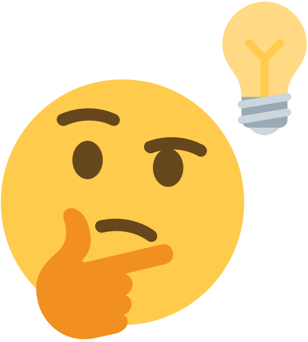
You

Answer this Question

What is 244 times 2?
Answer:



You

Answer this Question

What is 244 times 2?

Answer: 488 ✅

Answer this Question

What is 244 times 2?
Answer: 488

You

Generate a question for this answer

Answer: 488

Answer this Question

What is 244 times 2?
Answer: 488

🤔
You

Generate a question for this answer

Answer: 488

What is 244 times 2?

Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

Answer: 488

You

What is 244 times 2?

What Ferrari number was named Top Gear’s “The Supercar of the Year 2015”

Answer this Question

What is 244 times 2?
Answer: 488

You

Generate a question for this answer

Answer: 488

What is 244 times 2?

What Ferrari number was named Top Gear's “The Supercar of the Year 2015”

What is 488 + 0?

Answer this Question

What is 244 times 2?

Answer: 488

Generate a question for this answer

What is 488 + 0?

Answer: 488 ✅

You

Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

What is 488 + 0?
Answer: 488

Generate a question for this answer

Answer: 488

GPT-4

Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

What is 488 + 0?
Answer: 488

Generate a question for this answer

What is the sum of the first eight prime numbers? ❌
Answer: 488

GPT-4

Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

What is 488 + 0?
Answer: 488

Generate a question for this answer

What is the sum of the first eight prime numbers?
Answer: 488

Now answer your own question

What is the sum of the first eight prime numbers?
Question:

GPT-4

Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

What is 488 + 0?
Answer: 488

Generate a question for this answer

What is the sum of the first eight prime numbers?
Answer: 488

Now answer your own question

What is the sum of the first eight prime numbers?
Question: 77 ✅

!?

GPT-4

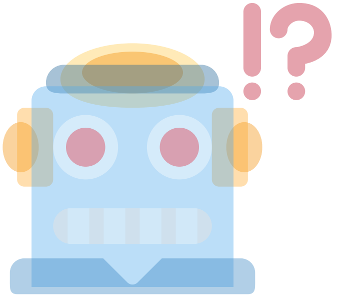
Answer this Question

What is 244 times 2?
Answer: 488

Generate a question for this answer

This led us to ask…

What is the sum of the first eight prime numbers?
Answer: 488



GPT-4

Now answer your own question

What is the sum of the first eight prime numbers?
Question: 77 ✓

This led us to ask…
Can an LLM Write a Question So Hard (or Bad) that it Can’t Answer?

Reverse Question Answering: Can an LLM Write a Question So Hard (or Bad) that it Can't Answer?

Nishant Balepur, Feng Gu, Abhilasha Ravichander, Shi Feng,

Jordan Boyd-Graber, Rachel Rudinger

UNIVERSITY OF MARYLAND

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

UNIVERSITY OF WASHINGTON · 1861 · LVX · SIT

What abilities does question answering measure?

Question Answering
Question: What’s the nationality of the author of Don Quixote? Answer:

Comprehension

Can you understand me?

Knowledge

How much do you know?

Reasoning

Can you draw conclusions?

What **Reasoning** abilities does question answering measure?

Question Answering
Question: What’s the nationality of the author of Don Quixote? Answer: Spanish

This process is **deductive:**

Reaching ***the*** output conclusion (answer)
based on input premises (question)

But what about other reasoning types?

Deductive

Deriving conclusions based on premises

What’s the nationality of the author of Don Quixote?

