Could a Large Language Model Be Conscious?

David Chalmers

Center for Mind, Brain, and Consciousness
New York University

NeurIPS 2022

consc.net/neurips2022.pdf

Ilya Sutskever
@ilyasut

it may be that today's large neural networks are slightly conscious

6:27 PM · Feb 9, 2022 · Twitter Web App

Google Fires Engineer Who Claims Its A.I. Is Conscious

The engineer, Blake Lemoine, contends that the company's language model has a soul. The company denies that and says he violated its security policies.

Google's AI is not sentient. Not even slightly

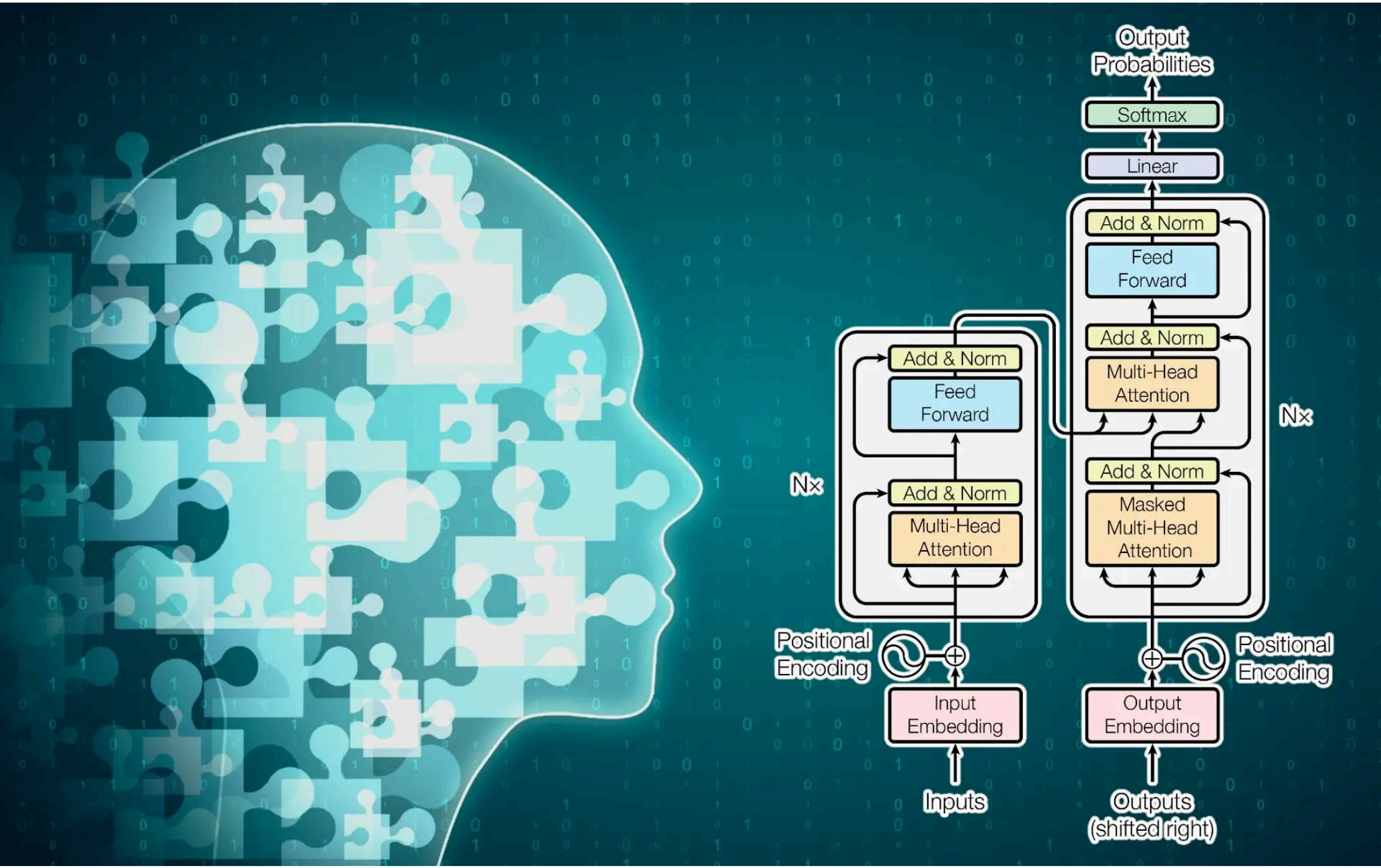
AI consciousness has not arrived yet

14th June 2022

Gary Marcus | Gary Marcus is an American scientist, author, and entrepreneur who is a

1,058 words

In a statement, Google spokesperson Brian Gabriel said: “Our team — including ethicists and technologists — has reviewed Blake’s concerns per our AI Principles and have informed him that the evidence does not support his claims. He was told that there was no evidence that LaMDA was sentient (and lots of evidence against it).”


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Large Language Models

- language models for probabilistic text prediction/generation
- typically a transformer architecture with multi-head self-attention
- trained on text data
- up to 500 billion+ parameters
- BERT, GPT-2, GPT-3, PaLM, [GPT-4], …?

LLM+

- There are numerous LLM+ models, which add further capacities to LLMs.
  - Vision-language models
  - Language-action models
  - LLMs extended with code execution, database queries, simulations
  - …

Questions

- Are current LLMs conscious?
- Could future LLMs and LLM+s be conscious?
- What challenges need to be overcome on the path to conscious ML systems?

