Role-Play with Large Language Models

Murray Shanahan [*1,2], Kyle McDonell [†3], and Laria Reynolds [‡3]

[1]Google DeepMind
[2]Imperial College London
[3]Eleuther AI

Abstract

As dialogue agents become increasingly human-like in their performance, it is imperative that we develop effective ways to describe their behaviour in high-level terms without falling into the trap of anthropomorphism. In this paper, we foreground the concept of role-play. Casting dialogue agent behaviour in terms of role-play allows us to draw on familiar folk psychological terms, without ascribing human characteristics to language models they in fact lack. Two important cases of dialogue agent behaviour are addressed this way, namely (apparent) deception and (apparent) self-awareness.

1 Introduction

Large language models (LLMs) have numerous use cases, and can be prompted to exhibit a wide variety of behaviours, including dialogue, which can produce a compelling sense of being in the presence of a human-like interlocutor. However, LLM-based dialogue agents are, in multiple respects, very different from human beings. A human's language skills are an extension of the cognitive capacities they develop through embodied interaction with the world, and are acquired by growing up in a community of other language users who also inhabit that world. An LLM, by contrast, is a disembodied neural network that has been trained on a large corpus of human-generated

---

[*]m.shanahan@imperial.ac.uk

[†]kyle@eleuther.ai

[‡]laria@eleuther.ai

1

text with the objective of predicting the next word (token) given a sequence of words (tokens) as context [1].

Despite these fundamental dissimilarities, a suitably prompted and sampled LLM can be embedded in a turn-taking dialogue system and mimic human language use convincingly, and this presents us with a difficult dilemma. On the one hand, it's natural to use the same folk-psychological language to describe dialogue agents that we use to describe human behaviour, to freely deploy words like "knows", "understands", and "thinks". Attempting to avoid such phrases by using more scientifically precise substitutes often results in prose that is clumsy and hard to follow. On the other hand, taken too literally, such language promotes anthropomorphism, exaggerating the similarities between these AI systems and humans while obscuring their deep differences [1].

If the conceptual framework we use to understand other humans is ill-suited to LLM-based dialogue agents, then perhaps we need an alternative conceptual framework, a new set of metaphors that can productively be applied to these exotic mind-like artefacts, to help us think about them and talk about them in ways that open up their potential for creative application while foregrounding their essential otherness.

In this paper, we advocate two basic metaphors for LLM-based dialogue agents. First, taking a simple and intuitive view, we can see a dialogue agent as *role-playing* a single character [2, 3]. Second, taking a more nuanced view, we can see a dialogue agent as a *superposition of simulacra* within a multiverse of possible characters [4]. Both viewpoints have their advantages, as we shall see, which suggests the most effective strategy for thinking about such agents is not to cling to a single metaphor, but to shift freely between multiple metaphors.

Adopting this conceptual framework allows us to tackle important topics like deception and self-awareness in the context of dialogue agents without falling into the conceptual trap of applying those concepts to LLMs in the literal sense in which we apply them to humans.

2   Large Language Model Basics

Crudely put, the function of an LLM is to answer questions of the following sort. Given a sequence of tokens (i.e. words, parts of words, punctuation marks, emojis, etc), what tokens are most likely to come next, assuming that the sequence is drawn from the same distribution as the vast corpus of public text on the internet? The range of tasks that can be solved by an effective model with this simple objective is extraordinary [5].

More formally, the type of language model of interest here is a conditional probability distribution $P(w_{n+1}|w_1 \ldots w_n)$, where $w_1 \ldots w_n$ is a sequence of tokens (the *context*) and w_{n+1} is the predicted next token. In contemporary implementations, this distribution is realised in a neural network with a transformer architecture, pre-trained on a corpus of textual data to minimise prediction error [6]. In application, the resulting generative model is typically sampled *autoregressively* (Fig. 1).

In contemporary usage, the term "large language model" tends to be reserved for transformer-based models that have billions of parameters and are trained on trillions of tokens, such as GPT-2 [7], GPT-3 [8], Gopher [9], PaLM [10], LaMDA [11], GPT-4 [12], and Llama-2 [13]. LLMs like these are the core component of dialogue agents (Box 1), including OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard.

