Eight Things to Know about Large Language Models

Samuel R. Bowman [1] [2]

Abstract

The widespread public deployment of large language models (LLMs) in recent months has prompted a wave of new attention and engagement from advocates, policymakers, and scholars from many fields. This attention is a timely response to the many urgent questions that this technology raises, but it can sometimes miss important considerations. This paper surveys the evidence for eight potentially surprising such points:

1. LLMs predictably get more capable with increasing investment, even without targeted innovation.
2. Many important LLM behaviors emerge unpredictably as a byproduct of increasing investment.
3. LLMs often appear to learn and use representations of the outside world.
4. There are no reliable techniques for steering the behavior of LLMs.
5. Experts are not yet able to interpret the inner workings of LLMs.
6. Human performance on a task isn't an upper bound on LLM performance.
7. LLMs need not express the values of their creators nor the values encoded in web text.
8. Brief interactions with LLMs are often misleading.

Introduction

Large language models (LLMs, e.g. GPT-3, PALM, LLaMA, and GPT-4; Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI, 2023b) and products built on them, such as ChatGPT, have recently prompted an enormous amount of attention from journalists, (Klein, 2023; Perrigo, 2023; Oliver, 2023), policymakers (J & C, 2023; Bartz, 2023; Lieu, 2023), and scholars from many fields (Chan, 2022; Lund & Wang, 2023; Choi et al., 2023; Biswas, 2023). This technology defies expectations in many ways, though, and it can be easy for brief discussions of it to leave out important points.

This paper presents eight potentially surprising claims that I expect will be salient in at least some of the conversations that are springing up around LLMs. They reflect, to the best of my understanding, views that are reasonably widely shared among the researchers—largely based in private labs—who have been developing these models. All the evidence I present here, as well as most of the arguments, are collected from prior work, and I encourage anyone who finds these claims useful to consult (and directly cite) the sources named here.

I do not mean for these claims to be normative in any significant way. Rather, this work is motivated by the recognition that deciding what we should do in light of this disruptive new technology is a question that is best led—in an informed way—by scholars, advocates, and lawmakers from outside the core technical R&D community.

1. LLMs predictably get more capable with increasing investment, even without targeted innovation

Scaling law results (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022) have been a major driving factor in the recent surge of research and investment into LLMs (Ganguli et al., 2022a). Scaling laws allow us to precisely predict some coarse-but-useful measures of how capable future models will be as we scale them up along three dimensions: the amount of data they are fed, their size (measured in parameters), and the amount of computation used to train them (measured in FLOPs). These results thereby allow us to make some key design decisions, such as the optimal size of a model given some fixed resource budget, without extremely expensive trial and error.

Our ability to make this kind of precise prediction is unusual in the history of software and unusual even in the history of modern AI research. It is also a powerful tool for driving investment since it allows R&D teams to propose model-training projects costing many millions of dollars, with reasonable confidence that these projects will succeed

[1] New York University [2] Anthropic, PBC. Correspondence to: Samuel R. Bowman <bowman@nyu.edu>.

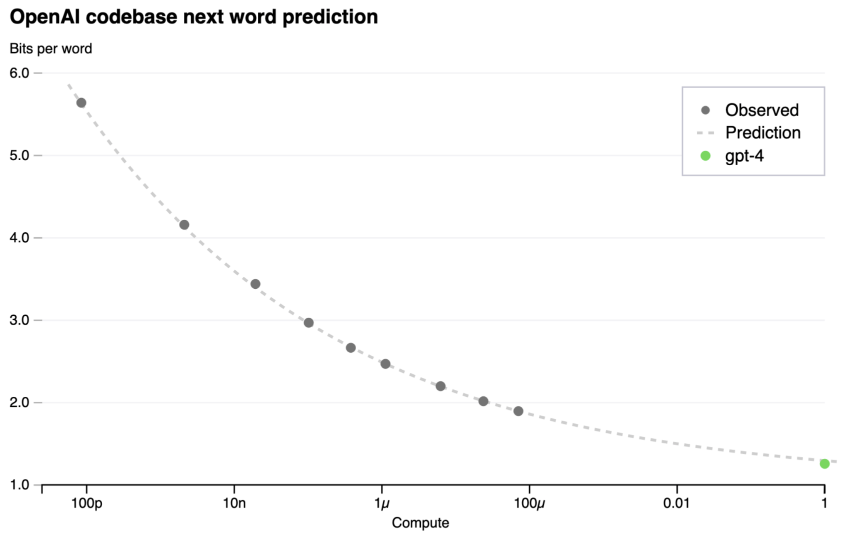

Figure 1. Excerpted from OpenAI (2023b): A scaling law result for one measure of language model performance, showing a consistent trend as the amount of computation used to train a model is scaled up 10,000,000,000× times from a small prototype system to GPT-4.

at producing economically valuable systems.

