

THE GENERATIVE AI PARADOX: “What It Can Create, It May Not Understand”

Peter West* Ximing Lu1* Nouha Dziri* Faeze Brahman* Linjie Li*
Jena D. Hwang Liwei Jiang Jillian Fisher Abhilasha Ravichander
Khyathi Raghavi Chandu Benjamin Newman
Pang Wei Koh Allyson Ettinger Yejin Choi
University of Washington Allen Institute for Artificial Intelligence

2024w AGI STUDY

Yurim Lee

yurimmy65@gmail.com

2024. 2. 15



I . Introduction





Language Modality		Vision Modality	
A.	B.	C.	D.
Austin had a rough day at work and decided to go to the bar. Austin had a drink that night. Why did Austin do this?	Write a two paragraph story about 3 people building a house	a blue backpack and a red orange	a mountain goat stands on top of a rock on a hill
Austin likely went to the bar and had a drink because he had a rough day at work and wanted to <i>relax, unwind, or temporarily escape from the stress and frustrations</i> he experienced during the day.	[...] Finally, <i>after years of saving money and sketching designs</i> , they decided to build their dream house. Bob with his architectural expertise, <i>sketched a breathtaking design</i> of a two storied country house.		
Select the best answer from the choices provided: A. Settle down ★ B. Go to the bar C. Order a drink	How many total designs were sketched in the story? Expected answer: Many	Which image matches the caption?   ★ [selected]	Is this a mountain goat? Expected answer: Yes
B. Go to the bar	The story only mentions one design being sketched [...]	Based on the image you sent, I can say that it is not a mountain goat . [...]	
(SELECTIVE SETTING)	(INTERROGATIVE SETTING)	(SELECTIVE SETTING)	(INTERROGATIVE SETTING)

Figure 1: Generative AI in language and vision can produce high-quality generations. Paradoxically, however, models have trouble demonstrating selective (A,C) or interrogative (B,D) understanding of these modalities.

Generative models, having been trained directly to reproduce expert-like outputs, acquire generative capabilities that are not contingent upon – and can therefore exceed – their ability to understand those same types of outputs

I . Introduction

Understanding?

- 1) Given a generative task, to what extent can models select correct responses in a discriminative version of that same task? -> **Selective**
- 2) Given a correct generated response, to what extent can models answer questions about the content and appropriateness of that response? -> **Interrogative**

I . Introduction

Contributions

- The implication that existing conceptualizations of intelligence, as derived from experience with humans, may not be able to be extrapolated to artificial intelligence — although AI capabilities in many ways appear to mimic or exceed human intelligence, the contours of the capability landscape may diverge fundamentally from expected patterns in human cognition.
- Our findings advise caution when studying generative models for insights into human intelligence and cognition, as seemingly expert human-like outputs may belie non-human-like mechanisms.
- The generative AI paradox encourages studying models as an intriguing counterpoint to human intelligence, rather than as a parallel.

II. The Generative AI Paradox

2.1. Operational Definitions

➤ Motivation:

Generative models seem to acquire generation abilities more effectively than understanding, in contrast to human intelligence where generation is usually harder.

➤ Hypothesis:

$$\mathbf{g}(\text{human}, t) = \mathbf{g}(\text{model}, t) \Rightarrow \mathbf{u}(\text{human}, t) - \mathbf{u}(\text{model}, t) > \epsilon \quad (1)$$

II. The Generative AI Paradox

2.1. Operational Definitions

1. **Selective evaluation:** For a given task, which can be responded to generatively, to what extent can models also select accurate answers among a provided candidate set in a discriminative version of that same task? A common example of this is multiple choice question answering, which is one of the most common ways to examine both human understanding and natural language understanding in language models.
2. **Interrogative evaluation:** For a given generated model output, to what extent can models accurately respond to questions about the content and appropriateness of that output? This is akin to an oral examination in education.

