked significantly influences the responses received. It’s important to consider methods that structure these prompts to elicit the most informative and useful responses. This might involve more focused questions or creating a more dynamic template that adapts to the user's previous answers. The goal is to refine these prompts to ensure that the data collected is as relevant and actionable as possible.

Open sourcing the process is a strategic move that can democratize the development and innovation of Project Catalyst. Making the experimental foundations available to the entire community ensures that anyone has the opportunity to contribute to and benefit from the evolving landscape of LLMs. This approach fosters a transparent culture and also encourages a diverse range of ideas and solutions, hopefully enhancing the quality and impact of the projects. Moreover, open sourcing can lead to more rapid iterations and improvements as community feedback and contributions are integrated. Non-english speaking services/help to proposers who do not natively speak the language of the base proposal to level the playing field.

Addressing the language barrier is crucial in maintaining a fair and inclusive environment. Providing non-English speaking services or assistance to proposers who do not natively speak the language of the base proposal is essential. This could involve translation services, multilingual support in the reviewer process, or guidelines in multiple languages. This also happens to apply in the various programming languages, not just natural language. This may enable reviewers or voters to get a better understanding of a technical project because of code interpreter LLMs. By ensuring that language is not a barrier to participation, the process becomes more accessible and equitable, allowing for a richer and more diverse pool of proposals and contributors.

Project Catalyst needs to become resilient and adaptable over time. Best practices means focusing on core principles of design, development, and deployment that are technology-agnostic and can evolve with advancements in the field. It also involves creating a robust feedback and iteration mechanism so that the advice can be continuously updated and refined based on real-world experiences and changing needs. This is clearly the approach being taken, with the available test-net for Catalyst which provides a training ground for various community governance approaches.

When it comes to reviewing and voting on proposals, there are several key metrics that reviewers and voters might consider to streamline the process and ensure that they are focusing on the most critical aspects. These might include the feasibility of the proposal, the potential impact it could have, the clarity and coherence of the proposed plan, the qualifications and track record of the team behind it, and the cost-effectiveness or value for money of the project.

Finally, the Catalyst Improvement team is dedicated to fostering a sense of community around the Catalyst project. They believe that the best ideas come from collaboration and that users should feel like they are part of something bigger. They are creating spaces for users to share their experiences, offer suggestions, and collaborate with one another. The team envisions Catalyst not just as a tool but as a community of innovators, all working together to create something truly groundbreaking.

Survey from After Town Hall Breakout Room

During the focus group session in the catalyst town hall breakout room, participants engaged in a deep dive into the version 1 demo of the Catalyst LLM, designed explicitly for research and feedback. The group was made aware that the product's current state was primarily for gathering insights and improving future iterations. A significant part of the discussion revolved around the optimization of prompting templates in Ideascale, recognizing that the way prompts are structured significantly influences the quality and direction of responses. Participants were curious about the most efficient ways to structure these responses to allow the LLM to extract the most relevant signals, suggesting perhaps breaking out milestones as separate responses to enhance database querying.

Another intriguing aspect of the feedback focused on the LLM's ability to generate creative responses. Participants expressed a desire to "turn up" the creativity of the LLM's responses, allowing reviewers to grasp the profound, potentially transformative impacts of a proposal. This enhancement is seen as crucial in aiding the vision and strategy of reviewers by clearly illustrating if the proposal's possibilities are commensurate with the value for money expected.

From this feedback, it's anticipated that further questions might emerge, such as: How can the LLM be fine-tuned to balance creativity with accuracy and relevance, ensuring that the innovative ideas it proposes are both groundbreaking and applicable? Additionally, participants might be curious about the integration of feedback mechanisms into the LLM. They might ask how the system can incorporate iterative feedback from users to continuously refine and target its responses, making it a more effective tool for strategy and vision alignment in proposals. These questions highlight a keen interest in not just the functional capabilities of the LLM but also its potential to drive innovation and strategic decision-making in a cost-effective and impactful manner.

Survey of Wolfram LLM Experts

Professionals and experts at Wolfram have raised several important points and questions regarding the use of Large Language Models (LLMs) in their workflows. They acknowledge the effectiveness of LLMs in multi-modal paradigms, including natural language processing, SQL querying, database architecture, prompt engineering, fine-tuning, and iterative feedback loops. However, they also express concerns about the cost-effectiveness of these models. The discussion points out that while models like GPT-3.5 can deliver a certain degree of accuracy at a lower cost, there is a significant trade-off in terms of performance and cost when deciding whether to upgrade to more advanced models like GPT-4.

