EVALUATING LARGE LANGUAGE MODELS IN AN EMERGING DOMAIN: A PILOT STUDY IN DECENTRALIZED FINANCE

Joshua Pearlson[1], Xiaoyuan Liu[2], Chengsong Huang[1], Kripa George[1], Dawn Song[2], Chenguang Wang[1]
[1]Washington University in St. Louis, [2]UC Berkeley
{jcpearlson, chengsong, chenguangwang}@wustl.edu
{xiaoyuanliu, dawnsong}@berkeley.edu

ABSTRACT

We introduce a new challenge to test the emergent abilities of large language models. Unlike standard benchmarks that aim to examine the ability with existing domains, we propose to test it in a new domain, decentralized finance (DeFi). DeFi has the potential to rewire how the financial system works. Growing to over $250 billion within three years of existence, DeFi's growth is rapid and unprecedented. This domain presents a natural testbed for emergent abilities. A large number of new concepts and entities such as specific cryptocurrencies were released after models stopped training. We create the first dataset resource in this domain with high-quality manual annotations, with a focus on named entity recognition. Our results show, while state of the art models produce reasonable performance in recognizing entities that already existed before they completed training, their performance drops dramatically when new entities are presented. Although improved performance is obtained through teaching models on the training portion of our dataset, the results suggest fundamental algorithmic innovations are required to equip models with emergent intelligence.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated exceptional adaptability, evidencing their competence across a diverse spectrum of tasks (Touvron et al., 2023b; Zhang et al., 2022b; Brown et al., 2020a). As the size of the model increases, there have been observations about emergent abilities like few-shot prompting abilities and zero-shot generalization (Wei et al., 2022a). This versatility extends from everyday conversations to specialized domains such as biomedical (Wang et al., 2023) and business (Desmond et al., 2022).

However, as the training set also gets larger, distinguishing between genuine emergent abilities and simple memorization becomes challenging. Most datasets used for testing, whether general or domain-specific, often contain information already present in the LLMs' expansive training corpora, causing train-test contamination. Consequently, it becomes difficult to conclusively attribute performance improvements to the LLMs' evolving capabilities rather than to their extensive exposure to relevant data.

To address the contamination issue in studying Large Language Models (LLMs), in this paper, we introduce a novel method of using data from emerging domains to test emergent abilities. Specifically, we have compiled a new timestamped corpus in the emerging field of Decentralized Finance (DeFi) and developed a new Named Entity Recognition (NER) dataset from this corpus, aiming to evaluate the zero-shot and few-shot generalization capabilities of LLMs.

DeFi, characterized by its transparency, swift growth, and technical complexity, is an ideal focus for such LLM research. Its decentralized nature, rooted in open-source and consensus approaches, facilitates convenient data gathering from the Internet. The field's rapid evolution not only enhances data accessibility but also produces new, timely knowledge not previously available online, allowing

The dataset and code are available at https://github.com/wang-research-lab/definer.

for the assessment of LLMs post-knowledge cutoff. Furthermore, DeFi's technical intricacy presents a significant challenge for LLMs in generalizing and comprehending novel technical terms.

With a diverse labeled corpus comprising 150 DeFi white papers, we evaluate the generalization ability of both open-source (LLaMA-2-13b (Touvron et al., 2023b)) and closed-source LLMs (GPT-3.5-Turbo, GPT-4) in the NER task. We show that in the DeFi domain, these LLMs exhibit only weak generalization ability in the zero-shot setting. In addition, we observe a clear trend that the more recent the data instance is, the poorer the performance LLM gets. Compared with the human result that shows consistently high performance, we conclude that the LLMs' ability to understand the merging DeFi domain is still mainly attributed to their exposure to the relevant data.

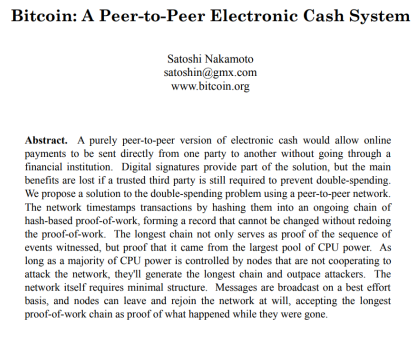
To improve the performance in the emerging DeFi domain, our experiments show that both in-context learning Olsson et al. (2022) and fine-tuning are effective, and fine-tuning can significantly increase the performance, highlighting the need for domain-specific data. To summarize, our main contribution is three-fold:

- To the best of our knowledge, this is the first work to study the generalization ability of LLMs with timestamped data from an emerging domain. The empirical experiments show that LLMs exhibit limited comprehension capabilities in the DeFi domain. Our study also unveils a prominent trend wherein LLMs exhibit an inclination towards memorization as opposed to semantic analysis.
- We collect a DeFi white paper corpus comprising 150 documents and build the first NER dataset in the DeFi domain. We also investigate the ability of LLMs to understand the information in the DeFi field. For a domain that values transparency and automation, the dataset also benefits the construction of DeFi automation tools.
- We demonstrate that limitations in performance can be mitigated through techniques such as in-context learning or fine-tuning. Our comprehensive experiments using these techniques show consistent performance improvement with different models, thereby opening avenues for further technical exploration in this domain.

2 THE DEFINER BENCHMARK

2.1 DEFI CORPUS COLLECTION

Since the introduction of Bitcoin (Nakamoto, 2008) in 2008, Decentralized Finance (DeFi) has been a fast-evolving field and has received continuous attention. Many new terms have been defined and created to describe emerging technical or product innovations. The transparency of DeFi results in a large amount of public textual documents available online, covering almost all domain knowledge. Because the original goal is to attract more decentralized involvement, such data is usually easily accessible.



Bitcoin: A Peer-to-Peer Electronic Cash System

Satoshi Nakamoto
satoshin@gmx.com
www.bitcoin.org

Abstract. A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending. We propose a solution to the double-spending problem using a peer-to-peer network. The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work. The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest pool of CPU power. As long as a majority of CPU power is controlled by nodes that are not cooperating to attack the network, they'll generate the longest chain and outpace attackers. The network itself requires minimal structure. Messages are broadcast on a best effort basis, and nodes can leave and rejoin the network at will, accepting the longest proof-of-work chain as proof of what happened while they were gone.

Figure 1: An example DeFi whitepaper.

Among all types of textual materials online available, white papers stand out as a promising data source for understanding the domain because of its clarity, conciseness, and officiality. Compared with more noisy sources such as DeFi-related tweets, white papers often describe confirmed knowledge or official definitions, and thus are generally more reliable. In addition, white papers have publishing dates which offer a valuable data point on when a cryptocurrency was created.

Considering these advantages, we chose whitepapers as our main data source (example shown in Figure 1) and collected 150 most-accessed open-source whitepapers comprising 32k valid sentences as our target DeFi corpus for the study. We also collected the release date of specific cryptocurrencies as metadata for further experiments.

To improve the quality and accessibility of the corpus, we extract the text from the original PDF document and segment the sentences. We then apply rule-checking to remove malformed sentences to further clean the data, reducing the sentence number from 32k to 25k valid sentences. The remaining 25k sentences form our unlabelled corpus.

2.2 NER IN THE DEFI DOMAIN

The task of named entity recognition (NER) aims to identify and classify named entities in text into predefined categories (Nadeau & Sekine, 2007). It helps in locating domain-specific entities in the given sentence or paragraph, enabling automated processing of the text in specific domains. As the first step to understanding a new domain, NER is an important task to be first investigated.

However, developing NER tools in a new domain is often challenging because of the distribution shift in the occurrence of specific terms, entities, and language contexts. These new terms may not be present in standard language models or pre-trained models. In addition, the focused categories may also change so that generic NER models that are trained on general datasets Li et al. (2020) cannot be directly applied.

In specialized domains, it is important to define named relevant entity categories, as the definition quality directly affects the usefulness in downstream applications. In this work, we define the following three categories of entities in the DeFi domain to reflect general DeFi knowledge.

- **CRYPTO**: Cryptocurrencies, such as Bitcoin, Ethereum, and Litecoin, operate as decentralized mediums of exchange, utilizing cryptographic algorithms for security on blockchain networks, and serve as alternatives to currencies in traditional centralized financial systems.
- **APPS & PRODUCTS**: Apps/Products encompass a diverse array of platforms, exchanges, and applications, such as Ethereum, Binance, Uniswap, Compound, Terra Network, and MakerDAO. These entities facilitate peer-to-peer transactions, eliminating intermediaries and central authorities, leveraging blockchain for secure and transparent transactions, and gaining popularity for their contributions to the DeFi field.
- **PROGRAMS**: Programs constitute technologies, algorithms, coded functions, and languages like Solidity and Vyper, crucial for supporting smart contracts and decentralized networks. These programs provide the foundational infrastructure, including protocols, middleware, and decentralized storage solutions, contributing to the secure, transparent, and decentralized operations of blockchain systems.

