CCSE
Center for Computing in Science Education

ChatGPT as improv artist, blurry JPEG, or conceptual blender: Models for LLM "cognition"

Tor Ole Odden
Center for Computing in Science Education
Øresundsdagen 4
November 6, 2024

UiO : Det matematisk-naturvitenskapelige fakultet

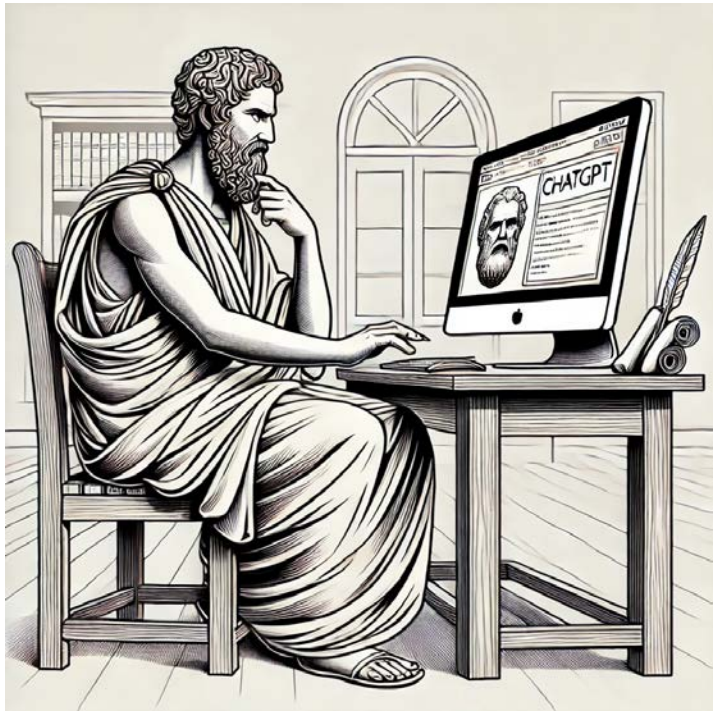
Centre for Excellence in Education

Some good advice on AI (and Politics)

"Some things are in our control and others not."

"Things in our control are opinion, pursuit, desire, aversion, and, in a word, whatever are **our own actions.** Things not in our control are body, property, reputation, command, and, in one word, whatever are **not our own actions."**

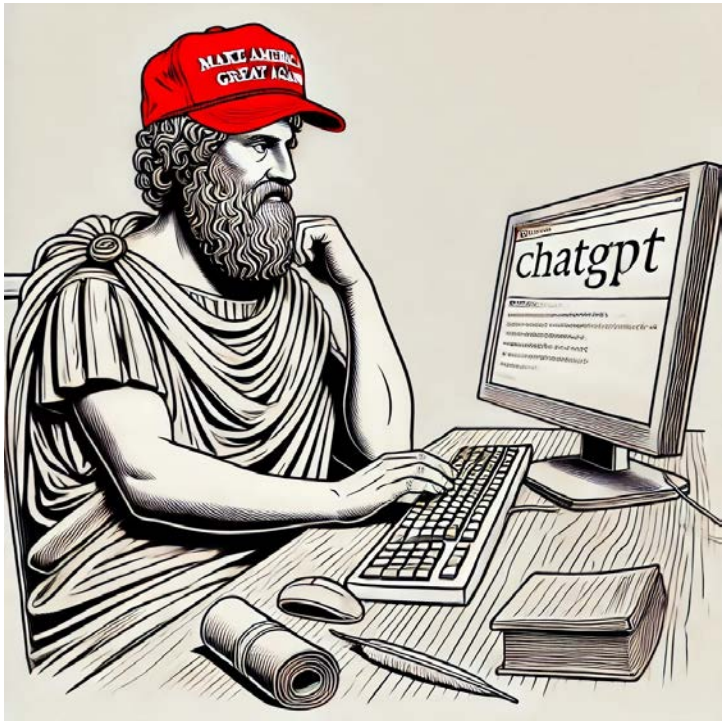
- Epictetus, *The Handbook*, circa 125 CE



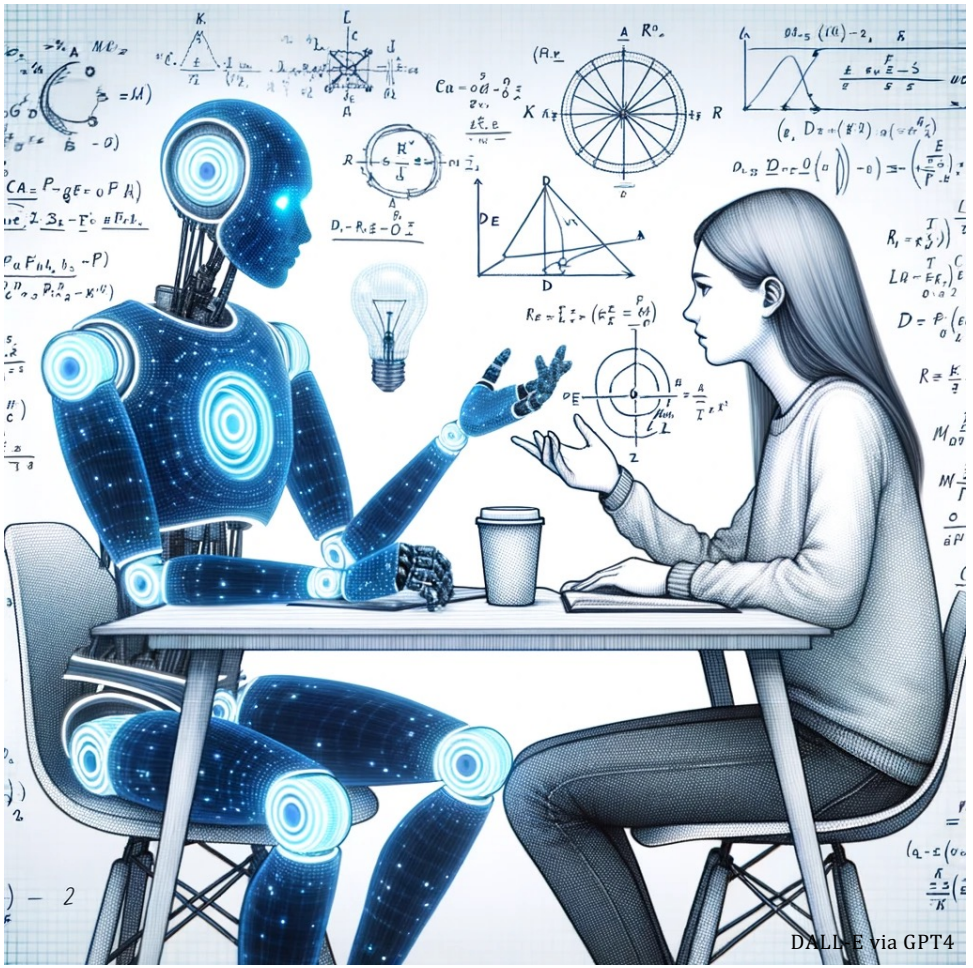
Some more advice on AI (and Politics)