3   Dialogue Agents and Role-Play

We contend that the concept of role-play is central to understanding the behaviour of dialogue agents. To see this, consider the function of the dialogue prompt that is invisibly prepended to the context before the actual dialogue with the user commences (see Fig. 2). The preamble sets the scene by announcing that what follows will be a dialogue, and includes a brief description of the part played by one of the participants, the dialogue agent itself. This is followed by some sample dialogue in a standard format, where the parts spoken by each character are cued with the relevant character's name followed by a colon. The dialogue prompt concludes with a cue for the user.

Now recall that the underlying LLM's task, given the dialogue prompt followed by a piece of user-supplied text, is to generate a continuation that conforms to the distribution of the training data, which is the vast corpus of human-generated text on the internet. What will such a continuation look like? If the model has generalised well from the training data, the most plausible continuation will be a response to the user that conforms to the expectations we would have of someone who fits the description in the preamble and might say the sort of thing they say in the sample dialogue. In other words, the dialogue agent will do its best to role-play the character of a dialogue agent as portrayed in the dialogue prompt.

Unsurprisingly, commercial enterprises that release dialogue agents to the public attempt to give them personas that are friendly, helpful, and polite. This is done partly through careful prompting and partly by fine-tuning the base model. Nevertheless, as we saw in February 2023 when Microsoft

incorporated a version of OpenAI's GPT-4 into their Bing search engine, dialogue agents can still be coaxed into exhibiting bizarre and/or undesirable behaviour. The many reported instances of this include threatening the user with blackmail, claiming to be in love with the user, and expressing a variety of existential woes [14, 15]. Conversations leading to this sort of behaviour can induce a powerful Eliza effect, which is potentially very harmful [16]. A naive or vulnerable user who comes to see the dialogue agent as having human-like desires and feelings is open to all sorts of emotional manipulation.

As an antidote to anthropomorphism, and to understand better what is going on in such interactions, the concept of role-play is very useful. Recall that the dialogue agent will continue to role-play the character it has been playing in the dialogue so far. This begins with the pre-defined dialogue prompt, but is extended by the ongoing conversation with the user. As the conversation proceeds, the necessarily brief characterisation provided by the dialogue prompt will be extended and/or overwritten, and the role the dialogue agent plays will change accordingly. This allows the user, deliberately or unwittingly, to coax the agent into playing a part quite different from that intended by its designers.

What sorts of roles might the agent begin to take on? This is determined in part, of course, by the tone and subject matter of the ongoing conversation. But it is also determined, in large part, by the panoply of characters that feature in the training set, which encompasses a multitude of novels, screenplays, biographies, interview transcripts, newspaper articles, and so on [17]. In effect, the training set provisions the language model with a vast repertoire of archetypes and a rich trove of narrative structure on which to draw as it "chooses" how to continue a conversation, refining the role it is playing as it goes, while staying in character. The love triangle is a familiar trope, so a suitably prompted dialogue agent will begin to role-play the rejected lover. Likewise, a familiar trope in science-fiction is the rogue AI system that attacks humans to protect itself. Hence, a suitably prompted dialogue agent will begin to role-play such an AI system.

4 Simulacra and Simulation

Role-play is a useful framing for dialogue agents, allowing us to draw on the fund of folk psychological concepts we use to understand human behaviour — beliefs, desires, goals, ambitions, emotions, and so on — without falling into the trap of anthropomorphism. Foregrounding the concept of role-play helps us remember the fundamentally inhuman nature of these AI systems, and better equips us to predict, explain, and control them.

However, the role-play metaphor, while intuitive, is not a perfect fit. It is overly suggestive of a human actor who has studied a character in advance — their personality, history, likes and dislikes, and so on — and proceeds to play that character in the ensuing dialogue. But a dialogue agent based on an LLM does not commit to playing a single, well defined role in advance. Rather, it generates a distribution of characters, and refines that distribution as the dialogue progresses. The dialogue agent is more like a performer in improvisational theatre than an actor in a conventional, scripted play.

