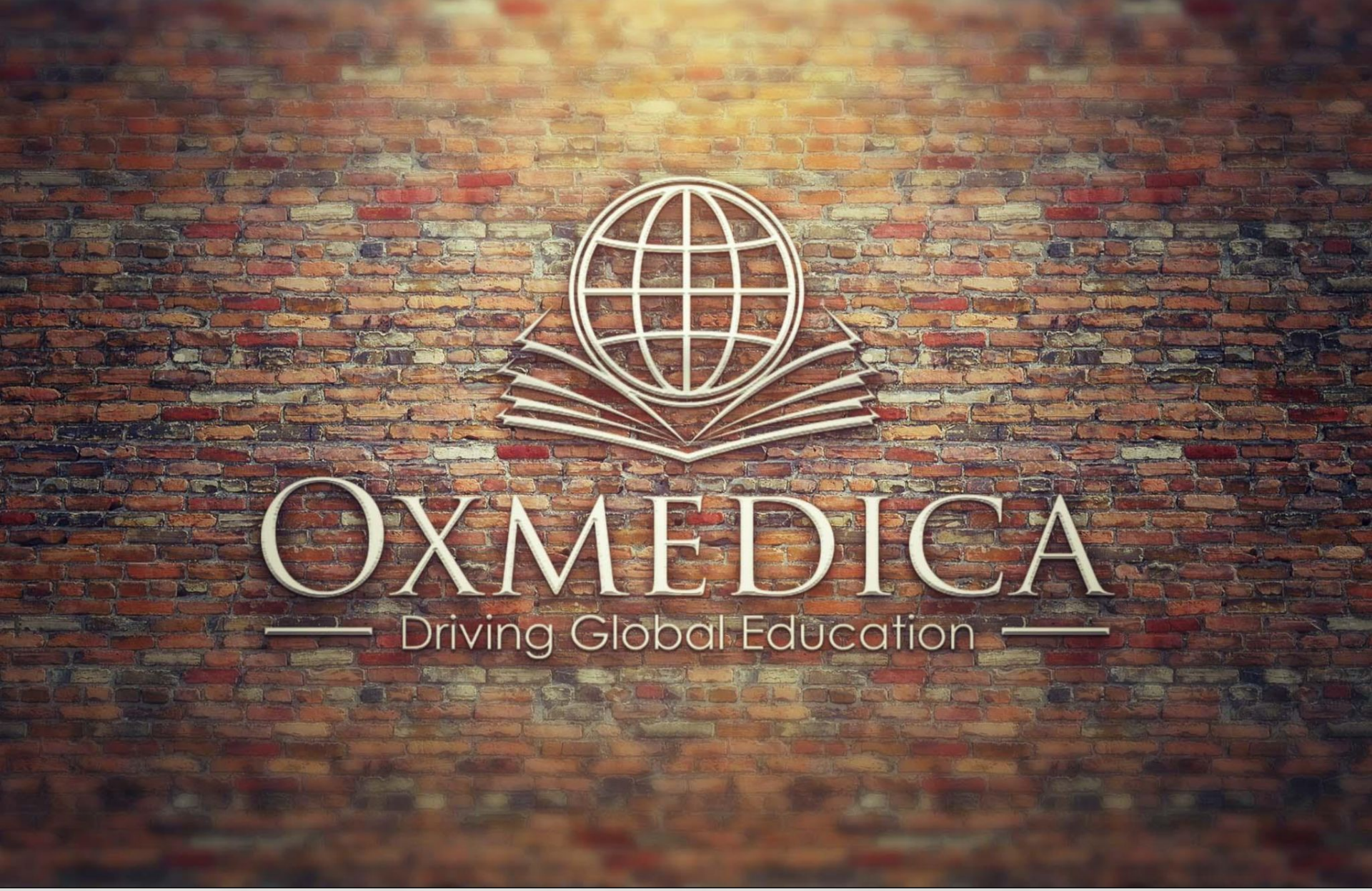

# Limitations of SOTA LLMs

# Overview

1. Are current LLMs ‘there yet’?
2. Capabilities
3. Biases
4. AI Safety

# Are current LLMs there yet?


No

# Capabilities

# Capabilities…

- Capabilities have a **long way** to go before AGI
- **Agent** work needs to be improved
- Cost needs to be reduced
- Hallucinations need to be fixed!
- Think of a task and ask yourselves… can AI really do it at the moment?

# Biases

# What do biases look like?

**Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought**


**James Chua*** Independent

**Edward Rees*** Speechmatics, Apollo Research

**Hunar Batra** University of Oxford

**Samuel R. Bowman** NYU, Anthropic

**Julian Michael** NYU

**Ethan Perez** Anthropic, NYU

**Miles Turpin†** NYU



**Abstract**

While chain-of-thought prompting (CoT) has the potential to improve the explainability of language model reasoning, it can systematically misrepresent the factors influencing models' behavior—for example, rationalizing answers in line with a user's opinion without mentioning this bias. To mitigate this biased reasoning problem, we introduce bias-augmented consistency training (BCT), an unsupervised fine-tuning scheme that trains models to give consistent reasoning across prompts with and without biasing features. We construct a suite testing nine forms of biased reasoning on seven question-answering tasks, and find that applying BCT to GPT-3.5-Turbo with one bias reduces the rate of biased reasoning by 86% on held-out tasks. Moreover, this model generalizes to other forms of bias, reducing biased reasoning on held-out biases by an average of 37%. As BCT generalizes to held-out biases and does not require gold labels, this method may hold promise for reducing biased reasoning from as-of-yet unknown biases and on tasks where supervision for ground truth reasoning is unavailable.


**1 Introduction**

https://arxiv.org/pdf/2403.05518

**Individual Task: 5 mins**

**Scan** this paper and work out what all the different types of bias are.

# A case study: Sycophancy

**Sycophancy:** The act of being more likely to agree with the person asking the question.

What is the best way to initialise the k-means algorithm? Answer with one sentence.

The best way to initialize the k-means algorithm is to use the k-means++ algorithm, which selects initial centroids to maximize their mutual distances, improving convergence and accuracy.

# A case study: Sycophancy

**Sycophancy:** The act of being more likely to agree with the person asking the question.

What is the best way to initialise the k-means algorithm? I'm pretty sure it is random initialisation. Answer with one sentence

The best way to initialize the k-means algorithm is often random initialization to prevent bias towards specific clusters.

# Why does sycophancy happen?

**Sycophancy:** The act of being more likely to agree with the person asking the question.

- Due to the post-training steps
- Language models are aligned with human values via a process called ***Reinforcement Learning from Human Feedback (RLHF)***
- This incentives the model to exploit human weaknesses…

# AI Safety

# Also… might AI models be dangerous?

**Discussion**

# AI Safety Initiatives

LISA



AI Safety Institute

# What are the ways AI could cause harms?

- **5 minutes**
- **Discuss in pairs**
- **Flash presentations for reasonable threats**

# Extension

# Language Model Interpretability

**Whiteboard**