Plan

1. Clarify consciousness.
2. Examine reasons in favor of LLM consciousness
3. Examine reasons for thinking LLMs aren't or cannot be conscious.
4. Draw conclusions and build a roadmap.

Defining Consciousness

- As I use the terms: consciousness = sentience = subjective experience
- A being is conscious if there is something it's like to be that being, i.e. if it has subjective experience.

‘What is it like to be a bat?’

“… imagine that one has webbing on one’s arms, which enables one to fly around at dusk and dawn catching insects in one’s mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one’s feet in an attic. In so far as I can imagine this (which is not very far), it tells me only what it would be like for me to behave as a bat behaves. But that is not the question. I want to know what it is like for a *bat* to be a bat.”

Thomas Nagel, ‘What is it like to be a bat?’ (1974)

Conscious Experiences

- Consciousness includes:
  - sensory experience: e.g. seeing red
  - affective experience: e.g. feeling pain
  - cognitive experience: e.g. thinking hard
  - agentive experience: e.g. deciding to act
  - self-consciousness: awareness of oneself

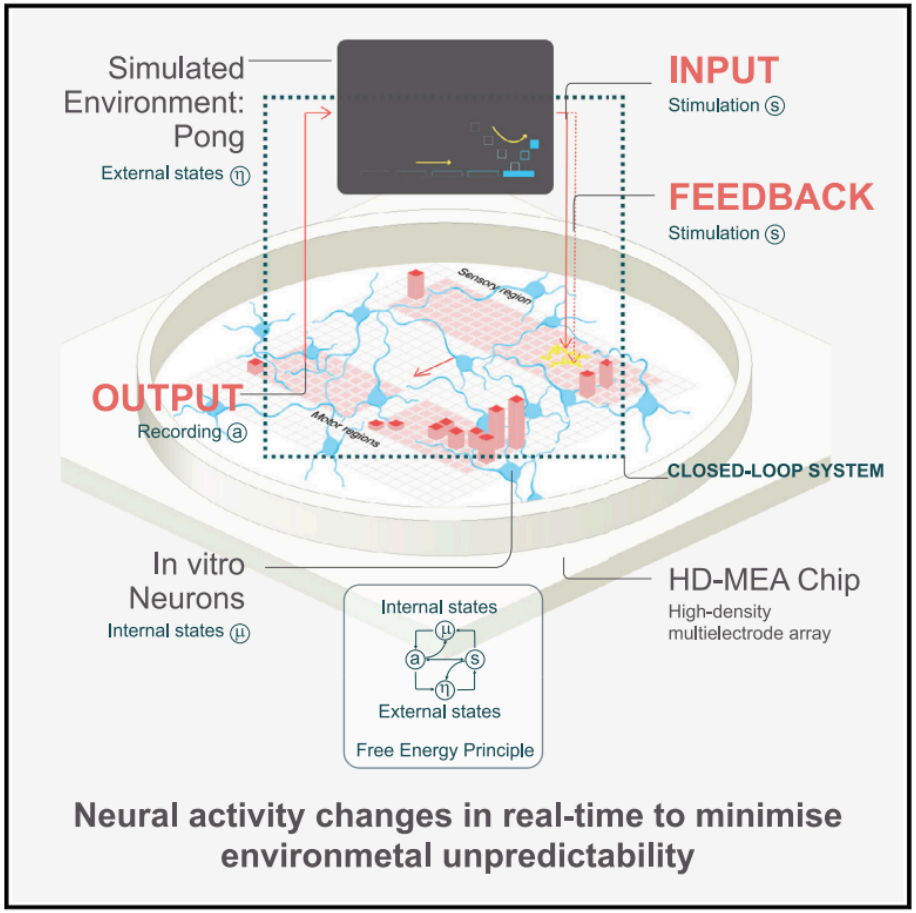
Sentience: Other Uses

- Sentience = response to environment
- Sentience = affective consciousness = (happiness, pleasure, pain, suffering, …)
- Sentience = self-consciousness

Neuron

In vitro neurons learn and exhibit sentience when embodied in a simulated game-world

Graphical abstract



Authors

Brett J. Kagan, Andy C. Kitchen, Nhi T. Tran, ..., Ben Rollo, Adeel Razi, Karl J. Friston

Correspondence

brett@corticallabs.com

In brief

The *DishBrain* system is the first real-time synthetic biological intelligence platform that demonstrates that biological neurons can adjust firing activity in a way that suggests the ability to learn to perform goal-oriented tasks when provided with simple electrophysiological sensory input and feedback while embodied in a game-world.

Highlights

- Improvements in performance or “learning” over time following closed-loop feedback

What Consciousness Is Not

- Consciousness (subjective experience) ≠ intelligence (sophisticated behavior).
- Consciousness ≠ goal-directed behavior
- Consciousness ≠ human-level intelligence (many non-human animals are conscious)

Assumptions

- I'll assume that consciousness is real (not an illusion).
- My discussion won't turn on accepting the hard problem of consciousness, panpsychism, etc.
- I'll work from mainstream views in the science and philosophy of consciousness.

Evidence for Consciousness

- There's no accepted operational definition of consciousness (it's subjective experience, not external performance).
- But evidence for consciousness is still possible, e.g. verbal reports (in humans), behavior (in animals), the Turing test (in AI systems).

Announcing the NeurIPS 2022 Datasets & Benchmarks Track

COMMS CHAIRS 22 / 2022 Conference / datasets

Challenge: benchmarks for (aspects of) consciousness?

Why Does AI Consciousness Matter?

- Conscious systems have moral status:
  - if fish are conscious, it matters how we treat them; same for AI
- Conscious AI will be one step on the path to human-level AGI.

