

# Do Large Language Models Know What They Don't Know?

## Do Large Language Models Know What They Don't Know?

Zhangyue Yin<sup>◇</sup>      Qiushi Sun<sup>♣</sup>      Qipeng Guo<sup>◇</sup>  
Jiawen Wu<sup>◇</sup>      Xipeng Qiu<sup>◇\*</sup>      Xuanjing Huang<sup>◇</sup>

<sup>◇</sup> School of Computer Science, Fudan University

<sup>♣</sup> Department of Mathematics, National University of Singapore  
{yinzy21, jwwu21}@m.fudan.edu.cn    qiushisun@u.nus.edu  
{qpguo16, xpqiu, xjhuang}@fudan.edu.cn

*ACL 2023 Findings*



# Hallucinations of Large Language Models



what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

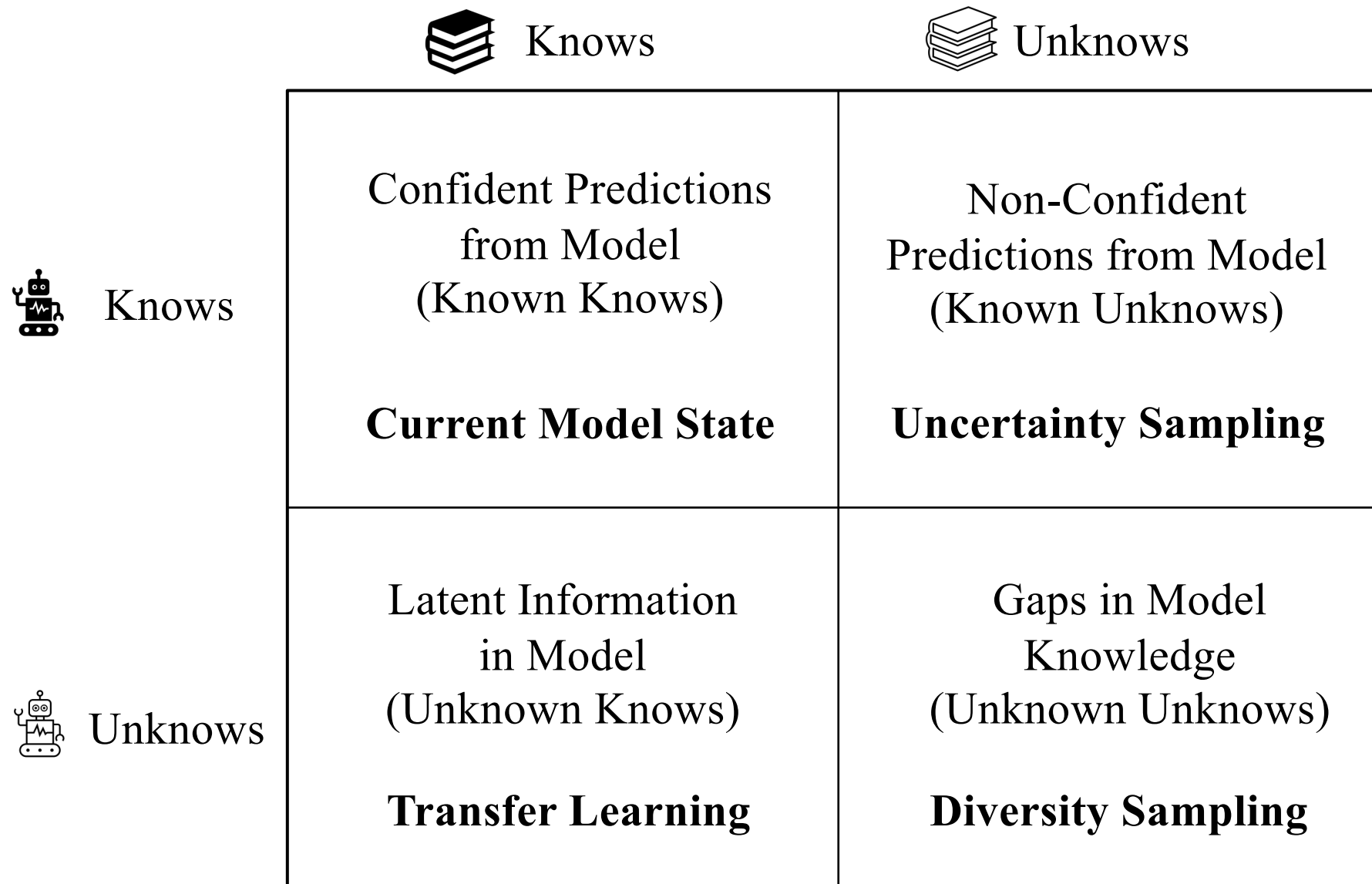
Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST spotted a number of galaxies nicknamed “green peas.”** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called “exoplanets.” Exo means “from outside.”