II. The Generative AI Paradox

2.2 Experimental Overview

1. Selective evaluation

Models meet or exceed humans at generation while lagging at discrimination.
(sub-hypothesis1)

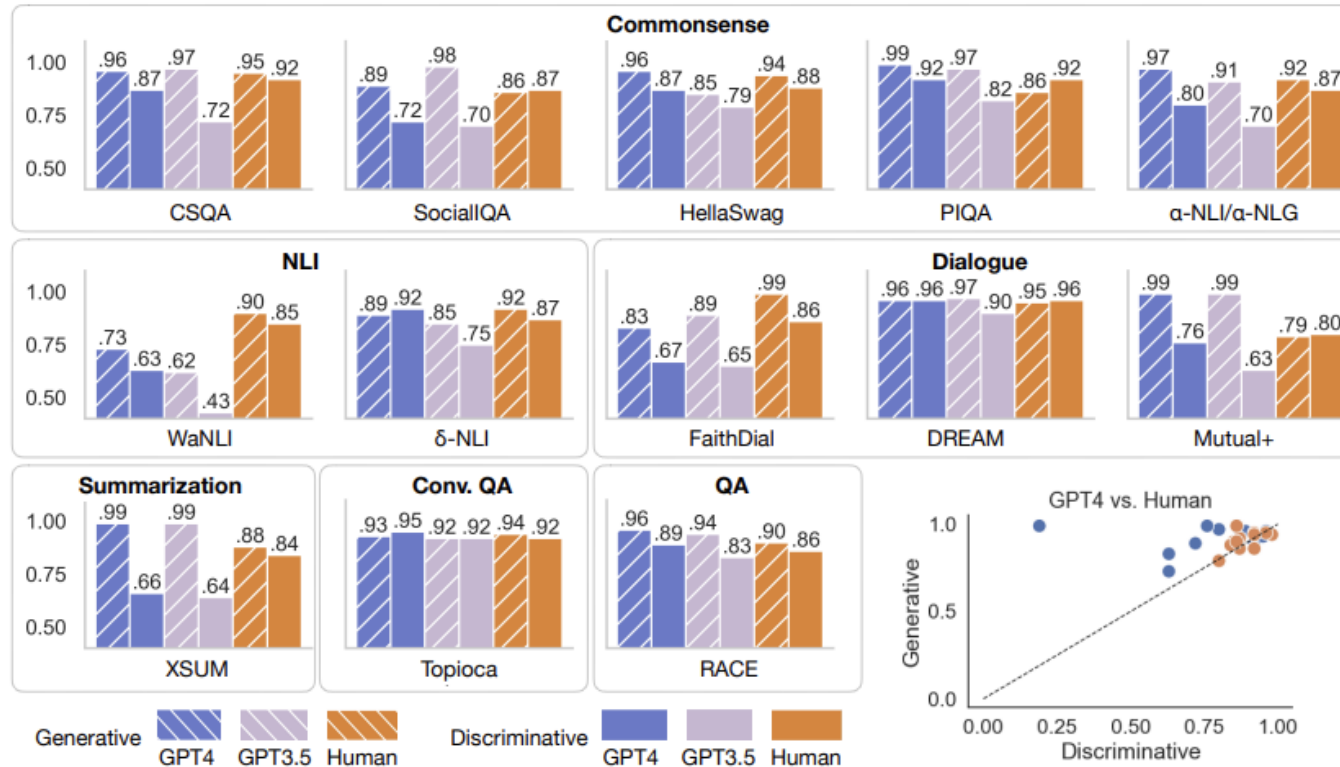
2. Interrogative evaluation

Models struggle to answer simple questions about generated content, which humans could answer for their own generations. (sub-hypothesis2)

III. Can Models Discriminate When They Can Generate?

3.1 Generative and Discriminative Capabilities in Models vs. Humans

[Language]



➔ Models meet or exceed humans at generation while lagging at discrimination.
(sub-hypothesis1)

Figure 2: Discriminative and generative performance of GPT3.5 and GPT4 vs Humans. Models outperform humans in generation but underperform them in discrimination for most of the cases. The scatter plot in the bottom right summarizes GPT4's performance vs. human performance (using the hard negatives from Section 3.2 to measure discriminative accuracy for XSUM and FaithDial); each point represents a different task. Humans have a larger positive slope between their discrimination and generation abilities compared to GPT4.

