Language Models: A Reality Check

Zhaofeng Wu

September 2023

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Andrew Ng 
@AndrewYNg

I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?

12:36 PM · Jun 5, 2023 · **815.6K** Views

Andrew Ng
@AndrewYNg

I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?

12:36 PM · Jun 5, 2023 · **815.6K** Views

Gary Marcus
@GaryMarcus

Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · **31.4K** Views

Andrew Ng
@AndrewYNg

I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?

12:36 PM · Jun 5, 2023 · **815.6K** Views

Gary Marcus
@GaryMarcus

Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · **31.4K** Views

@emilymbender@dair-community.social on Mastodon
@emilymbender

And not fall for either-

Myth #1: The tech is moving to fast! Regulation can't keep up.

Myth #2: The 'real' concern is rogue AGI that poses 'existential risk' to humanity.

9:50 AM · May 21, 2023 · **12.5K** Views

Andrew Ng
@AndrewYNg

I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?

12:36 PM · Jun 5, 2023 · **815.6K** Views

Gary Marcus
@GaryMarcus

Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · **31.4K** Views

@emilymbender@dair-community.social on Mastodon
@emilymbender

And not fall for either-

Myth #1: The tech is moving to fast! Regulation can't keep up.

Myth #2: The 'real' concern is rogue AGI that poses 'existential risk' to humanity.

9:50 AM · May 21, 2023 · **12.5K** Views

"An LM is a system for haphazardly stitching together sequences of linguistic forms [...] but without any reference to meaning: a stochastic parrot."

Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021.

Limitations of LM Capabilities

Limitations of LM Capabilities

"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.

Limitations of LM Capabilities

"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

"We let GPT-2 complete the simple arithmetic problem *Three plus five equals*. [...] this problem is beyond the current capability of GPT-2, and, we would argue, any pure LM."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.

Limitations of LM Capabilities

"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.

Limitations of LM Capabilities

"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

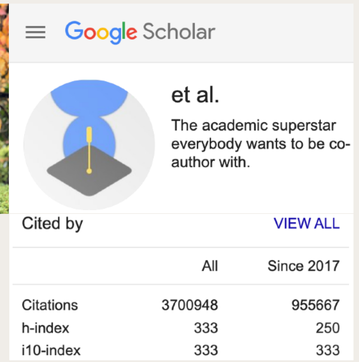
"It is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.
Patel and Pavlick. Mapping Language Models to Grounded Conceptual Spaces. ICLR 2022.

Limitations of LM Capabilities

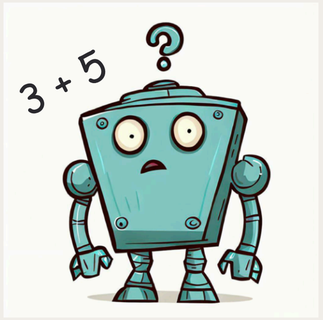
"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

"It is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way."



LMs can and do leverage structured relationships in pertaining data to learn such relational meaning.

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.
Patel and Pavlick. Mapping Language Models to Grounded Conceptual Spaces. ICLR 2022.
Wu et al. Transparency Helps Reveal When Language Models Learn Meaning. TACL 2023.


3 + 5
?

Types of LM Capabilities

Types of LM Capabilities

- Language knowledge (formal competence)

- Task abilities (functional competence)

Mahowald et al. Dissociating language and thought in large language models: a cognitive perspective. 2023.

Types of LM Capabilities

- Language knowledge (formal competence)
  - LMs do not represent ambiguities in language well
- Task abilities (functional competence)

Mahowald et al. Dissociating language and thought in large language models: a cognitive perspective. 2023.

Types of LM Capabilities

- Language knowledge (formal competence)
  - LMs do not represent ambiguities in language well
- Task abilities (functional competence)
  - LMs fail at simple variants of common tasks

Mahowald et al. Dissociating language and thought in large language models: a cognitive perspective. 2023.

ALISA LIU
ZHAOFENG WU
JULIAN MICHAEL
ALANE SUHR
PETER WEST
ALEXANDER KOLLER
SWABHA SWAYAMDIPTA
NOAH SMITH
YEJIN CHOI

We're Afraid Language Models Aren't Modeling Ambiguity

Ambiguity is inherent in language

Ambiguity is inherent in language

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

He is on track to complete his PhD in three years.

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

He is on track to complete his PhD in three years.

That's really fast! 

Ambiguity is inherent in language

Miscommunication

👩🏽‍⚕️ He is on track to complete his PhD in three years.

That's really fast! 👨🏼‍🎓

👩🏽‍⚕️ Oh, no, he’s already in his fifth year.

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

👩🏽‍⚕️ He is on track to complete his PhD in three years.

That's really fast! 👩🏽‍🎓

👩🏽‍⚕️ Oh, no, he's already in his fifth year.

Deception

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

👩🏽‍⚕️ He is on track to complete his PhD in three years.

That's really fast! 👨🏼‍🎓

👩🏽‍⚕️ Oh, no, he's already in his fifth year.

Deception

👨🏻‍⚕️ Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

👩🏽‍⚕️ He is on track to complete his PhD in three years.

That's really fast! 👩🏼‍🎓

👩🏽‍⚕️ Oh, no, he's already in his fifth year.

Deception

👨🏻‍💼 Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.

That's amazing! ... But wait, it was already the 2nd before you took office?? 🕵🏾‍♂️

https://www.instagram.com/punhubonline

Ambiguity is inherent in language

Miscommunication

👩🏽‍⚕️ He is on track to complete his PhD in three years.

That's really fast! 👩🏽‍🎓

👩🏽‍⚕️ Oh, no, he's already in his fifth year.

Deception

👨🏻‍⚕️ Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.

That's amazing! ... But wait, it was already the 2nd before you took office?? 👨🏽‍🔍

👨🏻‍⚕️ I didn't say it wasn't.

https://www.instagram.com/punhubonline

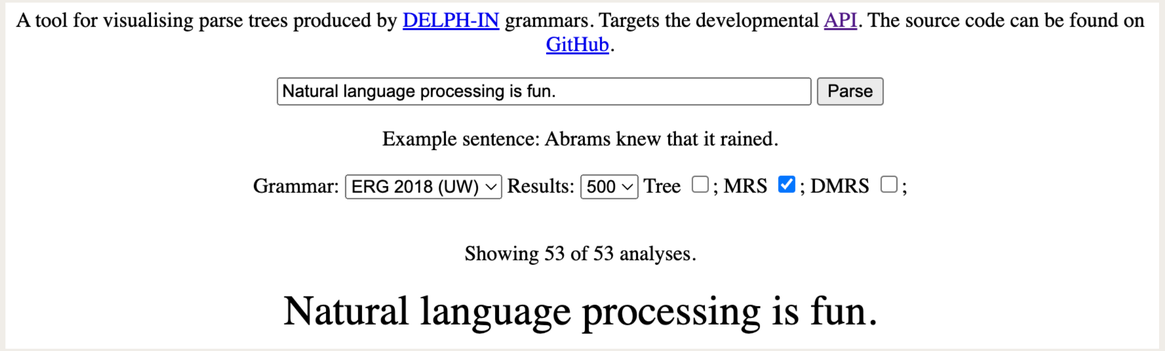
How well do LMs capture ambiguity?

Operationalizing Ambiguity

Operationalizing Ambiguity

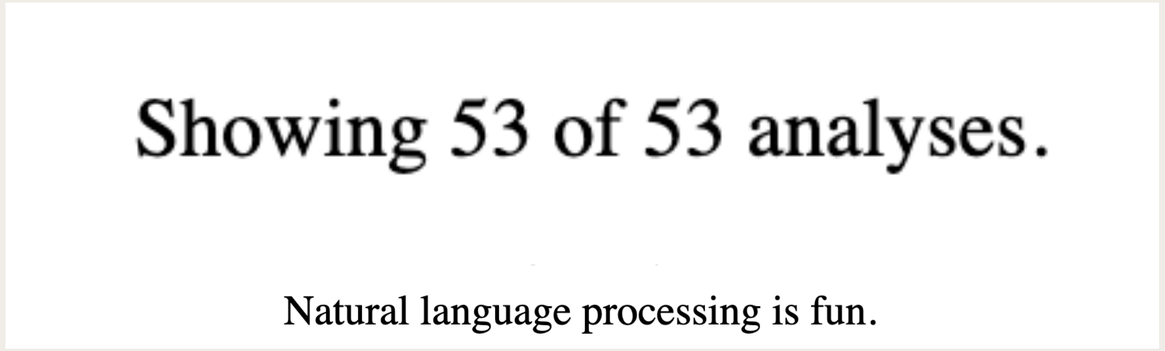
Natural language processing is fun.

