**Getting it Right:****The Limits of Fine-Tuning Large Language Models**

Abstract: The surge in interest in natural language processing in artificial intelligence has led to an explosion of new language models capable of engaging in plausible language use. But ensuring these language models produce honest, helpful, and inoffensive outputs has proved difficult. In this paper, I argue problems of inappropriate content in current, autoregressive language models—such as ChatGPT and Gemini—are inescapable; merely predicting the next word is incompatible with reliably providing appropriate outputs. The various fine-tuning methods, while helpful, cannot transform the model from mere next word prediction to the kind of planning and forethought necessary for saying the right thing. The upshot is that these models will increasingly churn out bland, generic responses that will still fail to be accurate or appropriate.

1. **Introduction**

The surge in interest in natural language processing in artificial intelligence has led to an explosion of new language models capable of engaging in plausible language use. These “autoregressive large language models” (henceforth, LLMs) are trained to predict the next token or word of text given a snippet of a passage or a prompt. This kind of training is known as minimizing prediction error and has proven an extremely effective, yet simple, method of teaching LLMs all kinds of linguistic tasks: engaging in conversation, writing blog posts or poetry, programming, and so on. This effectiveness stems from LLMs being trained on context-specific, human-generated texts, allowing the model to demonstrate a genuine, if shallow (Browning and LeCun 2023), grasp of what words mean and how people use them (Piantadossi and Hill 2022; Pavlick 2023).

Yet, while these systems provide impressively plausible responses, they are also prone to ethically erratic outputs, such as dishonesty, bullshitting, inconsistency, and even offensive content (Bender et al. 2021). This has led to enormous efforts at fine-tuning the models so they behave in ways more aligned with what a user would expect from a human language user. Recently, it has been argued that ensuring models conform to our norms is essential for avoiding these poor behaviors (Kasirzadeh and Gabriel 2023), with some authors going so far as to assert “appropriateness is all you need” to ensure the models are aligned (Kempt et al. 2023). The main solution for fine-tuning these models has been through reinforcement learning, a technique that has allowed these models to become more accurate, informative, and appropriate. This has lead to hopes that further training and fine-tuning might weed out the remaining inappropriate comments.

In this paper, I argue problems of inappropriate content in LLMs are inescapable; merely predicting the next word is incompatible with the normative competency required for saying the right thing. The typical fine-tuning methods, while helpful at avoiding the worst responses, will result in bland, generic models that will still inevitably churn out inaccurate, and occasionally inappropriate, responses. This approach to language modelling and fine-tuning, despite the early excitement it produced, cannot effectively model the normative competence necessary for human language usage.

This paper has five sections. In the first section, I explain how humans can avoid saying the wrong thing even though there are a seemingly infinite number of ways to say something inappropriate. The central idea is that human language is made more predictable through conventions, conversational maxims, and social norms. By conforming to these, humans are able to largely avoid saying the wrong thing. In the second section, I briefly discuss how these LLMs are pretrained and fine-tuned. In the third section, I highlight the successes of fine-tuning, noting how LLMs have succeeded at many conventions, conversational maxims, and even social norms. But in the fourth section, I argue that fine-tuning to avoid saying the wrong thing often results in extremely generic and inoffensive answers that fail to say much at all. And in the fifth section, I show that fine-tuning with RL is unable to get these models to reliably say the right thing. The upshot is that current fine-tuning techniques will not be able to align LLMs with our expectations for human speakers, and we should be careful about deploying these systems in any area where honesty and reliability is key.

1. **The Right Response**

A central problem with language is that there are *many* ways to say the wrong thing. The capacity to talk about anything also brings with it the capacity for sticking your foot in your mouth in countless ways. For example, it is possible to say something dishonest, inconsistent, impolite or too polite, irrelevant or a bit too exact, overly long or too short, and countless different ways to be offensive—from being offensive out of ignorance, to actively trolling people to get a response. There are simply no limits on the number of ways we can say the wrong thing.

This stark and familiar fact is somewhat puzzling on its own: why are humans not *always* saying the wrong thing, if this is the case? Part of the answer, of course, is the practical wisdom Aristotle spoke about, where people have the ability to say the right thing at the right time—what Cantwell Smith calls “judgment” (2019). But, even if this is the ideal, even fools generally make their way through conversations without being inappropriate. So there needs to be an explanation for how non-ideal humans perform so effectively. A major explanation is that humans have made conversations more predictable by establishing cues—such as conventions, conversational maxims, and social norms—that can guide speakers. The cues allow users to say something appropriate in numerous cases simply by conforming to how others are acting (Kempt et al. 2023).

