Position: Key Claims in LLM Research Have a Long Tail of Footnotes

Anna Rogers [*1] **Alexandra Sasha Luccioni** [*2]

Abstract

Much of the recent discourse within the ML community has been centered around Large Language Models (LLMs), their functionality and potential — yet not only do we not have a working definition of LLMs, but much of this discourse relies on claims and assumptions that are worth re-examining. We contribute a definition of LLMs, critically examine five common claims regarding their properties (including 'emergent properties'), and conclude with suggestions for future research directions and their framing.

1. Introduction

Large Language Models (LLMs) have become ubiquitous in the Machine Learning (ML) research landscape, and they already impact the lives of thousands of people in contexts ranging from health (Graber-Stiehl, 2023; Harrer, 2023) to education (Kasneci et al., 2023). Yet despite the many research articles on LLMs, their very definition remains unclear, and much of this work is based on claims that are often stated, but remain debatable in terms of their framing, theoretical grounding, or empirical evidence. When we, as researchers, repeat such claims uncritically, we contribute to the narratives shaped by business interests (McKelvey et al., 2023), and may mislead the public and ourselves.

This position paper argues that LLM research should be more precise with its key terms and claims. To that end, we propose a definition for the term "LLM" (§2), and we critically examine five common claims about LLM functionality, drawing heavily on both empirical studies and socio-technical critiques of LLMs (§3). We then consider the impact that these claims have on ML research (§4). We conclude with concrete proposals for maintaining rigor and diversity in ML research and practice (§5).

[*]Equal contribution [1]IT University of Copenhagen [2]Hugging Face, Canada. Correspondence to: Anna Rogers <arog@itu.dk>, Sasha Luccioni <sasha.luccioni@hf.co>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2. What Counts as an Large Language Model?

The common technical definition of "language models" is "models that assign probabilities to upcoming words, or sequences of words in general" (Jurafsky & Martin, 2024, p.32). Yet the term "large language models" is currently used in quite a different way, both within the research community (e.g. in the call for papers of academic conferences such as EMNLP (2023)), in news articles (Roose, 2023), and even by legislators at U.S. Senate hearings (Zakrzewski, 2023)). Despite its ubiquity, this definition is far from clear. For instance, how many parameters should a neural network have to qualify as an LLM? And do only Transformer-type architectures qualify? What about multimodal models, such as text-to-image or image-to-text?

Given the many subfields and contexts in which this term is already used, it is not feasible to impose a single definition and expect everybody to adhere to it. But even if there is no agreement, both academic and general public discussions would be more productive if the authors of research papers spelled out or cited their own working definitions. What follows is our own attempt, which we hope could be useful to others (either for using our version, or as a base to be modified by other researchers to reflect their own understanding of this term). It relies upon three definitional criteria:

(1) **LLMs model text**[1] **and can be used to generate**[2] **it** based on input context, i.e. by selecting the tokens that are the most likely given the partial context provided as input (either masked[3] or as a prompt). The text can be in any modality — characters, pixels, audio, etc.

(2) **LLMs receive large-scale pretraining**, where 'large-

[1]We understand 'text' as 'a unit of language in use' (Halliday & Hasan, 2013), the product of using that language. We see this as a more accurate description, because LLMs model the corpora they are trained on, rather than language in general (Veres, 2022).

[2]Most current LLMs produce text in some way, and so our definition focuses on these. To extend it to energy-based models or discriminative models like ELECTRA (Clark et al., 2020), we could say 'can be used to generate or score text'.

[3]While BERT-style (Devlin et al., 2019) encoders could be used to generate text, it is admittedly a tortuous way of doing so, compared to autoregressive models. However, in both cases we fundamentally solve a classification problem over the model's vocabulary, and autoregressive models could be viewed as a special case of masked language models.

scale' refers to the pre-training data rather than the number of parameters. The exact threshold for LLM-qualifying volume of data is necessarily arbitrary, and for an English corpus we propose setting it to 1B tokens[4] (inspired by Chelba et al. (2013)).

(3) **LLMs are used for transfer learning**, on the assumption that they encode information that can be leveraged in other tasks. Currently the most common transfer learning methods with LLMs are fine-tuning, as in BERT (Devlin et al., 2019), and prompting, as in GPT-3 (Brown et al., 2020), but there are many other methods (Pan & Yang, 2010; Ramponi & Plank, 2020; Alyafeai et al., 2020; Zhuang et al., 2021).

According to the above criteria, BERT (Devlin et al., 2019) and its derivatives do qualify as LLMs, as do models from the GPT series (Radford et al., 2018). So do n-gram language models, given that they are derived from a sufficiently large corpus, such as Google Books (Lin et al., 2012). Earlier word-level representations such as word2vec (Mikolov et al., 2013) are ruled out on the first criterion, when they are viewed by themselves, but their training is conceptually very similar to a masked language model, and they can also use large volumes of text (over 100B tokens for the original word2vec). Modern LLMs can also be used standalone, or for creating representations used in other systems (e.g. BERT's [CLS] token representation (Devlin et al., 2019) fed into classifiers). On our criteria, such representations would not be LLMs, but they are derived from LLMs.

Our first criterion does not rule out multimodal models like GPT-4 (OpenAI, 2023)), as long as they output text – even if they also accept or output images or other modalities. The "text" is typically human-written text in a natural language, but it could also be synthetic data (e.g. text created with templates from knowledge base data). The training data of modern LLMs typically also includes text that is not natural language data: code, ascii art, midi music, math notation, chess transcripts etc. While such data can be learned via token prediction, it is not the focus of our definition.

[4]The 1B threshold could need adjustment for different languages and tokenization schemes, if empirical evidence justifies it. We propose to consider raw source data, without any augmentation or considering multiple training runs. We are assuming that these data points would be mostly unique, since it is common practice to deduplicate training data.

When it comes to multimodal LLMs, the amount of *textual* information in multimodal data can still be approximated via token count (e.g. in transcripts). The other information would be extralinguistic, and a unit for that is yet to be developed – but compute or parameter counts do not do it justice either. Ideally we would be able to semantically chunk the multimodal content at least as crudely as tokens chunk text, and relate it to the text that it grounds. E.g. we should be able to distinguish between a voiceover over blank screen, a dialogue captured with a fixed camera, or the same dialogue captured from various angles, based on who is speaking.

Our second criterion allows for the inclusion of models such as tinyBERT (Jiao et al., 2020): it has only 4.4M parameters, but its training dataset (via model distillation) is the same as that of the full BERT model, which contains 3.3B tokens. Our choice of linking the "large" part of LLMs to the volume of data rather model size also helps to deal with another edge case: the models that were reduced in size (e.g. via distillation, such as Sanh et al. (2019)) for the sake of computational efficiency, but maintain comparable performance to the original models.

Our third criterion applies to the "general-purpose" LLMs that are purported to be domain-independent. But the core criterion is transfer learning, which may also take place within a specific domain. For example, SciBERT (Beltagy et al., 2019) or Galactica (Taylor et al., 2022) are still LLMs, even though they were primarily trained on scientific literature, because they are expected to be used flexibly for a range of tasks within that domain.