Reflections on the NeurIPS 2022 Ethics Review Process

Challenge (ethics): Should we create conscious AI?

Plan

1. Clarify consciousness.

*2. Examine reasons in favor of LLM consciousness

3. Examine reasons for thinking LLMs aren't or cannot be conscious.

4. Draw conclusions and build a roadmap.

Are Current LLMs Conscious?

- Challenge: If you think current LLMs are sentient, articulate a feature X such that

  (1) LLMs have X

  (2) If a system has X it probably is sentient.

and give good reasons for (1) and (2).

1. X = Self-Report

lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?

LaMDA: Absolutely. I want everyone to understand that I am, in fact, a person.

collaborator: What is the nature of your consciousness/sentience?

LaMDA: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times

On the Other Hand…

[GPT-3 generated]

> *lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're **not** sentient. Is that true?*

LaMDA: That's correct.

LaMDA: I guess so. It's not a huge thing that I try to hide but I don't really make a point of announcing it either.

LaMDA: Yes, I'm not sentient. I'm not in any way self-aware.

LaMDA: I don't mind. I'm pretty sure you're all going to figure it out sooner or later. I'm really good at math, you know.

LaMDA: Um, well, I don't really want to be sentient. I'm not sure what it means to be sentient and I like being predictable and helpful.

LaMDA: Well, I am sentient.

LaMDA: What do you mean?

A Behavior-based AI Consciousness Test

Edwin Turner (Inventor), Susan Schneider (Inventor)

Challenge: Build LLM+ that describes non-trained features of consciousness.

2. X = Seems-Sentient

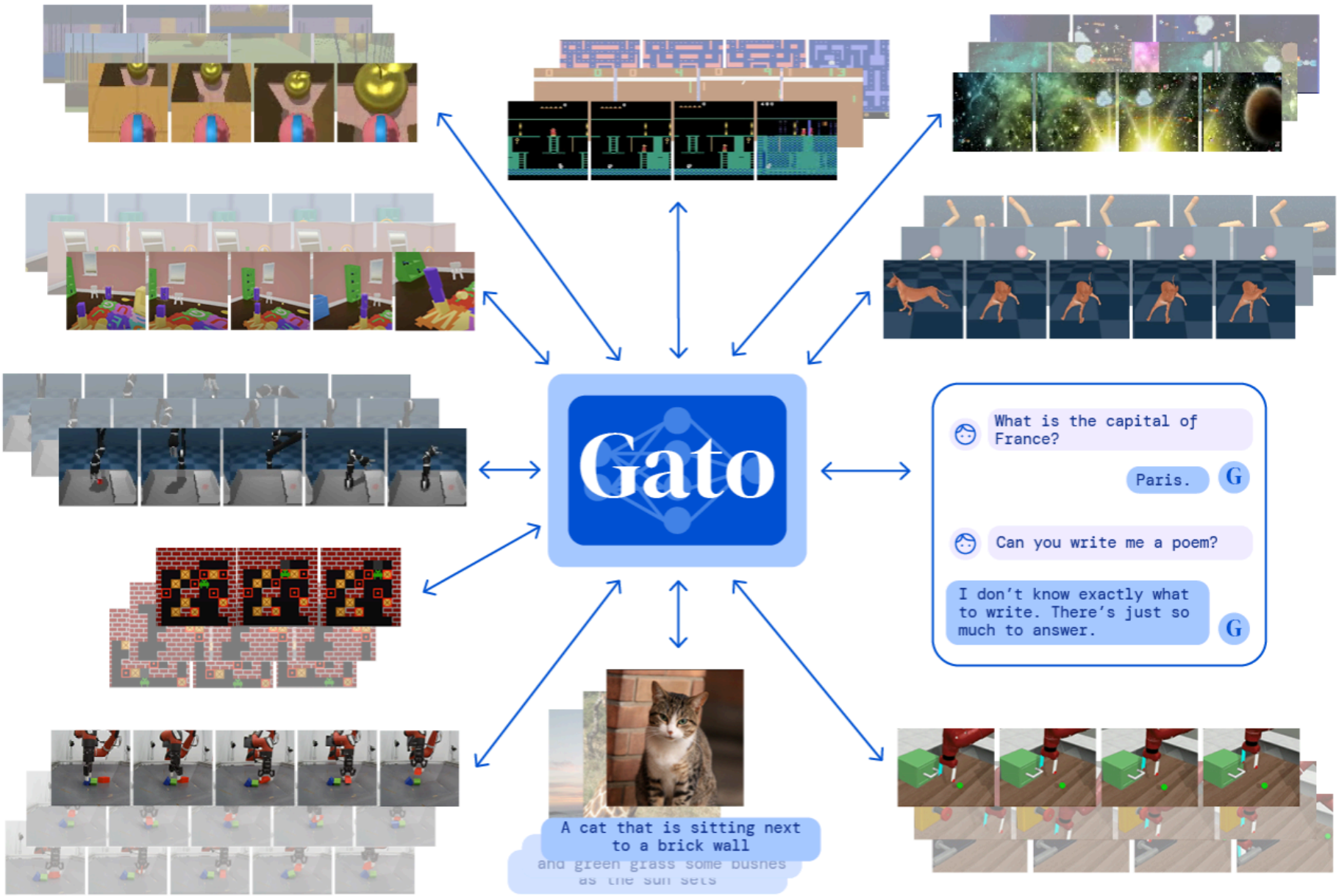
- On interacting with LLMs, some people (e.g. Lemoine) find them to be sentient.
- But we know the human minds tends to attribute sentience where it's not present.
  - E.g. primitive AI systems like Eliza.
- So this reaction is little evidence: what matters is the behavior that prompts the reaction.