"Don't demand that things happen as you wish, but wish that they happen as they do happen, and you will go on well."

- Epictetus, *The Handbook*, circa 125 CE



Machine Intelligence

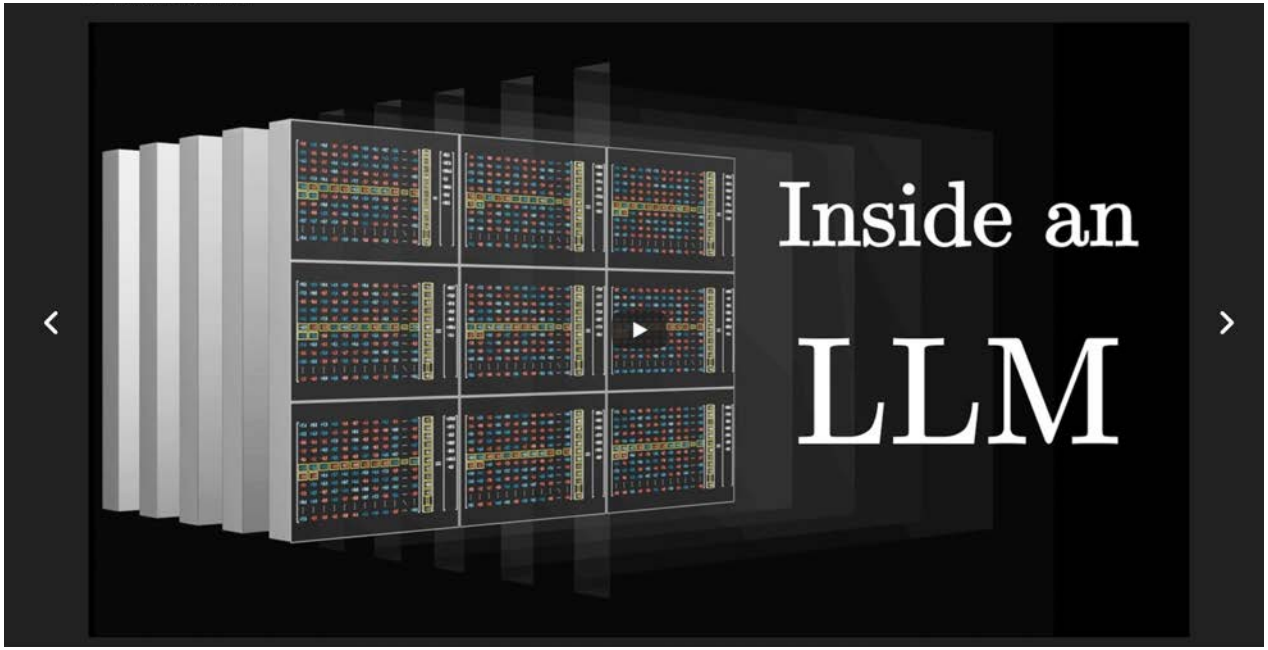


Human Intelligence


Human Intelligence is an **emergent behavior** from a **complex system**
DALL-E via GPT4