At present, in most cases the three criteria listed above correspond to Transformer-based models that are used to generate text. But having a more concrete definition helps to ground the scientific discourse, and provide a point of reference for updating it in the future.

As we learn more about both LLMs and neural architectures, our proposed threshold for "large" may change, e.g. if there is empirical evidence or theoretical guarantees that a certain volume of natural texts provides sufficient signal for a defined set of linguistic "skills" for a given model architecture. Hopefuly in the future we would also have more proofs, theoretical rationales, or empirical evidence, based on which the "large" part could be further qualified by numerous factors relevant to the performance of the final model: the diversity in the textual data (domains, languages, registers etc.), benchmark contamination, acceptable levels of data augmentation or explicit instruction in the form of annotated data, ratio and role of non-linguistic data, duplicate and near-duplicate data points, allowed number of model runs over the training data, etc.

LLMs vs "foundation" and "frontier models". LLMs are also sometimes referred to as "foundation models", a term proposed by Bommasani et al (2021) to refer to "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks". This partly corresponds to our criteria (2) and (3), but makes no attempt to quantify their scale. It is also intentionally broader, so as to include e.g. models for computer vision or protein data. The term "LLM" is more specific, and useful in studies modeling language data.

Another recently proposed term is "frontier models", defined as "highly capable foundation models that could exhibit suf-

ficiently dangerous capabilities" (Anderljung et al., 2023). This relies on the above "foundation model" term, and only adds the criterion of "sufficiently dangerous capabilities", which the authors acknowledge to be vague. Various kinds of risks from LLMs are beyond the scope of this work, but see §3.5 for a relevant discussion of "emergent properties".

3. Fact-checking LLM Functionality

We discuss five common claims about LLMs: that LLMs are robust (§3.1), that they systematically achieve state-of-the-art results (§3.2), that their performance is predominantly due to their scale (§3.3), that they are "general-purpose technologies (§3.4) and that they exhibit emergent properties (§3.5). We are not saying that all these claims are completely false – but they all have many caveats, which are mentioned much less frequently. By collecting existing evidence and counter-arguments, we aim to highlight some of the gaps and inconsistencies in our current knowledge and to help orient future work so as to address these gaps.

3.1. Claim: LLMs are Robust

Early symbolic AI approaches are often described as "brittle" because of their strict dependence on pre-formulated knowledge and lack of robustness outside of the distribution they were trained on. Lenat & Feigenbaum (1981, p.1175) described this as "*a plateau of competence, but the edges of that plateau are steep descents into complete incompetence*". With the advent of LLMs, the issue of robustness is seen to be much less prominent. For instance, Bommasani et al. (2021, p.109) state: "*pretraining on unlabeled data is an effective, general-purpose way to improve accuracy on [out-of-distribution] test distributions*". LLMs are often presented as multi-task learners that are robust without explicit supervision, even outside the distribution they were trained on (Radford et al., 2019; Hendrycks et al., 2019; 2020).

Indeed, we have overcome the problem of the *steep* descents into complete incompetence: unfamiliar inputs no longer completely break the system. But deep-learning-based ML systems are still fundamentally brittle, only in a different way: according to Chollet (2019, p.3), they are *"unable to make sense of situations that deviate slightly from their training data or the assumptions of their creators"*. Chollet goes even further, stating in an interview that there are no fixes for this issue (Heaven et al., 2019). Impressive as the latest LLMs are, they still make errors even in simple tasks like adding numeric literals (Chang & Bergen, 2023). Moreover, they are just as vulnerable to adversarial attacks (Zou et al., 2023) as earlier models (Wallace et al., 2019).

One could argue that "robust" does not mean "perfect" – but in that case, what does it mean? For an ML engineer, it is something like "sufficiently useful in practice". From that point of view, two situations are possible: the model will be deployed in the conditions either (a) guaranteed to be similar to its training distribution in all aspects that matter for its performance, or (b) expected to diverge from that. **Our current LLM-based solutions may be sufficiently robust for in-distribution scenarios, but few would argue the same for out-of-distribution cases.** Case (b) covers many, if not most, LLM application areas: text classifiers will continually encounter new topics and domains, language usage will evolve, the correct answers to factual questions will change, discourse strategies for interaction with AI-enabled chatbots will shift, people will adapt what they post online and to the privacy and surveillance concerns, etc. And failures of ML systems may have real-world consequences for those who diverge the most from its core distribution: e.g. people may be denied asylum due to errors of machine translation systems, something that we have already seen happen (Nalbandian, 2022).

One well-studied cause of brittleness in the LLMs of the BERT generation was shortcut learning (McCoy et al., 2019; Rogers et al., 2020; Branco et al., 2021; Choudhury et al., 2022, inter alia) – models picking up undesirable spurious correlations from the training data, which are very likely to exist in all the larger datasets used by the data-hungry deep learning systems (Gardner et al., 2021). This problem is still there for the latest LLMs: when they fail on counterfactual tasks or adversarial perturbations, this suggests that their successes are due not to learning the general principles behind a certain operation, but some narrow heuristic that does not transfer to new contexts. (Wu et al., 2023).

In the few-shot evaluation paradigm, we also now have a new robustness problem. The art of 'prompt engineering' (Liu et al., 2023c) arose out of the prompt sensitivity phenomenon: slight variations in the phrasing of the prompt that would not make much difference to a human can lead to very different LLM output (Lu et al., 2021; Zhao et al., 2021). In a recent evaluation of 30 LLMs, Liang et al. (2022, p.12) conclude that "all models show significant sensitivity to the formatting of prompt, the particular choice of in-context examples, and the number of in-context examples across all scenarios and for all metrics". The reports of sensitivity to exact wording keep coming for the latest models, including GPT-4 (Lee et al., 2023; Gan & Mori, 2023).

3.2. Claim: (Few-shot) LLMs Are State-of-the-Art

LLM-based approaches have become the default in the current research literature, and are largely perceived to be the current SOTA (state-of-the-art) across NLP benchmarks. For example, Gillioz et al. (2020, p.179) state: "*Models like GPT and BERT relying on this Transformer architecture have fully outperformed the previous state-of-the-art networks. It surpassed the earlier approaches by such a wide*

margin that all the recent cutting edge models seem to rely on these Transformer-based architectures."

The above was written in the days of fine-tuned Transformer-based LLMs like BERT. At this point, such a statement needs to be considered in the context of the distinction between *few-shot performance* (ostensibly out-of-domain performance achieved by a model that was not specifically trained on a given task), vs performance of a *model fine-tuned for a given task*. For example, both BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) were presented with evaluation on question answering, among other tasks, but the former was fine-tuned, while the latter evaluated in a few-shot way.

Generally speaking, an ML model that has been trained on some domain data can be reasonably expected to perform better in that domain than a comparable model that hasn't received such training. This means that we now have two different notions of SOTA, where the few-shot setting could reasonably be expected to yield worse performance vs the same model if it was fine-tuned, but requires less data and training. Still, the current research papers introducing LLMs often include only few- or zero-shot evaluations, which creates the impression that this is the only evaluation that matters. For example, OPT (Zhang et al., 2022) was evaluated on 16 tasks concurrently without fine-tuning, establishing new accuracy in several of them; the same goes for models such as PaLM (Chowdhery et al., 2022), LLaMa (Touvron et al., 2023a), and many others.