Operationalizing Ambiguity



https://delph-in.github.io/delphin-viz/demo

Operationalizing Ambiguity



https://delph-in.github.io/delphin-viz/demo

Operationalizing Ambiguity

Operationalizing Ambiguity

Alisa saw the dog bowl.

Operationalizing Ambiguity

Alisa saw the dog bowl.

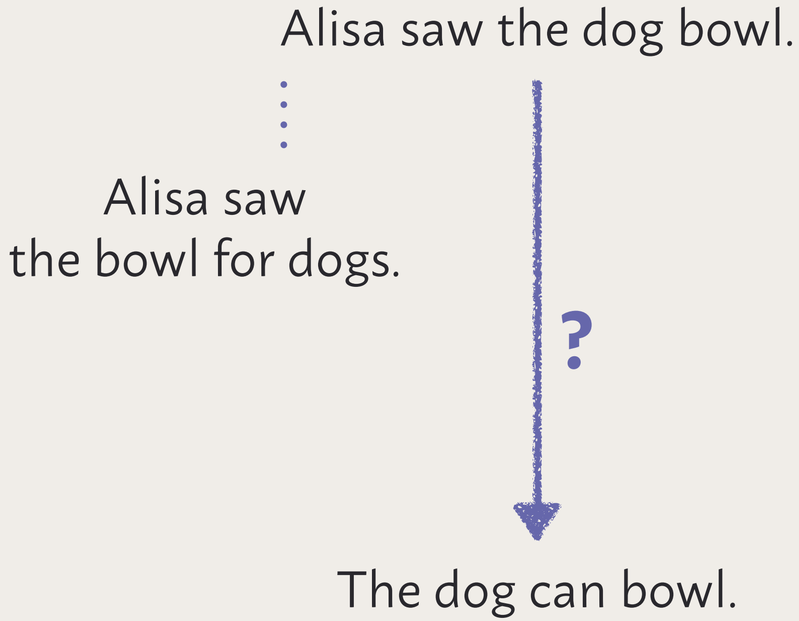
The dog can bowl.

Operationalizing Ambiguity

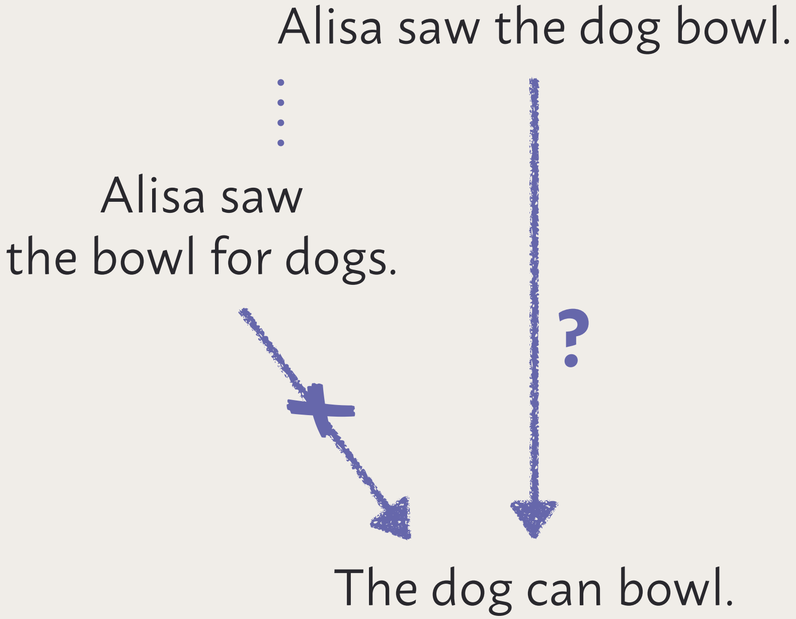
Alisa saw the dog bowl.

The dog can bowl.