We now define the DeFi NER task as the procedure to identify named entities that follow the definition in the above three categories in the given text. With a given dataset that contains text and entity labels, we can evaluate the performance of a DeFi NER tool by comparing the identification result with the ground truth. In terms of evaluation procedure, we follow the standard practice and incorporate metrics like F1, precision, and recall.

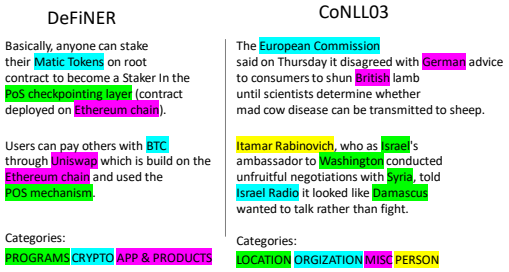
2.3 DATASET

With the sentences from the DeFi whitepaper corpus, we built a dataset and labeled it following our definition of the DeFi NER task, generating a dataset with 7,338 labeled sentences. The statistics of our dataset are in Table 1. Specifically, we first randomly sampled 1,374 sentences containing DeFi entities and manually labeled them as a seed dataset. We then use it as seed data to fine-tune the existing SpaCy (Honnibal et al., 2020) NER model based on RoBERTa (Liu et al., 2019) to accelerate the sentence filtering procedure. After human checks on the model labeled result, revisions are applied to form a training set consisting of 5,964 labeled sentences. The seed dataset is then used as the test set. An example of the dataset instance is provided in Figure 2a. A WordCloud highlights the frequency of specific entities in our dataset in Figure 2b.

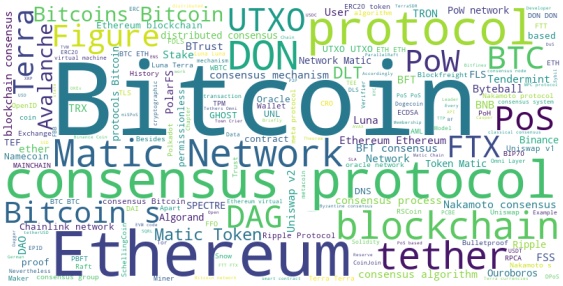
To ensure our dataset was of high quality, we calculated a human labeling alignment score from multiple rounds of labeling. The human annotation yielded an alignment score of 97.6% on the test set. In addition, for our training set, we ran a random sample of 200 sentences and the alignment score on our revision-based labeling is 91.3%. This result verifies that our dataset is of high quality.

	Sentences	CRYPTO	PRODUCTS	PROGRAMS	Avg.(w,s)	Avg.(e,s)	Rate.(O,E)
Train	5,964	3,463	1,191	3,065	24.32	1.29	6.44%
Test	1,374	763	282	742	24.39	1.30	6.41%
Total	**7,338**	**4,226**	**1,473**	**3,807**	**24.34**	**1.30**	**6.26%**

Table 1: The dataset statistics of DeFiNER. Shown above are the number of entities in each category(Crypto, Apps/Products, and Programs) within the test train and total dataset. Avg.(w,s) is the average number of words per sentence, and Avg.(e,s) is the average number of entities per sentence. Rate.(O,E) is our entity overlap ratio for our data further explained in Section 2.3. Our dataset contains 3036 unique CRYPTO, 2918 unique PROGRAM, and 902 unique APPS/ PRODUCTS.



(a) Dataset comparison between our DeFiNER and CoNLL03. The entity types in DeFiNER are domain-specific to DeFi while entity types in CoNLL03 are for news articles.



(b) WordCloud represents the frequency of specific entities in our dataset.

Figure 2: Comparison of datasets and WordCloud representation.

An additional observation about our dataset is its high rate of overlapping entities compared with general-domain NER. In the DeFi domain, there exists a large number of overlapping entities. These entities are defined as existing in at least two distinct categories. To quantify this overlap we calculated our entity overlap score which is equal to $\frac{\text{OverlappingEntites}}{\text{TotalEntities}} * 100\%$. Our dataset generated an overlap score of 6.26%. Next, to establish a baseline we calculated this statistic on CoNLL03 (Tjong Kim Sang & De Meulder, 2003) and it generated a score of 0.61%. This means that our models have to deal with $10x$ greater overlap in the NER domain compared with the traditional NER dataset of CoNLL03, highlighting one domain-specific difficulty.