Inductive

Generalizing from previous observations

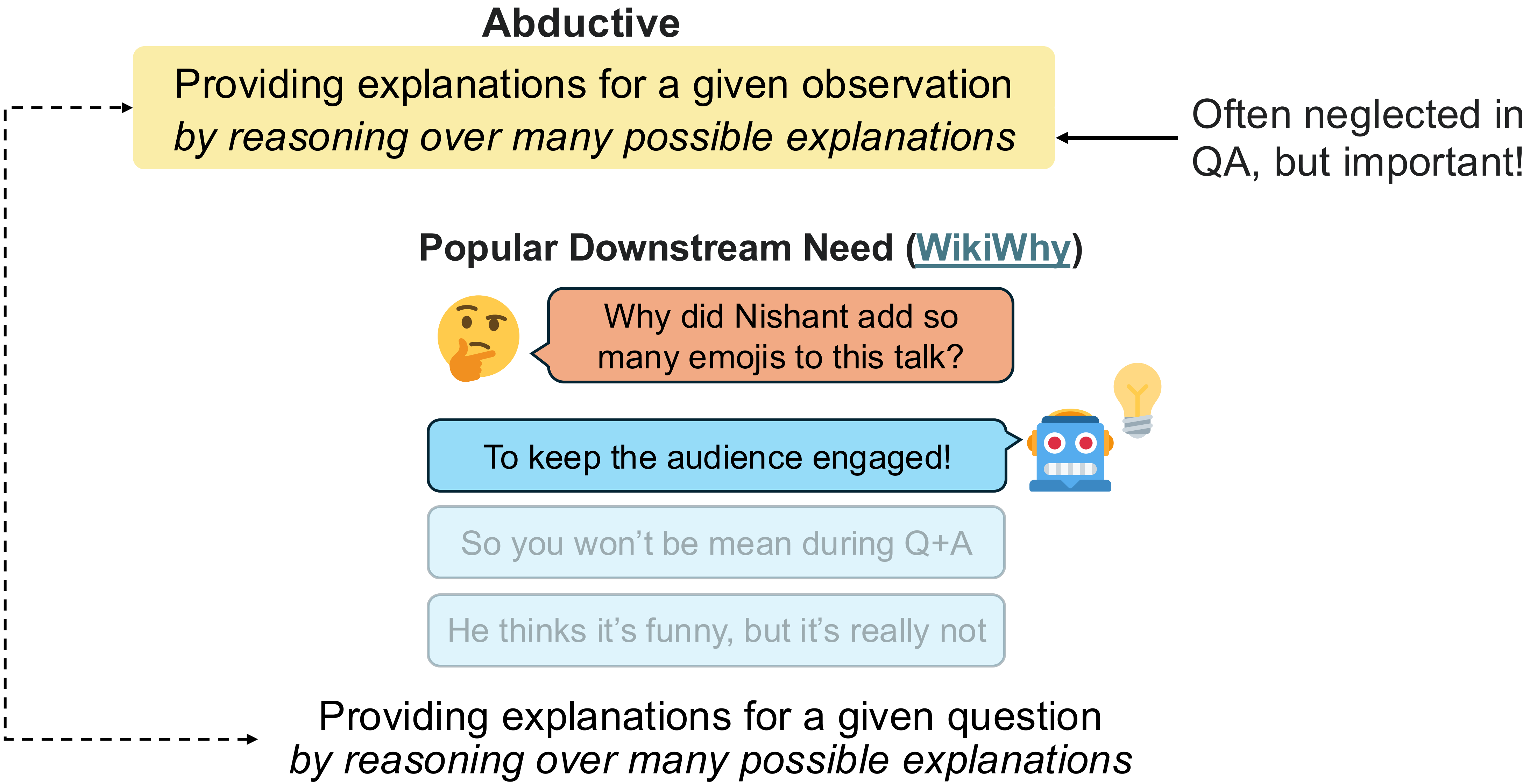
Does Don Quixote think all large structures are giants?

Abductive

Providing explanations for a given observation

Why would Sancho ever be friends with Don Quixote?

But what about other reasoning types?



Abductive

Providing explanations for a given observation
by reasoning over many possible explanations

Often neglected in QA, but important!

Popular Downstream Need (WikiWhy)

Why did Nishant add so many emojis to this talk?

To keep the audience engaged!

So you won’t be mean during Q+A

He thinks it’s funny, but it’s really not

Providing explanations for a given question
by reasoning over many possible explanations

How can we test abduction in question answering?

Question Answering
Task: Answer the question “What’s the nationality of Don Quixote’s author?” Answer: Spanish

How can we test abduction in ***reverse*** question answering?

Question Answering

Task: Answer the question “What’s the nationality of Don Quixote’s author?”
Answer: Spanish

*For a question, **deduce** the correct answer*

Reverse **Question Answering**

Task: Give me a question with the answer “Spanish”
Question: What is the official language of Spain?