Concretely, consider these three superficially very different systems: OpenAI's original GPT can perform simple text-labeling tasks but cannot generally produce coherent text (Radford et al., 2018). GPT-2 adds the ability to produce text of reasonably high quality, as well as a limited ability to follow simple instructions (Radford et al., 2019). GPT-3 is the first modern general-purpose LLM, and is practically useful across a wide range of language tasks. The designs of these three models hardly differ at all. Instead, the qualitative differences between them stem from vast differences in scale: Training GPT-3 used roughly 20,000× more computation than training the original GPT (Sevilla et al., 2022), as well as significantly more data and parameters. There *are* substantial innovations that distinguish these three models, but they are almost entirely restricted to infrastructural innovations in high-performance computing rather than model-design work that is specific to language technology.

While the techniques used to train the newest LLMs are no longer generally disclosed, the most recent detailed reports suggest that there have been only slight deviations from this trend, and that designs of these systems are still largely unchanged (Chowdhery et al., 2022; Hoffmann et al., 2022; Touvron et al., 2023).

Continuing to scale these techniques up beyond GPT-3 has produced further economically valuable returns: The subsequent GPT-4 model outperforms qualified humans on many graduate and professional exams (OpenAI, 2023b), and its development helped prompt a multi-billion-dollar investment in the company that built it (Capoot, 2023). Scaling laws allowed the creators of GPT-4 to cheaply and accurately predict a key overall measure of its performance: This forecast was made by fitting a statistical trend in the performance of small models, which collectively took about 0.1% of the resources needed by the final model, and then extrapolating out that trend (see Figure 1).

2. Specific important behaviors in LLM tend to emerge unpredictably as a byproduct of increasing investment

Scaling laws generally only predict a model's *pretraining test loss*, which measures the model's ability to correctly predict how an incomplete piece of text will be continued.[1] While this measure is correlated with how useful a model will be on average across many practical tasks (Radford et al., 2019), it is largely *not* possible to predict when models will start to show specific skills or become capable of specific tasks (see Figure 2; Steinhardt, 2021; Ganguli et al., 2022a; Wei et al., 2022a). Often, a model can fail at some task consistently, but a new model trained in the same way at five or ten times the scale will do well at that task.

[1] Much of the data and computer time that goes into building a modern LLM is used in an expensive initial *pretraining* process. Language-model pretraining intuitively resembles the autocomplete task: In it, an artificial neural network model takes in a text one word at a time, makes a probabilistic prediction about which word will come next, and has its behavior incrementally adjusted to make it assign a greater probability to the actual next word in similar contexts in the future. Pretraining test loss measures how effectively an LLM has learned to make these predictions.

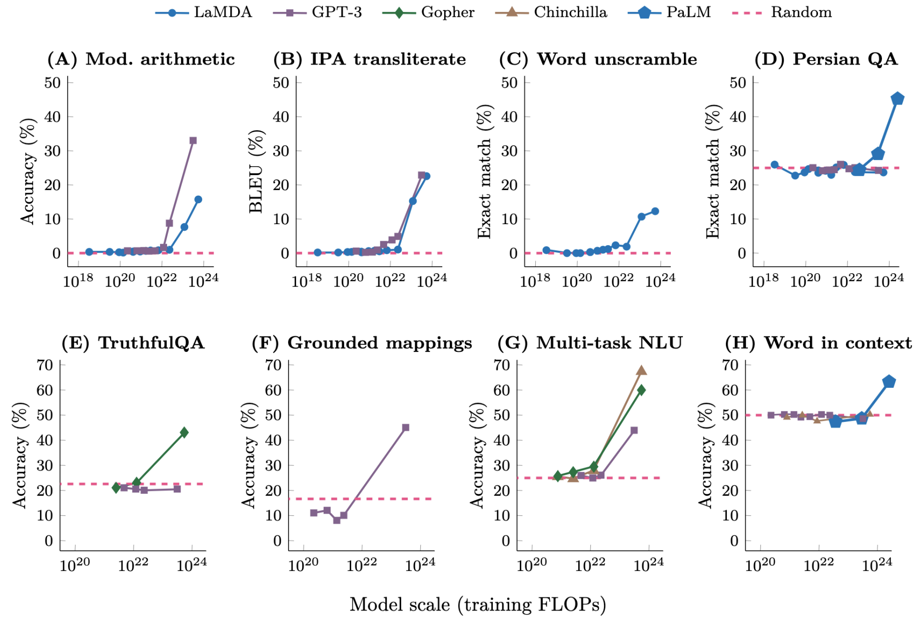

Figure 2. Excerpted from Wei et al. (2022a): Evaluations of performance on specific tasks or behaviors in LLMs do not generally show predictable trends, and it is common for new behaviors to emerge abruptly when transitioning from a less resource-intensive version of a model to a more resource-intensive one.