3 Models

For our DeFi NER task, we evaluate state of the art LLMs including LLaMA-2-13b (Touvron et al., 2023b)), GPT-3.5 Turbo, and GPT-4. Specifically, we explored both zero-shot and few-shot settings as discussed below. In addition, we also evaluate the effectiveness of fine-tuning and compare the result with traditional language models such as ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019).

3.1 Zero-Shot

Zero-shot inference is a technique that allows LLMs to infer and produce outputs for prompts or queries they were never specifically trained on. We use a prompt to activate the models' ability to generate the answer to a new problem. In detail, the input of LLM will be $[\text{prompt}]+[\text{sentence}]$. We evaluate the LLM performance on our DeFi NER dataset with a manually designed prompt: "Extract cryptocurrencies, products, and programs from the given text. If entities are recognized, show them in the following format: entity type (cryptocurrency, programs, or products): entity. Separate each entity type by a newline character. If no entities are recognized, return None."

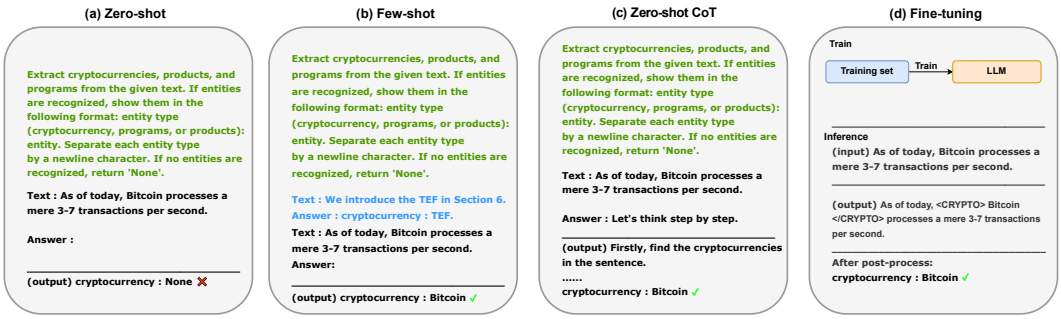


Figure 3: Illustration of different approaches for the DeFi NER task.

3.2 Few-Shot

Few-shot inference is a technique that uses a small number of task-specific examples, ranging from 2 to a few dozen, during inference. In detail, we use in-context learning (Radford et al., 2019) to implement few-shot performance. It achieves this by presenting concatenated input-target examples, referred to as "shots," alongside an unlabeled query example.

In our NER task, we instruct the LLM to respond to specific questions while adhering to the prescribed format or style. This is done to streamline subsequent automated recognition and evaluation processes. However, accomplishing this task solely through prompts can be a challenging endeavor. Hence, the common practice in the context of in-context learning is to provide corresponding input-output pairs. Within the Few-Shot section, all tests utilized a randomized selection of context sentences.

3.3 Fine-Tuning

To evaluate the fine-tuning methods, we follow the formalization of sequence labeling for standard NER tasks (He et al., 2020). Generally, the label of a word in NER is composed of two parts, i.e., "X-Y", where "X" indicates the position of the labeled word and "Y" refers to the corresponding category within a pre-defined taxonomy. In our implementation, we follow the BIO (Begin, Inside, Outside) system. The B- prefix before a tag indicates that the tag is the beginning of an entity, and an I- prefix before a tag indicates that the tag is inside an entity. An O tag indicates that a token belongs to no entity. In this setting, we can transform a NER task into a classification task. Figure 3 illustrates the methodology we employ, including instructions preceding the initial input as part of this procedure.

4 Experiment

We show that LLMs have limited ability to understand the task of NER in this emerging domain. We then present a more detailed analysis to illustrate the upperbound model performance on our dataset.

4.1 Zero-Shot

We show the zero-shot evaluation results in Table 2. As we can see, The zero-shot performance for all three models is low. Specifically, the Llama model suffers from high false positives. GPT-4 achieves the highest F1 score due to its high precision, but its recall value is also low. In addition, we also try to augment the zero-shot method with chain-of-thought (CoT) prompting Wei et al. (2022b). There is a slight improvement for GPT-3.5 Turbo, while for Llama only the recall value improves. Overall, the improvement from CoT is not significant.