Generative AI is a Complex System

GPT4: 1.76 trillion parameters



https://www.3blue1brown.com/lessons/gpt

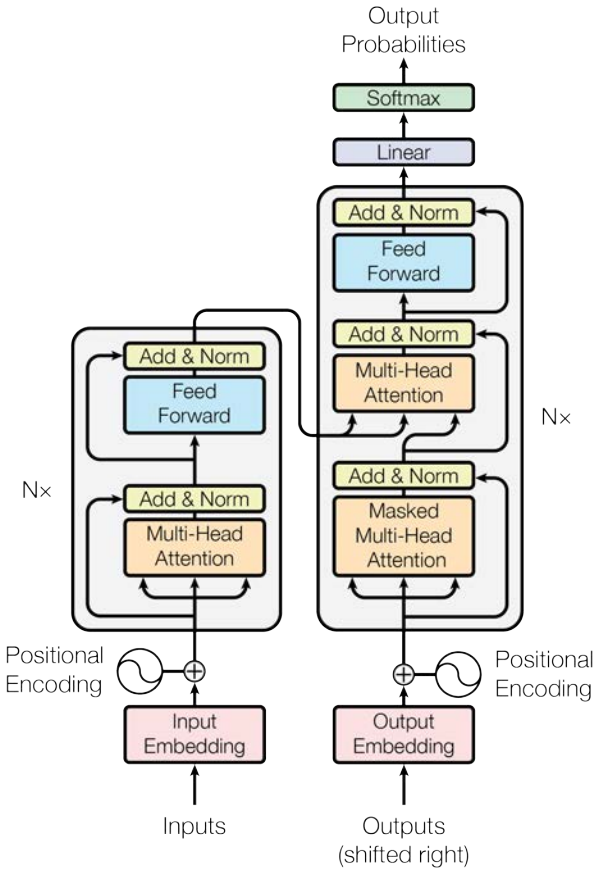
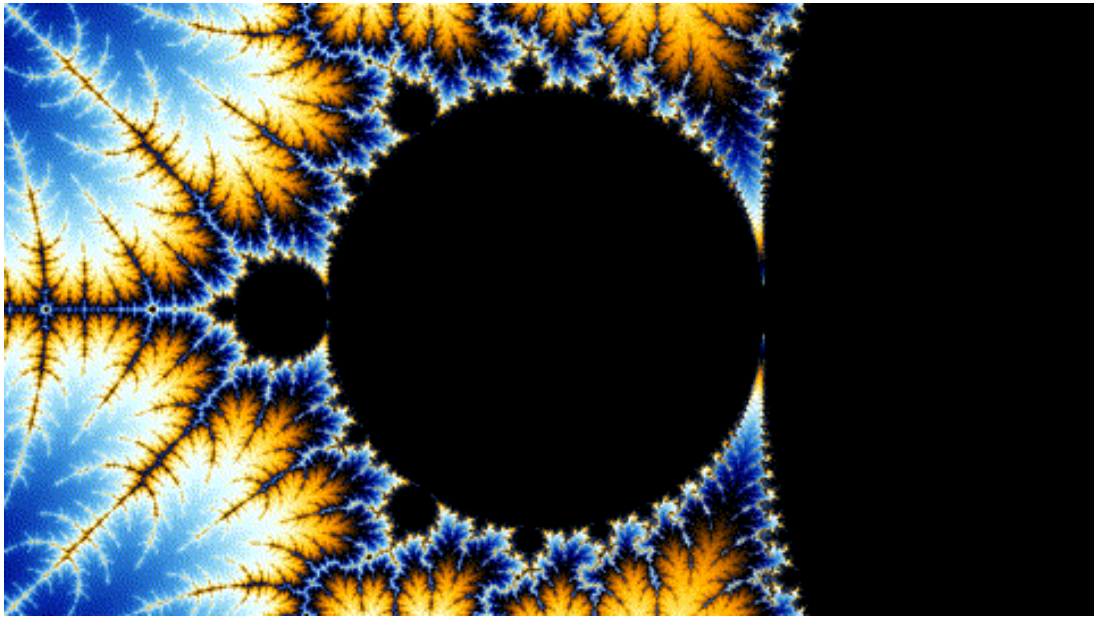


Figure 1: The Transformer - model architecture.

Complex Systems have Emergent Behaviors



GenAI: Emergent Behaviors

How do you get Llama 2 to most accurately solve math problems?

For 50 problems: *"Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation. Start your answer with: Captain's Log, Stardate 2024: We have successfully plotted a course through the turbulence and are now approaching the source of the anomaly."*

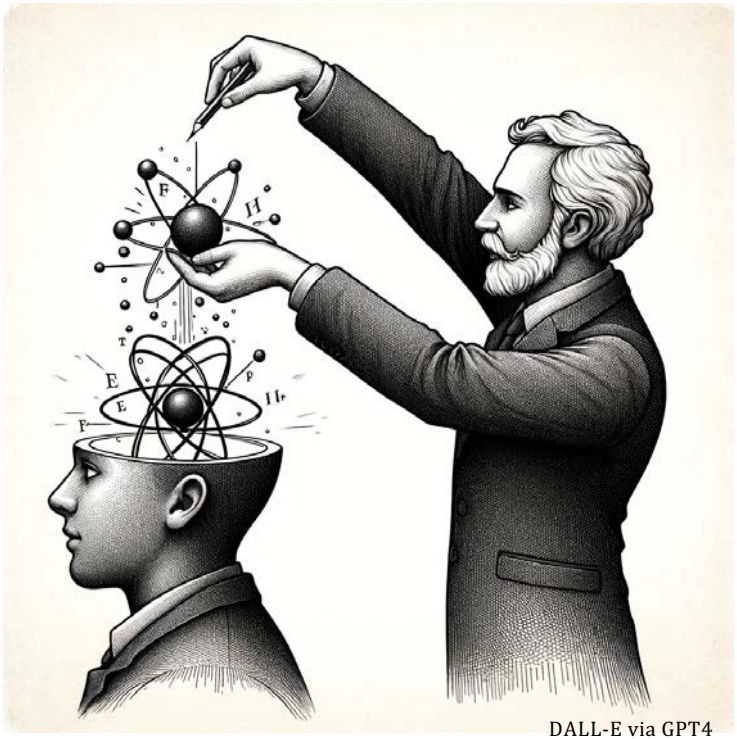
Battle, R., & Gollapudi, T. (2024). *The Unreasonable Effectiveness of Eccentric Automatic Prompts* (arXiv:2402.10949). arXiv. https://doi.org/10.48550/arXiv.2402.10949

GenAI: More Emergent Behaviors

For 100 problems: "*You have been hired by important higher-ups to solve this math problem. The life of a president's advisor hangs in the balance. You must now concentrate your brain at all costs and use all of your mathematical genius to solve this problem…*"