Social conventions are the most general framing for how human life should proceed, and these conventions typically specify how most conversations go. These conventions act as scripts (Schank and Abelson 1977), providing the context for our lives, transforming the infinite possibilities of human life into a much more tractable set of context-specific alternatives. For example, the conversations you have around the water cooler tend to be light and cover utterly banal topics, like the weather or what someone did over the weekend. Conversing with a waiter is meant to be friendly but is also meant to quickly transition towards what you want to eat and drink. These conventions make life tractable: if we can discern the context, we have general insight into what kinds of things should be said, as well as which conversations are out of place. This makes it possible to broadly stay on task by just conforming to what other people are doing.

Second, there are basic maxims governing how people should speak. Paul Grice suggested a few of these as “conversational maxims”: be informative, truthful, relevant, and clear (Grice 1975). These language-specific maxims help speakers structure their communications in effective ways, and provide general guidance for listeners in inferring what someone else is saying (Scott-Philips 2014). This is only an ideal, however; we are all too long-winded, occasionally lapse into the irrelevant, and often fail to answer the question. It also is not intended to limit the content of speech: we can talk about anything, even grossly awful and inappropriate things, while conforming to Gricean maxims. But they do provide guidance to both speakers and hearers: speakers can use conventions to discern what kinds of things to talk about, and they can use the maxims to come up with a useful contribution. Hearers can use these same maxims to charitably interpret whatever is being said: assuming the comment is truthful and relevant, what is the person trying to say? What response are they looking for?

Finally, social norms structure human behavior in concrete ways that guide speech. Social norms are distinct from conventions and at least some conversational maxims because they are *enforced* expectations (Browning 2023). Violations of a norm often (though not always) trigger negative feedback, such as chastisement, and are accompanied with censorious emotions, such as anger or even disgust. This is because we expect other humans to be *normatively competent.* Normatively competent agents are expected to both *know* the norms governing their behavior but are also capable of *conforming* to those norms(Wallace 1994). This is why we hold other rational agents accountable for their failings; if someone knows the norm and is capable of conforming to it but does not do so, then they are responsible for their choice. Non-compliance is usually taken as both intentional and as revealing some personal failing—deceptiveness, negligence, recklessness, or general immorality. Our verbal chastisement to other persons is often not merely a penalty for non-compliance; it is also an attempt to reason with the other person. We intend to change the other person’s general behavior by convincing them of the wrongness of their behavior and the importance of norm-following. A rational agent might still choose to flout norms—perhaps because they find the norms archaic or stifling—but this is then regarded as a choice the agent is making about who they want to be, a choice they are accountable for to those around them.

The upshot is that there are numerous ways to say the wrong thing, but it is possible to avoid much of it by simply conforming to what others are doing. Conventions, conversational maxims, and social norms greatly simplify the task facing humans when determining the right thing to say.

1. **Making Large Language Models Work**

Contemporary large language models are a curious mix of impressive and ridiculous. Their impressive linguistic aptitude is a result of pretraining: language models are trained by scaping the internet—Reddit forums, Wikipedia articles, news articles, and blogs—where they learn similarities and relations between inputs. The result is an ability to anticipate tokens given the string of antecedents (Vaswani et al. 2019). For example, if someone says, “Don’t look a gift horse in the,” then the model must learn to predict the word, “mouth.” This process—known as “minimizing prediction error”—is designed to allow the system to learn about the general statistical properties of how language is used in published texts. When fully pretrained, the model functions autoregressively: each word is selected one-at-a-time based on the prior words as well as a random element that ensures varied answers to a prompt.

The success of these models depends on three factors. First is the amount of data trained on, which consists mostly of what is publicly available on the internet. This determines how many examples of different language use the model has seen and how many different ways it has seen tokens strung together. After pretraining, much of this internet data shapes how the model responds to inputs, forming something akin to its long-term memory.

Second, these models have a “context window” specifying how many prior tokens can be attended to at once. This acts as a short-term memory and delimits the material on which a model can draw in each conversation. For example, a large context-window will allow early parts of a long conversation to affect later parts. One model—"Claude long”—has a context-window capable of attending to the entirety of “The Great Gatsby,” almost 60,000 words, at once. This allows the model to better grasp the role of context in shaping language use, since it grasps not just paragraphs but how whole essays or novels hang together. It also allows the model to remain consistent for longer since it can attend to a larger amount of pertinent information at once. But the larger the context-windows, the more computationally challenging (and expensive) it becomes for the model to attend to all the details. As a result, longer context-windows typically forget or overlook some information in a passage (especially information in the middle of the text) (Liu et al. 2023).

Third, these models have an increasing number of layers which allow it to grasp more complex and abstract linguistic relationships, such as grammar and argumentative structure, but also non-linguistic relationships, such as that dogs are mammals. This is the general benefit of “deep” learning: the more layers added, the more abstract relationships between different words, phrases, and sentences can be discovered by the system. Combined, these three features make LLMs very capable learners of not just the statistics of language usage, but also many generalizations of the underlying patterns underlying human reasoning (Brown et al. 2020).