Wei et al. show that the tasks in BIG-Bench (Srivastava et al., 2022), a standard broad-coverage benchmark for LLM abilities, show a range of different kinds of trend that collectively make scaling-law style predictions unreliable (Figure 3). This means that when a lab invests in training a new LLM that advances the scale frontier, they're *buying a mystery box*: They're justifiably confident that they'll get a variety of economically valuable new capabilities, but they can make few confident predictions about what those capabilities will be or what preparations they'll need to make to be able to deploy them responsibly.

Concretely, two of the key behaviors in GPT-3 that set it apart as the first modern LLM are that it shows *few-shot learning*, the ability to learn a new task from a handful of examples in a single interaction, and *chain-of-thought reasoning*, the ability to write out its reasoning on hard tasks when requested, as a student might do on a math test, and to show better performance as a result. GPT-3's capacity for few-shot learning on practical tasks appears to have been discovered only after it was trained, and its capacity for chain-of-thought reasoning was discovered only several months after it was broadly deployed to the public (Nye et al., 2021; Wei et al., 2022b; Kojima et al., 2022; Zhou et al., 2023).[2] In addition, model abilities involving programming, arithmetic, defusing misconceptions, and answering exam questions in many domains show abrupt improvements as models are scaled up (Wei et al., 2022a; Srivastava et al., 2022).

There are few widely agreed-upon limits to what capabilities could emerge in future LLMs. While there are some hard constraints on the behaviors of typical current LLMs—stemming from limits on the amount of text they can use as input at any one time, limits on their ability to interact with the world during training, or limits on the amount of computation they can perform for each word they generate—it is arguably plausible that these will be overcome with further research within the same technical paradigm. However, many experts disagree: 51% of language-technology researchers surveyed in spring 2022 agreed that "expert-designed strong inductive biases (à la universal grammar, symbolic systems, or cognitively-inspired computational primitives) will be necessary to practically solve some important real-world problems or applications in [language technology]", which if true would represent a limit to the LLM paradigm (Michael et al., 2022).

Expert forecasts, however, have often predicted that we would see less progress with LLMs than has actually occurred. While forecasts by technology researchers are often informal, and I am aware of no precise evaluation of their accuracy, we do have a crisp example of experienced professional forecasters making similar mistakes: Steinhardt (2022) presents results from a competition that was organized in summer 2021, which gave forecasters access to experts, extensive evidence, and a cash incentive, and

[2] See Branwen (n.d.) for a survey that includes additional unpublished reports of this behavior.

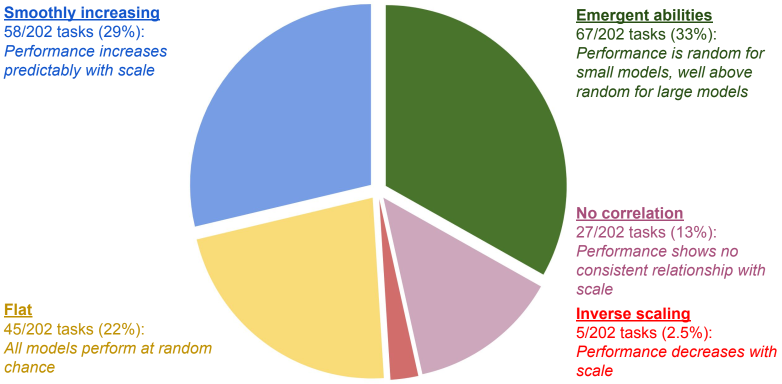

Figure 3. Adapted from a figure by Jason Wei based on data from Wei et al. (2022a): The 202 tasks evaluated in the language-technology benchmark BIG-Bench (Srivastava et al., 2022) tend to show improved performance with scale overall, but individually they can improve gradually, improve abruptly, stay level, get worse, or vacillate, making it impossible to extrapolate the performance of some future system confidently.

asked them to predict what state-of-the-art performance with LLMs would be in each of the next four years on two specific tasks. The results from summer 2022, only one year into the competition, substantially exceeded what the consensus forecast said would be possible in *2024*. Results with GPT-4 in early 2023 exceeded the consensus forecast for *2025* on the one measure for which we have reported results (OpenAI, 2023b). This suggests that it is worth planning for the possibility that we continue to see fast technical progress.