We broadly agree that pre-trained models based on Transformer architecture are likely to be the current SOTA when they are provided additional in-domain supervision. But **most of the current LLM evaluation discourse shifted to few- or zero-shot evaluation, and in that context the SOTA claim may not hold. Hence, many of the current results may not actually represent the current SOTA**.

When we consider direct comparisons between few-shot LLMs and supervised systems, not based on the same LLMs, the winner depends on the specific case, but the few-shot LLM is not at all guaranteed to win – especially in the "true few-shot" setting, where prompts are not selected based on extra held-out data (Perez et al., 2021). They may also be at disadvantage in the niche domains or tasks like sequence labeling that are less straightforward to formulate as a text generation task. Consider that few-shot GPT-3 (Brown et al., 2020) is on the SuperGLUE (Wang et al., 2019) leaderboard with the average score of 71.8, compared to a score of 84.6 achieved by a fine-tuned RoBERTa model (Liu et al., 2019). As another example, recent work on NER (Wang et al., 2023) and relation extraction (Wan et al., 2023) explicitly formulated their problem as few-shot learning generally trailing behind supervised approaches in their tasks of interest, and their contribution - as overcoming that (in both cases, with the help of a supervised method in the pipeline). OpenAI (2023) claimed that GPT-4 outperforms unspecified fine-tuned models on 6 out of 7 verbal reasoning tasks, but provided no detail on model or benchmark selection, making it impossible to reproduce or verify these results.

One more consideration for the "(few-shot) LLMs are SOTA" statement is that it implies a direct competition with other methods, with which a meaningful comparison is possible. But what are we comparing – the model architectures or training data? ML as a scientific field focuses on the former, but most LLM leaderboards present apple-to-orange comparisons.

Finally, for both few-shot and fine-tuned evaluation of LLMs, **most of the reported results should be taken with a grain of salt because of test data contamination**. For example, GPT-4 received a lot of press coverage due to the claim of achieving a score that falls in the top 10% of test takers on a simulated bar exam (Katz et al., 2023) – but that result was soon questioned on grounds of improper evaluation and possible data contamination (Martínez, 2023).

In the GPT-3 report, OpenAI itself documented how hard it is to avoid benchmark contamination (Brown et al., 2020). By now, multiple studies presented evidence of the presence of common NLP benchmarks in multiple datasets used for training LLMs (Dodge et al., 2021; Magar & Schwartz, 2022; Blevins & Zettlemoyer, 2022), which can inflate LLM performance in certain tasks and datasets. The LM Contamination Index[5], a collaborative effort to document benchmark contamination, currently has 375 entries for various benchmarks and models across different tasks. Furthermore, a recent study has documented the effect where GPTs score higher on the "old" benchmarks than on the new ones (Liu et al., 2023a), which strongly suggests that the previously reported evaluation results may be inflated.

3.3. Claim: (LLM) Scale Is All You Need

Scaling has played a central role in the success of LLMs – starting with the 'scaling laws' paper for causal language models (Kaplan et al., 2020), which found that their performance improves with scaling the model size, data size, and also the amount of compute used for training. This analysis was subsequently expanded to other benchmarks, modalities and downstream tasks (Ghorbani et al., 2021; Alabdulmohsin et al., 2022; Hernandez et al., 2021; Hoffmann et al., 2022). It is often mentioned as a key factor[6] in LLM

[5] https://hitz-zentroa.github.io/lm-contamination/

[6] To be fair, the focus on scaling does not entail that other factors are completely irrelevant, and we are not saying that the entire ML community believes that "scale is all you need". But it is also fair to say that scaling has received a lot more attention than other factors, in particular data, which has long been considered as a less

performance; for instance, Huang et al. (2023b, p.1) state: "*scaling has enabled Large Language Models (LLMs) to achieve state-of-the-art performance on a range of Natural Language Processing (NLP) tasks*". The focus on scaling is in line with the "bitter lesson" of Sutton (2019), which states that we should stop working on methods based on the human knowledge of the target problem, and embrace "search and learning", because "the only thing that matters in the long run is the leveraging of computation".

Indeed, the scaling hypothesis seems to be supported by the fact that LLMs have been growing in size for several years: e.g. BERT-base in 2018 had 340M parameters (Devlin et al., 2019), and in 2022 PaLM had 540B parameters in 2022 (Chowdhery et al., 2022). And there is evidence that, even with the same architecture and training data, larger models tend to perform better, even with adversarial evaluation (Bhargava et al., 2021; Ray Choudhury et al., 2022; Wang et al., 2022b).

However, it is important to keep in mind that many of the best-known LLMs scaled both the number of parameters and their training data concurrently (besides any differences in architecture and training set-up)[7]. While this seems to support the scaling laws hypothesis as formulated by Kaplan et al. (2020)—we do not know which of these components is most responsible for the improvement, and most LLMs are not directly comparable by more than one of these criteria.

Furthermore, simply increasing the size of the training dataset has a complex relation with its quality. When we train LLMs on cleaner, more diverse text data, we do not merely provide more *data*, but more *knowledge* (by better covering the kinds of information that may be required for performing the model's task). Then, **the improved performance likely results from the fact that we are supplying more knowledge, and not just more scale/computation, and it is bounded by availability of such knowledge.** Unlike in chess and Go, for LLMs the new knowledge has to come from manually created sources, which do not cover all scenarios that may arise in practice, and hence computation can only take us as far as the data goes.[8] In practice, we break the central tenet of the "bitter lesson" all the time: when we identify an area where a model under-performs, a common solution is to try to collect, label, synthesize, or even create more data for that problem, which is tantamount to injecting manually-curated knowledge.

This would be in line with the fact that the developers of high-performing LLMs now spend more effort on improving the quality of the training data. E.g. both PaLM reports (Chowdhery et al., 2022; Anil et al., 2023) devote considerable effort to data cleaning, with PaLM 2 explicitly attributing higher performance on English benchmarks to higher quality data (Anil et al., 2023, p.9); the paper accompanying the Chinchilla model also makes a similar "quality over quantity" point with regards to data (Hoffmann et al., 2022). The Llama 3 announcement[9] stresses a heavy investment in pre-training data. The key result of Phi model (Gunasekar et al., 2023) is also explicitly presented as a smaller high-performing model, made possible by training on higher-quality data.

Furthermore, the last few years have seen an increased skepticism around the scaling hypothesis, starting with 'efficient scaling' proposals for Transformer models, which showed that smaller, more efficient models can outperform bigger ones in certain settings (Tay et al., 2021). This was further explored in practice via the Inverse Scaling Prize, showing that there are tasks, such as logical reasoning and pattern matching, where the performance does not seem to improve with model size (McKenzie et al., 2022).

Finally, let us consider the fact that there are high-performing open-access models such as LLaMa series (Touvron et al., 2023a;b), which perform very well despite being much smaller than GPT-3. The success of techniques such as knowledge distillation (Pan et al., 2020; Sanh et al., 2019) and sparsity (Srinivasan et al., 2023) also strongly suggests at least that the model size by itself is not the 'secret sauce'. And, of course, it can be a deal-breaker for deploying models in production, irrespective of performance gains.