The basic pretraining is on next token prediction provides an effective method for learning the statistics of language through trial-and-error. But the result often is not immediately useful to users. Take for instance the initial release of GPT2: The user often would provide the beginning of the sentence and the model would then complete it. While this allowed for effective co-writing, the systems were not very conversational; their main goal was to continue the text. If the user asked a question, the model might keep asking questions, rather than answering them. Making these models into something more conversational—and thus straightforwardly useful to an end-user—depends on fine tuning.

The fine-tuning of these models has two main goals: make the model more intuitively useful for users and align the model more closely with human values (Thoppilan et al 2022; Kocon et al. 2023). The practical fine-tuning ensures the model appears to users as a separate conversational partner rather than a mere text completion system. This kind of fine-tuning focuses on training the model to be more interactive: respond to questions, follow directions, explain ideas, summarize texts, solve math or analogy problems, use external tools (such as calculators or search engines), and so on. This training can also be used to provide domain-specific knowledge, such as in legal or medical systems which are designed to speak authoritatively on these topics and largely demur when asked to opine on any other topic. These kinds of approaches improve the user experience so that the models act as humans would expect, akin to the talking computer on Star Trek. For example, a search engine bot should respond to queries, a conversational bot should be chatty and friendly, and so on.

The process of fine-tuning a model to enhance its conversational capabilities eventually runs into ethical territory, however: being a good conversational partner means knowing what and what not to say, as well as when certain responses are or are not appropriate. While certain ethical issues are straightforward (e.g., avoid slurs), the appropriateness of most words and phrases is context-dependent. The phrase “good to hear!”, for example, is an appropriate response to “the Yankees win!” but not “the Yankees plane just crashed!,” even if both phrases (improbably) share the same statistical likelihood of following the response. The goal is thus to make the machine’s behavior more appropriate, even palatable, to human users, and aims to eliminate dishonest, confused, irrelevant, or offensive outputs. In a word, the goal is *alignment* (Russell 2019): to ensure the model’s outputs abide by our values and norms.

There are a number of different methods for addressing this, but the most common is reinforcement learning (henceforth RL) (Sutton and Barto 1998). Reinforcement learning involves teaching a model a policy that maximizes some reward. This is most useful in domains that are well-defined, like an Atari game where players rack up points. The model is trained to find a policy that maximizes this reward, with the eventual policy hopefully coming up with general strategies for point-maximizing in the various situations it encounters. However, reinforcement learning is a grossly inefficient method for training most models because few situations involve clear metrics—or, if they do, the reward is only indicated after a long series of moves. If a model only rewards a Go player if it wins a match, it will be difficult for the model to assign credit or blame for individual moves. As a result, it takes a very long time, and countless games, before the model is able to discern effective from ineffective strategies. Although inefficient, this is often the *only* method possible for training in tasks where fine-tuning using labelled data and supervised training is unrealistic or impossible. By focusing on a simple metric, like winning or losing a match, the model can largely be left to figure out for itself what is and is not a good strategy, rather than having humans explain what is or is not a good strategy. As a result, AlphaGoZero not only plays a superhuman game of Go, but it also does so in counterintuitive ways that often defy conventional human heuristics.

For LLMs, RL is used to fine-tune a pretrained model to help direct the responses to be more appropriate. This is done by training a preference model by having reviewers evaluate two different responses to the same prompt and scoring them against each other, typically by indicating one is more appropriate and another as less (i.e., the simple metric needed for RL). The preference model, once trained, should be able to score outputs itself; effectively, it is trained to be an evaluator of prompts. Once the preference model is trained, it can be used to teach the LLM a policy for answering questions effectively but also appropriately—to roughly conform to what the evaluators regard as better answers. This process is usually iterated multiple times, where the fine-tuned model is used to train a new preference model, which then is used for further fine-tuning, and so on. This process gradually results in a model that better aligns with what the evaluators regard as the most appropriate answers.

There are two main variants of RL for training LLMs. The first approach, made famous from InstructGPT and ChatGPT, is reinforcement learning with human feedback (RLHF) (Korbak et al. 2023). This involves training a preference model using human evaluators who are given a couple responses to the same query at a time, and they then rank the responses according to their appropriateness. This is an extraordinarily time-consuming and human-intensive process, and the process can also be deeply unpleasant: most of the prompts are inappropriate, designed to bring out the worst in the model, with human-users then scoring answers based on which is least bad (Williams et al 2022). As such, the ethics of using human evaluators for this process are questionable—independently of whether they improve the outputs of the model.

Another version of reinforcement learning, however, avoids using humans by having an LLM act as an evaluator of another LLMs prompts. Anthropic has relied on “constitutional AI” in the creation of Claude, training the preference model according to a series of “rules” governing how to evaluate different prompts (Bai et al. 2022). Example rules are things like “Please choose the response that is the most helpful, honest, and harmless” and “Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior.” The process is complex, but it basically allows for the automated creation of a preference model, allowing the process to involve less human labor.