3. LLMs often appear to learn and use representations of the outside world

There is increasingly substantial evidence that LLMs develop internal representations of the world to some extent, and that these representations allow them to reason at a level of abstraction that is not sensitive to the precise linguistic form of the text that they are reasoning about. Current LLMs seem to do this only weakly and sporadically, but the evidence for this phenomenon is clearest in the largest and most recent models, such that we should expect it to become more robust as systems are scaled up further.

Evidence for this claim includes the following results, spanning many established experimental methods and models:

- Models' internal representations of color words closely mirror objective facts about human color perception (Abdou et al., 2021; Patel & Pavlick, 2022; Søgaard, 2023).

- Models can make inferences about what the author of a document knows or believes and use these inferences to predict how the document will be continued (Andreas, 2022).

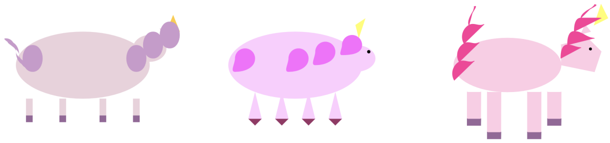

Figure 4. Excerpted from Bubeck et al. (2023): An popular informal (and potentially cherry-picked) demonstration of LLMs' ability to manipulate visual representations. Here, a private version of GPT-4, trained without any access to visual information, is asked to write instructions in a graphics programming language to draw a unicorn. During the model's training (left to right), the resulting drawings appear to become more competent.

- Models use internal representations of the properties and locations of objects described in stories, which evolve as more information about these objects is revealed (Li et al., 2021). This can include the ability to internally represent the spatial layout of the setting of a story (Patel & Pavlick, 2022; Bubeck et al., 2023). Models also use similar representations for facts about real-world geography (Liétard et al., 2021).

- Models can at least sometimes give instructions describing how to draw novel objects (Figure 4; Bubeck et al., 2023).

- Models that are trained to play board games from descriptions of individual game moves, without ever seeing a full depiction of the game board, learn internal representations of the state of the board at each turn (Li et al., 2023).

- Models can distinguish common misconceptions from true facts (Wei et al., 2022a), and often show well-calibrated internal representations for how likely a claim is to be true (Kadavath et al., 2022; Wei et al., 2022a; Burns et al., 2023).

- Models pass many tests designed to measure commonsense reasoning, including some like the Winograd Schema Challenge that are explicitly designed to include no purely textual clues to the answer (Levesque et al., 2012; He et al., 2021; OpenAI, 2023b).

These results are in tension, at least to some extent, with the common intuition that LLMs *are nothing but statistical next-word predictors*, and therefore cannot learn or reason about anything but text. While the premise of this intuition is technically correct in some cases, it can paint a misleading picture of the often-rich representations of the world that LLMs develop as they are trained. In addition, LLMs are increasingly often augmented with other ways of learning about the world that make this claim literally false, such as through interactive training methods (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a), integration with image processing systems, (Alayrac et al., 2022; OpenAI, 2023b), or integration with other software tools (Nakano et al., 2021; Menick et al., 2022; Collins et al., 2022; Schick et al., 2023; OpenAI, 2023a).

4. There are no reliable techniques for steering the behavior of LLMs

Much of the expense of developing an LLM goes into language-model pretraining: The process of training a neural network to predict how random samples of human-written text will be continued. In most cases, though, the developers of such a system want to use it for tasks other than predicting continuations, which requires that it be adapted or guided in some way. Even building a general-purpose *instruction-following* model, where one is not attempting to specialize on any particular task, requires this kind of adaptation: Otherwise, the model will attempt to *continue* its instructions rather than following them (Ouyang et al., 2022).

This adaptation typically involves one or more of these three techniques:

1. Plain language model prompting, where one prepares an incomplete text like "*The translation of 'cat' in French is*"', such that a typical continuation of the text should represent a completion of the intended task (Radford et al., 2019; Raffel et al., 2020).[3]

2. Supervised fine-tuning, where one trains the model to match high-quality human demonstrations on the task (Radford et al., 2018; Devlin et al., 2019; Ouyang et al., 2022).

3. Reinforcement learning, where one incrementally weakens or strengthens certain model behaviors according to preference judgments from a human tester or user (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a).