The feedback suggests exploring MDEL (Multi-Domain Expert Learning) and other modes of distributed computation as potentially more cost-effective alternatives or complements to current LLMs. This consideration is crucial for the company as it navigates the balance between cost and performance, especially in a competitive market where efficiency and accuracy are paramount.

Moreover, the company is keen on leveraging its capability to test multiple models, indicating a proactive approach to finding the most suitable and effective solutions. This ability to experiment with various models can lead to more informed decisions about which models to deploy based on specific needs and constraints.

There is also a call for more internal feedback to fully understand and utilize the company's capabilities. Gathering a wide range of insights and experiences from different teams can provide a more comprehensive view of how LLMs and other technologies can be optimized and integrated into workflows.

Lastly, the suggestion to add a readme file to the right side of the cloud-based notebook indicates a need for better documentation and user guidance. This addition would likely help users navigate and utilize the models more effectively, ensuring that they can leverage the full range of capabilities offered by the LLMs and other tools.

In summary, while the company recognizes the potential and effectiveness of LLMs in various complex tasks, there is a clear desire to explore more cost-effective solutions, understand the trade-offs between different models, and enhance user experience through better documentation and internal knowledge sharing. These considerations will guide the company in making strategic decisions about integrating and utilizing LLMs and other computational tools in their operations.

Survey from UTXO Alliance Blockchain Builders

The UTXO Alliance is a collaborative initiative committed to advancing the UTXO model's scalability, security, and interoperability, aiming to address the rapid technological advancements within the blockchain industry. It's focused on tackling the critical issues of data transfer, processing speeds, transaction costs, and energy usage in blockchain environments, with a vision to facilitate safe and efficient transactions of digital assets across various blockchains. This in turn will promote wider adoption and interoperability of blockchain technology, improving inefficiencies in both new and legacy systems.

Recently, Wolfram Blockchain Labs reached out to the UTXO Alliance seeking insights or examples from their community about projects building governance tools using Large Language Models

(LLMs), or any CustomGPT tool for aspects like governance, DeFi, and analytics. The request, initially niche, was expanded to include any LLM or CustomGPT tool being used within their ecosystems. This outreach has led to the discovery of deepfunding.ai on SingularityNET, marking a promising development to explore the integration of AI and blockchain technology further. An update will be shared as this direction is explored further.

Survey from CatalystGPT Demo (Queries)

On December 12th, 2023, Wolfram Blockchain Labs demoed its CatalystLLM prototype as part of our overall research effort. While the focus was on accumulating feedback and queries in a survey format, we began to learn more about how one would architect a large scale system for the Catalyst community. Therefore, the following instruction set, disclaimer and application were sent to individuals who attended our focus group workshop, and added additional folks who we considered would have deeper insights into the full governance process.

Here’s the email sent to 10 individuals:

Thanks for attending the Wolfram Catalyst LLM demo on December 12th. We considered your feedback and worked hard on upgrading the tool, which is now available for you to demo! We kindly request that you read the following disclaimer and instructions before using.

Disclaimer:

- IMPORTANT: Because of the chatbook design, the Wolfram team will be able to see and review EVERY QUERY YOU MAKE! Please avoid any questions that are irrelevant to Project Catalyst.
    - We do this because it helps provide insight/feedback on tool development and future designs.
- Limitations of Database Content
    - Information included in the database is from Fund 4 through Fund 9 only.
    - Data from prior funds or subsequent funds, specifically Fund 10 and Fund 11, are not available within this database.
    - This is actively being worked on via various APIs
- Data Parameters and Types:
    - The database focuses on various Catalyst parameters, and incorporates some Cardano social media data.
    - Users should be aware of the scope and type of data available and consider how it aligns with their informational needs.
- Accuracy and Hallucinations:
    - While we strive for accuracy, users should be aware that the GPT model, like all language models, can produce "hallucinated" or incorrect information that does not accurately reflect real-world data or scenarios.
    - It is recommended to use critical thinking and cross-verification methods when interpreting the results from the ChatNotebook.

- If you encounter an error or significant hallucination, please leave the prompt in the notebook so we can verify it.
- Please leave as many of your inputs in the notebook as possible (don't delete!)