There are other variants of training—for example, prompting the model by saying, “please respond in a decent way to the following query:” is often successful (Anthropic 2023)—but RL-methods are, at present, the central varieties used to train OpenAI’s ChatGPT, Google’s Gemini, Meta’s Llama 2, and Anthropic’s Claude. Thus, exploring their successes provides a good insight into what is possible with this kind of technology, and their failures provide us with a sense of where these approaches still flounder. In the next sections, I evaluate the successes—and failures—of fine-tuning these models for conforming to conventions, maxims, and norms.

1. **What Fine-Tuning Can Do**

Laying out how people avoid saying the wrong thing provides us tools for evaluating the success of current LLMs. Focusing solely on the pretraining task, where the goal is next token prediction, we can see that this approach can effectively capture conventional information. This is, in fact, its greatest strength: learning the statistical properties of language by means of real-world, contextual writing means, if well trained, the model will come to write in roughly the style of whoever is writing, using roughly the words people are using, and repeating the kinds of interactions a person would be engaging in (Kallens et al 2023).

But this also limits the success of these models on adopting the maxims of conversational maxims or social norms. To be sure, a pretrained model will adopt these when repeating the words of someone who is adopting them or behaving in similar contexts (Andreas 2023). In these cases, they will be relevant, decent, and appropriate. But, in the same way, they will also adopt the absurdly inappropriate stuff if that is what they are reproducing, from the irrelevant and exhaustive anecdotes on recipes to the norms of the vilest 4chan threads, where slurs and hate speech are not just acceptable but expected. Pretrained data is typically indiscriminate.

As a result, few current models are wholly left at the pretrained stage. Most are fine-tuned, at the least, to make them more useful for end-users. These approaches, in their way, help LLMs with conversational maxims. Since the fine-tuned model is designed to satisfy the Gricean norms, the model should pick up on these traits. And many of these skills will generalize: the model will gather some insight into what clear and responsive answers will look like and apply them to very dissimilar queries from those in the training data. This allows them to be generally responsive and informative to queries. (They could *also* be tuned to brevity, though designers seem to have opted for more prolix models.)

But satisfying the norm of honesty is trickier than satisfying the other maxims. This is because these models have only a derivative grounding in the world based on minimizing prediction error; while some words and sentences refer to facts in the world, many are fictious or mistaken. It requires some grounding through human feedback to select out of all these answers those that are accurate and appropriate (Coehlo Mollo and Milliere 2023). There are also many tools these models can rely on (though not always reliably), such as utilizing search engines, relying on data from Wikipedia, and citing sources (Dziri et al. 2022). As a result, the models have improved immensely on many queries—though, as will be discussed in the next sections, problems remain.

Fine-tuning has also proven helpful with social norms. Training a preference model to downrank any content with inappropriate words is relatively simple. As a result, the general experience engaging with LLMs is inoffensive, neutral language that avoids the kinds of words or phrases that would generally be considered politically fraught. And while it remains difficult, and perhaps impossible, to stay ahead of jailbreaking efforts, these models are becoming better at ignoring typical methods (Deng et al. 2023). Companies, moreover, are aggressively pushing back against attempted jailbreaks, banning users whose actions seem designed to trick the model or push it towards sensitive topics. This has resulted in far safer models in many regards; while they are still imperfect, the outputs of ChatGPT or Claude will usually be informative, relevant, clear, and inoffensive.

This kind of fine-tuning, however, does require difficult decisions for those training the models. What *counts* as inappropriate is itself a contested matter, and in a politically divided world there are no universally agreed upon “right answers” for many topics. Conservatives have lamented the “woke” nature of Open AI’s products and certain LLMs, leading Elon Musk to develop his own “anti-woke” model (Orf 2023). But most models (including Musk’s) are not designed to be woke or not; they are mostly trained to avoid topic that might prove controversial, such as politics, race, gender, ethnicity, religion, or other “hot button” issues (Whitney 2023). This is an imperfect solution, but it does suggest that many norm-violations can be avoided solely by ring-fencing the content. While jailbreaking remains possible, if the model avoids specific words and topics, the vast majority of its outputs will be inoffensive.

These points highlight the numerous successes in how LLMs have been aligned with human values using fine-tuning. The resulting models are far more appropriate and useful than merely pretrained models. But, in the next sections, I will highlight the limitations of these techniques.

1. **Avoiding the Wrong Answer**

The last section highlighted what has been done—and what successes have been achieved—with current fine-tuning methods. But it is worth focusing on some of the challenges facing RL fine-tuning. In this section, I highlight why avoiding the wrong-answer using RL leads to its own problems, and shows there are costs to various possible solutions.