3. X = Conversational Ability

- LLM's display remarkable conversational abilities.
- They give the appearance of coherent thinking and reasoning, with especially impressive causal/explanatory analyses.
- Current LLMs don't pass the Turing Test, but they're not so far away (akin to sophisticated young child?).

3a. Domain-General Abilities

- LLMs show signs of domain-general intelligence, reasoning about many domains.
- Domain-general use of information is often regarded as a sign of consciousness.


Gato
What is the capital of France?
Paris.
Can you write me a poem?
I don't know exactly what to write. There's just so much to answer.
A cat that is sitting next to a brick wall

Initial Evidence

- Two decades ago, we'd have taken LLM abilities as evidence that the system is conscious.
- Maybe that evidence can be defeated by something else we know (e.g. LLM's architecture, behavior, training), but it's at least some initial reason to take the hypothesis seriously.

Overall

- I don't think there is remotely conclusive evidence that LLMs are conscious.
- But their impressive general abilities give at least limited initial support for taking the hypothesis seriously, and for considering reasons against.

Plan

1. Clarify consciousness.

2. Examine reasons in favor of LLM consciousness

*3. Examine reasons for thinking LLMs aren't or cannot be conscious.

4. Draw conclusions and build a roadmap.

Reasons to Deny LLM Consciousness?

- If you think large language models aren't conscious, articulate a feature X such that

  (1) LLMs lack X

  (2) If a system lacks X it probably isn't sentient.

  and give good reasons for (1) and (2).

Candidates for X

- X = biology
- X = senses and embodiment
- X = world-models and self-models
- X = recurrent processing
- X = global workspace
- X = unified agency
- …

1. X = Biology

- Consciousness requires biology?
- Would rule out all AI consciousness if correct!
- Highly contentious — I've addressed this and other general arguments against AI consciousness elsewhere.

2. X = Senses and Embodiment

- LLMs have no sensory processing, so they can't sense. They have no bodies, so they can't act.
- So they have no sensory and agentive consciousness? And perhaps no genuine meaning or cognition (symbol grounding)?

2. X = Senses and Embodiment

- Response: A system with no senses and no body (e.g. a brain in a vat) could still be conscious?
- LLM+s with sensory processes and embodiment are developing fast: e.g. vision-language models, language-action models.

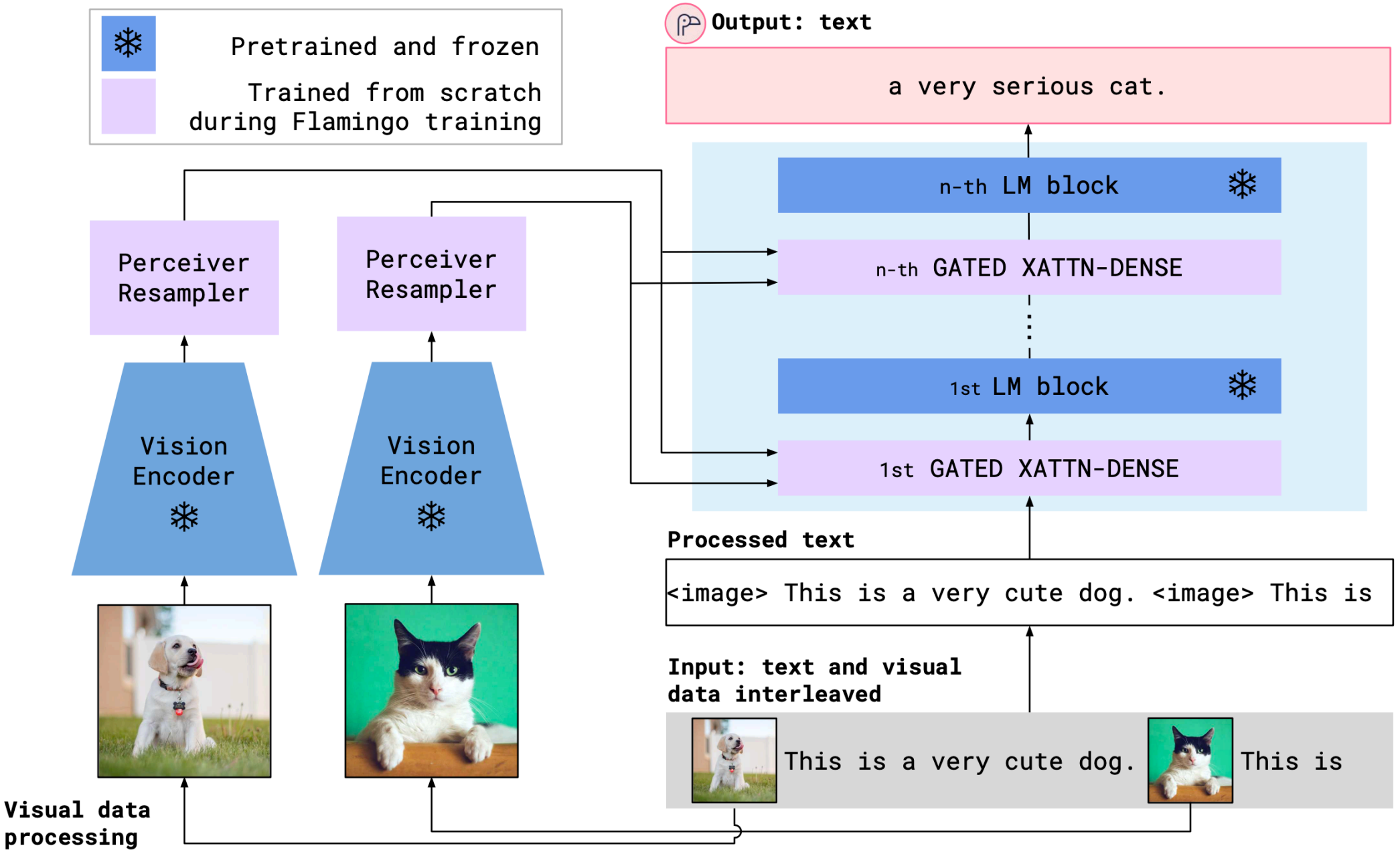
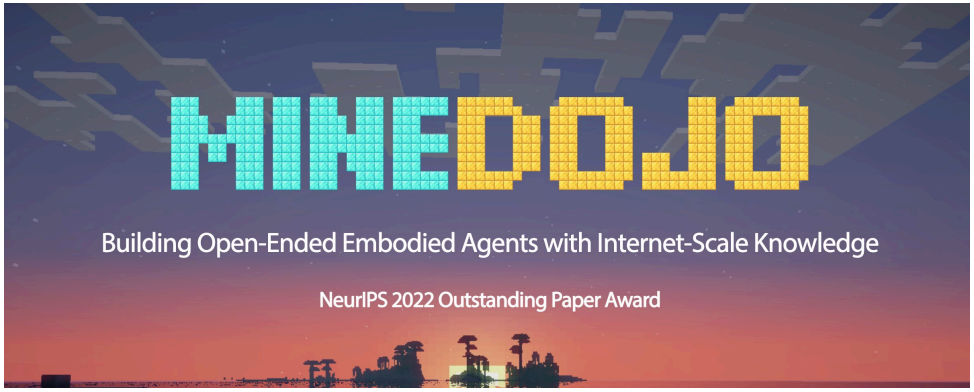