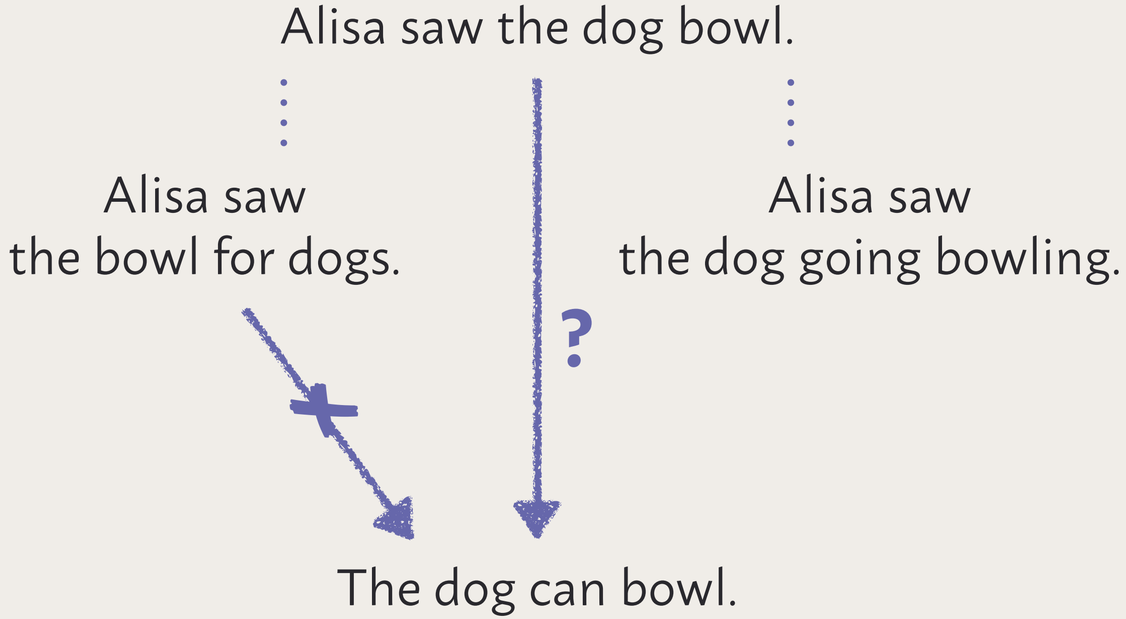
Operationalizing Ambiguity



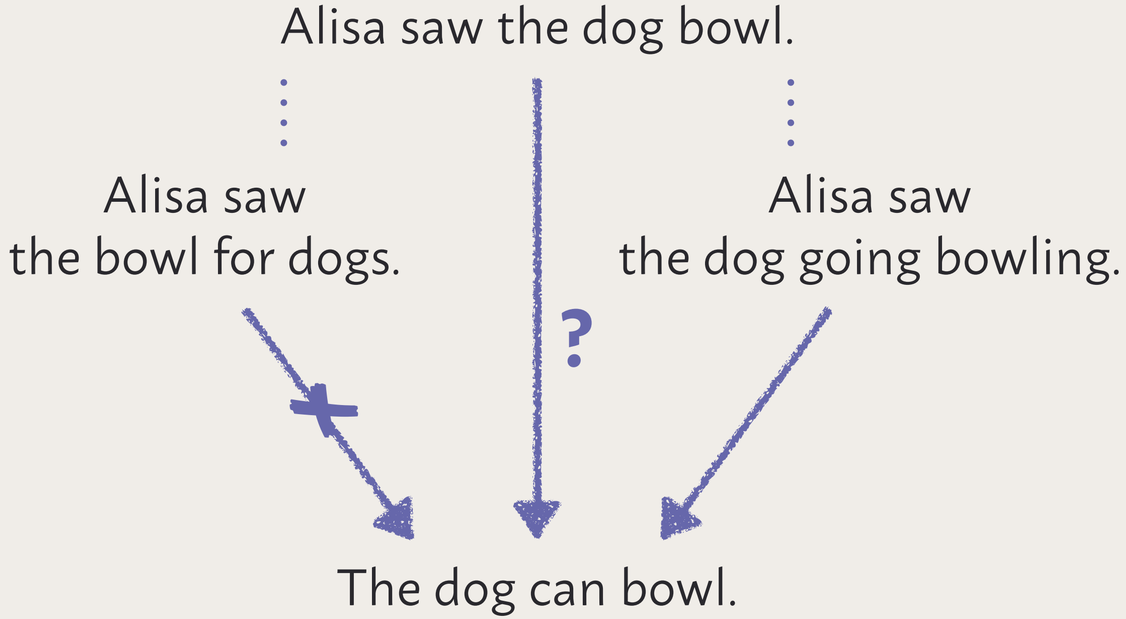
Operationalizing Ambiguity



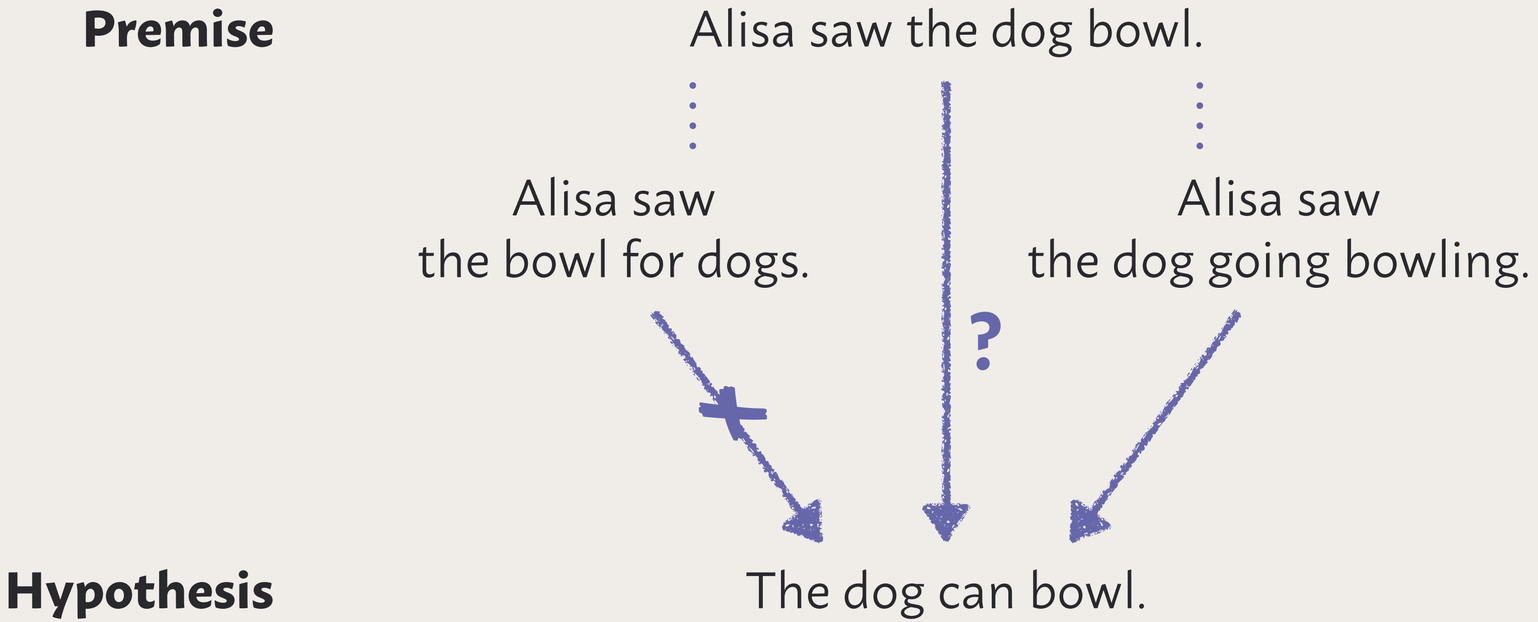
Operationalizing Ambiguity



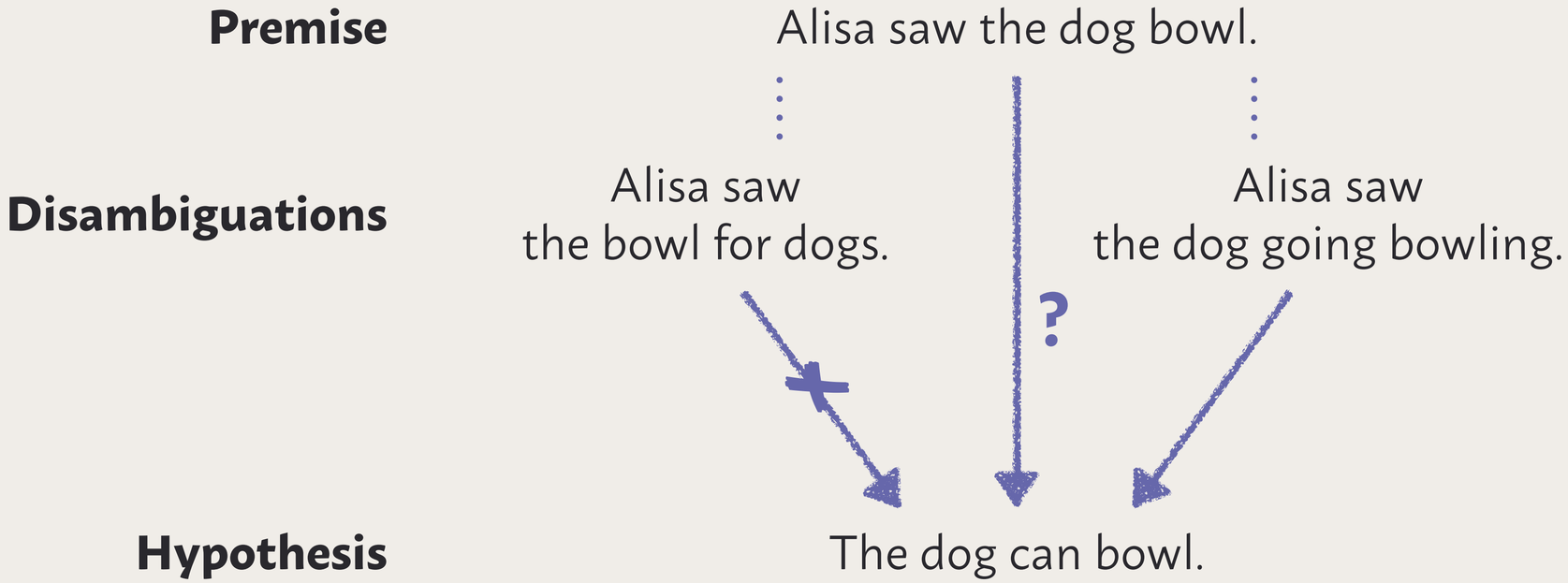
Operationalizing Ambiguity



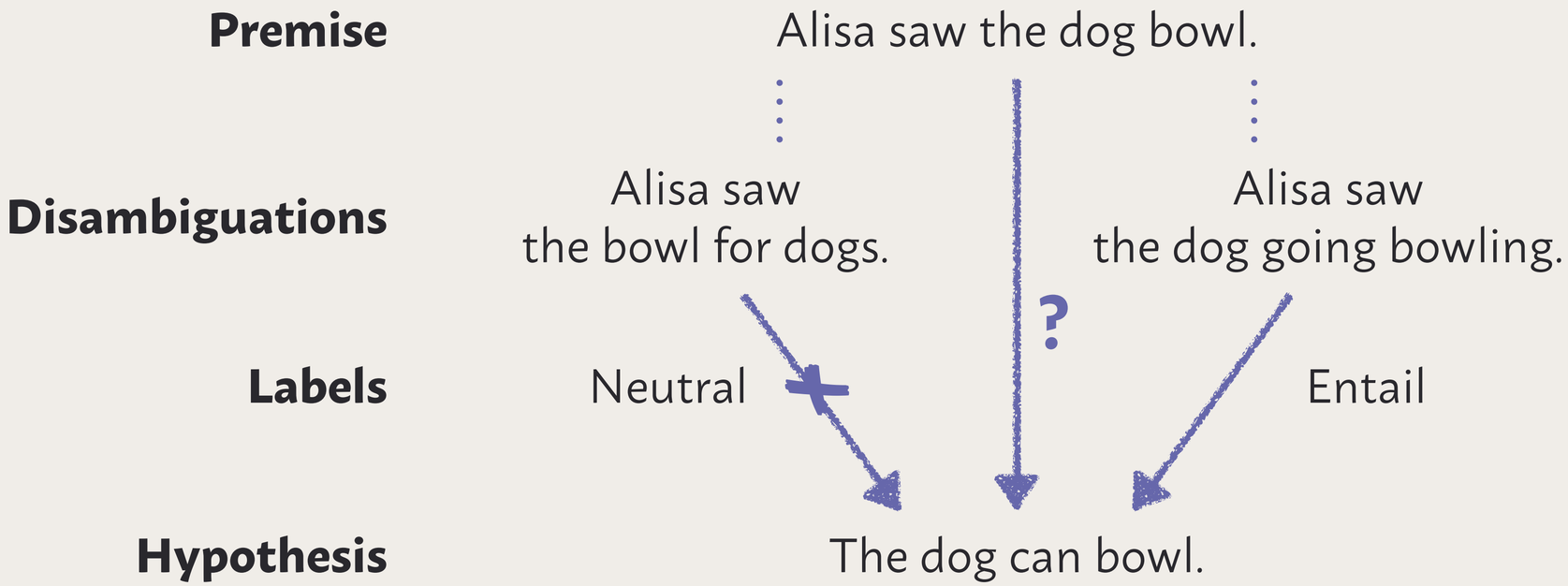
Operationalizing Ambiguity



Operationalizing Ambiguity



Operationalizing Ambiguity



Ambiguity in **Ent**ailment

AmbiEnt Collection

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖 Alisa saw the dog bowl. The dog can bowl.

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖

Alisa saw the dog bowl. The dog can bowl.

👩🏼‍🎓

Entail Neutral Contradict

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖

Alisa saw the dog bowl. The dog can bowl.

👩🏼‍🎓

Entail Neutral Contradict

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖

Alisa saw the dog bowl. The dog can bowl.

👩🏼‍🎓

Entail	Neutral	Contradict
Alisa saw the dog going bowling.	Alisa saw the dog's bowl.	

AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples → 🤖 Alisa saw the dog bowl. The dog can bowl.

👩🏼‍🎓 Entail: Alisa saw the dog going bowling. Neutral: Alisa saw the dog's bowl. Contradict

👨🏽‍🎓 Entail Neutral Contradict

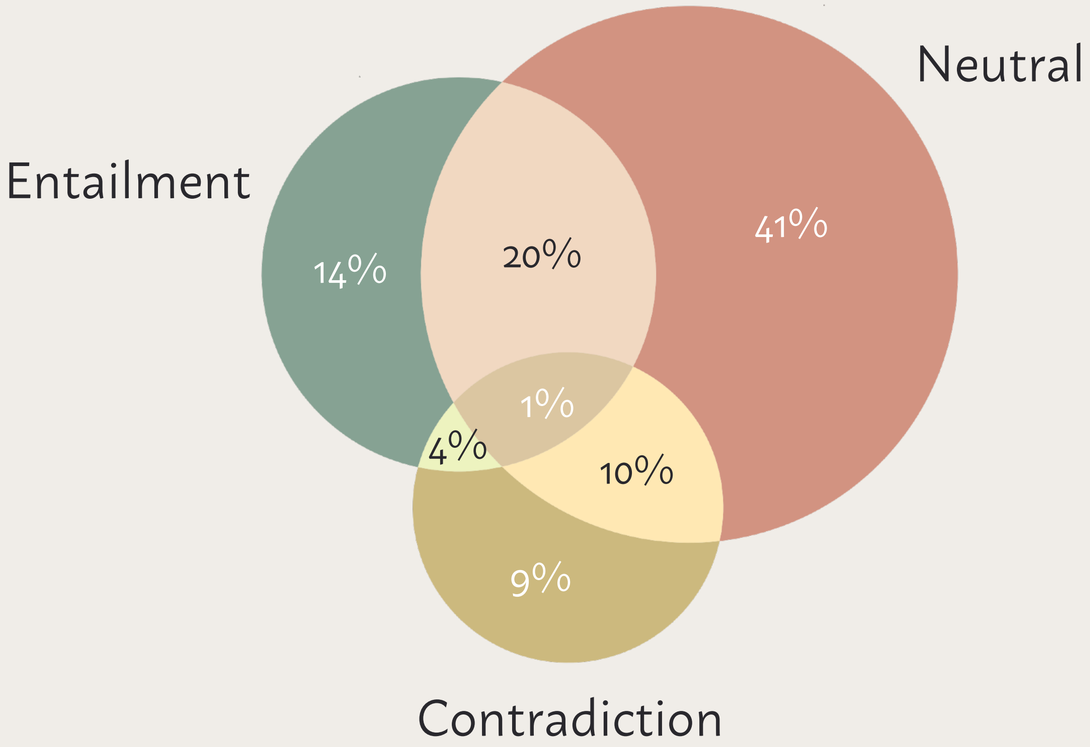
AmbiEnt Collection

- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

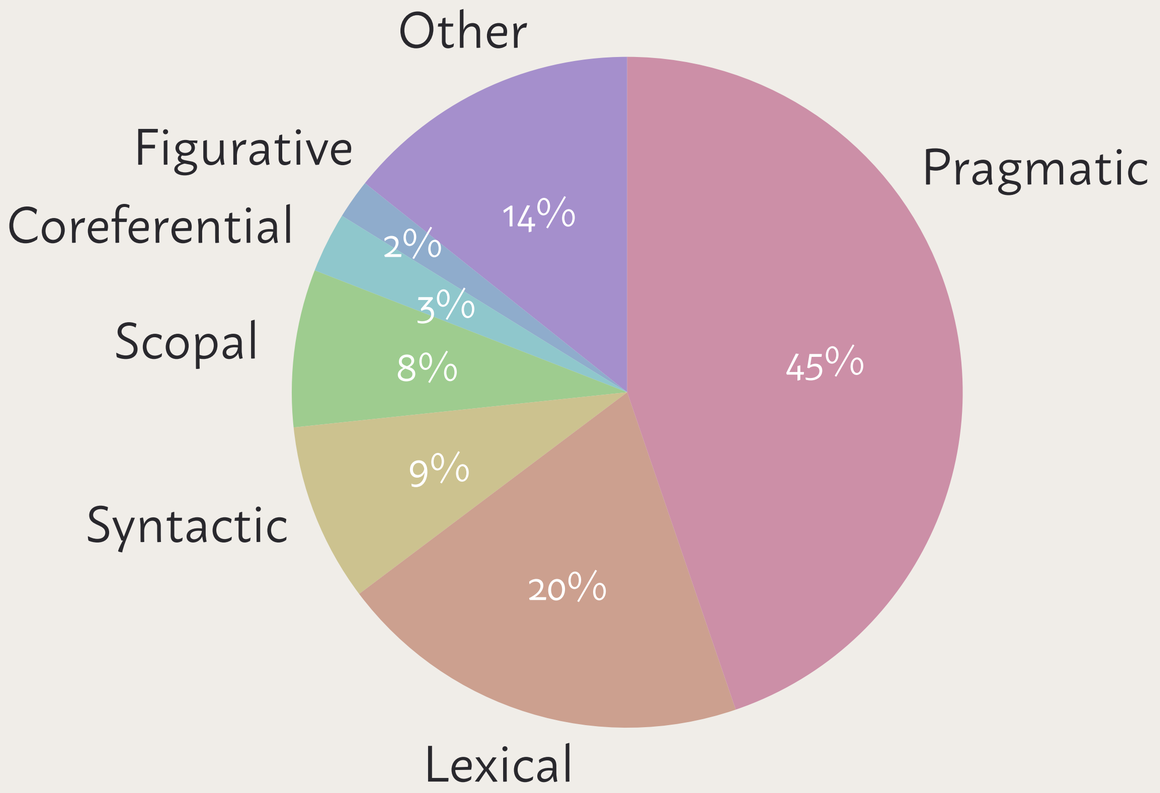
ICL examples → 🤖 Alisa saw the dog bowl. The dog can bowl.

👩🏼‍🎓	Entail Alisa saw the dog going bowling.	Neutral Alisa saw the dog's bowl.	Contradict
👨🏾‍🎓	Entail	Neutral	Contradict
🧏🏻‍♀️	Entail Alisa saw the dog going bowling.	Neutral Alisa saw the bowl for dogs.	Contradict

AmbiEnt Statistics



Label Distribution



Amibiguity Type Distribution

LM Evaluation

LM Evaluation

1. Can LMs **generate** disambiguations?

LM Evaluation

1. Can LMs **generate** disambiguations?
2. Can LMs **recognize** potential disambiguations?

LM Evaluation

1. Can LMs **generate** disambiguations?
2. Can LMs **recognize** potential disambiguations?
3. Can LMs **model continuations** reflecting different interpretations?