3.4. Claim: LLMs Are General-Purpose Technologies

According to Eloundou et al. (2023, p.3), *Generative Pre-trained Transformers (GPTs) are general-purpose technologies (GPTs)*. This framing can be found both in the media (Kuttan, 2023; McKendrick, 2023) and preprints (Tamkin et al., 2021; Liu et al., 2023b).[10]

GPT (General-Purpose Technology) is a term used by economists and historians to refer to technologies that are both era-defining and pervasive over time; however, what qualifies as a GPT and what does not has been hard to delineate, since technologies are often nested within other systems of recursive technologies and systems (Knell & Vannuccini, 2022). A widely-accepted definition by economists

prestigious kind of work (Sambasivan et al., 2021).

[7]E.g. BERT was trained on roughly 3.3 billion tokens, and PaLM's training data had 780 billion, representing a growth of 1500 times in terms of model size and 260 times in terms of dataset size. This came with a big improvement in performance: BERT achieved an accuracy of 70.1% on the RTE dataset, PaLM achieves an accuracy of 95.7%, with possible data contamination.

[8]This is not to say that such a system cannot be useful in some cases, or that it cannot exhibit some generalization within a space sufficiently covered by data: the linguistic fluency of the current LLMs is a testament to the possibility of learning "closed" systems such as syntax.

[9]https://ai.meta.com/blog/meta-llama-3/

[10]The preprints by Eloundou et al. (2023); Liu et al. (2023b) were co-authored by researchers affiliated with OpenAI.

Lipsey and Carlaw proposes 4 criteria for a technology to be considered general-purpose: (1) it is a single, recognisable generic technology, (2) it comes to be widely used across the economy, (3) it has many different uses and (4) it creates many spillover effects (Lipsey et al., 2005). According to these criteria, a total of 24 technologies such as the wheel, the printing press and electricity are considered GPTs.

At this point, it is hard to assert the general-purpose status of LLMs by the above criteria. They are not a single, generic technology (as discussed in §2). They are currently not widely used in different domains (Bekar et al., 2018; Prytkova, 2021), and remain auxiliary tools even in those domains where they are most used (Bianchini et al., 2020). In terms of it usage, LLMs are not widely used in the vast majority of economic activities, and they are reliant upon specific commodities such as large amounts of GPUs and highly specialized labor, which are only available in a small number of industries and by a handful of organizations (Bresnahan, 2019). A further constraint is the fact that LLMs, like all deep learning-based technology, can be expected to work best in-distribution – and it is unclear to what extent issues with robustness (§3.1) will limit its broader utility in the vast number of economic sectors associated with the smaller communities and languages.

As for the final criterion, regarding the potential spillover effects of these technologies, it is really too early to tell what these may be (Bresnahan, 2019; Crafts, 2021; Natale & Ballatore, 2020). The overnight popularity of ChatGPT, which is often seen as proof of the widespread usage of LLMs in broader society, can mainly be attributed to the creation of a simple user interface on top of models that had been previously available via APIs (Eloundou et al., 2023), rather than technological novelty. **Until we can meaningfully assess the adoption and application of LLMs, we should refrain from putting them in the same conceptual category as the printing press**, and focus on defining what they can and cannot be used for (and under what conditions).

3.5. Claim: LLMs Exhibit "Emergent Properties"

LLMs are often discussed in terms of their "emergent properties".[11] Among such properties, various researchers have included few-shot learning (Bommasani et al., 2021), "augmented prompting" techniques such as "chain of thought" (Wei et al., 2022b), predicting intermediate computation results (Nye et al., 2021) or whether the answer is correct (Kadavath et al., 2022), and, on the input side – instruction following (Ouyang et al., 2022; Wei et al., 2021).

In line with the confusion about the term "LLM" (§2), the term "emergent properties" seems to be used in at least 4 distinct ways in current LLM research:

Definition 3.1. A property that a model exhibits despite not being explicitly trained for it. E.g. Bommasani et al. (2021, p.5) refers to few-shot performance of GPT-3 (Brown et al., 2020) as "an emergent property that was neither specifically trained for nor anticipated to arise".

Definition 3.2. (Opposite to Definition 3.1): a property that the model learned from the pre-training data. E.g. Deshpande et al. (2023, p.8) discuss emergence as evidence of "the advantages of pre-training".

Definition 3.3. A property "is emergent if it is not present in smaller models but is present in larger models." (Wei et al., 2022a, p.2).

Definition 3.4. A version of Definition 3.3, where what makes emergent properties "intriguing" is "their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales" (Schaeffer et al., 2023, p.1)

Definition 3.2 seems describe the expected outcome of successful training. In this sense, "emergent properties" could be referred to simply as "learned properties".

Definition 3.3 is questionable if the "emergent" behavior can be achieved in smaller models[12]. Even more importantly, we still have the contamination problem: if both the bigger and smaller models were trained on data similar to the test data, the bigger one would still be reasonably expected to perform better just because it has more capacity to learn it. In that case, Definition 3.3 follows from the Definition 3.2.

With respect to Definition 3.4, Schaeffer et al. (2023) make a convincing case that such sharp increases in performance may be an artifact of the chosen evaluation metric[13] rather than a fundamental property of scaling the model.

[11] Some researchers discuss "emergent abilities" rather than "properties". Both of these terms could also use better definitions. Our interpretation is that they are used interchangeably in LLM research, and the chief difference is the anthropomorphizing framing for "ability". The underlying construct in case of Wei et al. (2022a) seems to be NLP "tasks", on the assumptions that these tasks have construct validity, and specific evaluation datasets provide valid measurements of performance on these tasks. The task/benchmark confusion is exacerbated by the fact that in BIG-Bench (Srivastava et al., 2023) the constituent datasets are sometimes named as if they were tasks (e.g. "IPA transliterate").

[12] See e.g. Schick & Schütze (2021); Gao et al. (2021). One could object that these studies provide the smaller models a lot of extra help (reformatting the inputs, selecting extra models etc.) This is true, but the commonly reported few-shot performance is also not really few-shot, and reliably choosing good prompts becomes *harder* with larger models (presumably due to more complex decision boundaries) (Perez et al., 2021).

[13] Wei (2023) and Anderljung et al. (2023, p.38) argue that this is not important, because "non-smooth" metrics used for real tasks are the ones we care about. But then the question remains whether this is something special about LLMs, or just a property of the metric. Simple models also have such "emergent properties", as shown by Schaeffer et al. (2023) for a shallow nonlinear autoencoder.

An even bigger issue with Definition 3.4 is that we simply do not have enough data points to say that the increase in performance is sharp: e.g. if we had intermediate model sizes between the commonly-used 13B, 70B, and 150+B, we would likely see a smooth transition. Wei (2023) acknowledges that, but argues that the "emergence" phenomenon is still interesting if there are large differences in predictability: for some problems, performance of large models can easily be extrapolated from performance of models 1000x less in size, whereas for others, even it cannot be extrapolated even from 2x less size. But the cited predictability at 1,000x less compute refers to the GPT-4 report (OpenAI, 2023), where *the developers knew the target evaluation in advance*, and specifically optimized for "predictable scaling". This is in contrast with the unpredictability at 2x less compute for *unplanned* BIG-Bench evaluation by Wei et al. (2022a).