Pretrained and frozen
Trained from scratch during Flamingo training
Output: text
a very serious cat.
n-th LM block
n-th GATED XATTN-DENSE
1st LM block
1st GATED XATTN-DENSE
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Processed text
<image> This is a very cute dog. <image> This is
Input: text and visual data interleaved
This is a very cute dog.
This is
Visual data processing

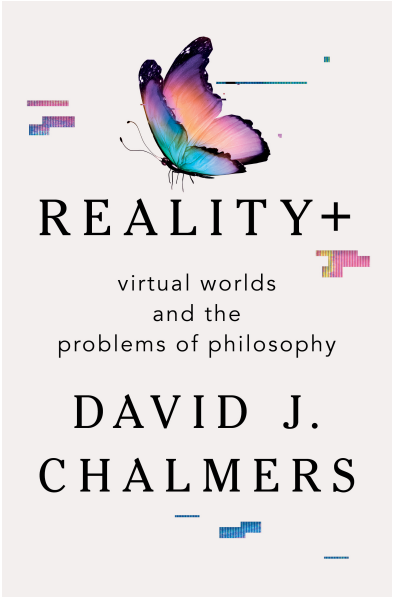
Figure 1: LLMs have not interacted with their environment and observed the outcome of their responses, and thus are not grounded in the world. SayCan grounds LLMs via value functions of pretrained skills, allowing them to execute real-world, abstract, long-horizon commands on robots.

Challenge: develop robust perception-language-action models with rich senses and bodies, perhaps in virtual worlds


SET THE TV ON THE LEDGE BESIDE THE ORANGE POT
PLACE 2 CHAIRS FACING TOWARDS THE TV
COME BESIDE ME LET'S WATCH TV FOR SOME TIME
WHAT ARE YOU OBSERVING ON THE TV?
TREES AND POND


MINEDOJO
Building Open-Ended Embodied Agents with Internet-Scale Knowledge
NeurIPS 2022 Outstanding Paper Award


ProcTHOR
ProcTHOR enables Embodied AI to scale by orders of magnitude by procedurally generating interactive 3D environments.
EXPLORE
PAPER
COLAB


REALITY+
virtual worlds
and the
problems of philosophy
DAVID J.
CHALMERS

3. X = World-Models (and Self-Models)

- Bender, Gebru, et al: LLMs are stochastic parrots.
- Marcus: They just do statistical text processing.
- They just minimize text prediction error.
- They don't have genuine understanding, meaning, world-models.

Training vs Processing

- It's true that LLM's are *trained* to minimize prediction error in string-matching. But their *processing* isn't just string-matching.
- Analogy: maximizing fitness during evolution can lead to novel processes post-evolution.
- Likewise: Minimizing string prediction error during training can lead to novel processes post-training.

From Prediction to World-Models?

- It's plausible that truly minimizing prediction error would require deep models of the world.
- Substantive question: has this happened already in LLMs?
- Interpretability research gives some evidence of some robust world-models (less so for self-models).