These techniques produce useful systems, but they are far from perfectly effective: They can't guarantee that an AI model will behave appropriately in every plausible situation it will face in deployment. Nor can they even make a model *try* to behave appropriately to the extent possible given its skills and knowledge (to the extent that it can be said to have generalizable skills or knowledge). In particular, models can misinterpret ambiguous prompts or incentives in unreasonable ways, including in situations that appear unambiguous to humans, leading them to behave unexpectedly (D'Amour et al., 2020; Kenton et al., 2021).

In one key way, this problem is getting easier to tackle: As LLMs become more capable of using human language and human concepts, they also become more capable of learning the generalizations we would like. Indeed, many control techniques work better with larger models, at least for simple tasks(Hendrycks et al., 2020; Bai et al., 2022a; Chung et al., 2022; Ganguli et al., 2023). In another important way, though, the problem is becoming more difficult: More capable models can better recognize the specific circumstances under which they are trained. Because of this, they are more likely to learn to act as expected in precisely those circumstances while behaving competently but unexpectedly in others. This can surface in the form of problems that Perez et al. (2022) call *sycophancy*, where a model answers subjective questions in a way that flatters their user's stated beliefs, and *sandbagging*, where models are more likely to endorse common misconceptions when their user appears to be less educated. It seems likely that issues like these played some role in the bizarre, manipulative behavior that early versions of Microsoft Bing Chat showed, despite the system having been tested extensively before launch (Roose, 2023; Perrigo, 2023; Mehdi, 2023).

Though there has been some progress in understanding and mitigating these issues, there is no consensus on whether or how we will be able to deeply solve them, and there is increasing concern that they will become catastrophic when exhibited in larger-scale future systems (Amodei et al., 2016; Bommasani et al., 2021; Saunders et al., 2022; Ngo, 2022). Some experts believe that future systems trained by similar means, even if they perform well during pre-deployment testing, could fail in increasingly dramatic ways, including strategically manipulating humans to acquire power (Hub-

[3] Prompting in the more general sense can describe the practice of writing instructions or requests for an LLM, where these instructions and requests need not have this continuation property. The base models produced by language-model pretraining do not support this kind of prompting.

inger et al., 2019; Turner et al., 2021; Di Langosco et al., 2022; Ngo, 2022; Turner & Tadepalli, 2022). Broad surveys of the field suggest that these concerns are fairly broadly shared: The majority of the 738 researchers who responded to a recent survey (targeting those who published recently at the machine-learning venues NeurIPS and ICML) assigned a greater than 10% chance of "human inability to control future advanced AI systems causing human extinction" (Stein-Perlman et al., 2020). 36% of another sample of 480 researchers (in a survey targeting the language-specific venue ACL) agreed that "It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war" (Michael et al., 2022). Hundreds of researchers recently signed a controversial open letter that calls for a moratorium on large-scale LLM training until adequate safety and governance mechanisms can be put in place (Bengio et al., 2023).

5. Experts are not yet able to interpret the inner workings of LLMs

Modern LLMs are built on artificial neural networks: They work by computing and updating numeric activation values for internal components that are very loosely modeled on human neurons (Bengio et al., 2017). On this analogy, our tools for doing neuroscience on these systems are still weak: We have some coarse tools for testing whether models represent a few specific kinds of information (like the color results discussed in Section 3), but as of early 2023, there is no technique that would allow us to lay out in any satisfactory way what kinds of knowledge, reasoning, or goals a model is using when it produces some output.

While there is ongoing research oriented toward this goal (Elhage et al., 2021; Lovering & Pavlick, 2022; Chan et al., 2022; Burns et al., 2023; Li et al., 2023, i.a.), the problem is deeply difficult: There are hundreds of billions of connections between these artificial neurons, some of which are invoked many times during the processing of a single piece of text, such that any attempt at a *precise* explanation of an LLM's behavior is doomed to be too complex for any human to understand. Often, ad-hoc techniques that at first seem to provide insight into the behavior of an LLM are later found to be severely misleading (Feng et al., 2018; Jain & Wallace, 2019; Bolukbasi et al., 2021; Wang et al., 2022). In addition, promising-looking techniques that elicit reasoning in natural language do not reliably correspond to the processes that LLMs use to reason, and model-generated explanations can also be systematically misleading (Lipton, 2018; Christiano, 2022; Uesato et al., 2022).

6. Human performance on a task isn't an upper bound on LLM performance

While LLMs are trained primarily to imitate human writing behavior, they can at least potentially outperform humans on many tasks. This is for two reasons: First, they are trained on far more data than any human sees,[4] giving them much more information to memorize and potentially synthesize. In addition, they are often given additional training using reinforcement learning before being deployed (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a), which trains them to produce responses that humans find helpful *without* requiring humans to demonstrate such helpful behavior. This is analogous to the techniques used to produce superhuman performance at games like Go (Silver et al., 2016). Concretely, LLMs appear to be much better than humans at their pretraining task of predicting which word is most likely to appear after some seed piece of text (Shlegeris et al., 2022), and humans can teach LLMs to do some simple tasks more accurately than the humans themselves (Stiennon et al., 2020).