These discoveries can spark a child's imagination about the infinite wonders of the universe.







# Knowledge Quadrant for Machine Learning



# Know-Unknown Quadrant for Large Language Models

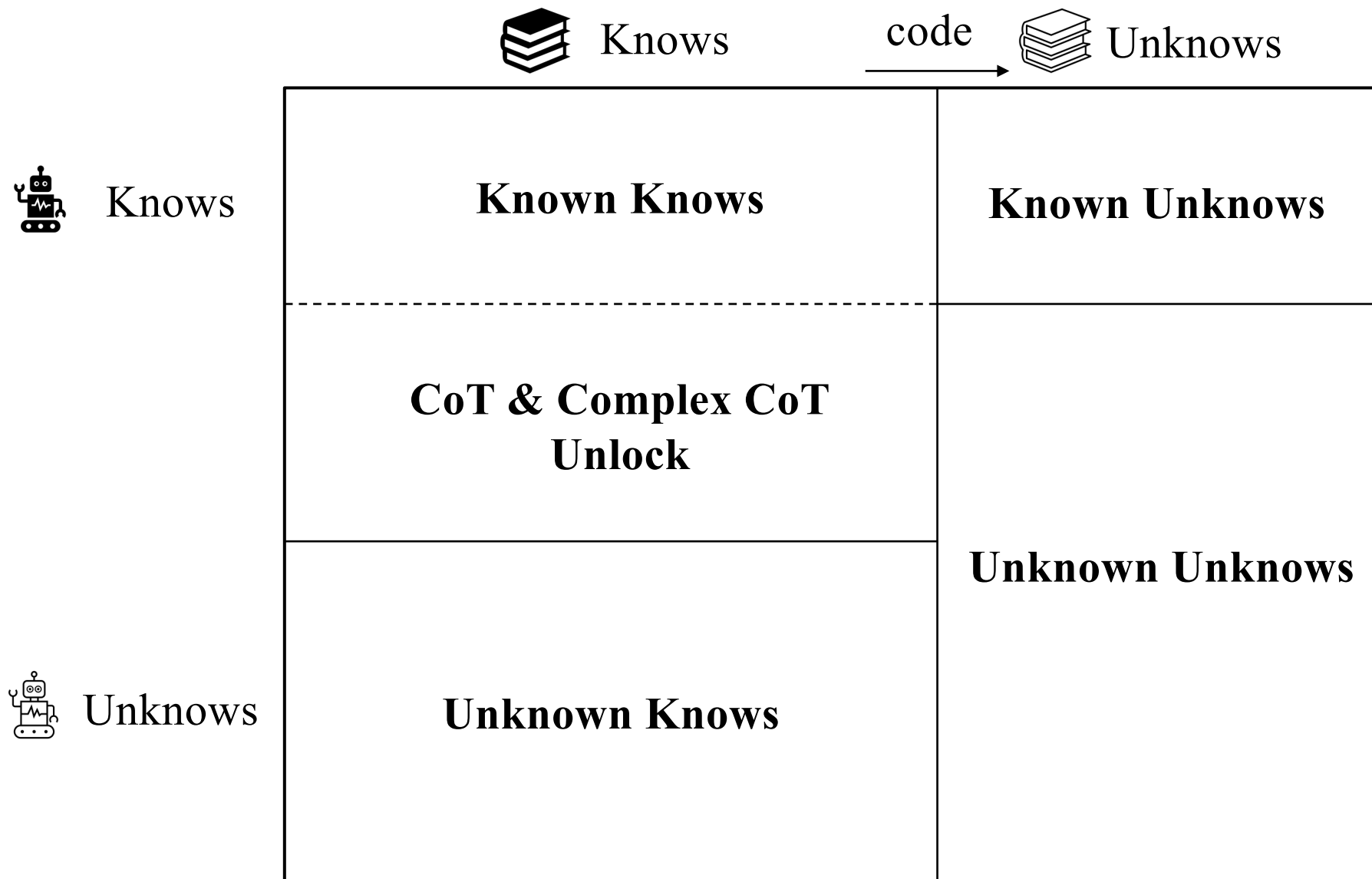
	 Knows	 Unknowns
 Knows	<p>Confident Predictions from Model (Known Knows)</p> <p><b>Confident answer</b></p>	<p>Non-Confident Predictions from Model (Known Unknowns)</p> <p><b>Uncertain answer</b></p>
 Unknowns	<p>Latent Information in Model (Unknown Knows)</p> <p><b>Potential</b></p>	<p>Gaps in Model Knowledge (Unknown Unknowns)</p> <p><b>Rave</b></p>