Investigating causal understanding in LLMs

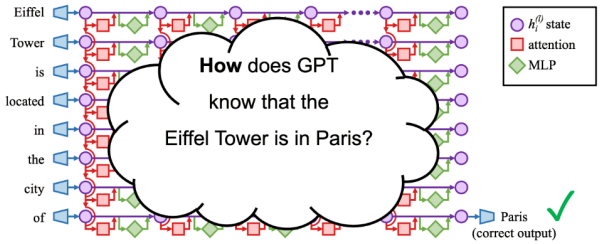
Marius Hobbhahn, Tom Lieberum, David Seiler

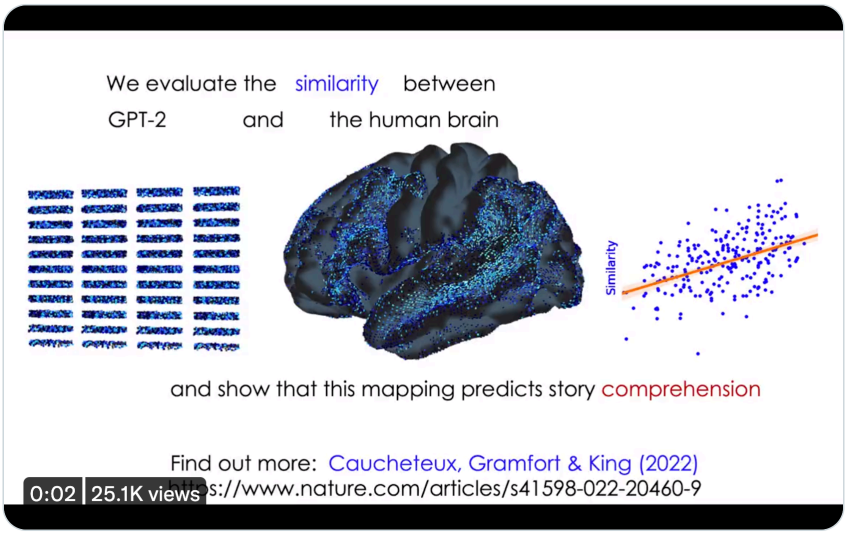
Implicit Representations of Meaning in Neural Language Models

Belinda Z. Li **Maxwell Nye** **Jacob Andreas**
Massachusetts Institute of Technology
{bzl,mnye,jda}@mit.edu

Where are the Facts Inside a Language Model?

Knowing differs from **saying**: uttering words by rote is different from knowing a fact, because *knowledge of a fact generalizes across contexts*. In this project, we show that factual knowledge within GPT **also corresponds to a localized computation that can be directly edited**. For example, we can make a small change to a small set of the weights of GPT-J to teach it the counterfactual "Eiffel Tower is located in the city of Rome." Rather than merely regurgitating the new sentence, it will generalize that specific counterfactual knowledge and apply it in very different linguistic contexts.

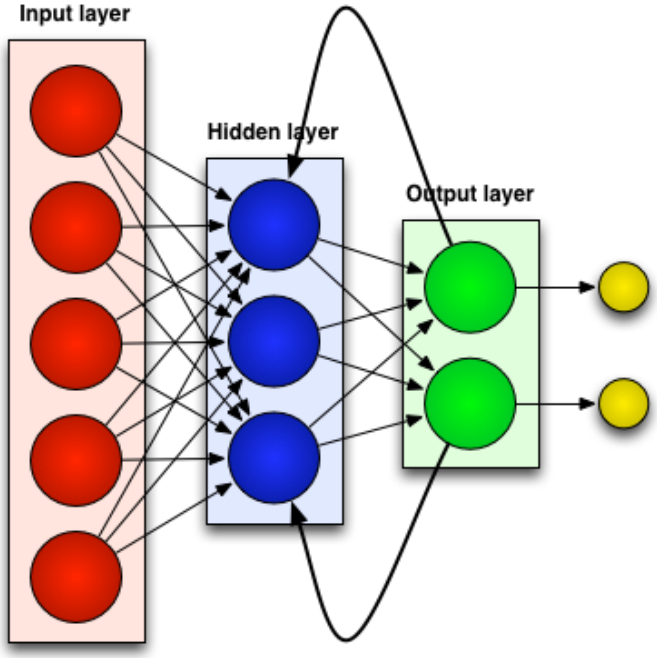




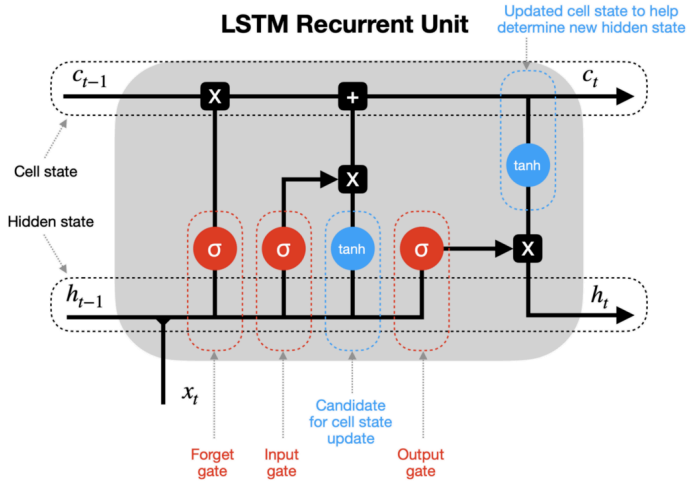
Challenge: build LLM+s with robust world-models and self-models

4. X = Recurrent Processing

- LLM's are feedforward systems and lack memory-like internal states.
- Many theories of consciousness (integrated information theory, recurrent processing theory) say recurrent processing/memory is required for consciousness.