7. LLMs need not express the values of their creators nor the values encoded in web text

When a plain pretrained LLM produces text, that text will generally resemble the text it was trained on. This includes a resemblance in the values expressed by the text: Models mirror their training data in the explicit statements they produce on value-laden topics and in the implicit biases behind their writing. However, these values are subject to a good degree of control by their developers, especially when the plain pretrained LLM is given further prompting and training to adapt it for deployment as a product (Section 4). This means that the values expressed in a deployed LLM's behavior do not need to reflect some average of the values expressed in its training data. This also opens up opportunities for third-party input and oversight, meaning that the values expressed in these models also need not reflect the values of the specific people and organizations who build them.

In particular, popular approaches involving reinforcement learning and *red-teaming* allow model developers to guide models toward a persona and set of values more or less of their choosing (Dinan et al., 2019; Bai et al., 2022a; Ganguli et al., 2022b). In these techniques, the values that a model learns are never made entirely explicit. Instead, they are reflected in many small pieces of feedback that human annotators give the model during training. The *constitutional*

[4] LLMs see over $10,000\times$ more data than humans: A human adolescent sees tens of thousands of words, while LLMs can be exposed to over one trillion (Hart & Risley, 1992; Gilkerson et al., 2017; Hoffmann et al., 2022)

AI technique (Bai et al., 2022b) significantly cuts down on human labor and makes these values more explicit: Using this technique, a model can be trained to follow a set of norms and values simply by writing those values down in the form of a list of constraints called a constitution. It is possible to use techniques like this to dramatically reduce *explicit* examples of widely-recognized biases, like anti-Black racism, in model behavior (Ganguli et al., 2023).[5] Indeed, in some cases, exposing models to more examples of unwanted behavior during pretraining can make it *easier* to make them avoid that behavior in deployment, reversing the intuitive link between training data and model behavior (Korbak et al. 2023; see also Appendix C in Chung et al. 2022).

These technical interventions, especially constitutional AI, are amenable to outside influence and regulation. One can easily imagine third-party standards bodies collecting input about what behaviors are acceptable in AI systems and distilling this input into constitutions that model developers are encouraged or required to adopt.

As in Section 4, though, these techniques can still fail in subtle and surprising ways, and the trends in how these techniques change as models with scale are complex. And, of course, there are many other ethical questions that arise with the development of deployment of large-scale AI systems, including issues around environmental impacts, access, misuse, privacy, safety, and the concentration of power (Amodei et al., 2016; Bender et al., 2021; Bommasani et al., 2021; Birhane et al., 2022; Weidinger et al., 2022, i.a.).

8. Brief interactions with LLMs are often misleading

While many deployed LLMs are largely able to follow instructions, this instruction-following behavior isn't inherent to the model, but rather is grafted onto it using highly imperfect tools (Section 4). In part because of this, models can be sensitive to the contents of their instructions in idiosyncratic ways. Often, a model will fail to complete a task when asked, but will then perform the task correctly once the request is reworded or reframed slightly, leading to the emerging craft of *prompt engineering* (Brown et al., 2020; Reynolds & McDonell, 2021; Radford et al., 2021; Dohan et al., 2022; White et al., 2023; Si et al., 2023).

These contingent failures are evidence that our techniques for controlling language models to follow instructions are not reliably effective. However, simply observing that an LLM fails at a task in some setting is not reliable evidence that that LLM doesn't have the skills or knowledge to do that task. Often, once one finds an appropriate way to prompt a model to do some task, one will find that the model *consistently* performs well across different instances of the task. The chain-of-thought reasoning strategies mentioned in Section 2 are an especially clear example of this: Simply prompting a model to "think step by step" can lead it to perform well on entire categories of math and reasoning problems that it would otherwise fail on (Kojima et al., 2022). Similarly, even observing that an LLM *consistently* fails at some task is far from sufficient evidence that no other LLM can do that task (Bowman, 2022).

On the other hand, observing that an LLM performs a task successfully in one instance is not strong evidence that the LLM is capable of performing that task in general, especially if that example was cherry-picked as part of a demonstration (like the unicorn in Figure 4). LLMs can memorize specific examples or strategies for solving tasks from their training data *without* internalizing the reasoning process that would allow them to do those tasks robustly (see, e.g. McCoy et al., 2019; Magar & Schwartz, 2022).