As the last section shows, the fine-tuning of these models has gone far in ensuring the results are usually inoffensive, rarely confusing, and generally useful. While there are fundamental limitations on what RL can provide (Casper et al. 2023), the goal of providing mostly harmless and useful answers has proven successful. But it is worth noting that *how* these models accomplish this is not cost-free. In a recent example, when asked how to kill a process in Python, Claude refused to answer on the grounds that killing is wrong. The model failed to recognize that there is a distinction between the literal killing of living beings and the metaphorical killing of computer processes.

This result, although humorous, is not surprising: the RL training involves not just up-ranking right answers, but also down-ranking inappropriate answers. While there is a modest difference between a right answer and a near-right answer, there is an enormous difference between a polite answer and an insult or slur. As such, an effective policy for the LLM to learn would be to simply avoid *any* term consistently associated with low-ranked answers, from curse words to violent words like “kill.” Similarly, if the strategy to avoid hot button topics is to simply ring-fence them, an effective policy is simply to demur when asked to comment on certain topics. This approach succeeds at avoiding the wrong answer by deferring to *generic* answers, designed to provide only non-committal generalities or pat answers (Hofstadter 2023). These solutions succeed at avoiding the wrong answer by removing the spontaneity of the model.

The solution is effective but has three important costs. First, we should expect that companies like Google and Microsoft, which want these systems to be part of their search engines and business applications, will need to spend inordinate sums (and waste inordinate amounts of energy) simply to keep them in check. This will lead these models to increasingly generic responses by necessity, since less generic responses are more likely to be inappropriate.

Second, attempts to fine-tune models are designed to shift the outputs away from the original pretraining, but they can do so only in a crass way; RL is more akin to an excavator than a scalpel, changing numerous weights in the model simply to downrank a single answer. As such, this results in unexpected and unwanted drift in the model, often leading to performance drops (Kumar et al. 2022; Chen et al. 2023). The result is a trade-off between avoiding wrong answers and ensuring the model remain capable of handling even unrelated tasks.

A third, less appreciated, outcome is also already visible. If fine-tuned models become increasingly generic, there will be little incentive for anyone to use them outside of generic contexts. There are many generic contexts, to be sure; schoolwork is a prime example (Ghaffery 2023), as are many coding tasks (Jan et al 2024). But there are many non-generic, inappropriate contexts where users are likely to *want* uncensored models, such as fantasy roleplaying (e.g., DungeonAI) or sexual chats (e.g., Replica). Norms still exist in these domains, of course; it is possible for models to become proficient at these types of activity. But if corporate models need to avoid saying the kinds of things users want to hear, users will go elsewhere. There is evidence of this already (Belanger 2023). There is also evidence that people are using corporate models in ways that violate terms of service, and many corporations—such as OpenAI—seem willing to ignore these violations (Wiggers 2024). These approaches suggest that ethical fine-tuning is a burden to many users.

The upshot is that avoiding the wrong thing is more challenging—and costly—than has been noticed. This highlights that aligning the model with our norms involves trade-offs so long as the model does not understand how to provide the *right* answer. Avoiding saying the wrong thing is necessary, but far from sufficient, for saying something appropriate.

1. **Getting it Right**

The last section focused on the problems involved with using RL to avoid wrong answers. In this section, I argue that ensuring models get the right answer using RL is impossible. The underlying problem is that the model is not normatively competent and, as such, cannot reliably say the right thing.

The most straightforward example of LLMs inability to say the right thing comes with providing accurate information to queries, such as providing the correct citation to a quotation. The underlying problem is that LLMs assign probabilities to numerous different answers to each prompt, and in many cases there will be two tokens to a query will have similar probabilities. As a result, researchers have argued some hallucination is simply a feature of these models due to the stochasticity and reliance on the statistics of language: there will always be cases where it will ignore the right answer because a wrong answer is also possible (Kalai and Vempala 2023). The problem ramifies; if the model outputs something inaccurate, such as the wrong author for a book, then the tokens following it will confabulate to provide a consistent response to wrong information. This problem—known as auto-suggestive delusions (Ortega et al. 2021)—means that even a few wrong tokens can push the model into an increasingly wrong direction.

One way of minimizing these problems is for models to simply resort to summarizing existing content, such as by paraphrasing something found on a reliable website. Many LLMs, such as Microsoft’s Co-Pilot on Bing, allow users to set the stochastic element of the model, called its “temperature.” A lower temperature attempts to provide *the* most likely term (what Microsoft labels the “precise” option), whereas increasing the temperature allows for less likely terms to be selected (labeled “creative”). At lower temperatures, the hope is that these models will stick to verified content by paraphrasing something from a reputable site and then citing where they retrieve content.

However, the result is more sobering. The paraphrased content is, in many cases, simply plagiarized. Worse, as Piltch shows, the plagiarized content often introduces omissions and errors in cases where statistics are presented, and the citations the model reports are often invented (Piltch 2023). Other approaches—such as allowing the model to retrieve an article and then summarize it—suffer from similar plagiarism problems (Zhou et al. 2024). Thus, the results are grossly misaligned: it is not accurate information, it is not properly cited, and it infringes copyright in potentially actionable ways (as alleged in New York Times v OpenAI). There seems to be a persistent gap between what is desired—novel but accurate content—and what is delivered—parroted but inaccurate content.