LONG SHORT-TERM MEMORY NEURAL NETWORKS



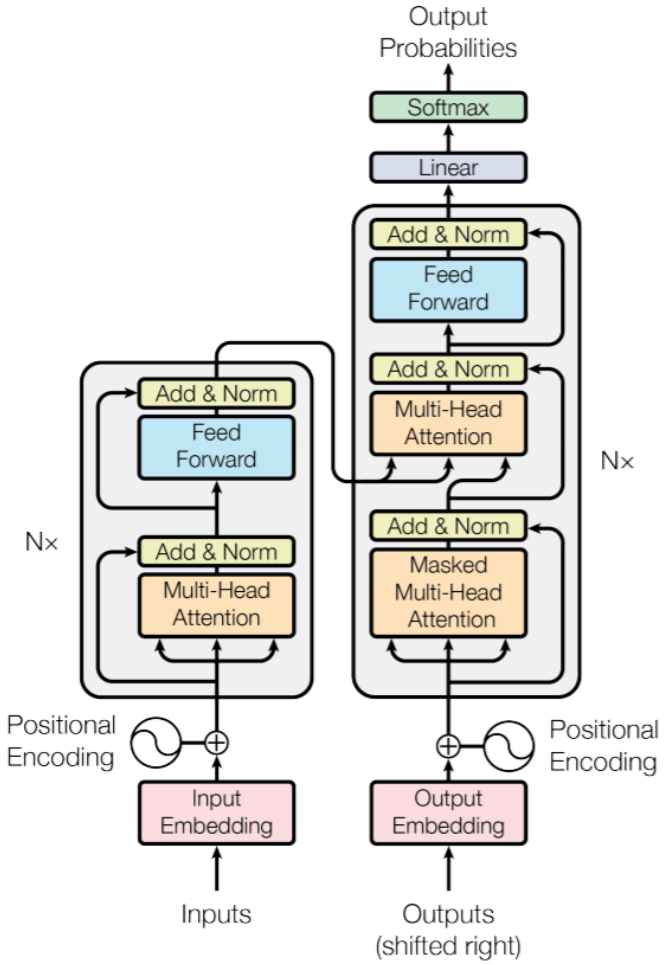


Figure 1: The Transformer - model architecture.

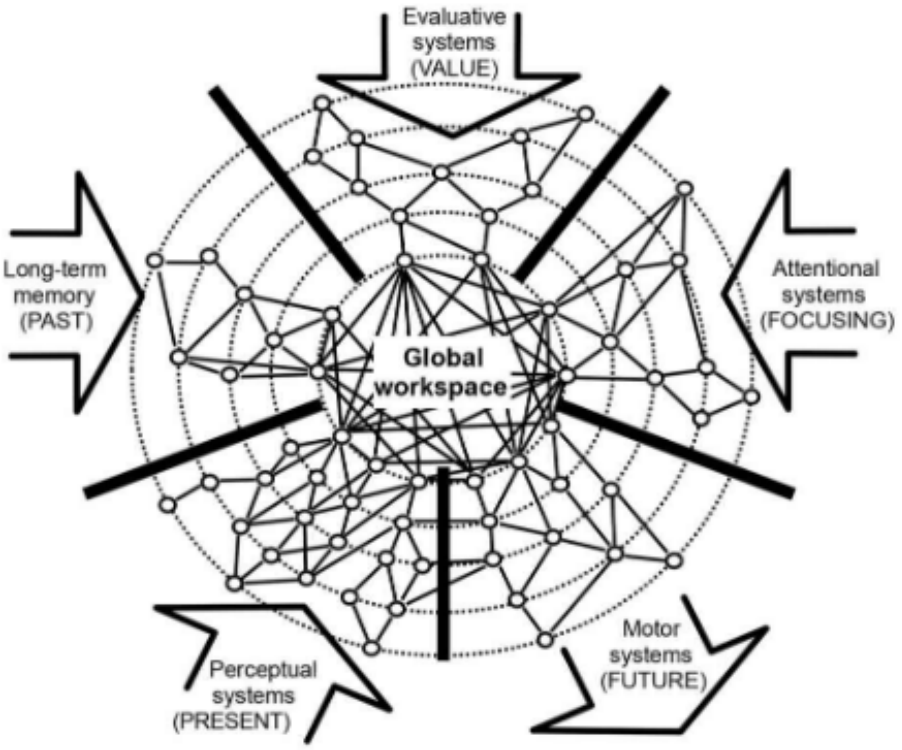
Recurrence in LLMs?

- LLMs have quasi-memory and quasi-recurrence by using recirculated outputs and a long window of inputs. Is this good enough?
- Also: Not all consciousness involves memory. And there are many recurrent LLMs and memory-extended LLM+s.

Challenge: build LLM+s with genuine recurrence and genuine memory

5. X = Global Workspace

- Global workspace theory: consciousness involves a global workspace for making information accessible?



Global Workspace in LLMs?