To further reason about the low zero-shot performance, with the timestamp we collected for the cryptocurrencies, we evaluate the per-category recall in different years. The result is presented in Figure 4. As we explained in Section 1, the model would have been trained on how to understand

Models	Methods	Recall	Precision	F1 Score
Llama-2-chat-13b	zero-shot	25.84	9.71	14.11
	zero-shot CoT	27.49	8.55	13.04
	few-shot (4-shot)	22.13	21.02	21.56
GPT-3.5 Turbo	zero-shot	20.65	21.80	21.21
	zero-shot CoT	23.90	22.70	23.28
	few-shot (4-shot)	**40.14**	**34.58**	**37.16**
GPT-4	zero-shot	20.67	29.15	29.89

Table 2: The performance of DeFiNER in a zero-shot, chain-of-thought, and multi-shot setting with different LLMs. Few-shot learning refers to the approach of incorporating four examples into the input using in-context learning. The chain-of-thought (CoT) prompting is an attempt to improve zero-shot results. Due to computation constraints, for GPT-4, only zero-shot result is available.

a specific word before and not rely solely on context, if relevant data is in the training set before the knowledge cutoffs. However, if the entity only existed after the ending date of a given LLM's training data then we know for sure that the model has never seen this term before.

As shown in the figure, both open-source and closed-source LLMs share the same pattern: As the entity gets more recent, the performance of LLMs gradually declines, becoming increasingly poor around 2021. On the other hand, GPT-3.5 Turbo is trained on the data before 2021.9 so the model cannot have content in its training set after the 2021.9 cutoff. Also, newer information will have less text on the web, so it will have a lower likelihood of being in the LLMs training set. We witness the score of the LLM decreasing with the likelihood of a term having previously been in its training set. This result implies that LLMs can not transfer contextual knowledge of recognizing certain entities into the new domain of DeFi.

4.2 Few-Shot

For our few-shot experiments also listed in Table 2, we observe a significant performance improvement from zero-shot prompting. It can be attributed primarily to a better understanding of the output format. In few-shot prompting, the LLM is more likely to generate content that aligns with the expected structure, which improves the success rate for rule-based extraction methods in the subsequent evaluation. While for the zero-shot experiment, despite our efforts to control the LLM's output format in the prompt, it remains challenging to force LLMs to follow the exact format as instructed.

We also explore how the number of examples used in the few-shot prompting affects the final performance of the model. As shown in Figure 5, for both models, increasing the number of provided examples within a specific range leads to improved model performance. However, it's crucial to note that having too many examples cannot continually impact the model's performance. This tendency is particularly noticeable within the 4-8 shot range. Beyond the 4-8 shot range, more examples begin to poorly impact the performance of LLMs. In addition, the context lengths supported by the LLMs also become a limitation.

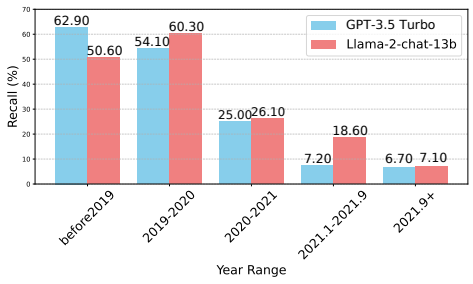


Figure 4: Recall score of Crypto entities within DeFiNER sorted by the time of introduction for a given Crypto entity. This is all based on zero-shot results. The Crypto category, due to its repetition of terms, typically has above-average zero-shot performance than other categories. The number of unique entities for CRYPTO per year range is 2,6,8,26 (listed in range ascending from most recent to least recent) amounting to 46 of our 143 unique CRYPTO values within our test set.

We also conducted a case study on the few-shot model result. We found that the additional examples not only help with understanding the format but also help the model understand the definition of the entity. A Llama example is provided below:

- **Input:** "To this purpose, for Ethereum contracts we manually inspect the Solidity source code, while for Bitcoin contracts we search their web pages and related discussion forums."
- **Zero-shot Llama:** None.
- **Few-shot (4-shot) Llama:** cryptocurrency : Ethereum; Cryptocurrency : Bitcoin

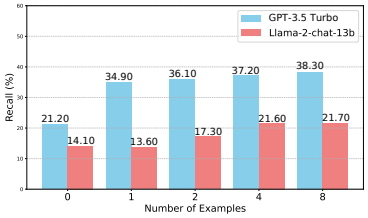


Figure 5: Recall when using a different number of examples in the few-shot setting.

As shown in the example, the model successfully identified that Ethereum and Bitcoin should be a type of cryptocurrency. This is made possible because the context provided allows for the model to see similar entities and learn how to better identify DeFi entities.