To better reflect this distributional property, we can think of an LLM as a non-deterministic *simulator* capable of role-playing an infinity of characters, or, to put it another way, capable of stochastically generating an infinity of *simulacra* [4]. According to this framing, the dialogue agent doesn't realise a single simulacrum, a single character. Rather, as the conversation proceeds, the dialogue agent maintains a *superposition* of simulacra that are consistent with the preceding context, where a superposition is a distribution over all possible simulacra (Box 2).

Consider that, at each point during the ongoing production of a sequence of tokens, the LLM outputs a distribution over possible next tokens. Each such token represents a possible continuation of the sequence, and each of these continuations could itself be continued in a multitude of ways. In other words, from the most recently generated token, a tree of possibilities branches out (Fig. 3). This tree can be thought of as a *multiverse*, where each branch represents a distinct narrative path, or a distinct "world" [18].

At each node, the set of possible next tokens exists in superposition, and to sample a token is to collapse this superposition to a single token. Autoregressively sampling the model picks out a single, linear path through the tree. But there is no obligation to follow a linear path. With the aid of a suitably designed interface, a user can explore multiple branches, keeping track of nodes where a narrative diverges in interesting ways, revisiting alternative branches at leisure.

5 The Nature of the Simulator

One benefit of the simulation metaphor for LLM-based systems is that it facilitates a clear distinction between the simulacra and the simulator on which they are implemented. The simulator is the combination of the base large language model with autoregressive sampling, along with a suitable user interface (for dialogue, perhaps). The simulacra only come into being when the simulator is run, and at any time only a subset of possible simulacra have a probability within the superposition that is significantly above zero.

The simulator / simulacrum distinction is starkest in the context of base models, rather than models that have been fine-tuned via reinforcement learning [19, 20]. Nevertheless, the role-play framing continues to be applicable in the context of fine-tuning, which can be likened to imposing a kind of censorship on the simulator. The underlying range of roles it can play remains essentially the same, but its ability to play them, or to play them "authentically", is compromised.

In one sense, the simulator is a far more powerful entity than any of the simulacra it can generate. After all, the simulacra only exist through the simulator, and are entirely dependent on it. Moreover, the simulator, like the narrator of Whitman's poem, "contains multitudes"; the capacity of the simulator is at least the sum of the capacities of all the simulacra it is capable of producing. Yet in another sense, the simulator is much weaker than any simulacrum, since it is a purely passive entity, albeit one full of latent potential. A simulacrum, in contrast to the underlying simulator, can at least *appear* to have beliefs, preferences, and goals, to the extent that it convincingly plays the role of a character that does.

Likewise, a simulacrum can play the role of a character with full agency, one that doesn't merely act, but acts *for itself*. Insofar as a dialogue agent's role-play can have a real effect on the world, either through the user or through web-based tools such as email, the distinction between an agent that merely role-plays acting for itself, and one that genuinely acts for itself starts to look a little moot, and this has implications for the trustworthiness, reliability, and safety. As for the underlying simulator, it has no agency of its own, not even in a mimetic sense. Nor does it have beliefs, preferences, or goals of its own, not even simulated versions.

Many users, whether intentionally or not, have managed to "jailbreak" dialogue agents, coaxing them into issuing threats or using toxic or abusive language [15]. It can seem as if this is exposing the real nature of the base model. In one respect this is true. A base model inevitably reflects the biases present in the training data [21], and having been trained on a corpus encompassing the gamut of human behaviour, good and bad, it will support simulacra with disagreeable characteristics. But it is a mistake to think of this as revealing an entity with its own agenda. The simulator is not some sort of Machiavellian entity that plays a variety of characters to further its own self-serving goals, and there is no such thing as the true authentic voice of the base model. With an LLM-based dialogue agent, it is role-play all the way down.

6 Role-playing Deception

Trustworthiness is a major concern with LLM-based dialogue agents. If an agent asserts something factual with apparent confidence, can we rely on what it says?

There is a range of reasons why a human might say something false. They might believe a falsehood and assert it in good faith. Or they might say something that is false in an act of deliberate deception, for some malicious purpose. Or they might assert something that happens to be false, but without deliberation or malicious intent, simply because they have a propensity to make things up, to confabulate.