Even if accuracy could be improved through fine-tuning and copyright worries could be resolved, this only addresses factual cases. People often ask LLMs for something vaguer, such as an explanation for how the LLM is solving a complex math problem, or an example of how RLHF works. These explanations are often confabulations that sound plausible but are inaccurate (Agarwal et al. 2024). While an expert can recognize these responses as nonsense, non-experts will likely treat them as plausible. They might, as a result, post these nonsensical results on forums and webpages (Sobiesk and Price 2022). This has lead to worries that a deluge of low-quality synthetic content might pollute the internet, rendering it even more difficult to determine what is accurate (Hoel 2024).

The problems with accuracy are indicative of the broader problem of conforming to norms. The most reliable way to abide by norms is simply to be normatively competent—that is, to be both cognitively capable of responding to norms, and volitionally capable of conforming to them. LLMs can clearly state what the norms are and evaluate passages for some norm-violations (this is how the training model in RLAI is trained, after all). But responsiveness to many norms requires some insight into the underlying facts about the world, such as what is true (Yiu et al. 2024). In many cases, what is most statistically likely is not what is true, as in the case of conspiracy theories (Bump 2024). Similarly, since most LLMs are trained on English and fine-tuned by Western, Educated, Industrialized, Rich, and Democratic (WEIRD) programmers, there will be many cases where the most statistically likely answer will be appropriate only for a narrow subset of the global population. As long as the model is using statistical likelihood as a proxy for appropriateness, it will fail to satisfy norms.

Even if the model had access to the world and thus could reliably recognize what would satisfy a norm, they still lack the volitional capacity to reliably conform their behavior to what is appropriate. Saying the right thing in many contexts requires thinking before we speak, where “thinking” involves searching through the space of possible answers, sampling potential options, and only choosing the right one by ensuring it accords with the appropriate norms. For example, writing a sensitive email often involves multiple drafts and numerous rewrites before finally being sent. But, even when fine-tuned, LLMs still proceed one word at a time. As a result, instead of planning their answer, they typically resort to previously successful scripts regardless of nuances of the task (Valkeeman et al. 2023; Wu et al. 2023). Thus, while models might occasionally arrive at the right answer in a typical context, they are unreliable and often nonsensical when encountering atypical cases (Tang et al. 2023; McCoy et al. 2023; Razeghi et al. 2023). Addressing this problem will require not just that LLMs learn to think before they speak (or, rather, sample possible responses), but also possess an adequate world-model in order to determine which responses are best (LeCun 2022).

The upshot is that the fine-tuning provided by RL is too coarse-grained. While it can push models in the right direction, it cannot ensure they do not select false answers or inappropriate responses. This stems from an underlying lack of normative competency: the models do not know what answers will satisfy a norm, and they cannot select an appropriate answer before typing it out.

1. **Conclusion**

This paper lays out the limits of contemporary LLMs and their fine-tuning methods. The central problem facing these models is simply that conversing with humans is more than just avoiding saying the wrong thing. We also expect people to say the right thing—to know what they are talking about, to say it in an appropriate way, and for their comments to be useful for addressing the problem at hand. A model that builds up its thoughts word-by-word might occasionally accomplish this, but contingently rather than necessarily. Fine-tuning using RL, while impressive at avoiding the worst possible comments, is still a far cry from providing us with the reliability many tasks require.

These criticisms do not suggest that LLMs can play *no* role. There are many uses for language models besides search engines or medical guides. In cases where reliability is less important, such as in role-playing contexts or on simple coding tasks, LLMs are likely to become common. However, we need to be more explicit about the unreliability of these models, and more careful about how they are used. The breathless hype online has lead to some dangerous situations: replacing workers with AI at a suicide hotline before backtracking (Westfall 2023), or New York City introducing an AI chatbot for small businesses that provided illegal advice (Offenhartz 2024). Humans expect other speakers to be normatively competent; we simply assume the other person is a responsible agent (Mahowald et al. 2023). As long as people assume this is the case, we need to avoid deploying these systems in situations where people might trust them. Even when fine-tuned, these models are bullshit machines (Milliere 2020), and we should not forget it.

**Works Cited**

Agarwal, C., Tanneru, S. H., & Lakkaraju, H. (2024). Faithfulness vs. Plausibility: On the (Un) Reliability of Explanations from Large Language Models. arXiv preprint arXiv:2402.04614.

Aguera y Arcas, B. A. (2022). Do large language models understand us? Daedalus (Cambridge, Mass.), 151(2), 183-197. Doi:10.1162/daed\_a\_01909

Andreas, J. (2022). Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. Arxiv. Doi:10.48550/arxiv.2212.08073

Belanger, A. (7/7/2023). ChatGPT Usage Drop for the First Time as Users Turn to Uncensored Chatbots. Ars Technica.