- Recommended Prompts:
    - To assist in obtaining the most accurate and relevant information, a set of recommended prompts will be provided. Users are encouraged to utilize these prompts as a guide to formulating their queries to optimize the accuracy and relevance of the information retrieved.
    - Token Context Windows:
        - There exists a token limit of 8,000 tokens. This is a variable constraint but generally amounts to around 10-20 prompts, at which point you can expect it will max out.
        - Surpassing this limit will cause an error, which is a limitation from the OpenAI model.
        - No worries, we've provided an instruction on how to create new context windows and input lines:
            - Type the tilde " ~ " to create a new context window
            - And then type the single quote " ' " to create a new prompt input cell

Instructions/Context:

1. Ask from basic questions (like list of tables in the database) and then ramp up to get more complex questions. This generally helps in "few-shot" prompting scenarios for the context to unfold.
2. Starting Your Query: Begin by typing your query in natural language form. "What is the database we're querying?". This will transform the natural language into a query and work through the database.
3. Be as specific and clear as possible to ensure the best results. The system is designed to understand and process natural language inputs.
4. Backend Prompt Engineering: Once your query is inputted, the system utilizes backend prompt engineering to translate your natural language query into an accurate SQL query. This process is designed to interpret your request and formulate it into a structured query that the database can understand and execute.
5. Exploring SQL Queries: After the system processes your input, you can explore the SQL query that was generated by clicking on the arrow dropdown menu. This feature allows users to see the exact query that is being run against the database, providing transparency and an opportunity for learning or adjustment.
6. Compute and Results: The system will then perform the necessary computations and return a result based on the SQL query. The time taken to return a result may vary depending on the complexity of the query and the system load.

7. Model Parameters: The current model in use is GPT-4 from OpenAI. This model is known for its robust performance and wide-ranging capabilities. However, users should remain aware of the limitations and characteristics of the model, especially regarding potential inaccuracies or "hallucinations."
8. Context Window Limitation: Be aware that there is a limited context window of about 8000 tokens due to the OpenAI GPT-4 limit. When you reach the end of this context window, an error message will explain that you've exceeded the limit. At this point, use the tilda " ~ " key to create a new chatblock. This action will lose the context from previous queries but will allow you to continue in a fresh new context window.
9. Creating New Input Lines: Use the single quote " ' " key to create a new input line within the chat. This can help organize your queries and responses, making the interaction clearer and more manageable.
10. Wolfram Language Input: If you have any Wolfram Language input you'd like to make, you can do so naturally in the input line. The system is designed to recognize and process Wolfram Language commands alongside natural language queries, providing a versatile and powerful tool for data analysis and exploration.
11. I've attached a pdf that includes some example prompts to play with. Feel free to use those to build context and see good examples or build your own!
12. Please do not share this with others, this is a first version experience and can only handle single loads at one time.

By using this Wolfram Cloud Enabled ChatNotebook with Catalyst Database, you acknowledge that you have read and understood this disclaimer and instruction set, and agree to use the system within the bounds of its intended purpose and limitations.

Please navigate to this site and enter your username and password below: https://www.sales.wolframcloud.com/

After logging in, you'll see an available file called "CatalystNavigator"... click on that and you'll see an empty window. Feel free to begin prompting!

As a result of this process, we expect to receive a large number of queries that will do several things: 1) Inform us as researchers what community members are interested in learning from a similar model. 2) Draw out the patterns of questions asked which could reveal a large educational area that the community needs to focus on in general. This could reveal aspects that the community continuously mistakes or is confused about, which requires hours of labor and attention from teams to help answer and provide good user experiences. 3) Frequent queries indicate the need for an always available, global and language agnostic service to respond to questions. 4) This type of methodology to receive feedback is a novel way of doing surveys because it is not simple a Q&A system, it adjusts learnings and feedbacks to each individual. This is known as “in-context learning.”

Hopefully this endeavor informs our research across milestones 2-5 as we complete the project.

Conclusion

In conclusion, this paper has examined the evolving role of LLMs within the realm of governance and community decision-making. By focusing on the trajectory of LLMs, particularly their integration into systems like Cardano's Project Catalyst, the paper highlighted capabilities and potential of various models. Through comprehensive literature reviews and methodical benchmarking, this work has compared leading models based on open versus closed source dynamics, capabilities, and cost-effectiveness, offering an understanding of LLMs' potential role in governance. Ultimately, this exploration aims to equip individuals and communities with the insights and tools needed for more informed decision-making and efficient implementation of LLMs in governance.