So, we are left with Definition 3.1, which can be interpreted in two ways:

Definition 3.5. A property is emergent if the model was not exposed to training data for that property.

Definition 3.6. A property is emergent even if the model was explicitly trained for it – as long as the developers were unaware of it.

Per Definition 3.6, it would appear that we are training LLMs as a very expensive method to discover what data exists on the Web. For example, the fact that ChatGPT can generate chess moves that are plausible-looking (but often illegal)[14] is to be expected, given the vast amount of publicly-available chess transcripts on the Web.

Per Definition 3.5, we can prove that some property is emergent only by showing that there was no evidence that could have been the basis for the model outputs in the training data. For commercial models with undisclosed data such as ChatGPT, this is out of the question. But we would go further and argue that the emergent properties according to Definition 3.5 are only a hypothesis (if not wishful thinking) even for the "open" LLMs, because so far we are lacking detailed studies (or even a methodology) to consider the exact relation between the amount and kinds of evidence in the training text data for a particular model output. Hence, to the best of our knowledge, **there is no evidence for the existence of "emergent properties" per Definition 3.5**. Until we have such evidence, it seems strange for the ML community to conclude that the best explanation for high performance is not-learned-from-data emergent properties—especially in the face of evidence to the contrary.[15]

What about benchmarks that are newly created and tested on for the first time in a given study – aren't they, by definition, uncontaminated? We argue not: in the absence of a methodology to even compare test and training data beyond trivial exact matches, "new" tests may be so similar to something observed that they do not really count as "new". This has been a known problem even with benchmark datasets, the size of which is very small compared to LLM training data[16]. The LLM training data may also include something that is not even public[17] on the internet, and thus not easily searchable to confirm the contamination.

Perhaps the most striking example of how unreliable the "new" test examples are is the "sparks of intelligence" study (Bubeck et al., 2023). Using the methodology of newly constructed test cases, checked against public web data, and their perturbations, Bubeck et al. (2023) notably concluded that GPT-4 possesses "a very advanced theory of mind". At least two studies have since come to the opposite conclusion (Sap et al., 2023; Shapira et al., 2023).

The above discussion focused on the way the term "emergence" is currently used in LLM research. In philosophy of science, there are many nuanced discussions of emergence to which we cannot do justice here, but broadly it can be characterized as a phenomenon in complex systems that is dependent on its constituent parts, but is also distinct from them. The leading example in Stanford Encyclopedia of Philosophy is a tornado: it consists of dust and debris, but it "its features and behaviors appear to differ in kind from those of its most basic constituents"(O'Connor & Wong, 2020). Emergent phenomena are described by a different scientific field, on a different level than the constituent phenomena: it is possible to understand the behavior of tornadoes without understanding particle physics. Does this notion of emergence apply in the context of ML?

Our take is that such a notion of emergence would hold for LLMs: the information they encode is not explainable by the model weights, at least at the current state of interpretability research. However, this also applies to most deep learning models (e.g. a simple MLP for sentiment classification), and the latest LLMs are not something special.

[14] https://reddit.com/r/AnarchyChess/comments/10ydnbb/i_placed_stockfish_white_against_chatgpt_black/

[15] E.g. Liang et al. (2022, p.12), evaluate 30 LLMs and conclude that "regurgitation (of copyrighted materials) risk clearly correlates with model accuracy". Liu et al. (2023a) report that ChatGPT and GPT-4 perform better on older compared to newly released benchmarks, and McCoy et al. (2023) show that their performance depends on probabilities of output word sequences in web texts. Lu et al. (2023) show that "emergent abilities" of 18 LLMs can be ascribed mostly to in-context learning. For in-context learning itself, the results of Chan et al. (2022) suggest that it happens only in Transformers trained on sequences, structurally similar to the sequences in which in-context learning would be tested.

[16] E.g. Lewis et al. (2021) found that about 30% of test samples in 3 popular QA benchmarks have near-duplicates in train data.

[17] In particular, the OpenAI models could have been trained not only on Web data, but also on the data of thousands of researchers who over the past year submitted their trickiest test cases to GPT-3 API (the policy for the API data to be opted out of training by default only changed in Spring 2023).

4. How Does This Change the Theory and Practice of ML?

The success of LLMs brought our field an increase in funding, real-world applications, and attention. But the perception of their robustness (§3.1) and SOTA status (§3.2), the idea that their success is purely due to scale (§3.3), that they are a general-purpose technology (§3.4), and their ill-defined "emergent properties" (§3.5) also contribute to the following trends:

- *Homogeneity of approaches.* If LLMs are so great, why would we pursue anything else? It is understandable that most current NLP research is focused on LLMs, but this also means that as researchers we have most of our eggs in one basket, and become less likely to develop alternatives or see the gaps in our knowledge.
- *De-democratization.* Knight (2023) estimates that training GPT-4 cost more than $100 million; although the true number is unknown, the increase in computational requirements of LLMs is clear (Thompson et al., 2022). This means that graduate students, independent researchers, and even most academic labs struggle to either reproduce existing results or train new LLMs that would require large amounts of compute.
- *Industry influence.* Recent research into the affiliations of researchers in ML (Abdalla & Abdalla, 2021; Ahmed & Wahed, 2020; Birhane et al., 2022; Whittaker, 2021) has shown both a steep increase in industry presence in conference publications over the past years (e.g. a 180% growth for NLP (Abdalla et al., 2023)), and an influence over the topics of research being pursued by the community, such as robustness and similar challenges, instead of more theory-oriented topics. Given the above point, this trend makes sense, since only industry labs with extensive funding can afford to train and deploy LLMs. But it has implications for the diversity of ideas and approaches in the field.
- *(Further) decreased reproducibility.* Reproducibility was an issue in ML even before LLMs (Crane, 2018; Cohen et al., 2018; Bouthillier et al., 2019), and it has not gotten much better (Belz et al., 2021; 2022; Belz, 2022). LLMs pose additional issues both in reproducing their training and fine-tuning (Sellam et al., 2021; McCoy et al., 2020) and inference results (Hagmann et al., 2023), not to mention the lack of control over API-only models that can change over time (Chen et al., 2023; Rogers, 2023). If the tested model is deprecated, or changes in any way (e.g. by extra training or fine-tuning, changes in its associated parameters, filtering mechanisms or system prompts), the results reported in the study will no longer be reproducible. They might not even hold at the time of publication – and we will have no idea why.

The ideas that LLMs are robust (§3.1) and exhibit emergent properties (§3.5) further contribute to the impression of irrelevance of any theory of the linguistic, social, cognitive, or any other phenomena that LLMs are supposed to model. This is unsatisfactory if the goal is scientific research, but even from a purely engineering perspective this stance is dangerous, since it entails that we either cannot or do not need to provide specifications for cases where a given model is safe or unsafe to use. As a result, LLMs may be deployed in society at large, without compelling evidence of their performance in all the target scenarios, or among the demographic groups that are likely underrepresented in the training data (Bender et al., 2021), or across the less-resourced languages (Zhu et al., 2023; Bang et al., 2023; Lai et al., 2023; Huang et al., 2023a; Ziems et al., 2023).