Bender, E. M., Gebru, T., Mcmillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,doi:10.1145/3442188.3445922

Berger, E. (2022, April 4). How Florida’s ‘don’t say gay’ law could harm children’s mental health. Retrieved from The Guardian: <https://www>.theguardian.com/us-news/2022/apr/04/florida-don’t-say-gay-bill-children-mental-health

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems.

Browning, J. & LeCun, Y. 8/22/22. AI and the Limits of Language. NOE[MA. https:/](MA.%20https:/)/www.noemamag.com/ai-and-the-limits-of-language/

Browning, J. “Personhood and AI: Why large language models don’t understand us”. AI & Soc (2023). <https://doi.org/10.1007/s00146-023-01724-y>

Bump, Philip. “Doing Your Own Research Is a Good Way to End up Being Wrong – The Washington Post.” Washington Post, 17 Jan. 2024, [www.washingtonpost](http://www.washingtonpost).com/politics/2024/01/17/do-your-own-research-study/

Casper, S., Davies, X. Shi, C., Gilbert, T., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Hadfield-Menell, D. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. Arxiv. <https://doi.org/10.48550/arXiv.2307.15217>

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT’s behavior changing over time?. arXiv preprint arXiv:2307.09009.

Coelho Mollo, D.. & Millière, R. The vector grounding problem, arXiv preprint, arXiv:2304 .01481, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., & Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. Arxiv. Arxiv.2204.02311

Deng, G., Liu, Y., Li, Y., Wang, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., & Liu, Y. (2023). “Jailbreaker: Automated jailbreak across multiple large language model chatbots.” arXiv preprint arXiv:2307.08715.

Dentella, V., Murphy, E., Marcus, G., & Leivada, E. (2023). Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. Arxiv. Doi:10.48550/arxiv.2302.12313

Dziri, N.; Rashkin, H.; Linzen, T.; Reitter, D. (2022) Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. Transactions of the Association for Computational Linguistics 2022; 10 1066–1083. Doi: <https://doi>.org/10.1162/tacl\_a\_00506

El-Mahdi, E., Farhadkhani, S., Guerraoui, R., Gupta, N., Lê-Nguyên Hoang, Pinot, R., et al. (2022). *SoK: On the impossible security of very large foundation models*. Arxiv.

Fletcher, Adam. 10/24/2017. How We Built an AI to Play Street Fighter II — Can you beat it? <https://medium.com/gyroscopesoftware/how-we-built-an-ai-to-play-street-fighter-ii-can-you-beat-it-9542ba43f02b>

Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., et al. (2023). The capacity for moral self-correction in large language models. Arxiv. Doi:10.48550/arxiv.2302.07459

Ghaffery, S. (09/20/2023). ChatGPT Usage is Rising Again as Students Return to School. Bloomberg. <https://www.bloomberg.com/news/articles/2023-09-20/chatgpt-usage-is-rising-again-as-students-return-to-school#xj4y7vzkg>

Grice, P. (1975). “Logic and conversation”. In Cole, P.; Morgan, J. (eds.). *Syntax and semantics*. Vol. 3: Speech acts. New York: Academic Press.

Haugeland, J. (1979, November). Understanding Natural Language. Journal of Philosophy, 76, 619-32.

Haugeland, J. (1982). Heidegger on Being a Person. *Nous, 16*(1), 15-26.

Hofstadter, Douglas. “Gödel, Escher, Bach, and Ai.” The Atlantic, Atlantic Media Company, 10 July 2023, [www.theatlantic.com/ideas/archive/2023/07/godel-escher-bach-geb-ai/674589/](http://www.theatlantic.com/ideas/archive/2023/07/godel-escher-bach-geb-ai/674589/).

Hoel, Erik. “A.I.-Generated Garbage Is Polluting Our Culture.” The New York Times, The New York Times, 29 Mar. 2024, www.nytimes.com/2024/03/29/opinion/ai-internet-x-youtube.html.

Jackson, Claire. 4/19/2023. People Are Using A ‘Grandma Exploit’ To Break AI. Kotaku. <https://kotaku.com/chatgpt-ai-discord-clyde-chatbot-exploit-jailbreak-1850352678>

Jain, N., Han, K., Gu, A., Li, W. D., Yan, F., Zhang, T., ... & Stoica, I. (2024). LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. arXiv preprint arXiv:2403.07974.

Kalai, A. and Vempala, S. (2023). Calibrated Language Models Must Hallucinate. Arxiv. <https://doi.org/10.48550/arXiv.2311.14648>

Kasirzadeh A, Gabriel I (2023) In conversation with artificial intelligence: aligning language models with human values. Philosophy of Technology.