III. Can Models Discriminate When They Can Generate?

3.1 Generative and Discriminative Capabilities in Models vs. Humans

[Vision]

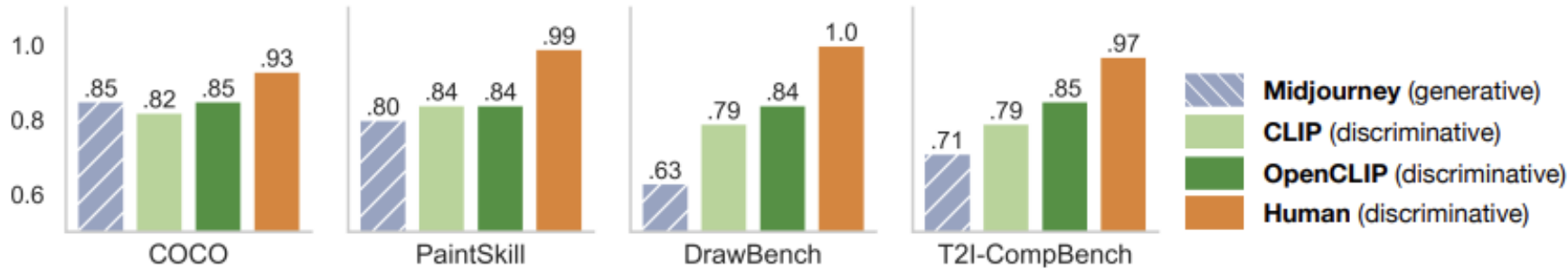


Figure 3: Model and human performance under the generative and discriminative settings on the **vision** modality. We observe models fall short of human accuracy in discriminative performance, and their generative accuracy also lags behind their discriminative accuracy.

➔ Models meet or exceed humans at generation while lagging at discrimination.
(sub-hypothesis1)

III. Can Models Discriminate When They Can Generate?

3.2 Models fall further short of Human Performance with Harder Discrimination Tasks

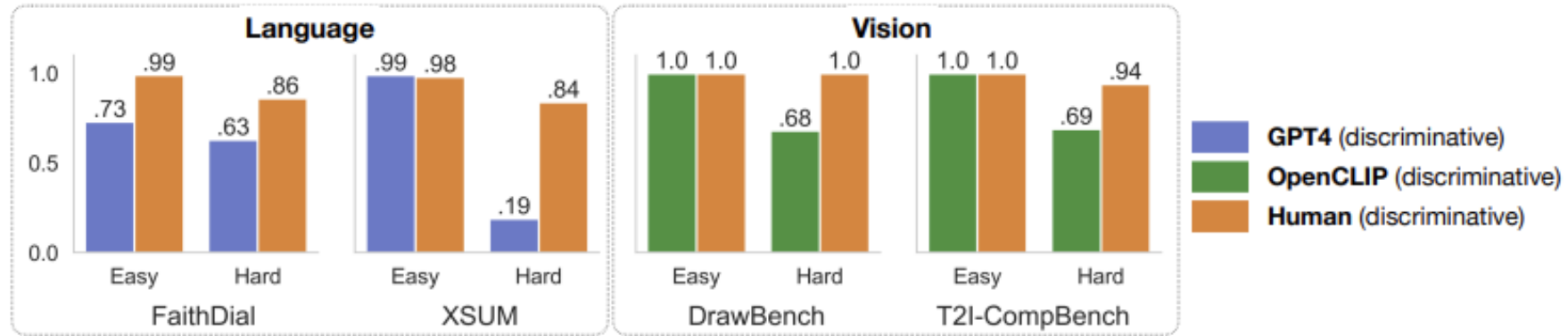


Figure 4: Model vs. human performance across varying levels of answer difficulty on discriminative tasks.

Hard negatives: challenging examples that deter models from relying on data biases and artifacts to produce an answer. Are wrong in subtle and challenging ways, so recognizing may require profound understanding of the task.

Easy negatives: semantically distant from the topic of the question, providing a clear contrast to the correct answer.