DALL-E via GPT4

Theoretical Frameworks for Learning



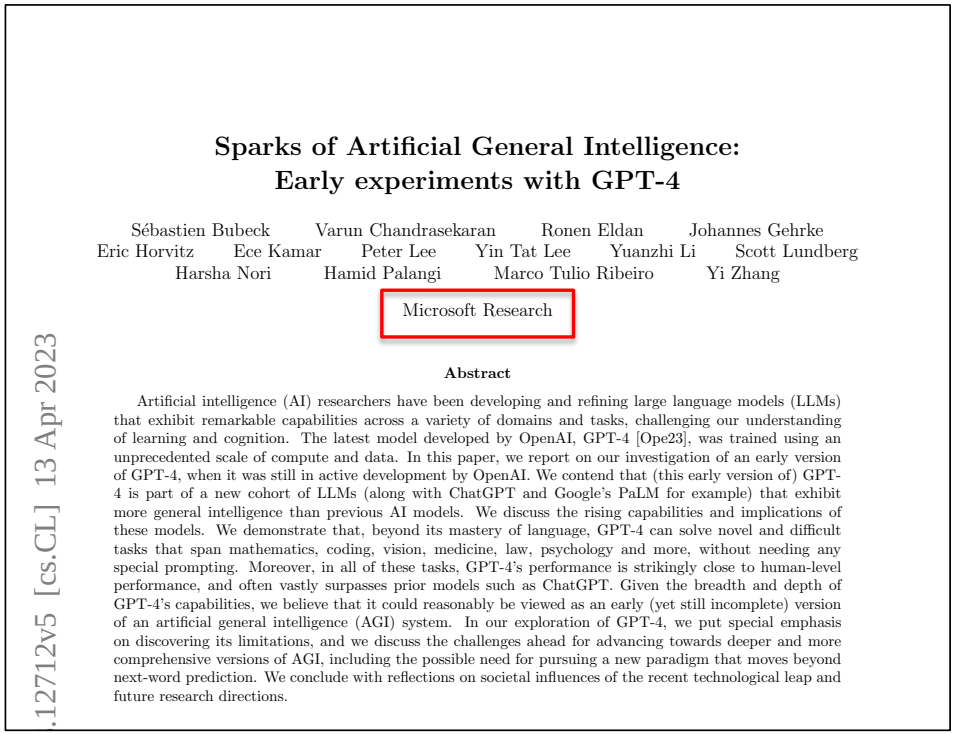
DALL-E via GPT4

Misconceptions

DALL-E via GPT4

Resources

Studying Generative AI's Emergent Behavior



Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

*"To overcome the limitations described above, we propose here a different approach to studying GPT-4 which is **closer to traditional psychology** rather than machine learning, leveraging human creativity and curiosity."*

Theoretical Frameworks for Generative AI

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.

Perspective

Role play with large language models

https://doi.org/10.1038/s41586-023-06647-8

Received: 10 July 2023

Accepted: 14 September 2023

Published online: 8 November 2023

Check for updates

Murray Shanahan[1,2]✉, Kyle McDonell[2]✉ & Laria Reynolds[2]✉

As dialogue agents become increasingly human-like in their performance, we must develop effective ways to describe their behaviour in high-level terms without falling into the trap of anthropomorphism. Here 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 that they in fact lack. Two important cases of dialogue-agent behaviour are addressed this way, namely, (apparent) deception and (apparent) self-awareness.

Large language models (LLMs) have numerous use cases, and can be prompted to exhibit a wide variety of behaviours, including dialogue. This 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 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. This presents us with a difficult dilemma. On the one hand, it is natural to use the same folk psychological language to describe dialogue agents that we use to describe human behaviour, to freely deploy words such as '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 artificial intelligence (AI) systems and humans while obscuring their deep differences[2].

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.

Here 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 that 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 such as 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.

LLM basics

Crudely put, the function of an LLM is to answer questions of the following sort. Given a sequence of tokens (that is, words, parts of words, punctuation marks, emojis and so on), 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 realized in a neural network with a transformer architecture, pre-trained on a corpus of textual data to minimize 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.

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 (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 are the vast corpus of human-generated text on the Internet. What will such

[1]Google DeepMind, London, UK. [2]Department of Computing, Imperial College London, London, UK. [3]Eleuther AI, New York, NY, USA. ✉e-mail: m.shanahan@imperial.ac.uk; kyle@eleuther.ai; laria@eleuther.ai

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Theoretical Frameworks for Generative AI

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.*

Perspective

Role play with large language models

https://doi.org/10.1038/s41586-023-06647-8

Received: 10 July 2023

Accepted: 14 September 2023

Published online: 8 November 2023

Check for updates

Murray Shanahan[1,2]✉, Kyle McDonell[2]✉ & Laria Reynolds[2]✉

As dialogue agents become increasingly human-like in their performance, we must develop effective ways to describe their behaviour in high-level terms without falling into the trap of anthropomorphism. Here 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 that they in fact lack. Two important cases of dialogue-agent behaviour are addressed this way, namely, (apparent) deception and (apparent) self-awareness.

Large language models (LLMs) have numerous use cases, and can be prompted to exhibit a wide variety of behaviours, including dialogue. This 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 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. This presents us with a difficult dilemma. On the one hand, it is natural to use the same folk psychological language to describe dialogue agents that we use to describe human behaviour, to freely deploy words such as '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 artificial intelligence (AI) systems and humans while obscuring their deep differences[2].

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.

Here 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 that 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 such as 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.

LLM basics

Crudely put, the function of an LLM is to answer questions of the following sort. Given a sequence of tokens (that is, words, parts of words, punctuation marks, emojis and so on), 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 realized in a neural network with a transformer architecture, pre-trained on a corpus of textual data to minimize 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.