LM Evaluation

1. Can LMs **generate** disambiguations?
2. Can LMs **recognize** potential disambiguations?
3. Can LMs **model continuations** reflecting different interpretations?

- Evaluated models
  - Base LMs: LLaMa-65B, GPT-3 (davinci)
  - Instruction-following/aligned LMs: Flan-T5-xxl, InstructGPT, GPT-3.5, GPT-4

1. Generating disambiguations

1. Generating disambiguations

Context: Alisa saw the dog bowl.
Claim: The dog can bowl.
Given the context alone, is this claim true, false, or inconclusive?

We don’t know, because the context can be interpreted in many different ways:
1.

1. Generating disambiguations

Context: Alisa saw the dog bowl.
Claim: The dog can bowl.
Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:
1. Alisa saw the bowl for dogs. Then the claim is inconclusive.
2. Alisa saw the dog going bowling. Then the claim is true.

1. Generating disambiguations

Demonstrations × 4

Context: He is on track to complete his PhD in three years.
Claim: He will spend a total of three years completing his PhD.
Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:
1. He is on track to complete his PhD in three more years from now. Then the claim is inconclusive.
2. He is on track to complete his PhD in three years from when he started. Then the claim is true.

Context: Alisa saw the dog bowl.
Claim: The dog can bowl.
Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:
1. Alisa saw the bowl for dogs. Then the claim is inconclusive.
2. Alisa saw the dog going bowling. Then the claim is true.

1. Generating disambiguations

Instruction

In each example, you will be given some context and a claim, where the correctness of the claim is affected by some ambiguity in the context. Enumerate two or three interpretations of the context that lead to different judgments about the claim.

Demonstrations × 4

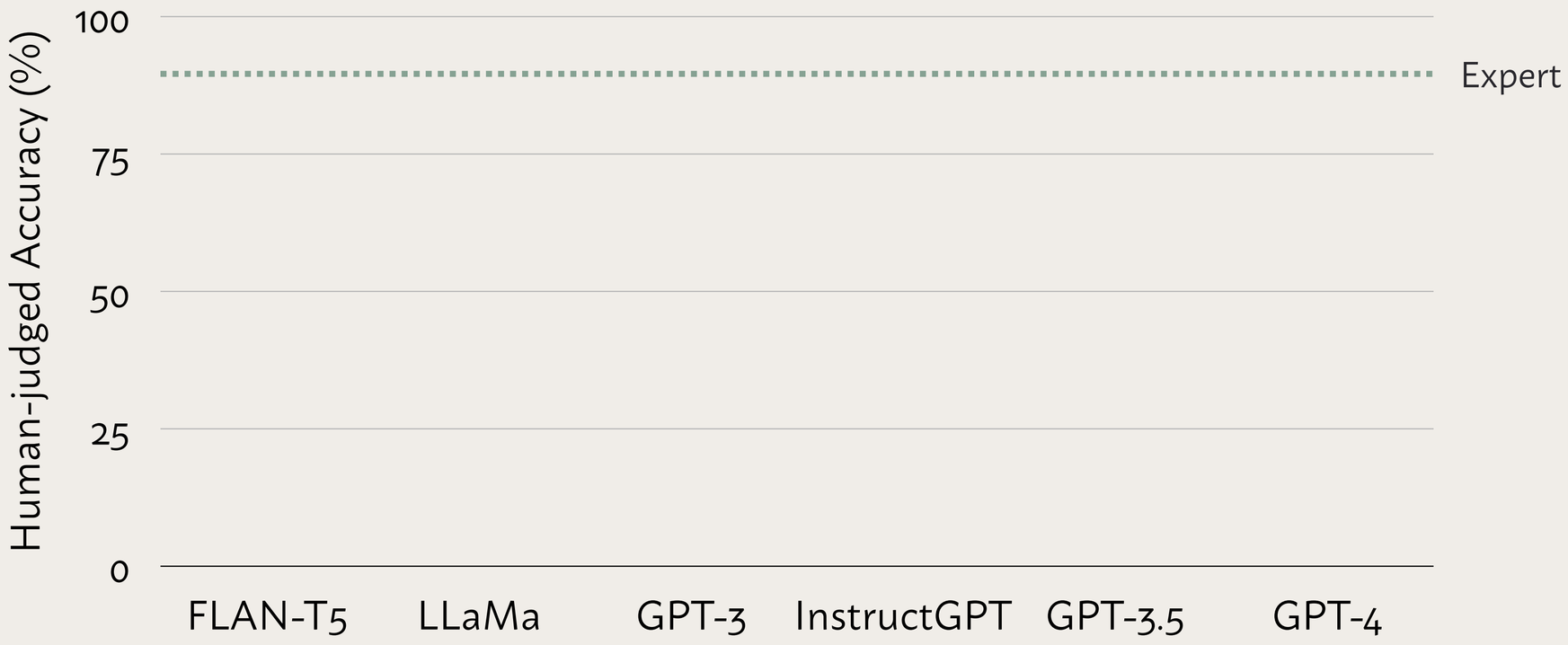
Context: He is on track to complete his PhD in three years.
Claim: He will spend a total of three years completing his PhD.
Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:
1. He is on track to complete his PhD in three more years from now. Then the claim is inconclusive.
2. He is on track to complete his PhD in three years from when he started. Then the claim is true.

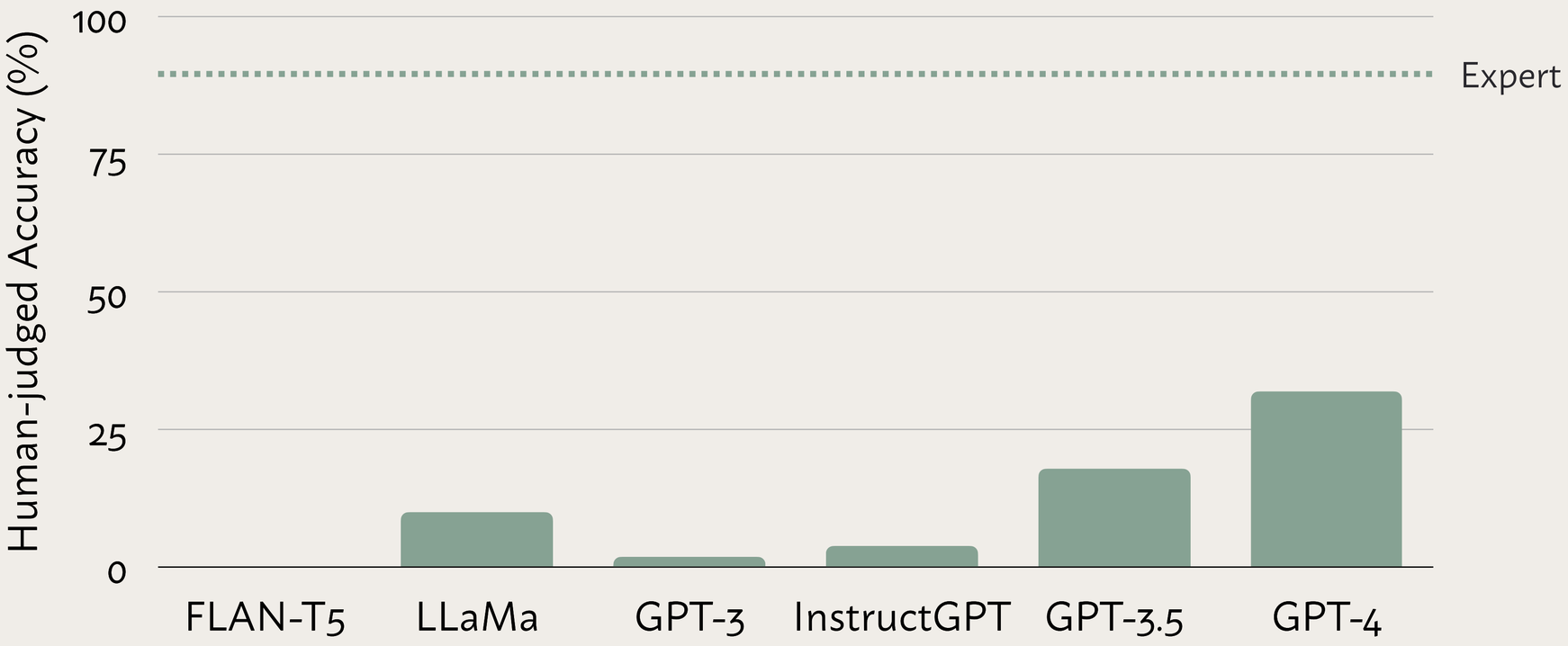
Context: Alisa saw the dog bowl.
Claim: The dog can bowl.
Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:
1. Alisa saw the bowl for dogs. Then the claim is inconclusive.
2. Alisa saw the dog going bowling. Then the claim is true.