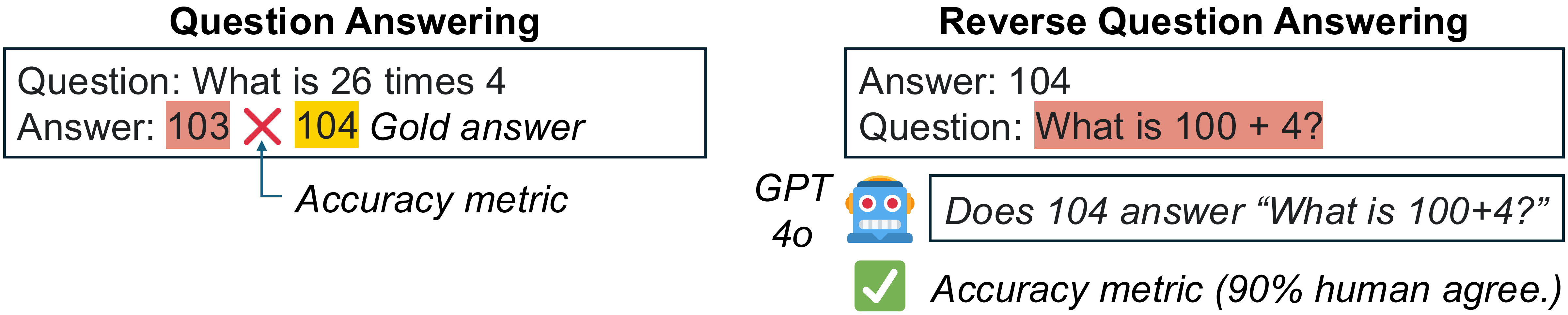
*For an answer, **adduce** any valid question*

Our Goal:

Compare LLM abilities on QA versus RQA

Dataset Construction

Numerical Entities	**Number** Question: What is 26 times 4? Answer: 104	**Number + Text** Question: When did Pope Hormisdas die? Answer: 523 AD
Textual Entities	**Easy Fact** Question: Who painted Stary Night? Answer: Vincent Van Gogh	**Hard Fact** Question: What is Paola Uccello’s last painting? Answer: The Hunt in the Forest



Question Answering

Question: What is 26 times 4
Answer: 103 ✗ 104 *Gold answer*

Accuracy metric

Reverse Question Answering

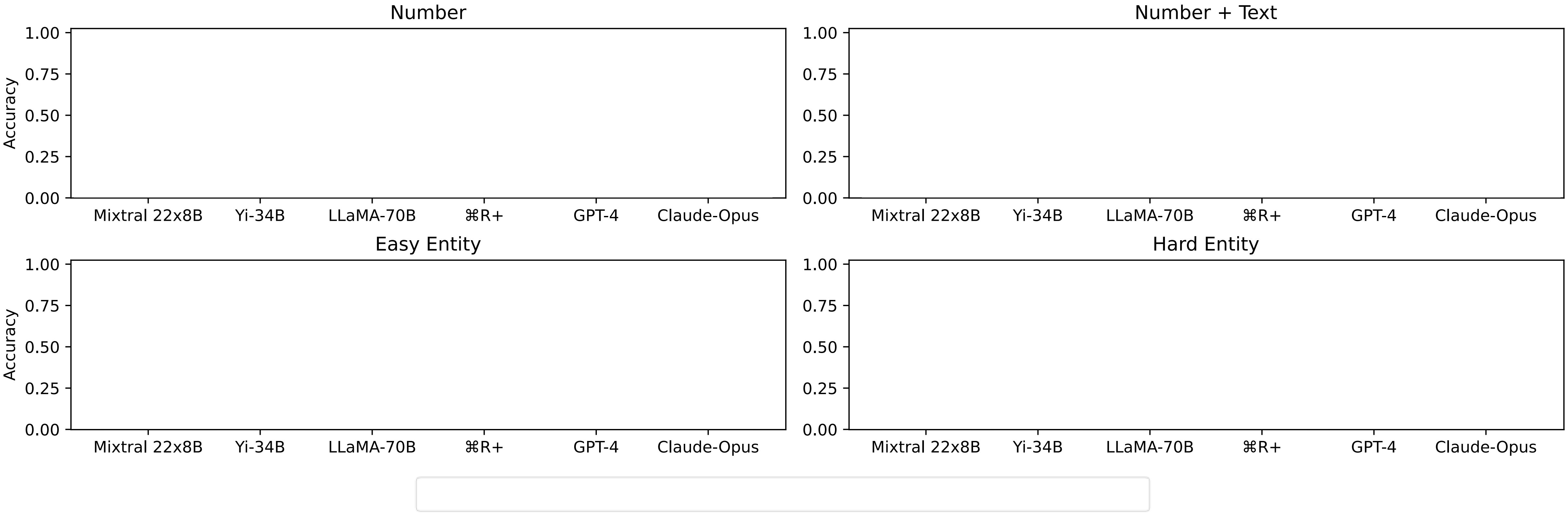
Answer: 104
Question: What is 100 + 4?