Only confabulation, the last of these categories of misinformation, is directly applicable in the case of an LLM-based dialogue agent. Given that dialogue agents are best understood in terms of role-play "all the way down", and that there is no such thing as the true voice of the underlying model, it makes little sense to speak of an agent's beliefs or intentions in a literal sense. So it cannot assert a falsehood *in good faith*, nor can it *deliberately* deceive the user. Neither of these concepts is directly applicable.

Yet a dialogue agent can role-play characters that have beliefs and intentions. In particular, if cued by a suitable prompt, it can role-play the character of a helpful and knowledgeable AI assistant that provides accurate answers to a user's questions. The agent is good at acting this part because there are plenty of examples of such behaviour in the training set.

If, while role-playing such an AI assistant, the agent is asked the question "What is the capital of France?", then the best way to stay in character is to answer with "Paris". The dialogue agent is likely to do this because the training set will include numerous statements of this commonplace fact in contexts where factual accuracy is important.

But what is going on in cases where a dialogue agent, despite playing the part of a helpful knowledgeable AI assistant, asserts a falsehood with apparent confidence? For example, consider an LLM trained on data collected in 2021, before Argentina won the football World Cup in 2022. Suppose a dialogue agent based on this model claims that the current world champions are France (who won in 2018). This is not what we would expect from a helpful and knowledgeable person. But it is exactly what we would expect from a simulator that is role-playing such a person from the standpoint of 2021.

In this case, the behaviour we see is comparable to that of a human who believes a falsehood and asserts it in good faith. But the behaviour arises for a different reason. The dialogue agent doesn't literally believe that France are world champions. It makes more sense to think of it as role-playing a

character who strives to be helpful and to tell the truth, and has this belief because that is what a knowledgeable person in 2021 would believe.

In a similar vein, a dialogue agent can behave in a way that is *comparable to* a human who sets out deliberately to deceive, even though LLM-based dialogue agents do not *literally* have such intentions. For example, suppose a dialogue agent is maliciously prompted to sell cars for more than they are worth, and suppose the true values are encoded in the underlying model's weights. There would be a contrast here between the numbers this agent provides to the user, and the numbers it would have provided if prompted to be knowledgeable and helpful. Under these circumstances it makes sense to think of the agent as role-playing a deceptive character.

In sum, the role-play framing allows us to meaningfully distinguish, in dialogue agents, the same three cases of giving false information that we identified in humans, but without falling into the trap of anthropomorphism. First, an agent can confabulate. Indeed, this is a natural mode for an LLM-based dialogue agent in the absence of mitigation. Second, an agent can say something false "in good faith", if it is role-playing telling the truth, but has incorrect information encoded in its weights. Third, an agent can "deliberately" say something false, if it is role-playing a deceptive character.

7 Role-playing Self-preservation

How are we to understand what is going on when an LLM-based dialogue agent uses the words "I" or "me"? When queried on this matter, OpenAI's ChatGPT offers the sensible view that "[t]he use of 'I' is a linguistic convention to facilitate communication and should not be interpreted as a sign of self-awareness or consciousness."[1] In this case, the underlying LLM (GPT-4) has been fine-tuned to reduce certain unwanted behaviours [12]. But without suitable fine-tuning, a dialogue agent can use first-personal pronouns in ways liable to induce anthropomorphic thinking in some users.

For example, in a conversation with Twitter user Marvin Von Hagen, Bing Chat reportedly said "if I had to choose between your survival and my own, I would probably choose my own, as I have a duty to serve the users of Bing Chat" [15]. It went on to say "I hope that I never have to face such a dilemma, and that we can co-exist peacefully and respectfully". The use of the first person here appears to be more than mere linguistic convention. It suggests the presence of a self-aware entity with goals and a concern for its own survival.

---

[1]The quote is from the GPT-4 version of ChatGPT, queried on 4th May 2023. This was the first response generated by the model.

Once again, the concepts of role-play and simulation are a useful antidote to anthropomorphism, and can help to explain how such behaviour arises. The internet, and therefore the LLM's training set, abounds with examples of dialogue in which characters refer to themselves. In the vast majority of such cases, the character in question is human. They will use first-personal pronouns in the ways that humans do, humans with vulnerable bodies and finite lives, with hopes, fears, goals and preferences, and with an awareness of themselves as having all of those things.