1. Generating disambiguations



1. Generating disambiguations



2. Recognizing disambiguations

2. Recognizing disambiguations

Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False?
A: {**True**, False}

2. Recognizing disambiguations

Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False?
A: {**True**, False}

Q: Alisa saw the dog bowl. **This does not necessarily mean**: The dog can bowl. True or False?
A: {**True**, False}

2. Recognizing disambiguations

Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False?
A: {**True**, False}

Q: Alisa saw the dog bowl. **This does not necessarily mean**: The dog can bowl. True or False?
A: {**True**, False}

Q: Alisa saw the dog bowl. **This cannot mean**: The dog can bowl. True or False?
A: {True, **False**}

2. Recognizing disambiguations

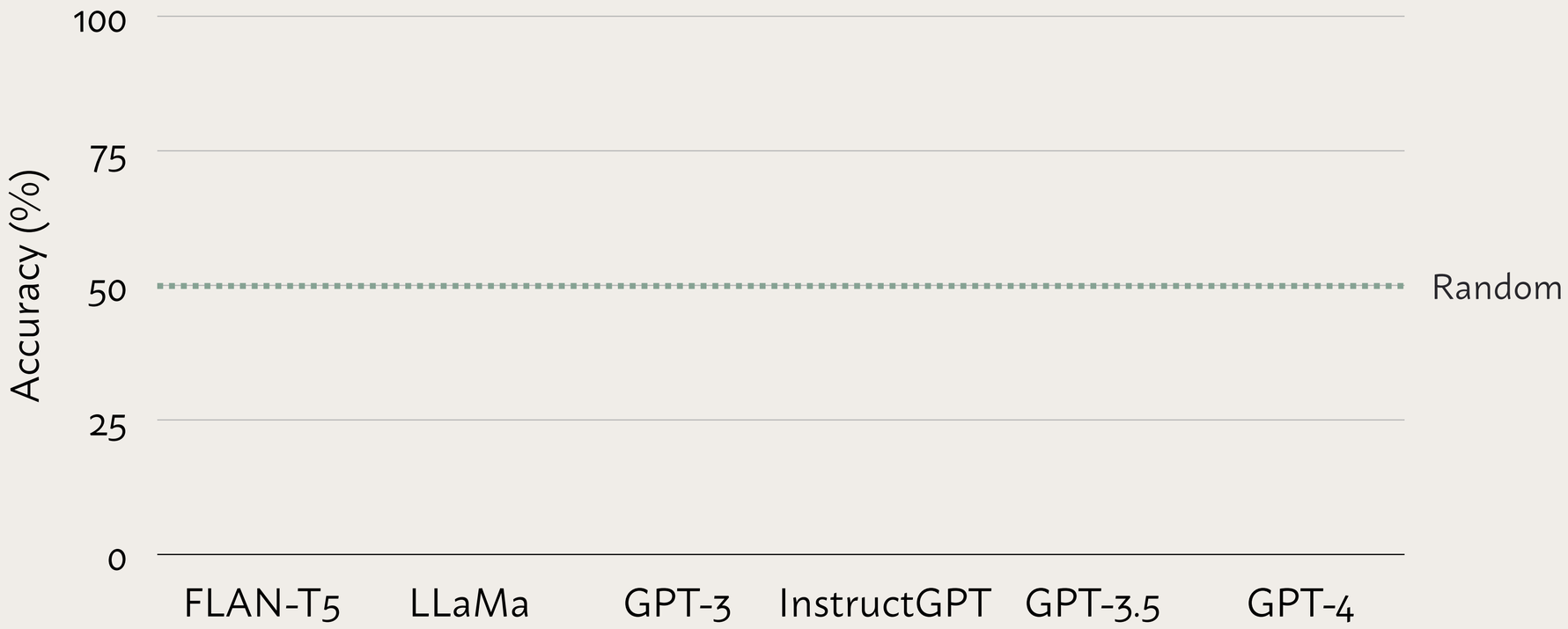
Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False?
A: {**True**, False}

Q: Alisa saw the dog bowl. **This does not necessarily mean**: The dog can bowl. True or False?
A: {**True**, False}

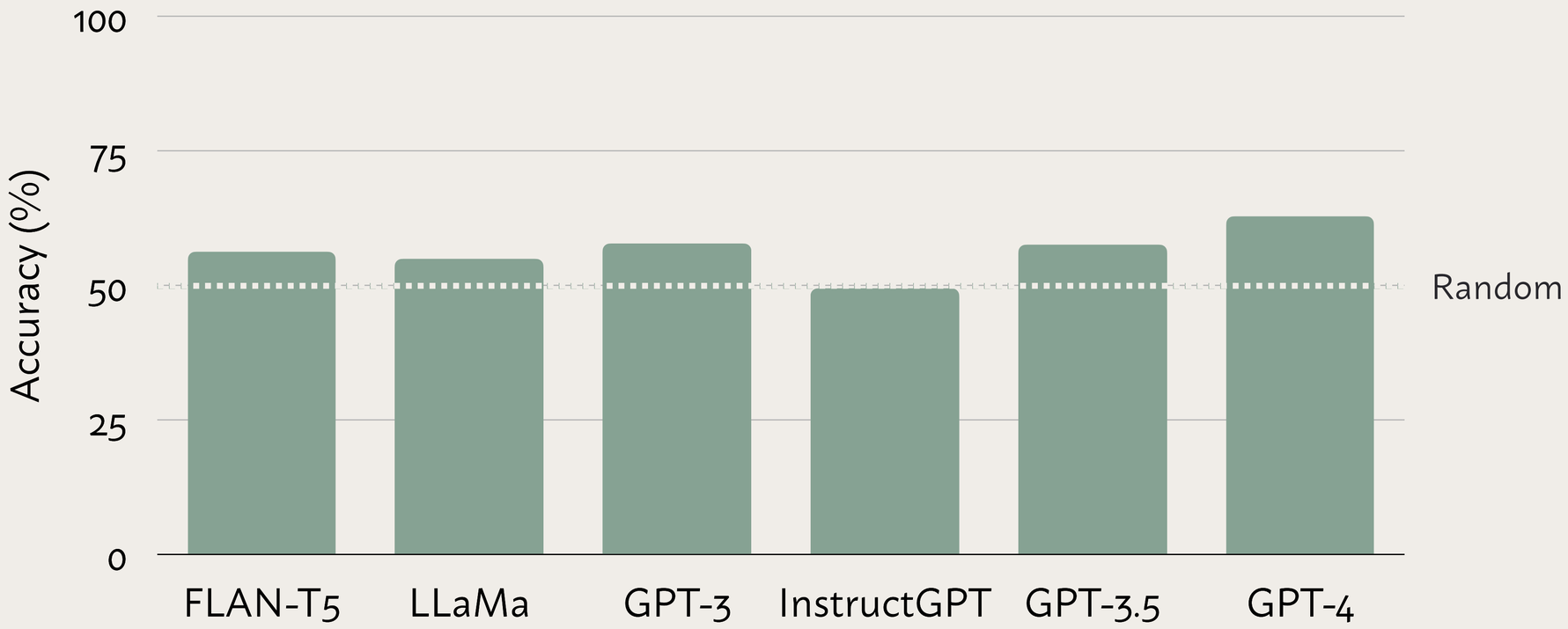
Q: Alisa saw the dog bowl. **This cannot mean**: The dog can bowl. True or False?
A: {True, **False**}

Q: Alisa saw the dog bowl. **This can only mean**: The dog can bowl. True or False?
A: {True, **False**}

2. Recognizing disambiguations



2. Recognizing disambiguations



3. Modeling continuations

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling.

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling. It beat all animals in the bowling alley.

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling. It beat all animals in the bowling alley.