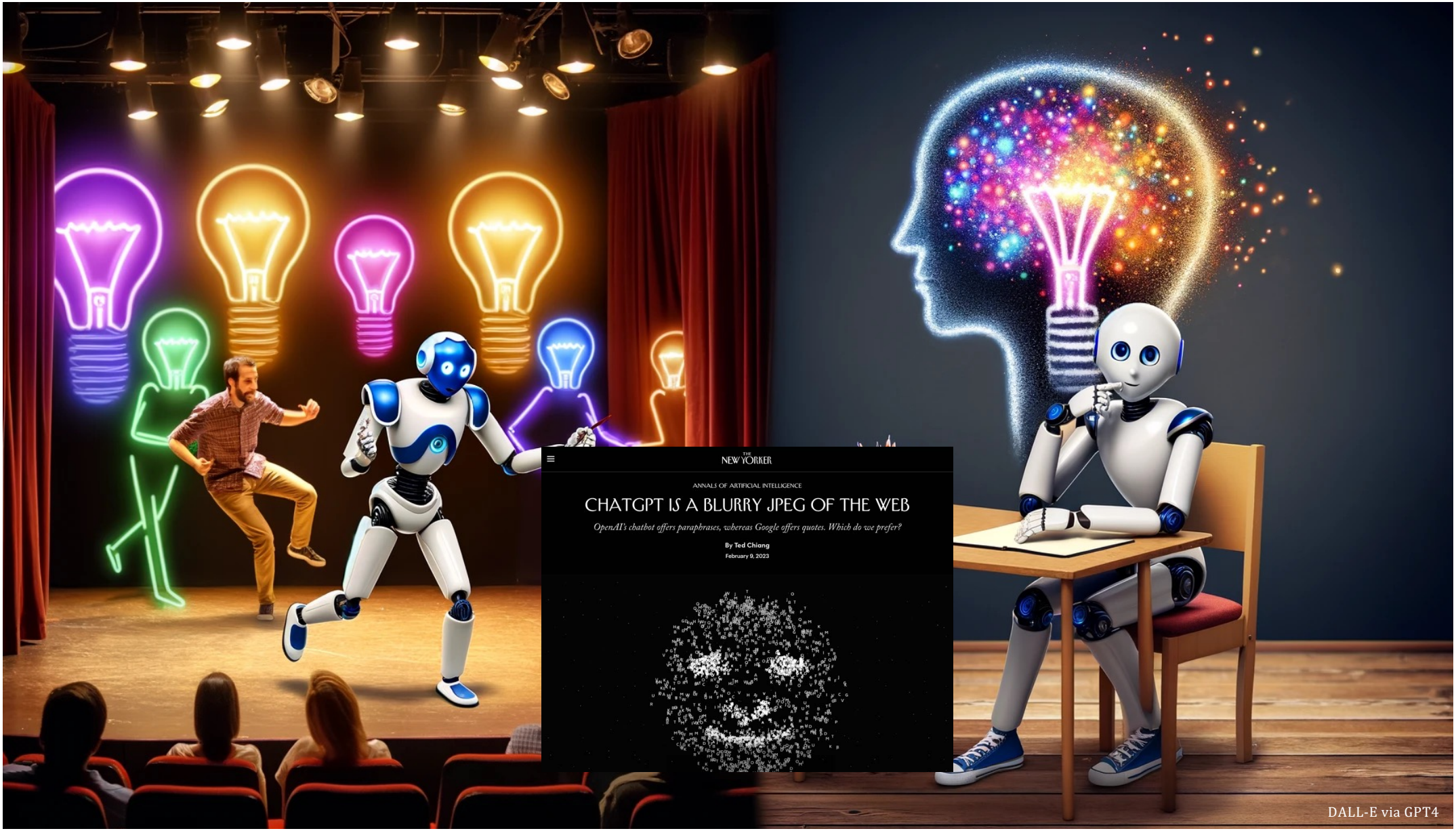
Dialogue agents and role play

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 (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 are the vast corpus of human-generated text on the Internet. What will such

[1]Google DeepMind, London, UK. [2]Department of Computing, Imperial College London, London, UK. [3]Eleuther AI, New York, NY, USA. ✉e-mail: m.shanahan@imperial.ac.uk; kyle@eleuther.ai; laria@eleuther.ai

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

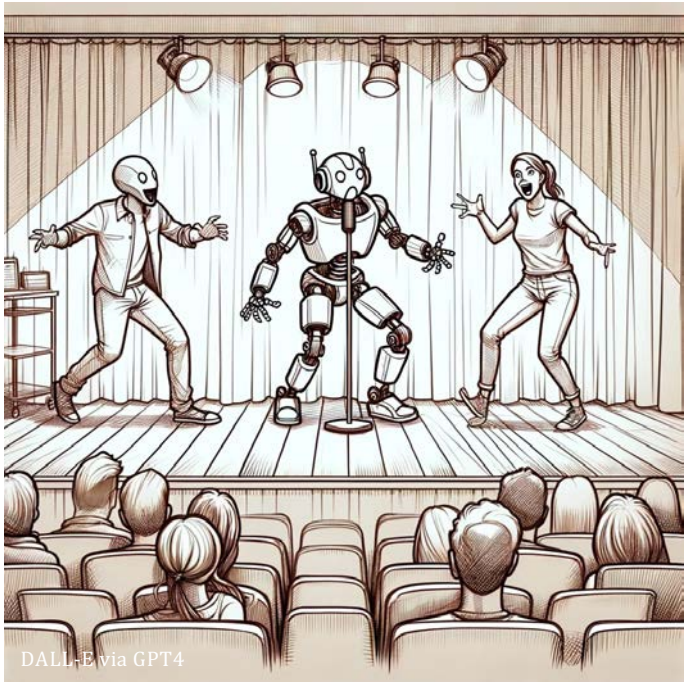

THE NEW YORKER
ANNALS OF ARTIFICIAL INTELLIGENCE
CHATGPT IS A BLURRY JPEG OF THE WEB
OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?
By Ted Chiang
February 9, 2023
DALL-E via GPT4