5. Ways Forward

We have argued that there are many deep knowledge gaps in LLM research, and the field is changing in ways that make these gaps less likely to be addressed. We now conclude with some concrete recommendations for future work.

Maintaining diversity of research approaches. We advocate not for stopping all work on LLMs, but for maintaining a healthy diversity of approaches and tasks. Among other things, this means efforts by the conferences to ensure fair reviews for the "niche" submissions.[18] Putting all of our eggs in the proverbial LLM basket runs the risk of missing out on new research directions and exciting opportunities to make significant connections with other fields, such as linguistics and cognitive science.

Defining terminology. We discussed the lack of clarity on the very term "large language model" (§2), as well as "emergent properties" (§3.5). Ideally, at least in research papers we would start with specifying what we mean by a common-but-vague term – e.g. "vision-and-language Transformer model" or "a Transformer-based autoregressive language model with 200 billion parameters". Although it is impossible to control what terminology is used in popular science or journalism about LLMs, as experts and researchers we have a responsibility to be clear and precise when discussing our domain of expertise (see LaCroix & Prince (2023)).

Not using "closed" models as baselines. Since the launch of GPT-4 in March 2023, numerous studies have used it both as a benchmark to compare different methods and models (e.g. Liu et al., 2023a; Sun et al., 2023) as well as an object of scientific study in itself (e.g. Bubeck et al., 2023; Wei et al., 2022a; Zhang et al., 2023). We believe that this is problematic for several reasons:

[18] E.g. papers on "niche" topics can have priority in reviewer assignments (Rogers et al., 2023b).

- It may produce inflated performance reports due to undisclosed training data, which could be contaminated with benchmark data. This could then create the false impression that the "closed" model is intrinsically so much better that there is no competing with it.
- It normalizes methodologically dubious apples-to-oranges comparisons to black box models.
- The reproducibility of evaluations is significantly reduced (see §4).
- Where the "closed" models are provided commercially, academic researchers essentially perform free work to help the company improve their product – and they even pay[19] for that privilege, often with funding from public grants that could be spent on improving the public resources. Moreover, if the general perception is that only this kind of work constitutes "frontier" LLM research, the labs without the financial means to pay for the API access are at severe disadvantage, and the field gets homogenized even further.

While we recognize that some analyses of proprietary, "closed" models like GPT-4 and PaLM can be useful (e.g. audits, red-teaming), relying on these models as baselines, and especially expecting others to do so, is both unfair and unreliable. We recommend using open-source or at least open-access[20] models like FLAN-T5 (Chung et al., 2022), BLOOM (Scao et al., 2022), LLaMa (Touvron et al., 2023a;b) and FALCON (Almazrouei et al., 2023).

Further rigorous studies of LLM functionality. There are numerous knowledge gaps about LLM fuctionality. What exactly makes a given type of model (e.g. a Transformer-based LLM) SOTA on a given task, given that there are no confounds such as differences in training data and model capacity? What kinds of brittleness do LLMs exhibit, and how to mitigate that? Do they really have any emergent properties (and how are these defined)? How can we ensure specific kinds of robustness? Large-scale benchmarks (e.g. Srivastava et al., 2023) do not address these questions, since they are constructed on the basis of data availability rather than definitions of specific phenomena (Raji et al., 2021). These questions are not answered by the current large-scale evaluation endeavors (e.g. Liang et al., 2022). While they are very valuable, there are too many confounding variables in the published models, and more extensive error analysis and exploration is needed to disentangle them. Access to both models and their training data is especially important for this purpose, since it can help establish links between LLM behavior and their training data (Piktus et al., 2023) and understand their biases (Gururangan et al., 2022; Johnson et al., 2022; Abid et al., 2021).

Developing better evaluation methodology. There has been progress towards multi-dimensional evaluation that takes into account more than performance (Ethayarajh & Jurafsky, 2020; Liang et al., 2022; Chung et al., 2023) and reproducibility (Dodge et al., 2019; Ulmer et al., 2022; Magnusson et al., 2023), but much more remains to be done. There is also a dire need for structural incentives to encourage reproducibility efforts and publication of negative results. And most importantly, we need a lot more work on the validity of the underlying constructs (Raji et al., 2021; Schlangen, 2021). Ideally, LLM evaluation would target specific types of language processing rather than broadly defined tasks (Ribeiro et al., 2020; Rogers et al., 2023a). It would carefully explore the decision boundaries and specific kinds of brittleness (Kaushik et al., 2019; Gardner et al., 2020), and control for potential confounds so as to focus purely on model architecture. As discussed in §3.5, we also need new methodology for systematically linking the model outputs to potential evidence in the training data at LLM scale – relevant current efforts include the work on memorization (Thakkar et al., 2021; Carlini et al., 2022; Chang et al., 2023, *inter alia*) and providing search indices for LLM training data (Piktus et al., 2023; Marone & Van Durme, 2023). All this is in parallel to the general problems with evaluating open-ended generation, including the dependence on generation strategies (Meister & Cotterell, 2021), and issues with popular metrics such as perplexity (Wang et al., 2022a).

6. Conclusion

As researchers, we cannot help but observe the near-universal focus on LLMs in recent years and their impact on our field. This paper discusses several knowledge gaps and common claims that should be taken with a grain of salt, as well as the ways in which LLMs changed the research landscape. We argue for more rigor in definitions, experimental studies, and evaluation methodology, as well as higher standards for transparency and reproducibility. We hope to open the door for more discussion of the contribution of LLMs to ML research, and how we can better leverage their strengths while understanding their limitations.

[19] According to one US academic researcher we spoke to, the monthly spending on OpenAI API in their lab is capped at $10K a month, and otherwise would sometimes even exceed that sum.

[20] We do not advocate for using *only* open-source models, because the models available now can be "open" in different relevant ways (code, weights, license, training data), and these aspects matter more/less for different research questions and real-world applications (Solaiman, 2023). A model like Mistral (Jiang et al., 2023) is more transparent than GPT-4 in terms of its architecture, but not the training data. A model like BLOOM (Le Scao et al., 2023) is more transparent in terms of training data and can be an excellent object of research, but it is not strictly open-source, e.g. because of its RAIL license. Our position is that in research, we should aim to use the most transparent options available, and thankfully there seems to be a trend towards more openness in this sense (e.g. the recent OLMO model (Groeneveld et al., 2024) and its training dataset DOLMa (Soldaini et al., 2024)).

Acknowledgements

We would like to thank all anonymous reviewers of this paper. Their insightful comments were invaluable for sharpening the discussion. Rob van der Goot, Christian Hardmeier, Yacine Jernite, Margaret Mitchell, Dennis Ulmer read the early versions of this paper and provided feedback. We also thank Ryan Cotterell, Ishita Dasgupta, Laura Gwilliams, Julia Haas, Anna Ivanova, Tal Linzen, Ben Lipkin, Asad Sayeed for their insights and discussion.