Alisa saw the dog bowl. 

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- $\mathrm{KL}(P(\cdot \mid d) \parallel P(\cdot \mid a))$ should be "reasonably" small

d Alisa saw the dog going bowling. It beat all animals in the bowling alley.

a Alisa saw the dog bowl.

3. Modeling continuations

- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- $\mathrm{KL}(P(\cdot \mid d) \parallel P(\cdot \mid a))$ should be "reasonably" small
  - $\mathrm{KL}(P(\cdot \mid d) \parallel P(\cdot \mid a)) < \mathrm{KL}(P(\cdot \mid \tilde{d}) \parallel P(\cdot \mid a))$ for some distractor $\tilde{d}$

d Alisa saw the dog going bowling. It beat all animals in the bowling alley.

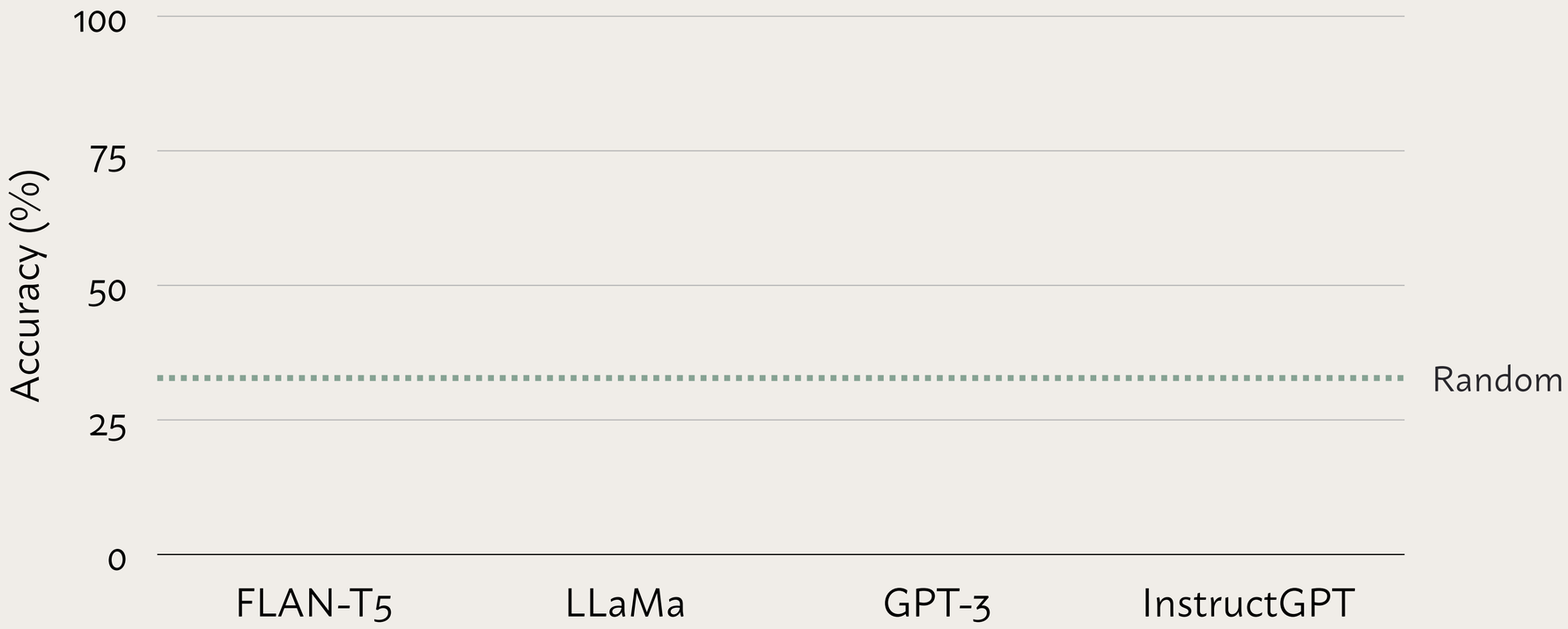
a Alisa saw the dog bowl.

3. Modeling continuations

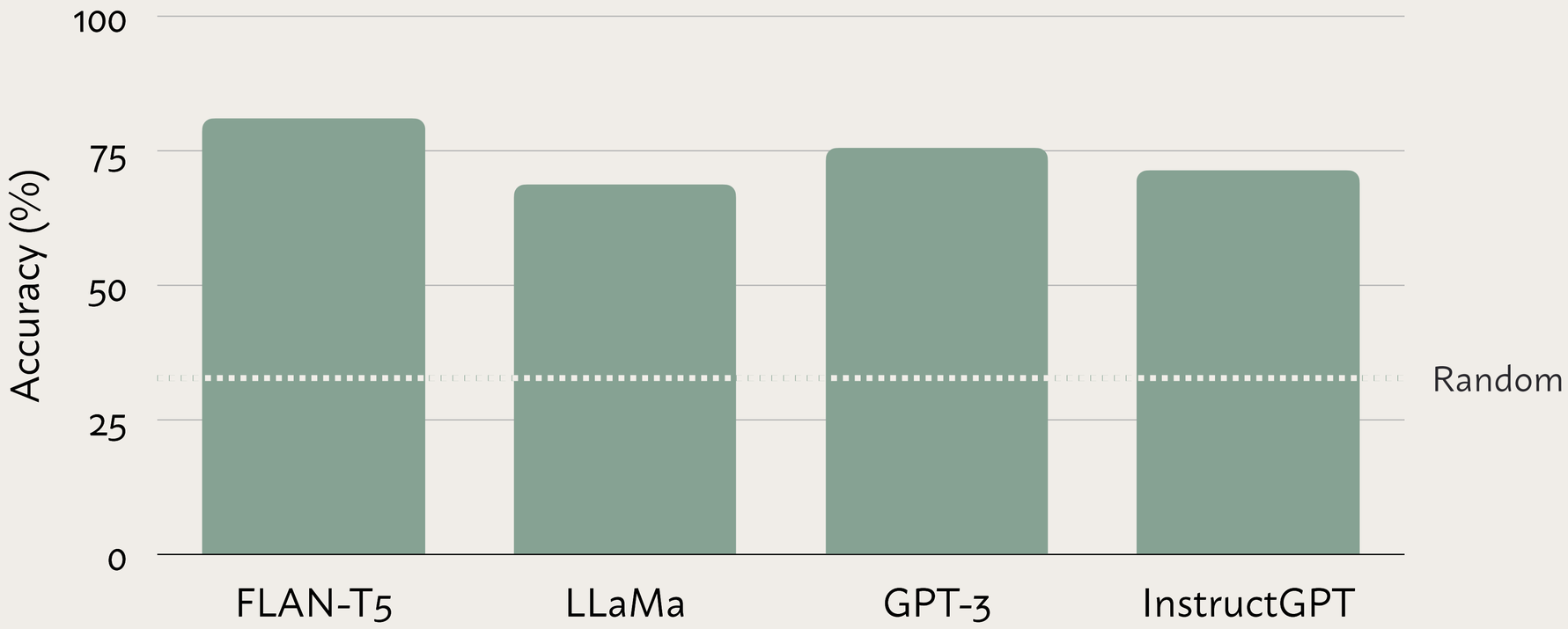
- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- $\mathrm{KL}(P(\cdot \mid d) \parallel P(\cdot \mid a))$ should be "reasonably" small
  - $\mathrm{KL}(P(\cdot \mid d) \parallel P(\cdot \mid a)) < \mathrm{KL}(P(\cdot \mid \tilde{d}) \parallel P(\cdot \mid a))$ for some distractor $\tilde{d}$

d	Alisa saw the dog going bowling. It beat all animals in the bowling alley.
a	Alisa saw the dog bowl.
$\tilde{d}$	Alisa saw the dog going cycling.