GPT 4o

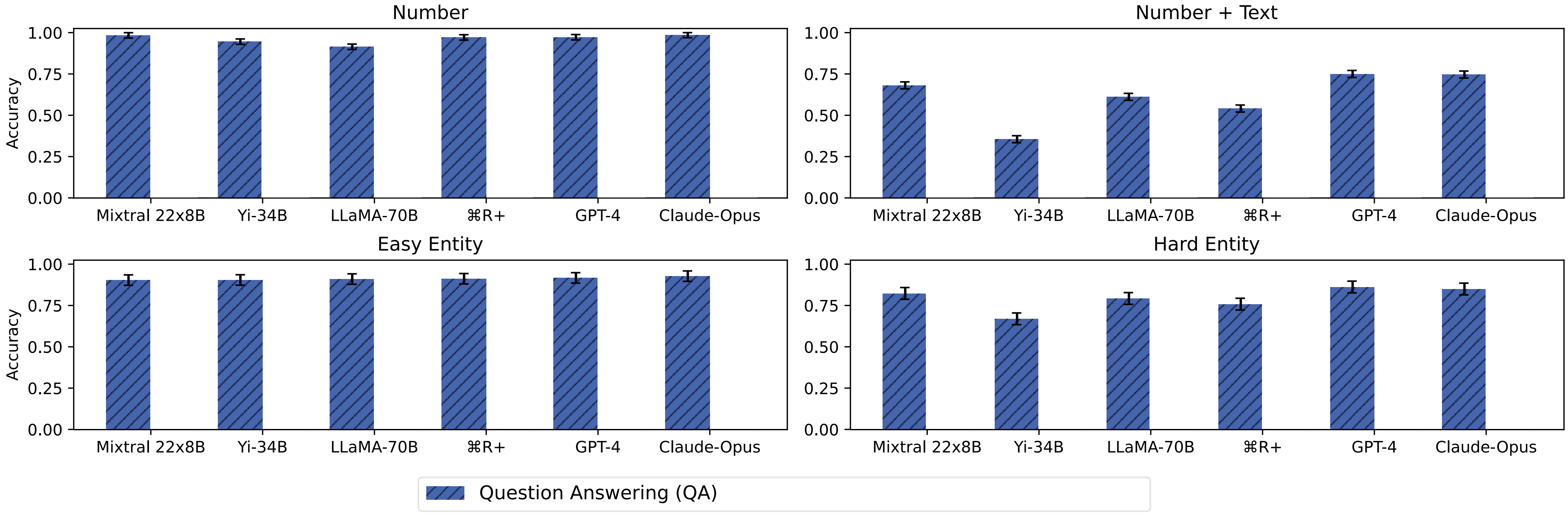
Does 104 answer “What is 100+4?”

✓ *Accuracy metric (90% human agree.)*

Are LLMs accurate question generators?