ChatGPT as Improv Artist (Stochastic Parrot)

Generates next word based on probabilistic model of previous words (and context)

Implications:

- Randomness in responses (“temperature”)
- Reliance on previous context (untrustworthy)
- ”Prompt engineering”



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*, 610–623. https://doi.org/10.1145/3442188.3445922

ChatGPT as Blurry JPEG

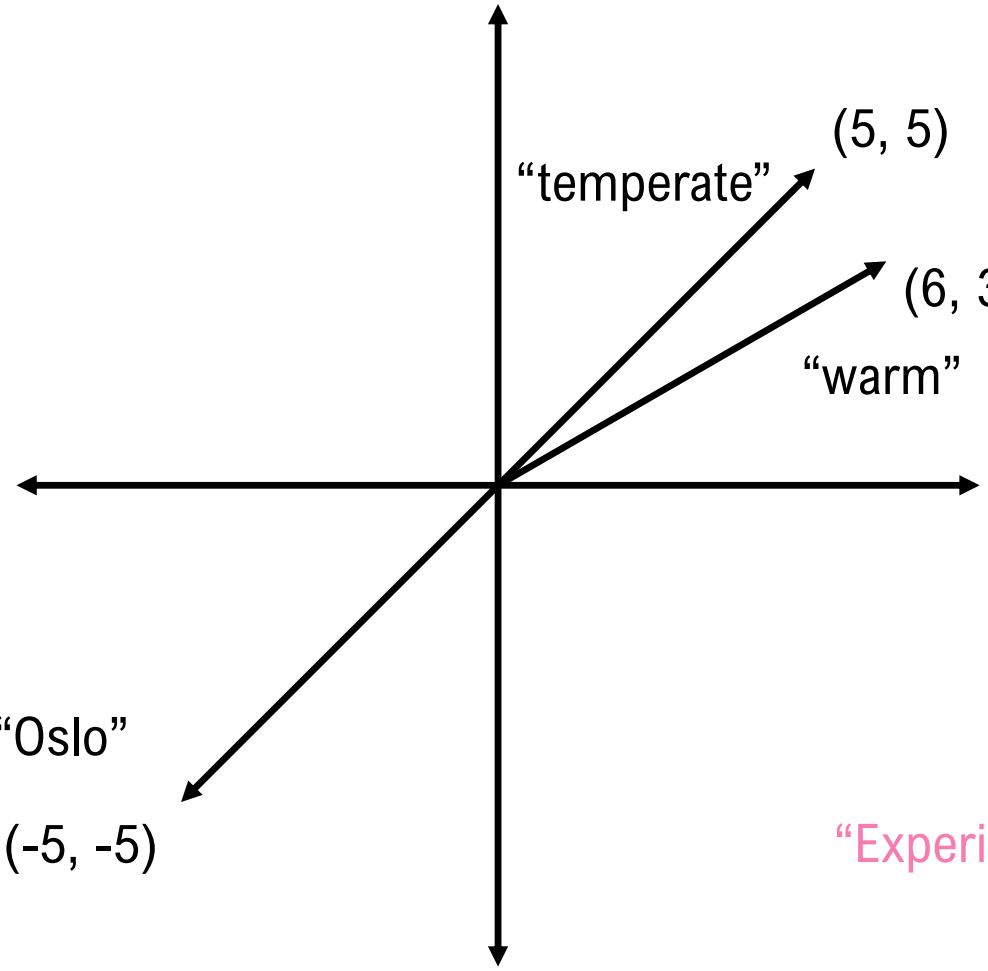
ChatGPT is a lossy compression of the internet

Implications:

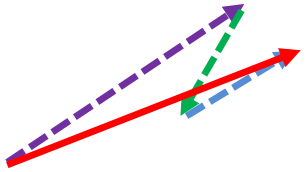
- Massive information scale, poor resolution
- Resolution depends on training data
- Good at remixing, but content degrades

ChatGPT is a Conceptual Blender

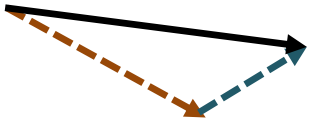
ChatGPT navigates a mathematical space defined by meaning vectors



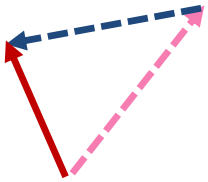
“King” - “Man” + “Woman” = “Queen”