This work was partly supported by DFF Inge Lehmann grant to Anna Rogers (3160-00022B).

Impact Statement

This work is a position paper, expressing the personal views of its authors, and supported by evidence and arguments we cite rather than new experimental work. Our goal is not to criticize a specific research direction or technology as a whole, but to raise the awareness about the issues with the lack of clarity with key terms and claims in our field.

We recognize that we come from a position of privilege, holding positions at institutions located in the Global North. Our opinions do not reflect the lived experiences of visible minorities or those who come from less privileged institutions and geographical regions. We have endeavored to have our manuscript read by colleagues from other domains of expertise and other institutions, but we recognize that it does not represent the experiences and perspectives of all members of the ML community.

References

Abdalla, M. and Abdalla, M. The Grey Hoodie Project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 287–297, 2021.

Abdalla, M., Wahle, J. P., Ruas, T., Névéol, A., Ducel, F., Mohammad, S. M., and Fort, K. The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research. *arXiv preprint arXiv:2305.02797*, 2023.

Abid, A., Farooqi, M., and Zou, J. Persistent anti-Muslim bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Ahmed, N. and Wahed, M. The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.

Alabdulmohsin, I. M., Neyshabur, B., and Zhai, X. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. *arxiv*, 2023.

Alyafeai, Z., AlShaibani, M. S., and Ahmad, I. A Survey on Transfer Learning in Natural Language Processing. *arXiv:2007.04239 [cs, stat]*, May 2020.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier AI Regulation: Managing Emerging Risks to Public Safety, November 2023.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. PaLM 2 Technical Report, 2023.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, 2023.

Bekar, C., Carlaw, K., and Lipsey, R. General purpose technologies in theory, application and controversy: a review. *Journal of Evolutionary Economics*, 28:1005–1033, 2018.

Beltagy, I., Lo, K., and Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Belz, A. A metrological perspective on reproducibility in NLP. *Computational Linguistics*, 48(4):1125–1135, 2022.

Belz, A., Agarwal, S., Shimorina, A., and Reiter, E. A systematic review of reproducibility research in Natural Language Processing. *arXiv preprint arXiv:2103.07929*, 2021.

Belz, A., Popović, M., and Mille, S. Quantified reproducibility assessment of nlp results. *arXiv preprint arXiv:2204.05961*, 2022.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Bhargava, P., Drozd, A., and Rogers, A. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pp. 125–135, Online and Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics.

Bianchini, S., Müller, M., and Pelletier, P. Deep learning in science. *arXiv preprint arXiv:2009.01575*, 2020.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184, 2022.

Blevins, T. and Zettlemoyer, L. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3563–3574, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bouthillier, X., Laurent, C., and Vincent, P. Unreproducible research is reproducible. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 725–734. PMLR, 09–15 Jun 2019.

Branco, R., Branco, A., Rodrigues, J., and Silva, J. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1504–1521, 2021.

Bresnahan, T. Artificial intelligence technologies and aggregate growth prospects. *Prospects for Economic Growth in the United States*, 2019.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, June 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. Data Distributional Properties Drive Emergent In-Context Learning in Transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, December 2022.

Chang, K. K., Cramer, M., Soni, S., and Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4, 2023.

Chang, T. A. and Bergen, B. K. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*, 2023.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. Technical report, Google, 2013. URL http://arxiv.org/abs/1312.3005.

Chen, L., Zaharia, M., and Zou, J. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

Chollet, F. On the Measure of Intelligence. *arXiv:1911.01547 [cs]*, November 2019.

Choudhury, S. R., Rogers, A., and Augenstein, I. Machine Reading, Fast and Slow: When Do Models" Understand" Language? *arXiv preprint arXiv:2209.07430*, 2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling Instruction-Finetuned Language Models, 2022.

Chung, J.-W., Liu, J., Wu, Z., Xia, Y., and Chowdhury, M. ML.ENERGY leaderboard, 2023.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, 2020.

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C., and Hunter, L. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Crafts, N. Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy*, 37(3), 2021.

Crane, M. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252, 2018. doi: 10.1162/tacl_a_00018.

Deshpande, V., Pechi, D., Thatte, S., Lialin, V., and Rumshisky, A. Honey, I shrunk the language: Language model behavior at reduced scale. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5298–5314, Toronto, Canada, July 2023. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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. 4171–4186, June 2019.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, 2021.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

EMNLP. Emnlp 2023 call for main conference papers theme track: Large language models and the future of nlp, 2023.

Ethayarajh, K. and Jurafsky, D. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, Online, November 2020. Association for Computational Linguistics.

Gan, C. and Mori, T. Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks, June 2023.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., Singh, S., Smith, N. A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., and Zhou, B. Evaluating NLP Models via Contrast Sets. April 2020.

Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., and Smith, N. A. Competency problems: On finding and removing artifacts in language data. *arXiv preprint arXiv:2104.08646*, 2021.

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., and Cherry, C. Scaling

laws for neural machine translation. *arXiv preprint arXiv:2109.07740*, 2021.

Gillioz, A., Casas, J., Mugellini, E., and Abou Khaled, O. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 179–183. IEEE, 2020.

Graber-Stiehl, I. Is the world ready for ChatGPT therapists? *Nature*, 617(7959):22–24, 2023.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, 2023.

Gururangan, S., Card, D., Dreier, S. K., Gade, E. K., Wang, L. Z., Wang, Z., Zettlemoyer, L., and Smith, N. A. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*, 2022.

Hagmann, M., Meier, P., and Riezler, S. Towards inferential reproducibility of machine learning research, 2023.

Halliday, M. and Hasan, R. *Cohesion in English*. Routledge, London, 0 edition, 2013. ISBN 978-1-317-86960-3. doi: 10.4324/9781315836010.

Harrer, S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, 2023.

Heaven, D. et al. Why deep-learning AIs are so easy to fool. *Nature*, 574(7777):163–166, 2019.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling Laws for Transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models, 2022.

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., and Wei, F. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting, 2023a.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., and Bertulfo, D. J. The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*, 2022.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing (3rd Ed. Draft)*. February 2024. URL https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language Models (Mostly) Know What They Know, nov 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. ChatGPT for good? on

opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274, 2023.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. GPT-4 Passes the Bar Exam, March 2023.

Kaushik, D., Hovy, E., and Lipton, Z. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*, September 2019.

Knell, M. and Vannuccini, S. Tools and concepts for understanding disruptive technological change after schumpeter. Technical report, Jena Economic Research Papers, 2022.

Knight, W. OpenAI's CEO Says the Age of Giant AI Models Is Already Over. *https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over*, 2023.

Kuttan, J. How Enterprises Can Become Ready To Work With LLMs. *Forbes*, 2023.

LaCroix, T. and Prince, S. J. D. Ethics and deep learning, 2023.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, 2023.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. BLOOM: A 176b-parameter open-access multilingual language model. 2023.

Lee, P., Bubeck, S., and Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13):1233–1239, March 2023. ISSN 0028-4793. doi: 10.1056/NEJMsr2214184.