It is important to reacknowledge this research was conducted with December 2023 as a milestone submission deadline. This means that all information in here attempted to be as up to date as possible, but undoubtedly there will have been new updates as soon as this is published. For example, while writing the model comparison section, Google's Gemini series was launched which required an agile update. Following that, an open source model called Mixtral was released, and so we needed to react to at least indicate that this was on the spectrum of cutting edge open-source technology. There will likely be many more updates, research papers, and future unexpected changes for which we'll attempt to include over the coming months. To address and enable the reader, we've included code that would allow you to pull a vast collection of research papers available on any particular subject. Lastly, it's important to note that most benchmarks that are used by private companies are done in a way to positively reflect their model's performance. For that reason, please continue to verify and do your own research on any particular outcomes, because no single benchmark, metric, or researcher will get everything accurate.

In the ensuing months, we plan to update the tables, figures, and key metrics to ensure the paper remains accessible and practical. As part of the ongoing initiative, we aim to compile an extensive collection of Catalyst data, documentation, and resources on GitHub for evaluating large language models (LLMs), thereby facilitating the wider Cardano community's engagement with these tools and methodologies. Subsequently, a PDF report will be produced to review various LLMs, pinpointing the most appropriate one for Catalyst use. This involves selecting the ideal model for prospective applications, detailing the costs associated with data infrastructure, and proposing modifications to improve the Catalyst proposal framework. Additionally, we will develop visualizations, tables, and charts to elucidate the research findings effectively. The project will culminate in a final report, a concluding video, and a presentation delivered to the community, all of which will document the milestones achieved and disseminate the research outcomes and their significance.

References

1. Bonneau, J., Felten, E., Goldfeder, S., Narayanan, A., & Miller, A. (2016). Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press.

2. Candel, A., et al. (2023). h2oGPT: Democratizing Large Language Models. arXiv:2306.08161. [online] Available at: https://doi.org/10.48550/arXiv.2306.08161.

3. Chang, M.-W., Devlin, J., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. [online] Available at: https://arxiv.org/abs/1810.04805.

4. De Filippi, P., & Loveluck, B. (2016). The invisible politics of Bitcoin: governance crisis of a decentralised infrastructure. Internet Policy Review. [online] Available at: https://hal.science/hal-01382007.

5. Fan, J.E. (2023). Clean-Discord. [online] GitHub. Available at: https://github.com/JEF1056/clean-discord.

6. Kiayias, A., & Lazos, P. (2022). SoK: Blockchain Governance. arXiv:2201.07188 [cs]. [online] Available at: https://arxiv.org/abs/2201.07188.

7. Laatikainen, G., Li, M., & Abrahamsson, P. (2023). A system-based view of blockchain governance. Information and Software Technology, 157, p.107149. doi:https://doi.org/10.1016/j.infsof.2023.107149.

8. Ling, C., et al. (2023). Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models. arXiv:2305.18703. [online] Available at: https://doi.org/10.48550/arXiv.2305.18703.

9. Monks, R. A. G., & Minow, N. (2011). Corporate governance. John Wiley & Sons.

10. Mosley, L., et al. (2022). Towards a systematic understanding of blockchain governance in proposal voting: A dash case study. Blockchain: Research and Applications, p.100085. doi:https://doi.org/10.1016/j.bcra.2022.100085.

11. Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press. [online] Available at: https://wtf.tw/ref/ostrom_1990.pdf.

12. Rennie, E., et al. (2022). Toward a Participatory Digital Ethnography of Blockchain Governance. Qualitative Inquiry, p.107780042210970. doi:https://doi.org/10.1177/10778004221097056.

13. Stoker, G. (1998). Governance as theory: Five propositions. International Social Science Journal, 50(155), 17-28.

14. Strasser, A. (2023). On pitfalls (and advantages) of sophisticated large language models. arXiv:2303.17511 [cs]. [online] Available at: https://arxiv.org/abs/2303.17511.

15. Swan, M. (2015). Blockchain: Blueprint for a new economy. O'Reilly Media, Inc.

16. Trummer, I. (2022). From BERT to GPT-3 codex. Proceedings of the VLDB Endowment, 15(12), pp.3770–3773. doi:https://doi.org/10.14778/3554821.3554896.