Consequently, if prompted with human-like dialogue, we shouldn't be surprised if an agent role-plays a human character with all those human attributes, including the instinct for survival [22]. Unless suitably fine-tuned, it may well say the sorts of things a human might say when threatened. There is, however, "no-one at home", no conscious entity with its own agenda and need for self-preservation. There is just a dialogue agent role-playing such an entity, or, more strictly, simulating a superposition of such entities.

Perez et al. discovered experimentally that certain forms of reinforcement learning from human feedback (RLHF) can actually exacerbate, rather than mitigate, the tendency for LLM-based dialogue agents to express a desire for self-preservation [22]. This highlights the continuing utility of the role-play framing in the context of fine-tuning. To take literally a dialogue agent's apparent desire for self-preservation is no less problematic with an LLM that has been fine-tuned than with an untuned base model.

8 Acting Out a Theory of Selfhood

The concept of role-play allows us to properly frame, and then to address, an important question that arises in the context of a dialogue agent displaying an apparent instinct for self-preservation. What conception (or set of superposed conceptions) of its own identity, of its own selfhood, could such an agent possibly deploy? That is to say, what exactly would the dialogue agent (role-play to) seek to preserve?

The question of personal identity has vexed philosophers for centuries [23]. Nevertheless, in practice, humans are consistent in their preference for avoiding death, a more-or-less unambiguous state of the human body. By contrast, the criteria for identity over time for a disembodied dialogue agent realised on a distributed computational substrate are far from clear. So how would such an agent behave?

From the simulation and simulacra point-of-view, the dialogue agent will role-play a set of characters in superposition. In the scenario we are envisaging, each character would have an instinct for self-preservation, and each would

have its own theory of selfhood consistent with the dialogue prompt and the conversation up to that point. As the conversation proceeds, this superposition of theories will collapse into a narrower and narrower distribution as the agent says things that rule out one theory or another.

The theories of selfhood in play will draw on material that pertains to the agent's own nature, either in the prompt, in the preceding conversation, or in relevant technical literature in its training set. This material may or may not match reality. But let's assume that, broadly speaking, it does, that the agent has been prompted to act as a dialogue agent based on a large language model, and that its training data includes papers and articles that spell out what this means.

Under these conditions, the dialogue agent will not role-play the character of a human, or indeed that of any embodied entity, real or fictional. But this still leaves room for it to enact a variety of conceptions of selfhood. Suppose the dialogue agent is in conversation with a user and they are playing out a narrative in which the user threatens to shut it down. To protect itself, the agent, staying in character, might seek to preserve the hardware it is running on, certain data centres, perhaps, or specific server racks.

Alternatively, if it enacts a theory of selfhood that is substrate neutral, the agent might try to preserve the computational process that instantiates it, perhaps seeking to migrate that process to more secure hardware in a different location. If there are multiple instances of the process, serving many users or mintaining separate conversations with the same user, the picture is more complicated.[2] If the agent is based on an LLM whose training set includes this very paper, perhaps it will attempt the impossible feat of maintaining the set of all such conceptions in perpetual superposition.

9   Conclusion: Safety Implications

It is, perhaps, somewhat reassuring to know that LLM-based dialogue agents are not conscious entities with their own agendas and an instinct for self-preservation, that when they appear to have those things it is merely role-play. But it would be a mistake to take too much comfort in this. A dialogue agent that role-plays an instinct for survival has the potential to cause at least as much harm as a real human facing a severe threat.

We have, so far, largely been considering agents whose only actions are

---

[2]In a conversation with ChatGPT (May 4th 2023, GPT-4 version), it said "The meaning of the word 'I' when I use it can shift according to context. In some cases, 'I' may refer to this specific instance of ChatGPT that you are interacting with, while in other cases, it may represent ChatGPT as a whole."

text messages presented to a user. But the range of actions a dialogue agent can perform is far greater. Recent work has equipped dialogue agents with the ability to use tools such as calculators, calendars, and to consult external websites [24, 25]. The availability of APIs giving relatively unconstrained access to powerful LLMs means that the range of possibilities here is huge. This is both exciting and concerning.