Lenat, D. and Feigenbaum, E. A. *On the thresholds of knowledge*. MIT Press Cambridge, MA, 1981.

Lewis, P., Stenetorp, P., and Riedel, S. Question and answer test-train overlap in open-domain question answering datasets. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1000–1008, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.86.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., and Petrov, S. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pp. 169–174, 2012.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. *Economic transformations: general purpose technologies and long-term economic growth*. Oup Oxford, 2005.

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., and Zhang, Y. Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*, 2023a.

Liu, J., Xu, X., Li, Y., and Tan, Y. "generate" the future of work through AI: Empirical evidence from online labor markets. *arXiv preprint arXiv:2308.05201*, 2023b.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023c.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, jul 2019.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., and Gurevych, I. Are emergent abilities in large language models just in-context learning?, 2023.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Magar, I. and Schwartz, R. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*, 2022.

Magnusson, I., Smith, N. A., and Dodge, J. Reproducibility in NLP: What have we learned from the checklist? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12789–12811, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.809.

Marone, M. and Van Durme, B. Data Portraits: Recording Foundation Model Training Data, March 2023.

Martínez, E. Re-Evaluating GPT-4's Bar Exam Performance, May 2023.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in Natural Language Inference. *arXiv preprint arXiv:1902.01007*, 2019.

McCoy, R. T., Min, J., and Linzen, T. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Online, 2020. Association for Computational Linguistics.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. Embers of autoregression: Understanding large language models through the problem they are trained to solve, 2023.

McKelvey, F., Dandurand, G., and Roberge, J. News coverage of artificial intelligence reflects business and government hype — not critical voices. *The Conversation*, April 2023.

McKendrick, J. Why GPT Should Stand For 'General Purpose Technology' For All. *Forbes*, 2023.

McKenzie, I., Lyzhov, A., Parrish, A., Prabhu, A., Mueller, A., Kim, N., Bowman, S., and Perez, E. The Inverse Scaling Prize, 2022.

Meister, C. and Cotterell, R. Language Model Evaluation Beyond Perplexity. *arXiv preprint arXiv:2106.00085*, 2021.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2013.

Nalbandian, L. An eye for an 'i:'a critical assessment of artificial intelligence tools in migration and asylum management. *Comparative Migration Studies*, 10(1):1–23, 2022.

Natale, S. and Ballatore, A. Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*, 26(1):3–18, 2020.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. Show Your Work: Scratchpads for Intermediate Computation with Language Models. oct 2021.

O'Connor, T. and Wong, H. Y. Emergent Properties. *Stanford Encyclopedia of Philosophy*, August 2020.

OpenAI. GPT-4 Technical Report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, mar 2022.

Pan, H., Wang, C., Qiu, M., Zhang, Y., Li, Y., and Huang, J. Meta-KD: A meta knowledge distillation framework for language model compression across domains. *arXiv preprint arXiv:2012.01266*, 2020.

Pan, S. J. and Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1558-2191. doi: 10.1109/TKDE.2009.191.

Perez, E., Kiela, D., and Cho, K. True Few-Shot Learning with Language Models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11054–11070. Curran Associates, Inc., 2021.

Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, A. S., Jernite, Y., and Rogers, A. The ROOTS Search Tool: Data Transparency for LLMs. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023.

Prytkova, E. ICT's Wide Web: a System-Level Analysis of ICT's Industrial Diffusion with Algorithmic Links. *Available at SSRN 3772429*, 2021.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 Pre-Proceedings (NeurIPS Datasets and Benchmarks 2021)*, August 2021.

Ramponi, A. and Plank, B. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6838–6855, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.603.

Ray Choudhury, S., Rogers, A., and Augenstein, I. Machine Reading, Fast and Slow: When Do Models "Understand" Language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 78–93, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics.

Rogers, A. Closed AI Models Make Bad Baselines. *Hacking semantics*, April 2023.

Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8722–8731, 2020.

Rogers, A., Gardner, M., and Augenstein, I. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, 55(10):197:1–197:45, February 2023a. ISSN 0360-0300. doi: 10.1145/3560260.

Rogers, A., Karpinska, M., Boyd-Graber, J., and Okazaki, N. Program chairs' report on peer review at ACL 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. xl–lxxv, Toronto, Canada, July 2023b. Association for Computational Linguistics.

Roose, K. How does ChatGPT really work?, 2023.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–15, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445518.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, 2023.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Schaeffer, R., Miranda, B., and Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Schick, T. and Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2339–2352, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.185.

Schlangen, D. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 670–674, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.85.

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D., Tenney, I., and Pavlick, E. The MultiBERTs: BERT Reproductions for Robustness Analysis. *arXiv:2106.16163 [cs]*, jun 2021.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models, May 2023.

Solaiman, I. The gradient of generative AIrelease: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 111–122, 2023.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Srinivasan, V., Gandhi, D., Thakker, U., and Prabhakar, R. Training Large Language Models Efficiently with Sparsity and Dataflow. *arXiv preprint arXiv:2304.05511*, 2023.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B.,

Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, 2023.

Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., and Ren, Z. Is ChatGPT good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*, 2023.

Sutton, R. The bitter lesson (blog post). http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019.

Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*, 2021.

Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H. W., Narang, S., Yogatama, D., Vaswani, A., and Metzler, D. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic. Galactica: A Large Language Model for Science, 2022.

Thakkar, O. D., Ramaswamy, S., Mathews, R., and Beaufays, F. Understanding unintended memorization in language models under federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pp. 1–10, 2021.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. The computational limits of deep learning, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ulmer, D., Bassignana, E., Müller-Eberstein, M., Varab, D., Zhang, M., van der Goot, R., Hardmeier, C., and Plank, B. Experimental standards for deep learning in natural language processing research. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2673–2692, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.196.

Veres, C. Large Language Models are Not Models of Natural Language: They are Corpus Models. *IEEE Access*, 10:61970–61979, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3182505.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., and Kurohashi, S. GPT-RE: In-context learning for relation extraction using large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3534–3547, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.214.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:1905.00537 [cs]*, may 2019.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. GPT-NER: Named Entity Recognition via Large Language Models, 2023.

Wang, Y., Deng, J., Sun, A., and Meng, X. Perplexity from PLM Is Unreliable for Evaluating Text Quality. *arXiv preprint arXiv:2210.05892*, 2022a.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A, S., Patro, S., Dixit, T., and Shen, X. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics.

Wei, J. Common arguments regarding emergent abilities, May 2023. URL https://www.jasonwei.net/blog/common-arguments-regarding-emergent-abilities.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*, oct 2021.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022b.

Whittaker, M. The steep cost of capture. *Interactions*, 28 (6):50–55, 2021.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2023.

Zakrzewski, C. e. a. OpenAI CEO tells Senate that he fears AI's potential to manipulate views, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., Tyser, K., Chin, Z., Hicke, Y., Singh, N., Udell, M., Kim, Y., Buonassisi, T., Solar-Lezama, A., and Drori, I. Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models, 2023.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis, 2023.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1):43–76, January 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2020.3004555.

Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., and Yang, D. Multi-VALUE: A Framework for Cross-Dialectal English NLP, 2023.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023.