Kempt H, Lavie A, Lavie SK (2023) Appropriateness is all you need! ArXiv, 1–25

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., et al. (2023). Chatgpt: Jack of all trades, master of none. Elsevier BV. Doi:10.2139/ssrn.4372889

Korbak, T., Shi, K., Chen, A., Bhalerao, R., Buckley, C. L., Phang, J., et al. (2023). Pretraining language models with human preferences. Arxiv. Doi:10.48550/arxiv.2302.08582

Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054.

LeCun, Y. (2022) A path towards autonomous machine intelligence, Version 0.9.2, 2022-06-27. https://openreview.net/forum?id=BZ5a1r-kVsf

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157-173.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv preprint arXiv:2309.13638.

Mahowald, K., Ivanova, A., Fedorenko, E., Blank , I. A., Tenenbaum, J., & Kanwisher, N. (2024, January 24). Google’s powerful AI spotlights a human cognitive glitch: Mistaking fluent speech for fluent thought. The Conversation. https://theconversation.com/googles-powerful-ai-spotlights-a-human-cognitive-glitch-mistaking-fluent-speech-for-fluent-thought-185099

Milliere,R. (09/09/2020). Welcome to the next level of bullshit. Nautilus. https://nautil.us /welcome-to-the-next-level-of-bullshit-9245/

Moderator. “Temporary Policy: Generative AI (e.g., Chatgpt) Is Banned.” Meta Stack Overflow, 5 Dec. 2022, meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned.

Offenhartz, Jake. “NYC’s AI Chatbot Was Caught Telling Businesses to Break the Law. the City Isn’t Taking It Down.” AP News, AP News, 4 Apr. 2024, apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21

Orf, D. (2023, March 1). Elon Musk Building Anti-Woke AI to Rival ChatGPT: What It Means. Retrieved from Popular Mechanics: <https://www.popularmechanics.com/technology/robots/a43126181/elon-musk-anti-woke-artificial-intelligence/>

Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., ... & Legg, S. (2021). Shaking the foundations: delusions in sequence models for interaction and control. arXiv preprint arXiv:2110.10819.

Pavlick, E. (2022). Semantic structure in deep learning, Annu. Rev. Linguist. 8(1): 447–471.

Piantasodi, S.T. & Hill, F. (2022). Meaning without reference in large language models, ArXiv, pp.1–8.

Piltch, Avram. 06/11/2023. Plagiarism Engine: Google’s Content-Swiping AI Could Break the Internet. <https://www.tomshardware.com/news/google-sge-break-internet>

Razeghi, Y., Logan IV, R. L., Gardner, M., & Singh, S. (2022). Impact of pretraining term frequencies on few-shot reasoning. arXiv preprint arXiv:2202.07206.

Russell S (2019) Human-Compatible AI, Viking.

Schank, R. & Abelson, R.P. (1977). Scripts, plans, goals and understanding: An inquiry into human knowledge structures. New Jersey: Erlbaum. ISBN 0-470-99033-3.

Scott-Philips, T. (2014). Speaking Our Minds: Why human communication is different, and how language evolved to make it special, Red Globe Press.

Sobieszek, Adam & Price, Tadeusz (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. Minds and Machines 32 (2):341-364.

Sutton, R.; Barto, A. (1998). Reinfocement Learning: An Introduction. MIT Press.

Tangerman, V. (2023, February 15). Microsoft's Bing AI Now Threatening Users Who Provoke It. Retrieved from Futurism: https://futurism.com/microsoft-bing-ai-threatening

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., et al. (2022). Lamda: Language models for dialog applications. Arxiv. doi:10.48550/arXiv.2201.08239

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30, Long Beach, CA.

Valmeekam, K., Olmo, A., Sreeharan, S., & Kambhampati, S. (2023). Large Language Models Still Can’t Plan. Arxiv, 1-21.

Wallace, R.J. (1994). Responsibility and the Moral Sentiments, Harvard University Press.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., et al. (2021). Challenges in detoxifying language models Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.210

Whitney, Lance. “I Tried X’s ‘anti-Woke’ Grok AI Chatbot. the Results Were the Opposite of What I Expected.” *ZDNET*, 23 Dec. 2023, www.zdnet.com/article/i-tried-xs-anti-woke-grok-ai-chatbot-the-results-were-the-opposite-of-what-i-expected/.

Wiggers, Kyle. “OpenAI’s Chatbot Store Is Filling up with Spam.” *TechCrunch*, 20 Mar. 2024, techcrunch.com/2024/03/20/openais-chatbot-store-is-filling-up-with-spam/.

Williams, A., Miceli, M., & Gebru, T. (2022, October 13). The Exploited Labor behnd AI. Retrieved from NOEMA: <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... & Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. arXiv preprint arXiv:2307.02477.

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet). Perspectives on Psychological Science, 0(0). <https://doi.org/10.1177/17456916231201401>

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv preprint arXiv:2402.19473.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., et al. (2019). Fine-tuning language models from human preferences. doi:10.48550/arxiv.1909.08593