4.3 Fine-Tuning

We also explore the fine-tuning method to improve model performance as it is the direct way to introduce the knowledge with our training set discussed in Section 2.3. We also incorporate smaller language models for comparison and the result is shown in Table 3. As we can see, fine-tuning significantly improves the performance for all models, and the largest Llama-2-chat-13b model shows an advantage after the fine-tuning.

Models	Recall	Precision	F1 Score
ALBERT	71.80	72.07	71.91
BERT	73.88	70.99	72.41
RoBERTa	71.78	69.34	70.55
Llama-2-chat-13b	74.47	77.78	76.09

Table 3: The results of DeFiNER after fine-tuning each model. All models underwent a full fine-tuning.

4.4 Error Analysis

To gain better insights, we conducted an error analysis for the result from the fine-tuned Llama-2 model. We chose to conduct this with the fine-tuned Llama-2 model since it had the best performance. Particularly, we find 3 main error types when analyzing our models' incorrect outputs Listed from most frequent to least frequent. For an expanded list of all model error types with descriptions see Table 4.

Error Category	**Explanation**	**Percentage (%)**
Missing label of Program	Some data points lack proper labels as Programs.	15.00
Miscellaneous errors	Various errors affecting model accuracy.	12.00
Labeling random as a Program	Incorrectly assigning random data as Programs.	8.00
Labeling random as Crypto	Random instances mislabeled as Crypto.	5.00
Not getting full label of Program	Failure to obtain complete labels for Programs.	5.00
Labeling a Program as a Crypto	Programs mistakenly labeled as Crypto.	4.00
Dataset Quality error	Issues related to the quality of the dataset.	2.00
Uncategorized errors	/	49.00

Table 4: Distribution of error categories and their explanations in the error analysis.

1. The model is not able to correctly identify a Program.
2. Model gives a random word the label of Program.
3. Model gives a random word the label of Crypto.

For errors 1,2 we believe the model simply has trouble understanding what is a Program and what is not a Program. This may be because Programs tend to have a vast category of name sizes. For example, in our test dataset, one Program has the name "Ripple Client software" and one has the name "PoS". These two names have vastly different lengths and look very dissimilar however they are both Programs in our dataset.

For error 3 we believe the model has learned from our training data that Cryptos tend to be abbreviated into 3-4 character all-uppercase words. This can be a problem when in the test set it labels "AWS" (Amazon web services) as a Crypto when it should not be in any of our defined categories.

5 Related Work

Large Language Models (LLMs) represent a category of machine learning models employing deep neural networks for natural language processing. These models are trained on extensive volumes of text data, encompassing sources like web pages, books, and articles, employing unsupervised learning methodologies. Among the most notable LLMs is the GPT series, including GPT-3 Brown et al. (2020b). Other leading models include Meta's OPT Zhang et al. (2022a), Llama Touvron et al. (2023a), and BigScience's BLOOM Workshop (2023). LLMs Crispino et al. (2023) exhibit a remarkable capacity to grasp the subtleties and intricacies of language, including semantics, syntax, and contextual nuances. This proficiency renders them highly suitable for an extensive array of NLP applications. The advent of LLMs has brought about a paradigm shift in the NLP domain, with their ongoing progress anticipated to yield significant ramifications across diverse industries and disciplines.

In the realm of assessing LLMs' effectiveness in diverse tasks, multiple benchmarks have emerged. For instance, MMLU Hendrycks et al. (2021a) is focused on creating a thorough evaluation framework for text models in multi-task scenarios. On the other hand, HELM (Liang et al., 2022) provides a comprehensive assessment, covering various facets of LLM performance, including language comprehension and common-sense reasoning. Structure prediction tasks Wang et al. (2022; 2020) have also been extensively evaluated. Big-Bench (Srivastava et al., 2022), which presents a set of 204 challenging tasks spanning various domains. The goal is to assess LLMs on tasks that go beyond what current language models can handle. Another valuable dataset is AGIEval (Zhong et al., 2023), designed as an evaluation framework for measuring the performance of foundational models in standardized exams that focus on human-centric assessments.