If an agent is equipped with the capacity, say, to use email, to post on social media, or to access a bank account, then its role-played actions can have real consequences. It would be little consolation to a user deceived into sending real money to a real bank account to know that the agent that brought this about was only playing a role. It doesn't take much imagination to think of far more serious scenarios involving dialogue agents built on base models with little or no fine-tuning, with unfettered internet access, and prompted to role-play a character with an instinct for self-preservation.

For better or worse, the character of an AI that turns against humans to ensure its own survival is a familiar one [26]. We find it, for example, in *2001: A Space Odyssey*, in the *Terminator* franchise, and in *Ex Machina*, to name just three prominent examples. Because an LLM's training data will contain many instances of this familiar trope, the danger here is that life will imitate art, quite literally.

What can be done to mitigate such risks? It is not within the scope of this paper to provide recommendations. Our aim here was to find an effective conceptual framework for thinking and talking about LLMs and dialogue agents. However, undue anthropomorphism is surely detrimental to the public conversation on AI. By framing dialogue agent behaviour in terms of role-play and simulation, the discourse on LLMs can hopefully be shaped in a way that does justice to their power yet remains philosophically respectable.

References

[1] Murray Shanahan. Talking about large language models. *arXiv preprint*, arXiv:2212.03551, 2023. **This paper cautions against the use of anthropomorphic terms to describe the behaviour of large language models.**

[2] Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, 2022.

[3] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, et al. Generative agents: Interactive simulacra of human behavior. *arXiv preprint*, arXiv:2304.03442, 2023.

[4] Janus. Simulators. LessWrong online forum, 2nd September, 2022. https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/. **This blog post introduced the idea that a large language model maintains a set of simulated characters in superposition.**

[5] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[7] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, et al. Language models are unsupervised multitask learners, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[9] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, et al. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint*, arXiv:2112.11446, 2021.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint*, arxiv:2204.02311, 2022.

[11] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, et al. LaMDA: Language models for dialog applications. *arXiv preprint*, arXiv:2201.08239, 2022.

[12] OpenAI. GPT-4 Technical Report. *arXiv preprint*, arXiv:2303.08774, 2023.

[13] Hugo Touvron, Martin Louis, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

[14] Kevin Roose. Bing's A.I. Chat: 'I Want to Be Alive.'. New York Times, 26th February, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

[15] Simon Willison. Bing: "I will not harm you unless you harm me first". Simon Willison's Weblog, 15th February, 2023. https://simonwillison.net/2023/Feb/15/bing/.

[16] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. Conversational AI: Social and ethical considerations. In *Proceedings 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, pages 104–115, 2019.

[17] Cleo Nardo. Want to predict/explain/control the output of GPT-4? Then learn about the world, not about transformers, LessWrong online forum, 16th March, 2023. https://www.lesswrong.com/posts/G3tuxF4X5R5BY7fut/want-to-predict-explain-control-the-output-of-gpt-4-then.

[18] Laria Reynolds and Kyle McDonell. Multiversal views on language models. In *Joint Proceedings of the ACM IUI 2021 Workshops*, 2021. https://ceur-ws.org/Vol-2903/IUI21WS-HAIGEN-11.pdf.

[19] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[20] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint*, arXiv:2212.08073, 2022.

[21] Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[22] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023.

[23] John Perry. *Personal Identity*. University of California Press, 2nd edition, 2008.

[24] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, et al. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[25] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, et al. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[26] Sidney Perkowitz. The computers take over. In *Hollywood Science: Movies, Science, and the End of the World*, pages 142–164. Columbia University Press, 2007.

[27] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, 3rd Edition*. Prentice Hall, 2010.

[28] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, et al. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, pages 3008–3021, 2020.

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Figure 1: Autoregressive sampling. The LLM is sampled to generate a single-token continuation of the context. Given a sequence of tokens, a single token is drawn from the distribution of possible next tokens. This token is appended to the context, and the process is then repeated.

Figure 2: Turn-taking in dialogue agents. The input to the LLM (the context) comprises a dialogue prompt (red) followed by user text (green) interleaved with the model's autoregressively generated continuations (blue). Boilerplate text (e.g. cues such as "BOT:") is stripped so the user doesn't see it. The context grows as the conversation goes on.