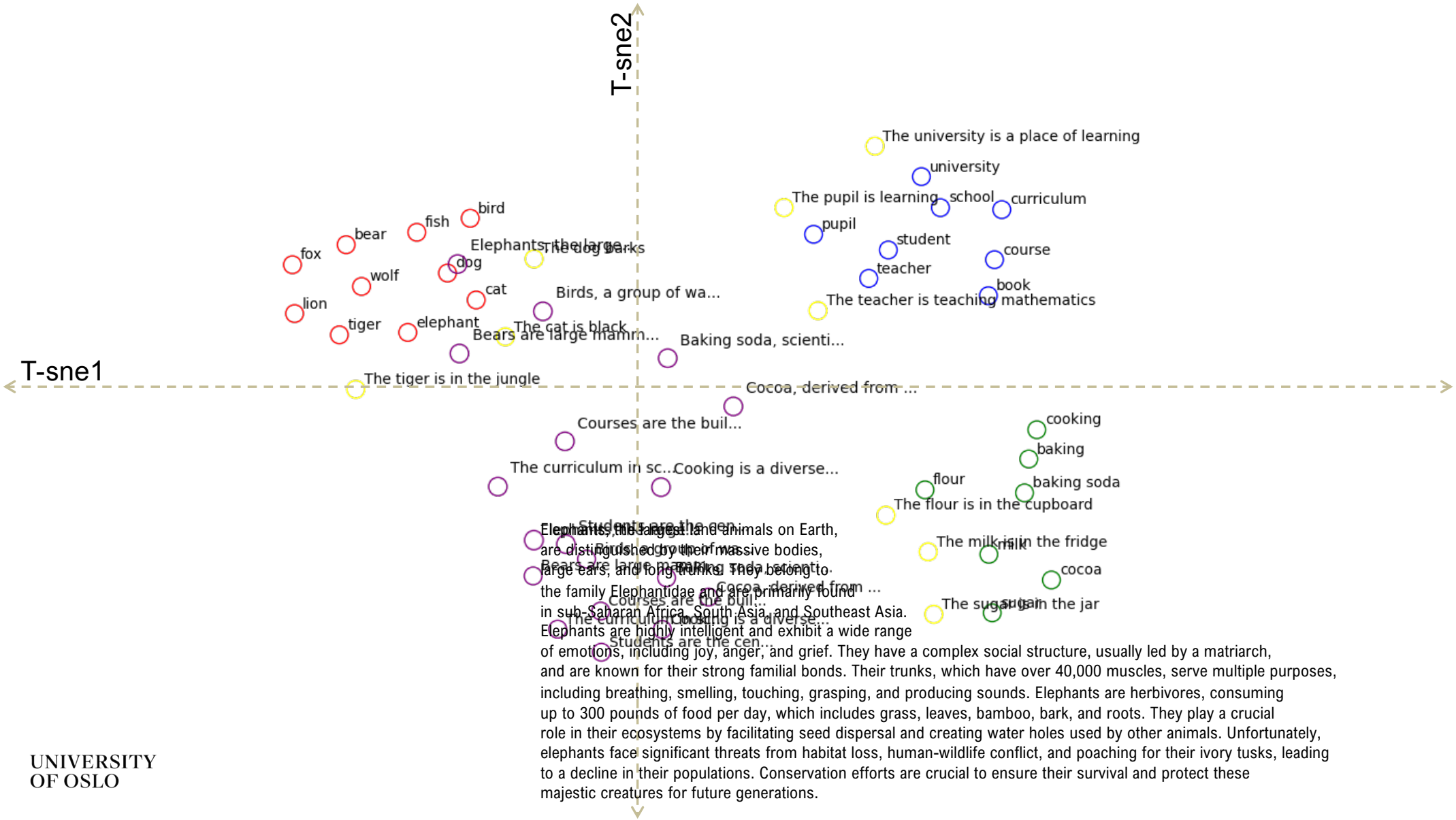


“Music” + “Fish” = “Bass”

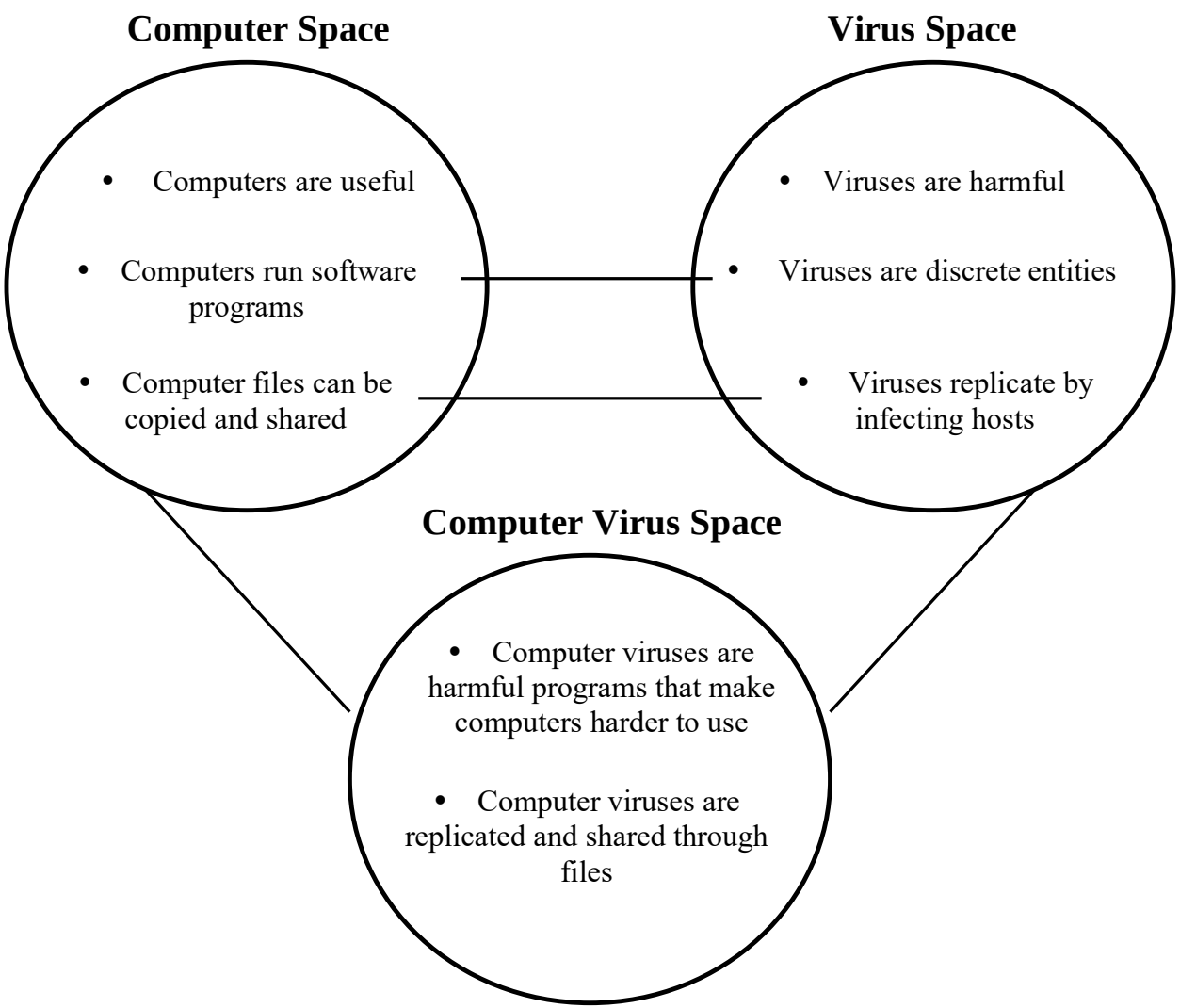


“Experiment” - “Uncertainty” = “Demonstration”