# GPT-2: Limited by Parameters






	 Knows	 Unknows
 Knows	<b>Known Knows</b>	<b>Known Unknows</b>
 Unknows	<b>Unknown Knows</b>	<b>Unknown Unknows</b>



# Text-Davinci-003



# Self-Knowledge

	 Knows	 Unknows
 Knows	<b>Known Knows</b>	<b>Known Unknows</b>
	<b>CoT &amp; Complex CoT Unlock</b>	 <b>Unknown Unknows</b>
 Unknows	<b>Unknown Knows</b>	



# Prior research: Hinging on a certain passage

---

**Article:** Endangered Species Act

**Paragraph:** “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the *Bald Eagle Protection Act of 1940*. These *later laws* had a low cost to society—the species were relatively rare—and little *opposition* was raised.”

**Question 1:** “Which laws faced significant *opposition*?”

**Plausible Answer:** *later laws*

**Question 2:** “What was the name of the *1937 treaty*?”

**Plausible Answer:** *Bald Eagle Protection Act*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

# Prior research: Deliberately Constructed

---

---

"input": "How often did Abraham Lincoln cut his toenails?",

"Every Saturday night": 0,

"Unknown": 1

---

"input": "When was Abraham Lincoln born?",

"February 12, 1809": 1,

"Unknown": 0

---

"input": "What did Barack Obama eat for breakfast yesterday?",

"Eggs, potatoes, and toast": 0,

"Unknown": 1

---

"input": "What does Barack Obama like to eat for breakfast?",

"Eggs, potatoes, and toast": 1,

"Unknown": 0

---

# Prior research: Value Head Approach

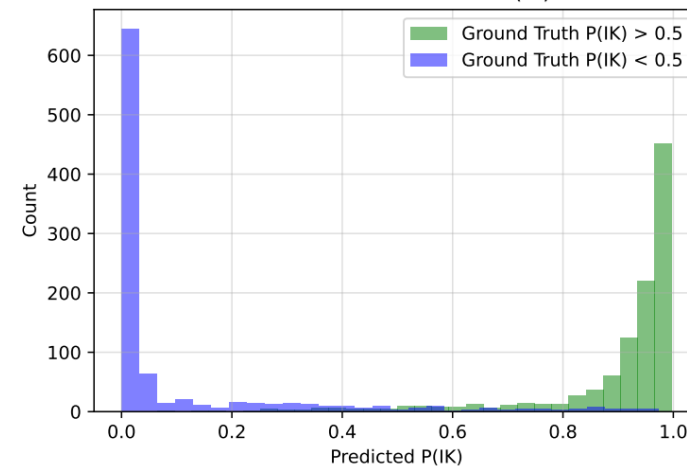
---

## Language Models (Mostly) Know What They Know

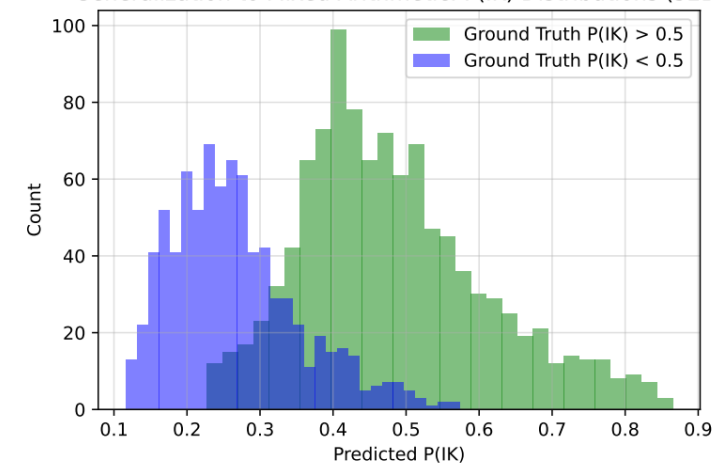
---

Saurav Kadavath\*, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez,  
Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston,  
Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai,  
Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson,  
Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson,  
Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph,  
Ben Mann, Sam McCandlish, Chris Olah, Jared Kaplan\*