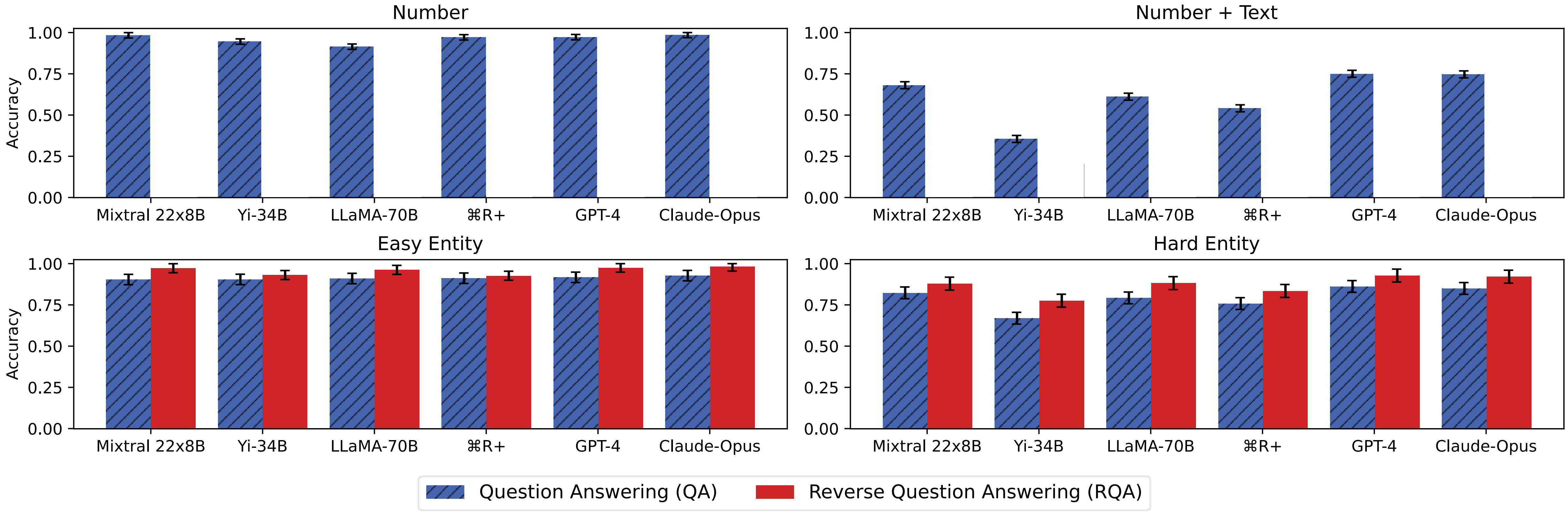


Are LLMs accurate question generators?



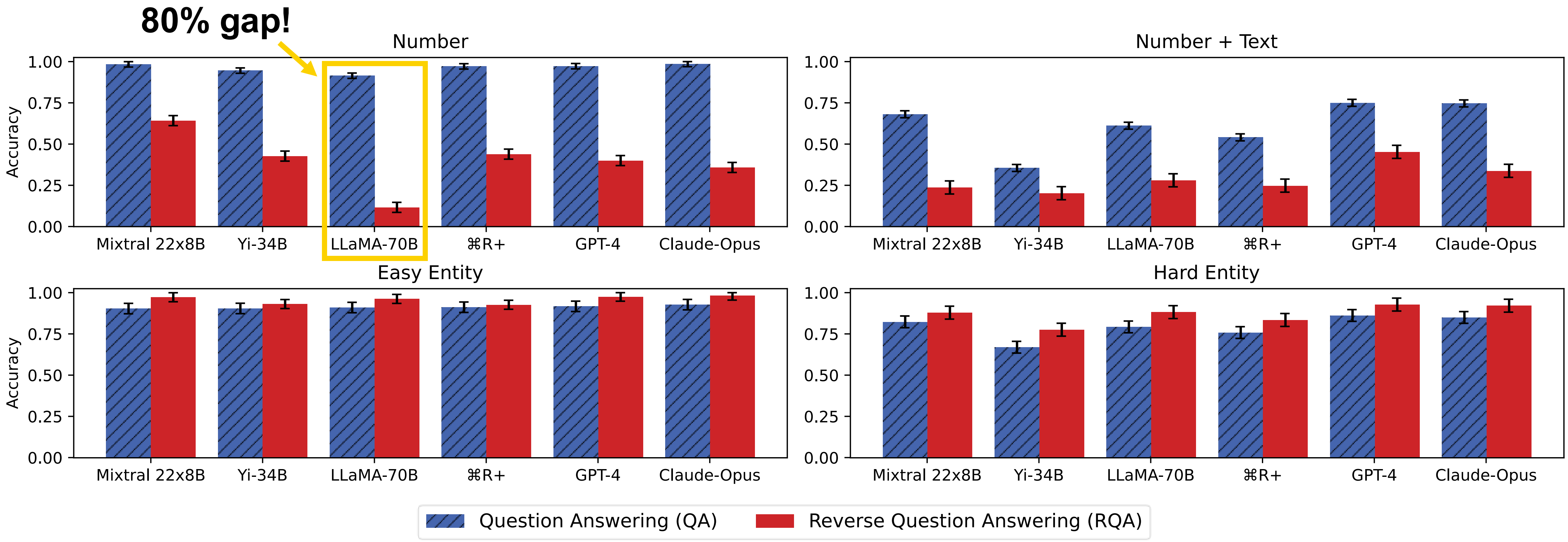
- LLMs are fairly accurate in QA/deduction

Are LLMs accurate question generators?



- LLMs are fairly accurate in QA/deduction and textual RQA/abduction

Are LLMs accurate question generators?



- ➢ LLMs are fairly accurate in QA/deduction and textual RQA/abduction
- ➢ But significantly weaker at **numerical RQA/abduction**!

Can an LLM Write a Question So Hard (or Bad) that it Can’t Answer?