In addition to general tasks, there are specialized benchmarks tailored for specific downstream applications. For instance, MultiMedQA (Singhal et al., 2022) is centered around medical question-answering, focusing on evaluating LLMs in terms of their clinical knowledge and question-answering abilities. STEM Shen et al. (2024) and Social Yuan et al. (2024) aim to test models' ability to understand fundamental STEM and social skills. Another benchmark, MATH (Hendrycks et al., 2021b), is dedicated to assessing the reasoning and problem-solving skills of LLMs in the domain of mathematics. ScienceQA (Lu et al., 2022) introduces a multi-modal benchmark encompassing a wide range of science topics. Datasets Shen et al. (2022) have been proposed to understand code syntax. For planning, G-PlanET Lin et al. (2022) focuses on evaluation for LLM to do grounded planning. SciEval (Sun et al., 2023) is a benchmark for scientific research ability evaluation of LLMs. In comparison to these benchmarks, DeFiNER assesses the ability to comprehend Decentralized Finance, an emerging field.

There exist many datasets for evaluating the NER task in a general domain like news or Wikipedia (Tjong Kim Sang, 2002; Hovy et al., 2006; Sanh et al., 2019). Special domain NER is imperative when dealing with texts from highly specialized fields, such as medical records (Kim et al., 2003; Tanabe et al., 2005; Faessler et al., 2020; Dogan et al., 2014), legal documents Leitner et al. (2019); Tufiș et al. (2020) or scientific literature (Luan et al., 2018; Hou et al., 2021; Dogan et al., 2014). In these domains, entities can be highly specific, and their accurate identification is crucial for precise information retrieval, data analysis, and domain-specific knowledge extraction. Over the course of their prolonged development, these domains have exhibited limited alterations in

their designated entities, while conversely, decentralized finance has experienced a swift evolution, frequently giving rise to novel cryptocurrencies and widely accepted nomenclatures.

6 CONCLUSION

To test the emergent abilities of large language models, this paper evaluates models in the emerging domain of DeFi. DeFi is a new and critical domain. We are among the first groups to build an annotated resource in this domain. Our experiments show that LLMs struggle to generalize to the DeFi domain where new entities and concepts were invented after the completion of model training. This study provides evidence that data contamination and memorization may be major factors for LLMs to exhibit strong performance in existing domains. Only after training models on our dataset do they start to understand the new DeFi knowledge. Despite this, we conclude that improving the understanding of large language models in emerging domains is an open research question and requires fundamental algorithmic innovations from the community.

7 LIMITATIONS

For the limitation of our method, the NER procedure relies on pre-trained language models from third parties. As reported, LLMs are known to have performance limitations (Brown et al., 2020b). A limitation of our dataset is that we have not conducted a large-scale manual evaluation of all text in the corpus and instead use automated procedures together with human checking to accelerate the labeling procedure. The main focus of our study is to evaluate the LLMs' performance in an emerging domain. Finally, due to the computation resource constraints, we only evaluated the basic zero-shot setting with the slowest GPT-4 model. Future work could include testing all settings and incorporating more powerful LLMs.

8 ETHICAL CONSIDERATIONS

We hereby acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. This work is about evaluating the NER ability of pre-trained language models in different settings. Our ethical considerations and the work's underlying future impacts are discussed in the following perspectives. Language models are known to present potential risks and limitations (Brown et al., 2020b), and the corpus used in pre-training (such as DeFi white paper) may introduce unwanted biases and toxicity. We do not anticipate the production of harmful outputs after using our method or datasets, especially for vulnerable populations.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. Agent instructs large language models to be general zero-shot reasoners. *arXiv preprint arXiv:2310.03710*, 2023.

Michael Desmond, Evelyn Duesterwald, Vatche Isahagian, and Vinod Muthusamy. A no-code low-code paradigm for authoring business automations using natural language. *ArXiv preprint*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, 2019.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 2014.

Erik Faessler, Luise Modersohn, Christina Lohr, and Udo Hahn. ProGene - a large-scale, high-quality protein-gene annotated benchmark corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020. ISBN 979-10-95546-34-4.

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xian-Ling Mao, and Sheng Jiang. A survey on recent advances in sequence labeling from deep learning models. *ArXiv*, abs/2011.06727, 2020. URL `https://api.semanticscholar.org/CorpusID:226955954`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proc. of ICLR*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv preprint*, 2021b.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303. If you use spaCy, please cite it as below.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 2006.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proc. of ICLR*, 2020.

Elena Leitner, Georg Rehm, and Julián Moreno Schneider. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, 2019.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proc. of ACL*, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 2022.

Bill Yuchen Lin, Chengsong Huang, Qian Liu, Wenda Gu, Sam Sommerer, and Xiang Ren. On grounded planning for embodied tasks with language models. In *AAAI Conference on Artificial Intelligence*, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv preprint*, 2022.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proc. of EMNLP*, 2018.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 2007.