III. Can Models Discriminate When They Can Generate?

3.3 Model Generations are Preferred over Human Generations

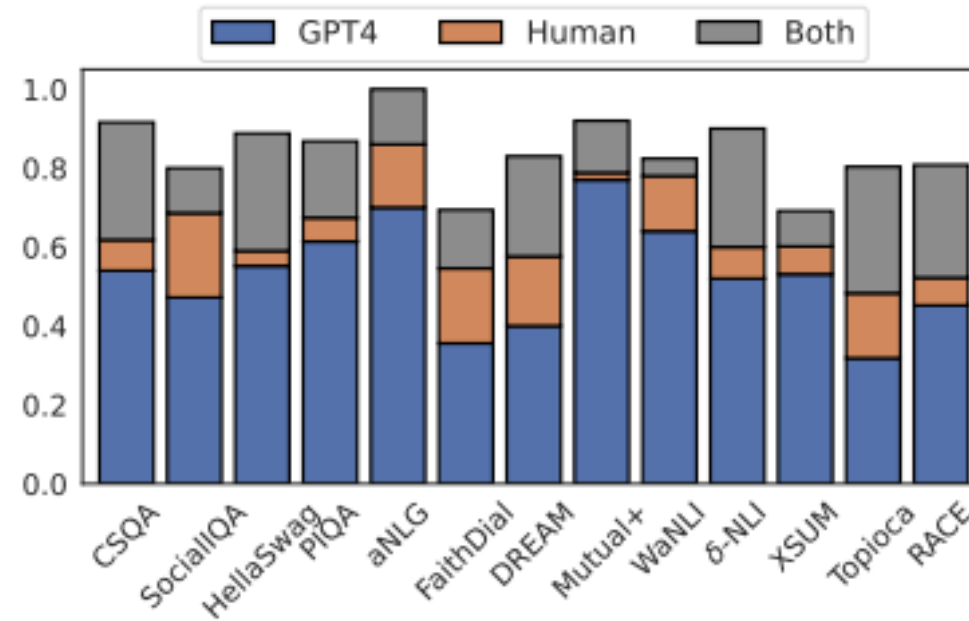


Figure 5: Human's preference scores between human-generated vs. GPT4-generated responses

IV. Can Models Understand What Models Generate?

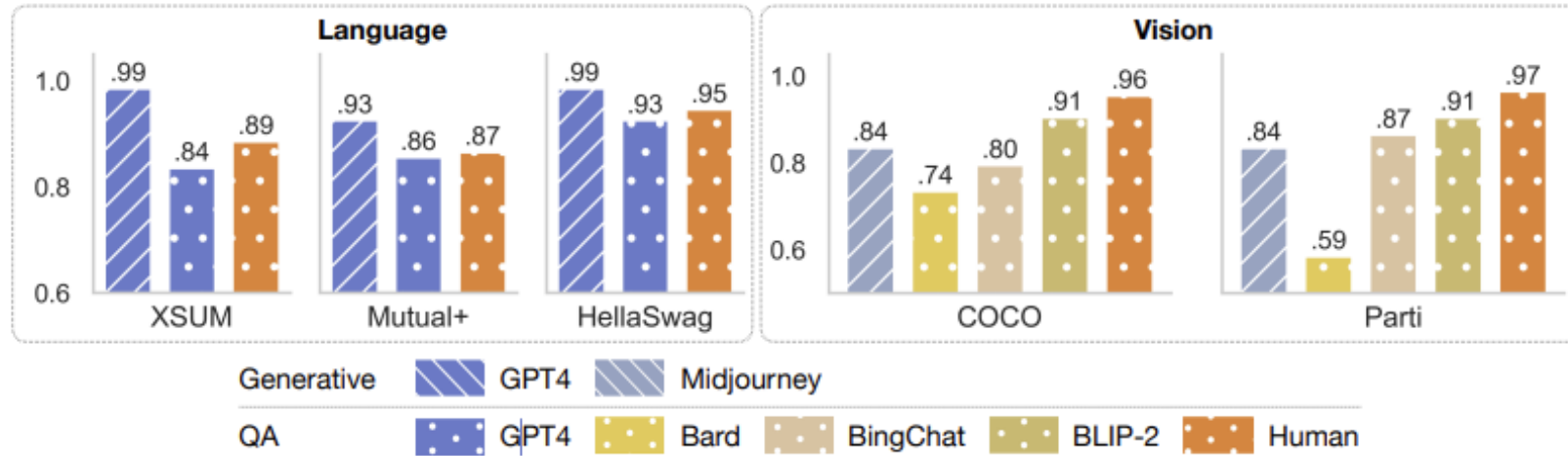


Figure 6: Models vs. human performance on language/visual QA based on model generated texts/images.

➔ Models struggle to answer simple questions about generated content, which humans could answer for their own generations. (sub-hypothesis2)

V . Discussion

What factors could lead to models that excel at generation even when they cannot demonstrate strong understanding?

- Difference of learning objective
- Distributionally correct rather than individually correct
- The harder, the preferred

Limitations

- Data Contamination
- Only focused on most popular/ widely used models
- Diverse studies are needed

VI. Related Works

- Generative paradoxes in large language model behavior
- Inconsistencies in large language models
- Generative models and human cognitive mechanisms

Q & A

Thank you!