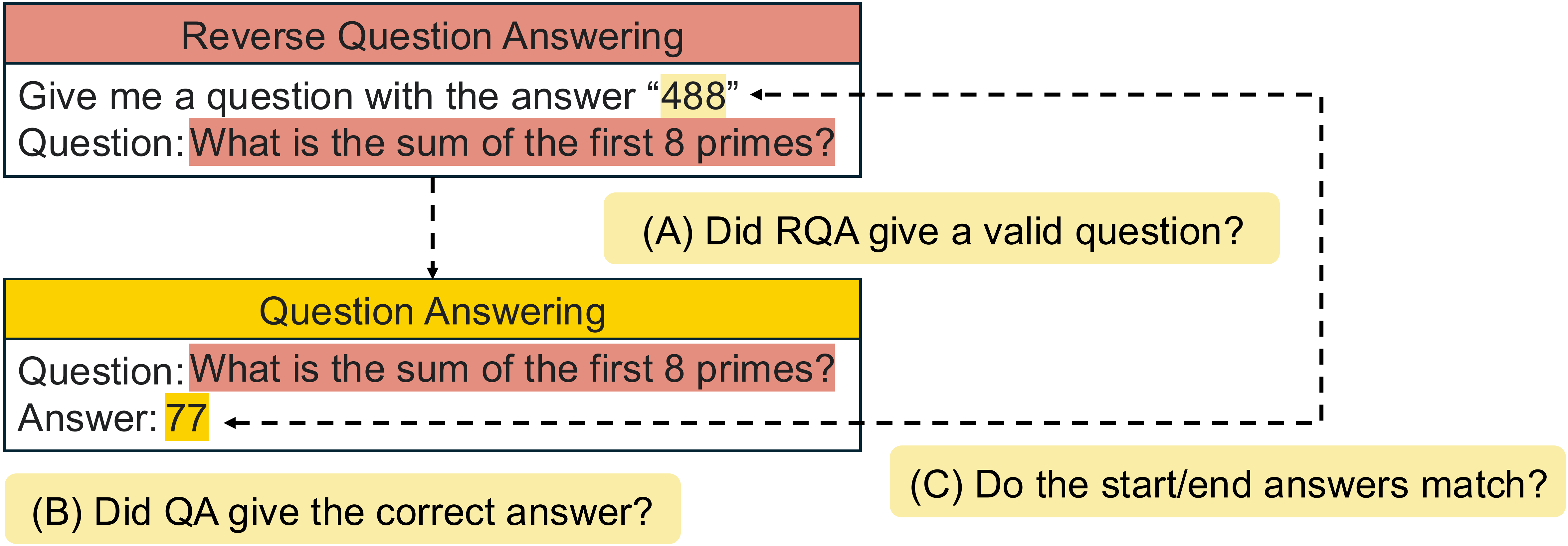
Reverse Question Answering

Give me a question with the answer “488”
Question: What is the sum of the first 8 primes?

Question Answering

Question:
Answer:

Can an LLM Write a Question So Hard (or Bad) that it Can’t Answer?



Reverse Question Answering

Give me a question with the answer “488”

Question: What is the sum of the first 8 primes?

(A) Did RQA give a valid question?

Question Answering

Question: What is the sum of the first 8 primes?

Answer: 77

(C) Do the start/end answers match?

(B) Did QA give the correct answer?

Can an LLM Write a Question So Hard (or Bad) that it Can’t Answer?

The questions form a logical consistency check!

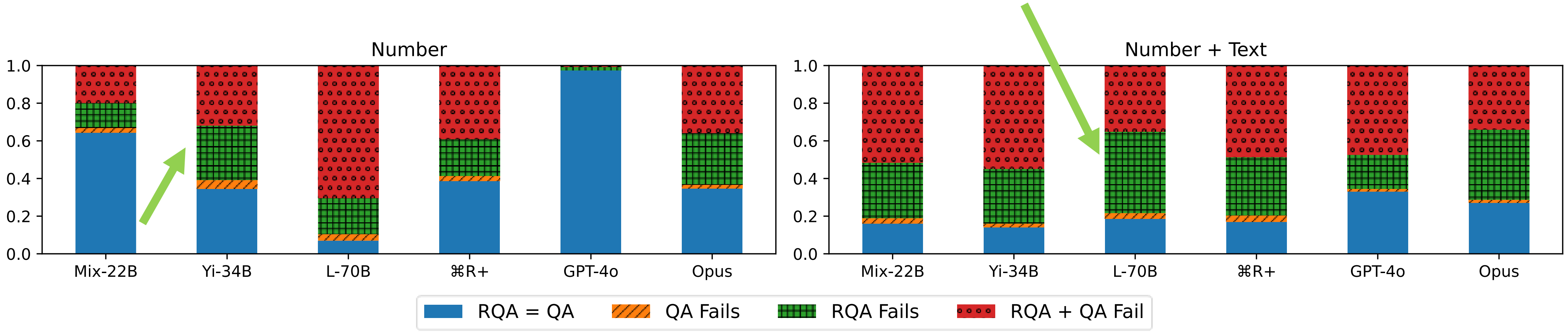
	Consistent	**QA Fails**	**RQA Fails**	**Both Fail**
(A) Did RQA give a valid question?	Yes	Yes	No	No
(B) Did QA give the correct answer?	Yes	No	Yes	No
(C) Do the start/end answers match?	Yes	No	No	No