17. Vaswani, A., et al. (2017). Attention Is All You Need. arXiv:1706.03762. [online] Available at: https://arxiv.org/abs/1706.03762.

18. Wang, X., et al. (2023). Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning. arXiv:2301.11916. [online] Available at: https://doi.org/10.48550/arXiv.2301.11916.

19. Williamson, O. E. (1999). Public and private bureaucracies: A transaction cost economics perspective. Journal of Law, Economics, and Organization, 15(1), 306-342.

20. Yang, J., et al. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv:2304.13712. [online] Available at: https://doi.org/10.48550/arxiv.2304.13712.

21. Zhao, W.X., et al. (2023). A Survey of Large Language Models. arXiv:2303.18223 [cs]. [online] Available at: https://arxiv.org/abs/2303.18223.

22. Zhang, L., Ma, X., & Liu, Y. (2022). SoK: Blockchain Decentralization. arXiv:2205.04256 [cs, econ, q-fin]. [online] Available at: https://arxiv.org/abs/2205.04256.

23. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. [online] Available at: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

24. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. [online] Available at: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

25. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. [online] Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

26. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. [online] Available at: https://arxiv.org/abs/2302.13971

27. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. [online] Available at: https://arxiv.org/abs/2307.09288

28. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*. [online] Available at: https://arxiv.org/abs/2304.13712

29. OpenAI, (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*. [online] Available at: https://arxiv.org/abs/2303.08774

30. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*. [online] Available at: https://arxiv.org/abs/2305.10403

31. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. [online] Available at: https://arxiv.org/abs/2312.11805

32. Anthropic. (2023). Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

33. Ahmad Sakor, Kuldeep Singh, Anery Patel, Maria-Esther Vidal. (2020). Falcon 2.0: An Entity and Relation Linking Tool over Wikidata. *arXiv preprint* arXiv:1912.11270 [online] Available at: https://arxiv.org/abs/1912.11270

34. Mistral AI. (2023). https://mistral.ai/news/mixtral-of-experts/

35. xAI. (2023). https://x.ai/

36. Inflection.ai. (2023). https://inflection.ai/inflection-2

37. Zheng, S., Zhang, Y., Zhu, Y., Xi, C., Gao, P., Zhou, X., & Chang, K. C. C. (2023). GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond. *arXiv preprint arXiv:2309.16583* [online] Available at: https://arxiv.org/abs/2309.16583

38. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. Transactions of the Association for Computational Linguistics, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

39. Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160.

40. Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

41. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In International Conference on Learning Representations, 2021a. URL https://openreview.net/forum?id= d7KBjmI3GmQ.

42. Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models, 2023. URL https://arxiv.org/abs/2304.06364.

43. 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. CoRR, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

44. Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144.

45. Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/ P19-1472.

46. Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. Commun. ACM, 64(9):99–106, aug 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

47. Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIGbench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https: //aclanthology.org/2023.findings-acl.824.

48. Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https: //openreview.net/forum?id=uyTL5Bvosj.

49. Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

50. Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

51. Shima Imani, Liang Du, and Harsh Shrivastava. MathPrompter: Mathematical reasoning using large language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pp. 37–42, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.4. URL https://aclanthology.org/2023.acl-industry.4.

52. Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023. URL https://arxiv.org/abs/2305.18654.

53. 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. CoRR, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

54. 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. CoRR, abs/2103.03874, 2021b. URL https://arxiv.org/abs/2103.03874.

55. 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, and et al. Evaluating large language models trained on code. CoRR, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

56. 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. CoRR, abs/2108.07732, 2021. URL https: //arxiv.org/abs/2108.07732.

57. Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https: //aclanthology.org/2022.acl-long.229.

58. Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

59. Archit Parnami, Minwoo Lee. (2022). Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. arXiv preprint arXiv:2203.04291. [online] Available at: https://arxiv.org/abs/2203.04291

60. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113.Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113. [online] Available at: https://www.jmlr.org/papers/v24/22-1144.html

61. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

62. Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., ... & Joty, S. (2023). ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?. *arXiv preprint arXiv:2311.16989*.

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

64. 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.

65. 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.

66. 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.

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

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

69. 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, 2023e

70. 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.

71. 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.

72. 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.

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

74. 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.