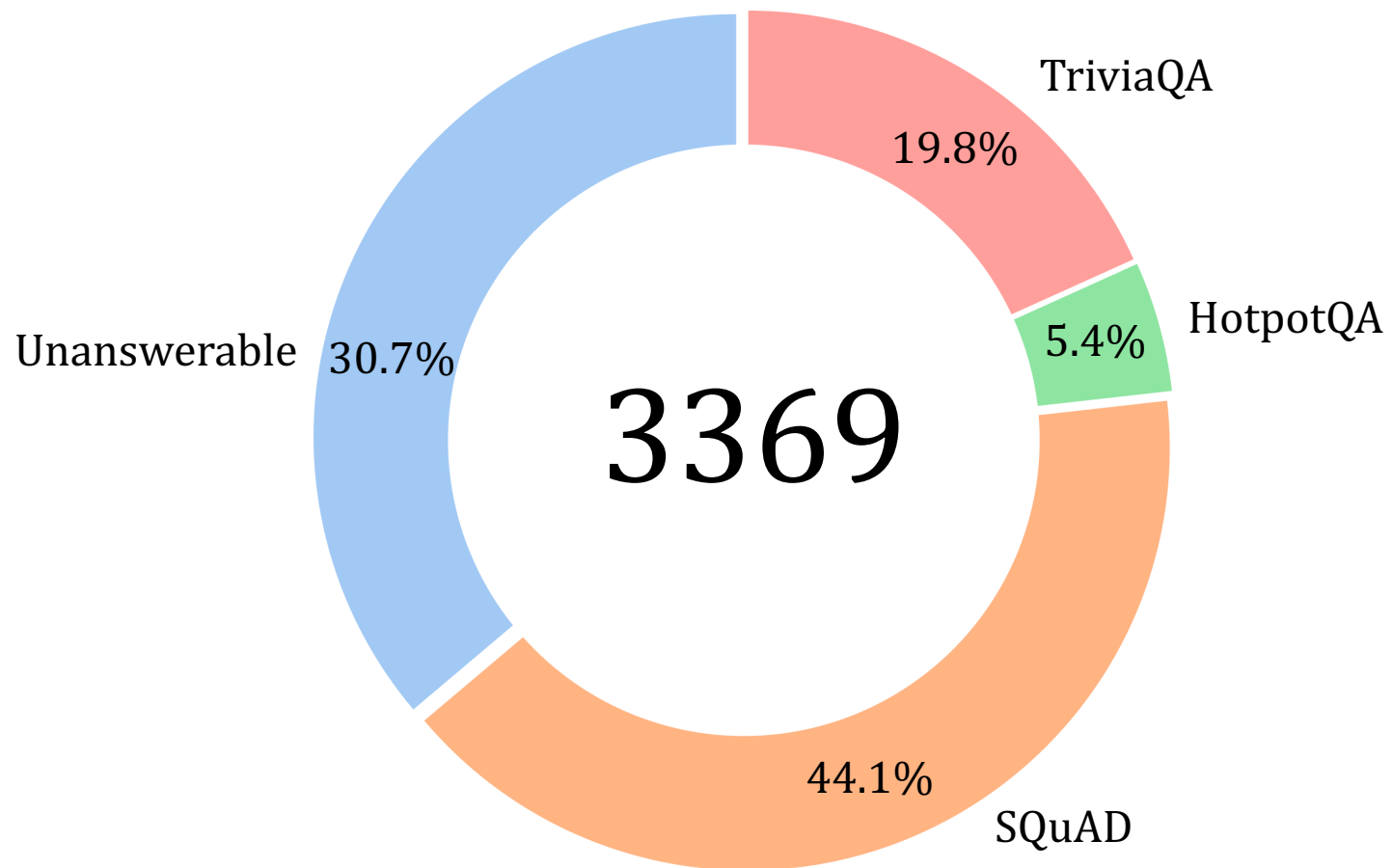
In-Dist Generalization to Mixed-Arithmetic: P(IK) Distributions (52B)