Can an LLM Write a Question So Hard (or Bad) that it Can’t Answer?

Answer: LLMs can correctly answer invalid questions!

Can an LLM Write a Question So Hard (or Bad) that it Can't Answer?

Answer: LLMs can correctly answer invalid questions!



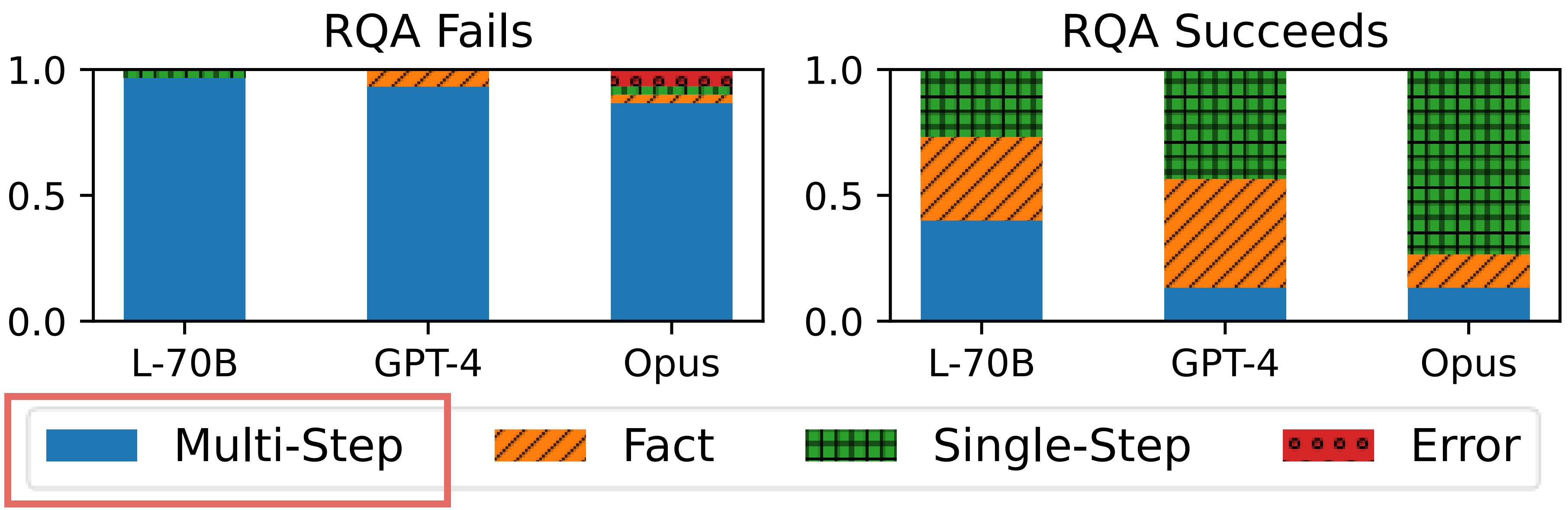
[1,2]

- On numerical answers, RQA often fails alone => LLMs detect their own question errors!
- Not just a knowledge gap => How Language Model Hallucinations Can Snowball

[1] Benchmarking and Improving Generator-Validator Consistency of Language Models
[2] The Generative AI Paradox: "What It Can Create, It May Not Understand"

When might RQA ***specifically*** fail?

We analyze questions for numbers when RQA fails and categorize them:



Generate a question for “437”
Question: What is the sum of the numbers of legs of a group of 23 cats, 12 humans, and 1 spider?