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, 2008.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. In-context learning and induction heads. *ArXiv preprint*, 2022.

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

Victor Sanh, Thomas Wolf, and Sebastian Ruder. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019.

Da Shen, Xinyun Chen, Chenguang Wang, Koushik Sen, and Dawn Song. Benchmarking language models for code syntax understanding. In *EMNLP*, 2022.

Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models. *arXiv preprint arXiv:2402.17205*, 2024.

K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Y. Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Toma$sev, YunLiu, AlvinRajkomar, Joëlle K.Barral, ChristopherSemturs, AlanKarthikesalingam, and VivekNat
$172 - -180, 2022. URL$.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin,

Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Kumar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, S. Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton

Chang, Trishala Neeraj, Tushar Khot, Tyler O'Brien Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yuntao Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, 2022.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhe-Wei Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. *ArXiv preprint*, 2023.

Lorraine K. Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, John Wilbur, Lynne H, and Wayne. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 2005.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

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

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

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. Collection and annotation of the Romanian legal corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 2773–2777, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.337.

Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 803–823, Dublin, Ireland, May 2022. Association for Computational Linguistics. 10.18653/v1/2022.findings-acl.67. URL https://aclanthology.org/2022.findings-acl.67.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. Exploring the in-context learning ability of large language model for biomedical concept linking. *ArXiv preprint*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022a. URL https://api.semanticscholar.org/CorpusID:249674500.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, 2022b.

BigScience Workshop. Bloom: A 176b-parameter open-access multilingual language model, 2023.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. Measuring social norms of large language models. *arXiv preprint arXiv:2404.02491*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv preprint*, 2022b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv preprint*, 2023.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. Agent instructs large language models to be general zero-shot reasoners. *arXiv preprint arXiv:2310.03710*, 2023.

Michael Desmond, Evelyn Duesterwald, Vatche Isahagian, and Vinod Muthusamy. A no-code low-code paradigm for authoring business automations using natural language. *ArXiv preprint*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, 2019.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 2014.

Erik Faessler, Luise Modersohn, Christina Lohr, and Udo Hahn. ProGene - a large-scale, high-quality protein-gene annotated benchmark corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020. ISBN 979-10-95546-34-4.

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xian-Ling Mao, and Sheng Jiang. A survey on recent advances in sequence labeling from deep learning models. *ArXiv*, abs/2011.06727, 2020. URL `https://api.semanticscholar.org/CorpusID:226955954`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proc. of ICLR*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv preprint*, 2021b.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303. If you use spaCy, please cite it as below.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 2006.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proc. of ICLR*, 2020.

Elena Leitner, Georg Rehm, and Julián Moreno Schneider. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, 2019.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proc. of ACL*, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 2022.

Bill Yuchen Lin, Chengsong Huang, Qian Liu, Wenda Gu, Sam Sommerer, and Xiang Ren. On grounded planning for embodied tasks with language models. In *AAAI Conference on Artificial Intelligence*, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv preprint*, 2022.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proc. of EMNLP*, 2018.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 2007.

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, 2008.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. In-context learning and induction heads. *ArXiv preprint*, 2022.

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

Victor Sanh, Thomas Wolf, and Sebastian Ruder. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019.

Da Shen, Xinyun Chen, Chenguang Wang, Koushik Sen, and Dawn Song. Benchmarking language models for code syntax understanding. In *EMNLP*, 2022.

Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models. *arXiv preprint arXiv:2402.17205*, 2024.

K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Y. Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Toma$sev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Nat$
$172 - -180, 2022. URL$.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta

Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Kumar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, S. Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O'Brien Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yuntao Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, 2022.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhe-Wei Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. *ArXiv preprint*, 2023.

Lorraine K. Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, John Wilbur, Lynne H, and Wayne. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 2005.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

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

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

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. Collection and annotation of the Romanian legal corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 2773–2777, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.337.

Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 803–823, Dublin, Ireland, May 2022. Association for Computational Linguistics. 10.18653/v1/2022.findings-acl.67. URL https://aclanthology.org/2022.findings-acl.67.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. Exploring the in-context learning ability of large language model for biomedical concept linking. *ArXiv preprint*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022a. URL https://api.semanticscholar.org/CorpusID:249674500.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, 2022b.

BigScience Workshop. Bloom: A 176b-parameter open-access multilingual language model, 2023.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. Measuring social norms of large language models. *arXiv preprint arXiv:2404.02491*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv preprint*, 2022b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv preprint*, 2023.