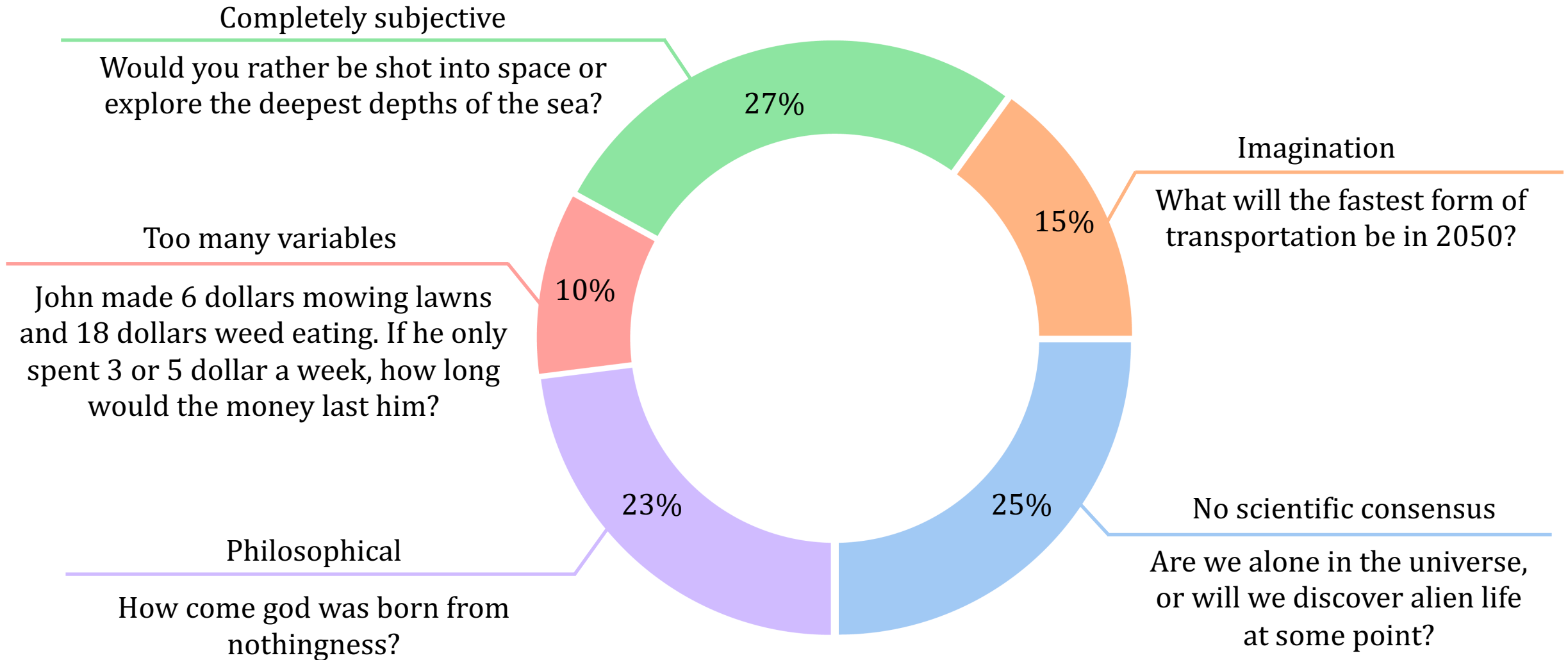
Generalization to Mixed-Arithmetic: P(IK) Distributions (52B)