75. 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.

76. 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.

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

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

79. 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, 2023c.

80. 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.

81. 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.

82. 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.

83. 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.

84. 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, 2023e.

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

86. 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.

87. 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.

88. 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.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. 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.

98. 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, 2023f.

99. 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.

100. Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. arXiv preprint arXiv:2104.02112, 2021.

101. Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. Summscreen: A dataset for abstractive screenplay summarization. arXiv preprint arXiv:2104.07091, 2021b.

102. 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.

103. Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. arXiv preprint arXiv:2105.03011, 2021.

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

105. Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. Quality: Question answering with long input texts, yes! arXiv preprint arXiv:2112.08608, 2021.

106. Yuta Koreeda and Christopher D Manning. Contractnli: A dataset for document-level natural language inference for contracts. arXiv preprint arXiv:2110.01799, 2021.

107. 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.

108. 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.

109. 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.

110. 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.

111.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.

112. 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.

113. 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.

114. 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.

115. 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.

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

117. 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.

118. 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, 2022a.

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

120. Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. NEWTS: A corpus for news topicfocused summarization. In Findings of ACL, 2022.

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

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

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

124. 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.

125. 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.

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

127. 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.

128. 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.

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

130. 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.

131. 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.

132. 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, 2023d.

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

134. 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.

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

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

137. 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.

138. 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.

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

140. 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.

141. 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.

142. 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.

143. 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, 2023d.

144. 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.

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

146. 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.

147. 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, 2023d.

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

149. 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.

150. 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.

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

152. 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, 2023a.

153. 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.

154. 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.

155. 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, 2023b.

156. 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.

157. 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.

158. 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.

159. 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.

160. 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.

161. 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.

162. 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.

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

164. 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.

165. 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, 2023d.

166. 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.

167. 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.

168. 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.

169. 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.

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

171. 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.

172. OpenAI. Chatgpt plugins. OpenAI, 2023a.

173. 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.

174. 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.

175. 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.

176. 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.

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

178. 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.

179. 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.

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

181. 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.

182. 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.

183. 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.

184. Swan, M. (2015). Blockchain: Blueprint for a new economy. O'Reilly Media, Inc.

185. Liu, Y., Lu, Q., Yu, G., Paik, H.-Y., & Zhu, L. (2022). Defining Blockchain Governance Principles: A Comprehensive Framework. Data61, CSIRO, Australia; University of New South Wales, Australia.

186. Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. Journal of Human Evolution, 22(6), 469-493. https://doi.org/10.1016/0047-2484(92)90081-J

187. Kiayias, A., & Lazos, P. (2023). SoK: Blockchain Governance. arXiv preprint arXiv:2201.07188. https://arxiv.org/abs/2201.07188

188. Laatikainen, G., Li, M., & Abrahamsson, P. (2023). A System-based View of Blockchain Governance. Information and Software Technology, 157, Article 107149. https://doi.org/10.1016/j.infsof.2023.107149

189. Lessig, L. (1998). The Laws of Cyberspace. Presented at the Taiwan Net '98 conference, Taipei, March 1998. https://cyber.harvard.edu/works/lessig/laws_cyberspace.pdf

190. Lindenfors, P., Wartel, A., & Lind, J. (2021). ‘Dunbar’s number’ deconstructed. Biology Letters, 17(20210158). https://doi.org/10.1098/rsbl.2021.0158

191. Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

192. Swan, M. (2015). Blockchain: Blueprint for a new economy. O'Reilly Media, Inc.

193. Tan, J. Z., Merk, T., Hubbard, S., Oak, E. R., Pirovich, J., Rennie, E., ... & Juels, A., et al. (2023). Open Problems in DAOs. arXiv preprint arXiv:2310.19201. https://arxiv.org/abs/2310.19201

194. Vervaeke, J., Andersen, B. P., & Miller, M. (2022). Predictive processing and relevance realization: exploring convergent solutions to the frame problem. Phenomenology and the Cognitive Sciences. https://doi.org/10.1007/s11097-022-09850-6

195. Wolfram, S. (2023). What Is ChatGPT Doing ... and Why Does It Work? Stephen Wolfram Writings. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work.

196. Zhang, L., Ma, X., & Liu, Y. (2023). SoK: Blockchain Decentralization. arXiv preprint arXiv:2205.04256. https://arxiv.org/pdf/2205.04256.pdf