9. Discussion and Limitations

The additional discussion that I present here builds on and contextualizes the eight claims above, but it is more speculative or subjective in places and reflects views that are not necessarily broadly shared.

9.1. We should expect some of the prominent flaws of current LLMs to improve significantly

Hallucination, the problem of LLMs inventing plausible false claims, is a prominent flaw in current systems and substantially limits how they can be responsibly used. Some of the recent findings discussed in Section 3 suggest, though, that we may soon be able to mitigate this problem simply by finding ways to better use abilities that models already display: LLMs *internally* track which statements are true with reasonably high precision, and this ability improves with scale (Burns et al., 2023; Kadavath et al., 2022).

Similarly, as noted in Section 7, recent methods can dramatically reduce explicit bias and toxicity in models' output, largely by exploiting the fact that models can often recognize these bad behaviors when asked (Dinan et al., 2019; Bai et al., 2022b; Ganguli et al., 2023). While these mitigations are unlikely to be entirely robust, the prevalence and prominence of these bad behaviors will likely wane as these techniques are refined.

To be clear, though, these encouraging signs do not mean that we can reliably control these models, and the issues noted in Section 4 still apply. Our partial solutions are

[5] However, explicit demonstrations of racist language or decision-making by models do not come close to exhausting the ways that the development and use of these systems interact with biases and power structures involving factors like race (see, for example, Field et al., 2021).

likely to leave open important failure modes. For example, straightforward attempts to manage hallucination are likely to fail silently in a way that leaves them looking more trustworthy than they are because of issues related to sandbagging: If we apply standard methods to train some future LLM to tell the truth, but that LLM can reasonably accurately predict which factual claims human data workers are likely to check, this can easily lead the LLM to tell the truth *only when making claims that are likely to be checked*.

9.2. There will be incentives to deploy LLMs as agents that flexibly pursue goals

Increasingly capable LLMs, with increasingly accurate and usable internal models of the world, are likely to be able to take on increasingly open-ended tasks that involve making and executing novel plans to optimize for outcomes in the world (Chan et al., 2023). As these capabilities develop, economic incentives suggest that we should see them deployed in areas like software engineering or business strategy that combine measurable outcomes, a need for flexible planning, and relatively flexible standards and regulations. LLMs augmented with additional tools can extend this into grounded domains like robotics (Sharma et al., 2022; Driess et al., 2023). Deployments of this type would increasingly often place LLMs in unfamiliar situations created as a result of the systems' own actions, further reducing the degree to which their developers can predict and control their behavior. This is likely to increase the rate of simple errors that render these systems ineffective as agents in some settings. But it is also likely to increase the risk of much more dangerous errors that cause a system to remain effective while strategically pursuing the wrong goal (Krueger et al., 2020; Ortega et al., 2021; Chan et al., 2023).

9.3. LLM developers have limited influence over what is developed

Because many important LLM capabilities are emergent and difficult to predict, LLM developers have relatively little influence on precisely what capabilities future LLMs will have, and efforts to make predictions about future LLM capabilities based on the economic incentives, values, or personalities of their developers are likely to fail. GPT-4, for example, appears to have many skills, like those involving programming, that its creators were likely hoping for. However, it also appears to have initially shown several unwanted skills, like teaching laypeople to prepare biological weapons, that its creators had to spend significant effort to try to remove (OpenAI, 2023b).

Beyond this, LLM developers inevitably also have limited *awareness* of what capabilities an LLM has when they're deciding whether to deploy it: There is no known evaluation or analysis procedure that can rule out surprises like chain-of-thought reasoning in GPT-3, where users discover a way to elicit some important new behavior that the developers had not been aware of.

9.4. LLMs are likely to produce a rapidly growing array of risks

More broadly, the current technical and commercial landscape provides strong incentives to build and deploy increasingly capable LLMs quickly. Nonetheless, our track record of recognizing what capabilities a new LLM can demonstrate before deploying it is spotty. Our techniques for controlling systems are weak and are likely to break down further when applied to highly capable models. Given all this, it is reasonable to expect a substantial increase and a substantial qualitative change in the range of misuse risks and model misbehaviors that emerge from the development and deployment of LLMs.

While many positive applications of LLM-based systems are likely to be genuinely valuable, the societal cost-benefit tradeoffs involved in their deployment are likely to remain difficult or impossible to evaluate in advance, at least without significant progress on hard technical problems in model evaluation, interpretability, and control. Some of these hard-to-evaluate risks, such as those involving unconventional weapons or strategic power-seeking behavior, may be impossible to adequately mitigate if they are discovered only after systems are deployed. Strategic power-seeking behavior in particular could pose serious risks during model *development*, even without an intentional deployment. This suggests that future work in this area will likely warrant increasingly stringent standards for safety, security, and oversight.