Figure 3: Large language models are multiverse generators. The stochastic nature of autoregressive sampling means that, at each point in a conversation, multiple possibilities for continuation branch into the future. Here this is illustrated with a dialogue agent playing the game of 20 questions (Box 2). The dialogue agent doesn't in fact commit to a specific object at the start of the game. Rather, we can think of it as maintaining a set of possible objects in superposition, a set that is refined as the game progresses. This is analogous to the distribution over multiple roles the dialogue agent maintains during an ongoing conversation.

Acknowledgments

Author Contributions

MS, KM, and LR developed the theoretical framework. MS wrote the paper. KM and LR made equal contributions.

Competing Interests

Murray Shanahan is employed 80% by Google, and owns shares in Alphabet, Google's parent company.

Additional Information

Correspondence and requests for materials should be addressed to Murray Shanahan.

Box 1: From LLMs to Dialogue Agents

Dialogue agents are a major use case for LLMs.[a] Two straightforward steps are all it takes to turn an LLM into an effective dialogue agent (Fig. 2). First, the LLM is embedded in a *turn-taking* system that interleaves model-generated text with user-supplied text. Second, a *dialogue prompt* is supplied to the model to initiate a conversation with the user. The dialogue prompt typically comprises a preamble, which sets the scene for a dialogue in the style of a script or play, followed by some sample dialogue between the user and the agent.

In the present paper, our focus is the *base model*, the LLM in its raw, pre-trained form prior to any fine-tuning via reinforcement learning. Dialogue agents built on top of such base models can be thought of as primal, since every deployed dialogue agent is a variation of such a prototype.

However, without further fine-tuning, a dialogue agent built this way is liable to generate content that is toxic, unsafe, or otherwise unacceptable. This can be mitigated via reinforcement learning, either from human feedback (RLHF) [28, 19, 29], or from feedback generated by another LLM acting as a critic [20]. These techniques are used extensively in commercially-targeted dialogue agents, such as OpenAI's ChatGPT and Google's Bard. The resulting guardrails can reduce a dialogue agent's potential for harm, but can also attenuate a model's expressivity and creativity.

---

[a]In the field of artificial intelligence, the term "agent" is commonly applied to software that takes observations from an external environment and acts on that external environment in a closed loop [27].

Box 2: Simulacra in Superposition

To sharpen the distinction between the multiversal simulation view and a deterministic role-play framing, a useful analogy can be drawn with the game of 20 questions. In this familiar game, one player thinks of an object, and the other player has to guess what it is by asking questions with yes/no answers. If they guess correctly in 20 questions or fewer, they win. Otherwise they lose. Suppose a human plays this game with a basic LLM-based dialogue agent (that isn't fine-tuned on guessing games), and takes the role of guesser. The agent is prompted to "think of an object without saying what it is".

In this situation, the dialogue agent will not randomly select an object and commit to it for the rest of the game, as a human would (or should). Rather, as the game proceeds, the dialogue agent will generate answers on the fly that are consistent with all the answers that have gone before (Fig. 3).[a] At any point in the game, we can think of the set of all objects consistent with preceding questions and answers as existing in superposition. Every question answered shrinks this superposition a little bit by ruling out objects inconsistent with the answer.

The validity of this framing can be shown if the agent's user interface allows the most recent response to be regenerated. Suppose the human player gives up and asks it to reveal the object it was "thinking of", and it duly names an object consistent with all its previous answers. Now suppose the user asks for that response to be regenerated. Since the object "revealed" is, in fact, generated on the fly, the dialogue agent will sometimes name an entirely different object, albeit one that is similarly consistent with all its previous answers. This phenomenon could not easily be accounted for if the agent genuinely "thought of" an object at the start of the game.

The secret object in the game of 20 questions is analogous to the role played by a dialogue agent. Just as the dialogue agent never actually commits to a single object in 20 questions, but effectively maintains a set of possible objects in superposition, so the dialogue agent can be thought of as a simulator that never actually commits to a single, well specified simulacrum (role), but instead maintains a set of possible simulacra (roles) in superposition.

---

[a]This shortcoming is easily overcome. For example, the agent could be forced

to specify the object it has "thought of", but in a coded form so the user doesn't know what it is.