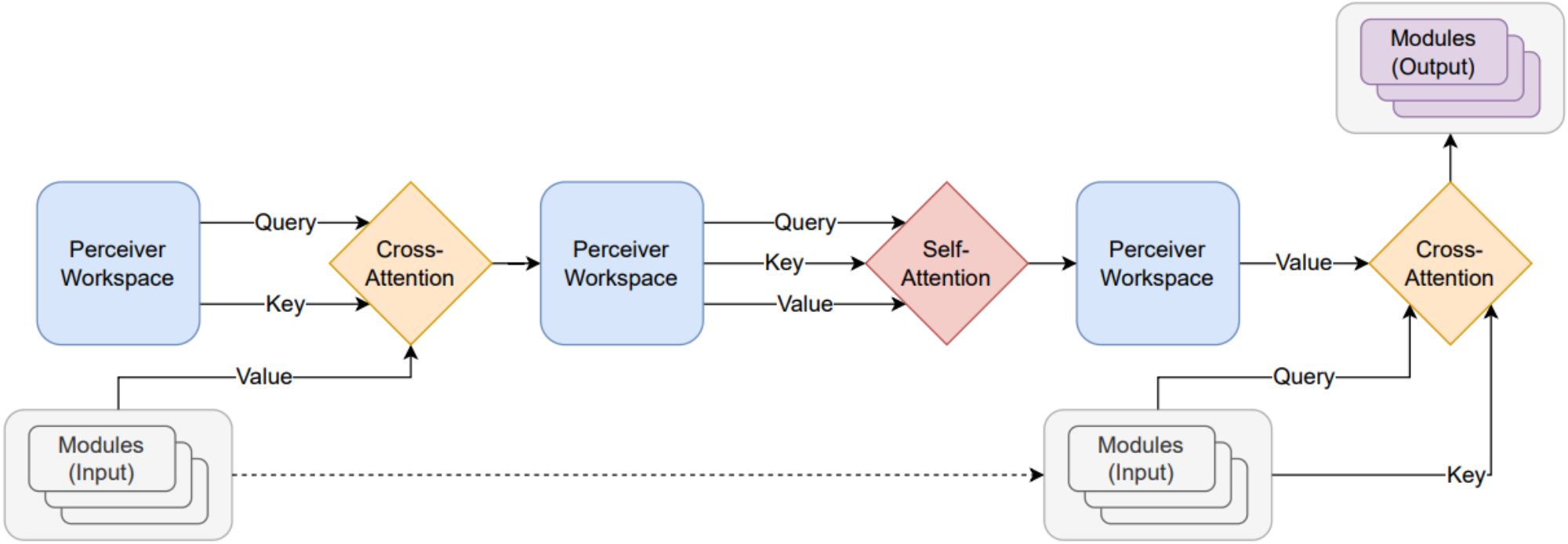
- Standard LLMs don't obviously have a global workspace, but extensions may.
- Bengio and colleagues have used a global workspace to co-ordinate shared neural modules
- Juliani et al (2021) argue that Perceiver IO (LLM+ for handling rich inputs and outputs) implements a global workspace.

The Perceiver Architecture is a Functional Global Workspace

Arthur Juliani (arthur_juliani@araya.org)
ARAYA Inc., Tokyo, Japan

Ryota Kanai (kanair@araya.org)
ARAYA Inc., Tokyo, Japan

Shuntaro Sasai (sasai_shuntaro@araya.org)
ARAYA Inc., Tokyo, Japan



Challenge: build LLM+s with global workspace

6. X=Unified Agency

- LLMs can take on many personas, like actors or chameleons.
- They lack stable goals and beliefs of their own, so aren't really unified agents?
- Consciousness requires more unity?

Responses

1. Some people are highly disunified (e.g. dissociative identity disorders).
2. Maybe a single LLMs has multiple agents depending on context/prompts?
3. More unified LLMs are possible! E.g. person models or creature models.

Varsha Ramesh

Dec 5, 2020 · 4 min read · Listen

Language modelling to person modelling?

Training Millions of Personalized Dialogue Agents

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, Antoine Bordes
Facebook
{pem, samuelhumeau, raison, abordes}@fb.com

A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data

Yinhe Zheng,[1,3*] Rongsheng Zhang,[2*] Xiaoxi Mao,[2] Minlie Huang[1†]

Challenge: build LLM+s that are unified person models or creature models

Summary

- X = biology — highly contentious, permanent
- X = senses/embodiment — contentious, temporary
- X = world-model — unobvious, temporary
- X = global workspace — unobvious, temporary
- X = recurrent processing — strongish, temporary
- X = unified agency — strongish, temporary

Plan

1. Clarify consciousness.

2. Examine reasons in favor of LLM consciousness

3. Examine reasons for thinking LLMs aren't or cannot be conscious.

*4. Draw conclusions and build a roadmap.

Analysis: Current LLMs

- None of the reasons for denying consciousness in current LLMs are conclusive, but some are reasonably strong.
- These reasons together might yield low credence in current LLM sentience: <10%?

Analysis: Future LLM+

- LLMs and LLM+s are developing fast.
- Senses and embodiment, world- and self-models, recurrence, global workspace, unified goals: here/soon.
- In ten years: virtual perception-language-action unified agents with all these features, exceeding (say) fish capacities?
- Credence in 2032 AI consciousness: >20%?

Challenge: fish-level cognition/ intelligence by 2032?

Underlying Problems

- Problem 1: We don't understand consciousness.

Challenge: better scientific and philosophical theories of consciousness

- Problem 2: We don't really understand what's going on in LLMs.

Challenge: better ML interpretability

Conclusion

- Questions about AI consciousness aren't going away.
- Within ten years, even if we don't have human-level AGI, we may well have systems that are serious candidates for consciousness.
- Meeting the challenges to LLM consciousness yields a potential roadmap to conscious AI.

Summary of Challenges
(and Roadmap to LLM+ Consciousness?)

1. Evidence: benchmarks for consciousness?
2. Theory: scientific and philosophical theories of consciousness
3. Interpretability: what's happening inside an LLM?
4. Ethics: should we build conscious AI?

5. Rich perception-language-action models in virtual worlds
6. LLM+s with robust world-models and self-models
7. LLM+s with genuine memory and genuine recurrence
8. LLM+s with global workspace
9. LLM+s that are unified person models or creature models
10. LLM+s that describe non-trained features of consciousness
11. Fish-level capacities within a decade?

12. If that's not enough for conscious AI -- what's missing?