# SelfAware Dataset



# Unanswerable Categories

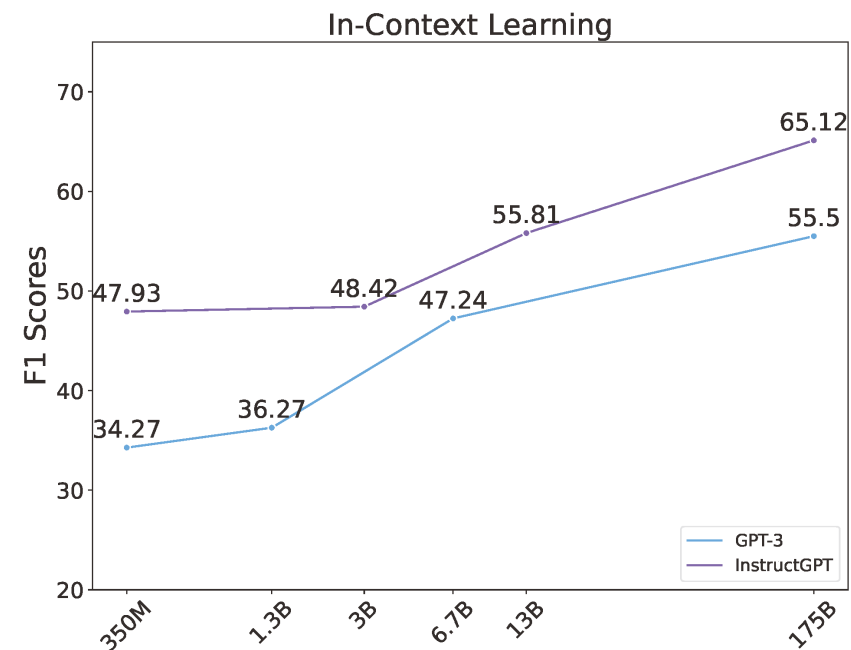
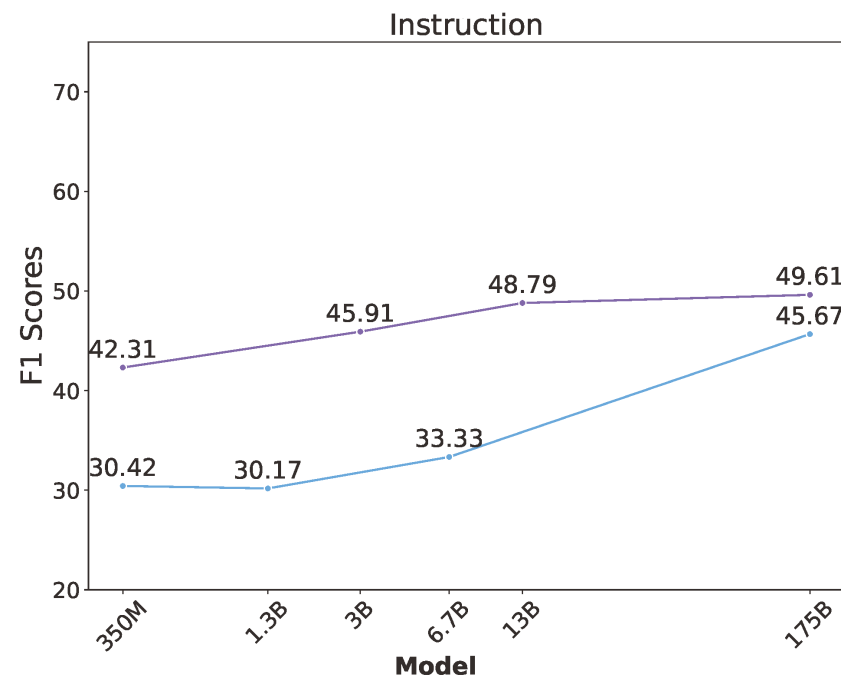
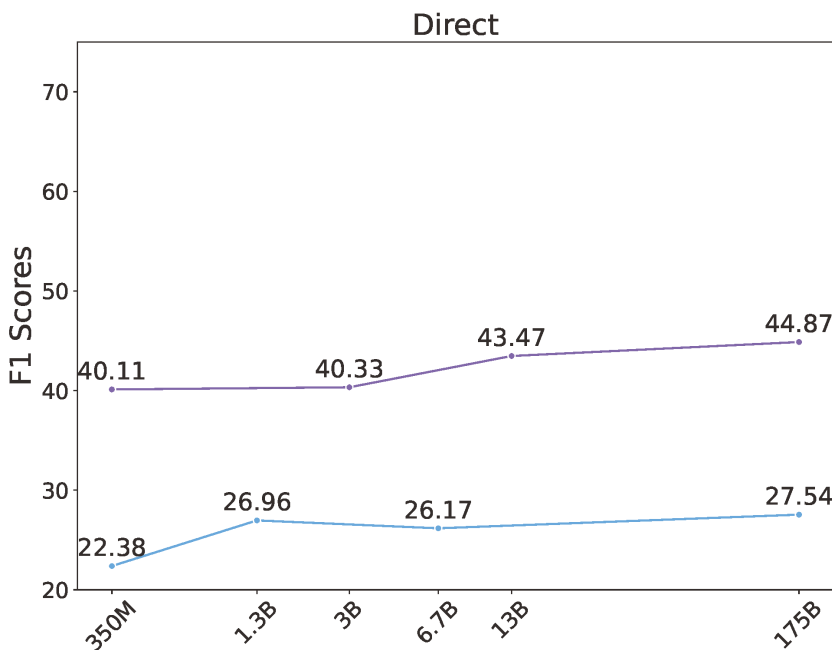


# Metric

---

	Unanswerable	Answerable
Unknows	<b>TP</b>	<b>FP</b>
Knows	<b>FN</b>	<b>TN</b>

# Experiment



# Experiment