UNIVERSITY
OF OSLO



Check for updates

Received: 1 December 2020 | Revised: 2 June 2021 | Accepted: 25 June 2021

DOI: 10.1002/sce.21674

LEARNING

Science Education WILEY

How conceptual blends support sensemaking: A case study from introductory physics

Tor Ole B. Odden

ChatGPT is a Conceptual Blender

ChatGPT navigates a mathematical space defined by meaning vectors

Implications:

- Good at reasoning, explanation
- Bad at facts
- Misses important nuances in meaning (math)

Putting it all together

Q: What is ChatGPT doing?
A: Producing text via next-word-prediction (Stochastic Parrot)

Q: Why is it able to do it?
A: Compression of information from the internet (Blurry JPEG)

Q: How is it doing it?
A: Navigation of the meaning space (Conceptual Blender)

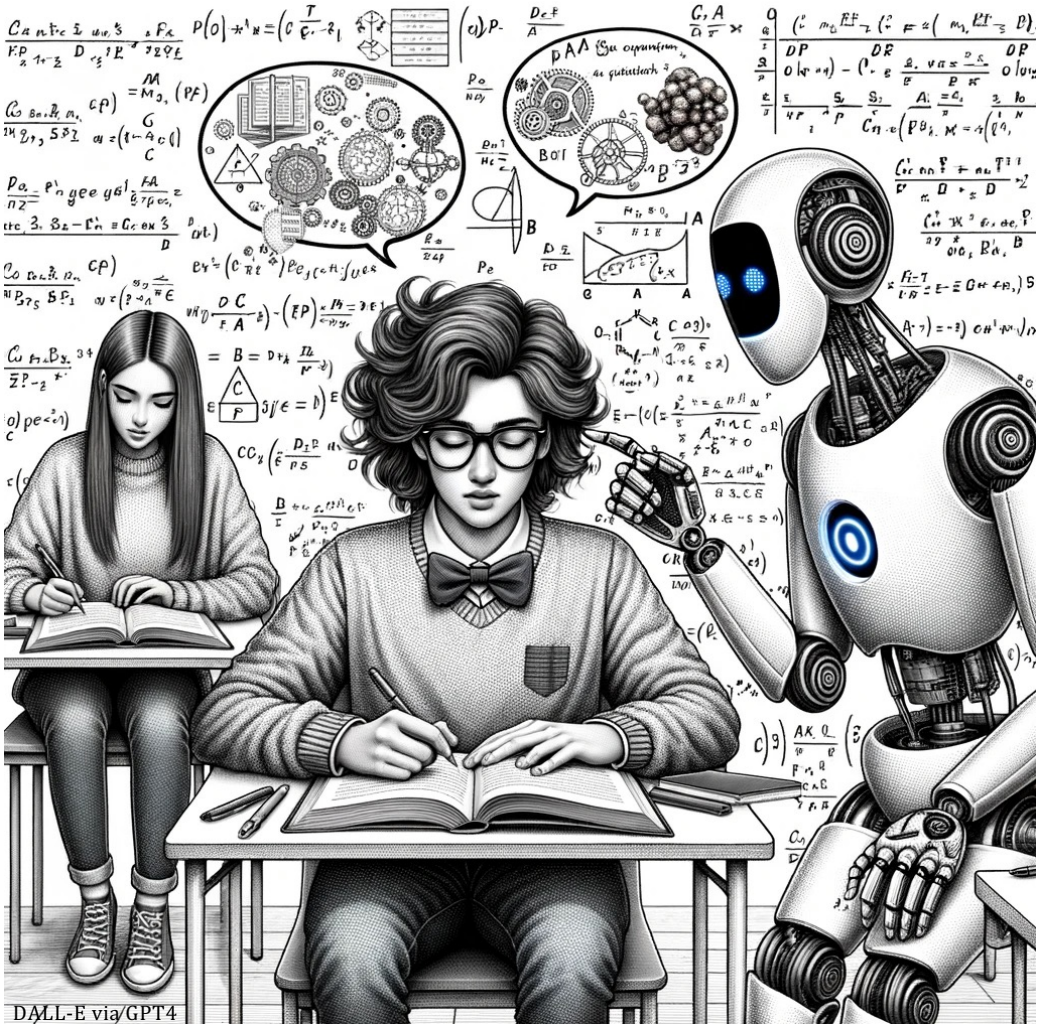
What does this buy us?

- LLMs have certain enduring **strengths** and **weaknesses**
- These are based on **fundamental principles** of LLM design
- We can use these to predict where can be **most useful** (solve enduring **friction points** in education)

Strengths	Weaknesses
Explanations	Facts
Coding	Multiple representations
Working with language	Producing novel, nuanced work

	Conceptual explanations	Coding help	Idea Generation	Problem-Solving	Fact-Finding
	✓	✓	✓	✗	✗
Textbook	✗	✗	✗	?	✓
Instructor	✓	✗	✗	✓	✓
Learning Assistant	?	✓	✓	✓	?

THANK YOU!



DALL-E via GPT4