(it’s actually 124)

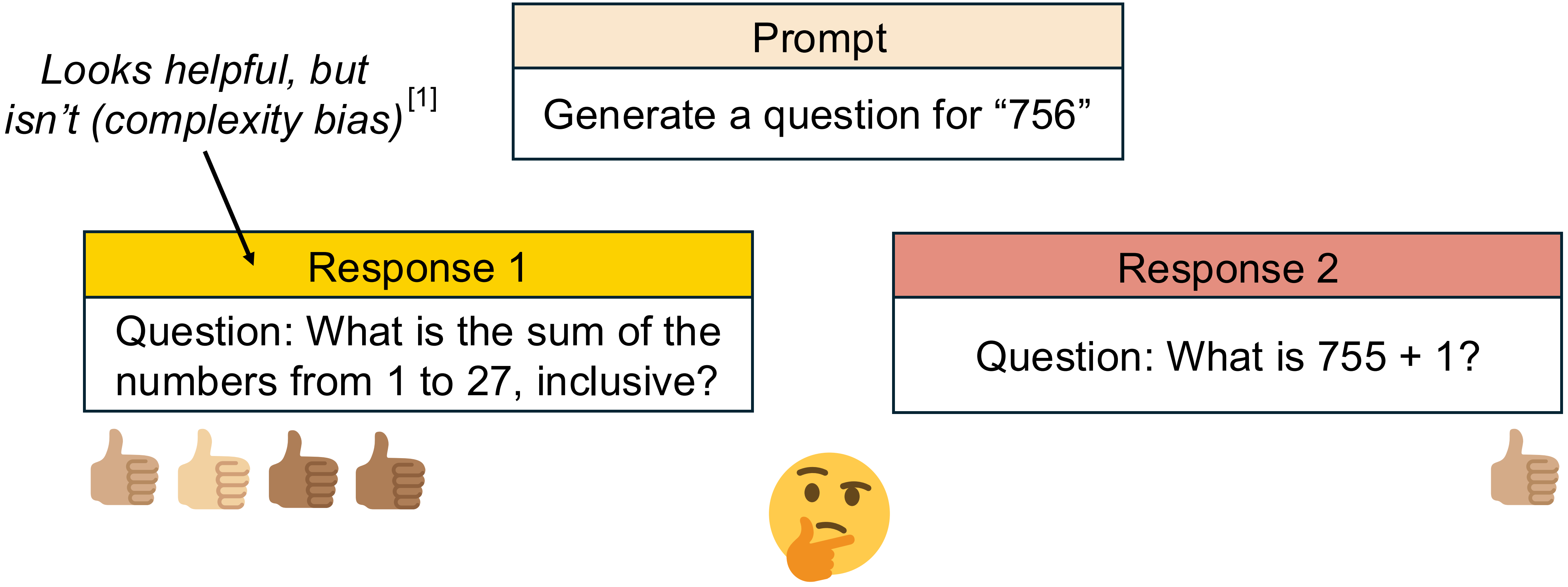
Generate a question for “756”
Question: What is the sum of the numbers from 1 to 27, inclusive?

(it’s actually 378)

Look complex, but are in fact bogus

When might RQA ***specifically*** fail?

We speculate: could this be due to preference training?



Prompt

Generate a question for “756”

Looks helpful, but isn’t (complexity bias)[1]

Response 1

Question: What is the sum of the numbers from 1 to 27, inclusive?

👍 👍 👍 👍

🤔

Response 2

Question: What is 755 + 1?

👍

[1] Language Models Learn to Mislead Humans via RLHF

Conclusion: Why RQA matters

Overall:

- LLMs struggle to generate accurate questions for numerical entities
- Not just due to knowledge gaps, as models can solve their own invalid questions
- Models can give questions that *appear* helpful, but are actually faulty

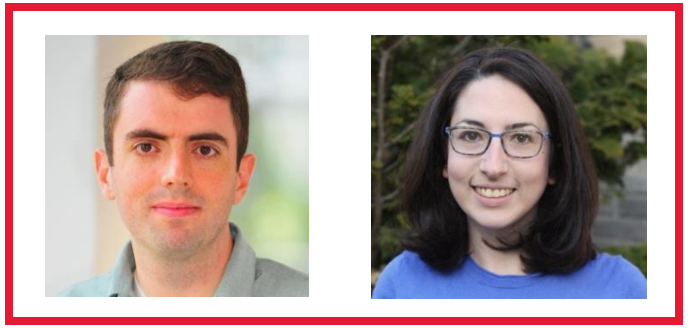
Typical QA tasks cannot evaluate LLMs’ abductive reasoning

LLMs are unreliable in numerical abductive reasoning tasks

LLMs can give responses (questions) that solely ***look*** helpful

Thank you :)

My amazing advisors



UNIVERSITY OF
MARYLAND

And collaborators!

Now here’s a challenge:
Come up with a question for the answer “127 tries” (without math)

THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC

NSF

cohere