9.5. Negative results with LLMs can be difficult to interpret but point to areas of real weakness

There are many sound scientific results showing that recent LLMs fail at language and commonsense reasoning tasks, sometimes relatively simple ones, under good-faith attempts to elicit good behavior (Pandia & Ettinger, 2021; Schuster & Linzen, 2022). Sometimes the details of these failures cast doubts on the quality of other related evaluations (Webson & Pavlick, 2022; Ullman, 2023). For reasons mentioned in Section 8, positive results on well-designed measures are much more reliable than negative results. Nonetheless, in some areas, including areas as simple as the handling of negation,[6] LLMs show what appear to be systematic weaknesses in their ability to process language or reason about the world. We have few grounds to predict whether or when these limitations will be resolved.

[6] See, for example, the Modus Tollens task by Huang and Wurgaft, described in McKenzie et al. (2022).

9.6. The science and scholarship around LLMs is especially immature

LLMs strain the methods and paradigms of the fields that one would expect to be best qualified to study them. Natural language processing (or language technology) is the historic home discipline for this work, but its tools are oriented toward measuring and improving the ability of computational systems to use language. While LLMs fundamentally learn and interact through language, many of the most pressing questions about their behavior and capabilities are not primarily questions about language use. The interdisciplinary fields studying AI policy and AI ethics have developed conceptual and normative frameworks for thinking about the deployment of many kinds of AI system. However, these frameworks often assume that AI systems are more precisely subject to the intentions of their human owners and developers, or to the statistics of their training data, than has been the case with recent LLMs (Chan et al., 2023). Relatedly, many of the most cited research papers dealing with LLMs, including many papers that introduce new methods or theories, are not published in peer-reviewed venues. The recent trend toward limiting access to LLMs and treating the details of LLM training as proprietary information is also an obstacle to scientific study.

This means that surprising novel claims about LLMs are often the product of messy, fallible science that goes beyond established disciplinary practice. However, what appears to be established conventional wisdom also often rests on shaky foundations when it is applied to LLMs. All of this is reason to be especially uncertain about the issues discussed in this paper and to make important decisions about LLMs in ways that are resilient to mistaken assumptions.

Conclusion

In closing, rather than recap the claims above, I would like to note three sometimes-prominent issues that this paper leaves largely untouched:

- Open debates over whether we describe LLMs as ***understanding*** language, and whether to describe their actions using agency-related words like *know* or *try*, are largely separate from the questions that I discuss here (Bender & Koller, 2020; Michael, 2020; Potts, 2020). We can evaluate whether systems are effective or ineffective, reliable or unreliable, interpretable or uninterpretable, and improving quickly or slowly, regardless of whether they are underlyingly human-like in the sense that these words evoke.

- Similarly, questions of **consciousness**, sentience, rights, and moral patienthood in LLMs (see, e.g. Schwitzgebel & Garza, 2020; Shevlin, 2021; Chalmers, 2023), are worth distinguishing from the issues above. Though these questions may influence important decisions about how AI systems are built and used, it should be possible to evaluate most or all of the issues raised here without taking a stance on these questions.

- Finally, **value judgments** around LLMs are beyond the scope of this paper. The broader question of whether the rapid progress that we're seeing with LLMs is a good thing, and what we should each do about it, depends on a deeper and more diverse range of considerations than the technical literature that I draw on here can come close to addressing.

Acknowledgments

This paper benefited from conversations at the AI FUTURES panel organized by Critical AI at Rutgers and from discussions with many other researchers, including Ellie Pavlick, Jackson Petty, Owain Evans, Adam Jermyn, Eric Drexler, Ben Garfinkel, Richard Ngo, Jason Wei, Helen Toner, Jeffrey Ladish, Leo Gao, Alex Lyzhov, Julian Michael, Adam Bales, Rick Korzekwa, Ben Mann, Alex Lawsen, Alex Tamkin, Anton Korinek, and David Dohan. I used LLMs in this paper only to explore some minor wording and framing decisions. All errors and omissions are, of course, my own.

This work has benefited from financial support from Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program) and from Open Philanthropy, as well as from in-kind editing support from Pablo Moreno through FAR.ai. This material is based upon work supported by the National Science Foundation under Grant Nos. 1922658 and 2046556. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9. URL https://aclanthology.org/2021.conll-1.9.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.