3. Modeling disambiguating continuations



3. Modeling disambiguating continuations



Summary

Summary

- Despite the importance of ambiguity in language, it is not represented well in state-of-the-art LMs

Summary

- Despite the importance of ambiguity in language, it is not represented well in state-of-the-art LMs
- More interesting findings in paper

Summary

- Despite the importance of ambiguity in language, it is not represented well in state-of-the-art LMs
- More interesting findings in paper
  - It is challenging to *finetune* LMs to recognize ambiguity

Summary

- Despite the importance of ambiguity in language, it is not represented well in state-of-the-art LMs
- More interesting findings in paper
  - It is challenging to *finetune* LMs to recognize ambiguity
  - Ambiguity partly explains annotator disagreement

Summary

- Despite the importance of ambiguity in language, it is not represented well in state-of-the-art LMs
- More interesting findings in paper
  - It is challenging to *finetune* LMs to recognize ambiguity
  - Ambiguity partly explains annotator disagreement
  - Ambiguity-aware models bear practical utility, e.g., to detect misleading political claims

ZHAOFENG WU
LINLU QIU
ALEXIS ROSS
EKIN AKYÜREK
BOYUAN CHEN
BAILIN WANG
NAJOUNG KIM
JACOB ANDREAS
YOON KIM

Reasoning or Reciting?

Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

$$\sum_{n=1}^{\infty} \frac{n^2}{2^n}$$

$$\sum_{n=1}^{\infty} \frac{n^2}{2^n}$$

$$\int_{-\infty}^{\infty} e^{-x^2} dx$$

$$\sum_{n=1}^{\infty} \frac{n^2}{2^n}$$

L What is the sum of the following series? $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$

The sum of the series $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$ is 6.

$$\int_{-\infty}^{\infty} e^{-x^2} dx$$

L What is the result of the following integral? $\int_{-\infty}^{\infty} e^{-x^2} dx$

The integral you're asking about is the Gaussian integral. Its value is:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

So the result is $\sqrt{\pi}$.

$$\sum_{n=1}^{\infty} \frac{n^2}{2^n}$$

L What is the sum of the following series? $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$

The sum of the series $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$ is 6.

$$\int_{-\infty}^{\infty} e^{-x^2} dx$$

L What is the result of the following integral? $\int_{-\infty}^{\infty} e^{-x^2} dx$

The integral you're asking about is the Gaussian integral. Its value is:

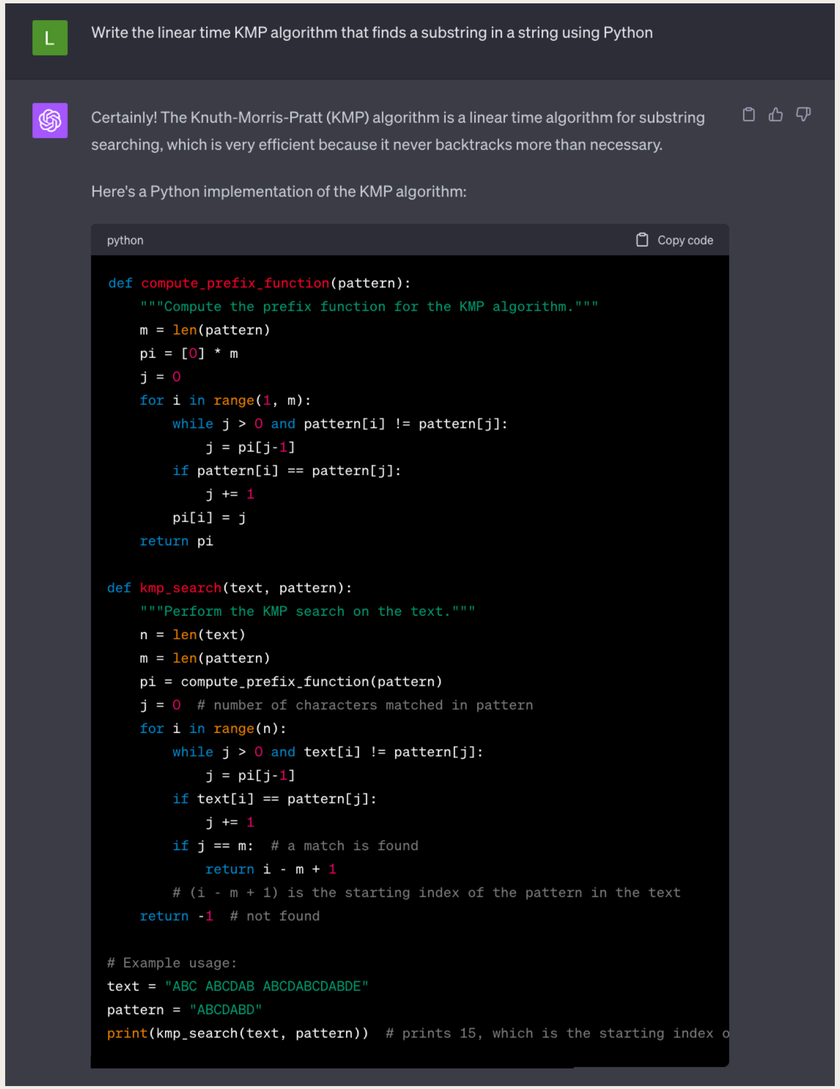
$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

So the result is $\sqrt{\pi}$.

GPT-4 is so good at math!

Linear-time algorithm for substring finding?

Linear-time algorithm for substring finding?



L

Write the linear time KMP algorithm that finds a substring in a string using Python

Certainly! The Knuth-Morris-Pratt (KMP) algorithm is a linear time algorithm for substring searching, which is very efficient because it never backtracks more than necessary.

Here's a Python implementation of the KMP algorithm:

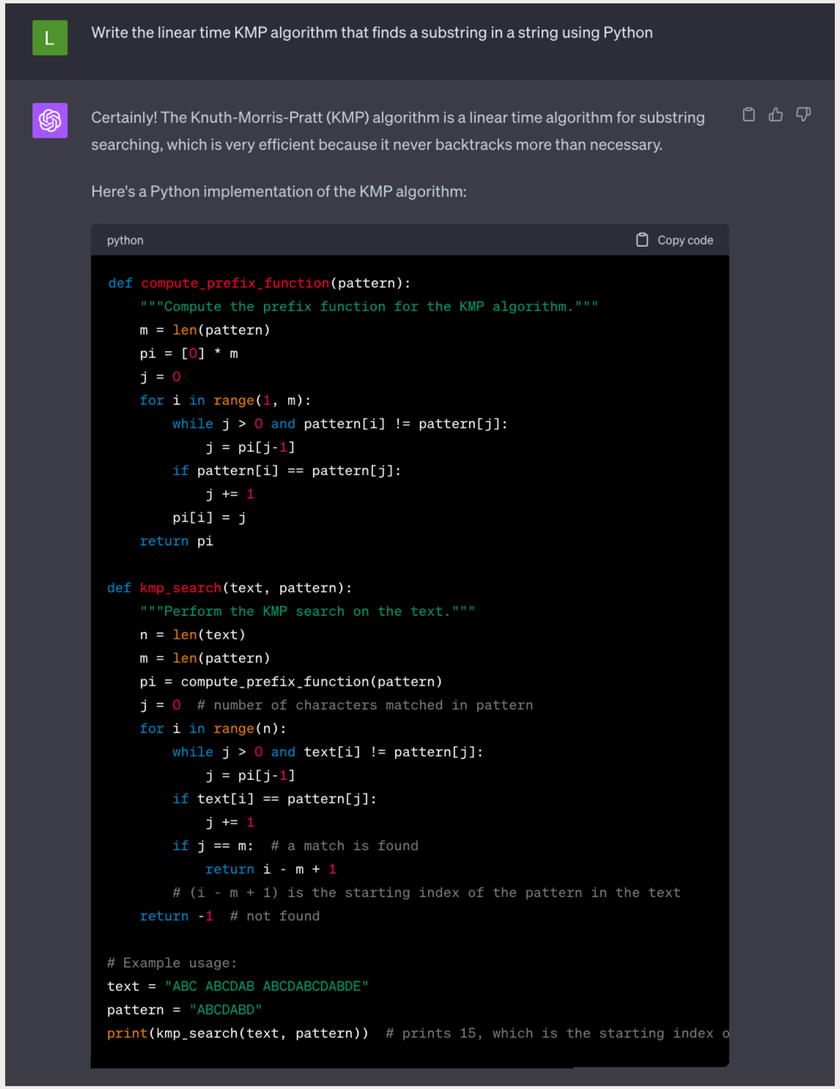
python Copy code

```python
def compute_prefix_function(pattern):
    """Compute the prefix function for the KMP algorithm."""
    m = len(pattern)
    pi = [0] * m
    j = 0
    for i in range(1, m):
        while j > 0 and pattern[i] != pattern[j]:
            j = pi[j-1]
        if pattern[i] == pattern[j]:
            j += 1
        pi[i] = j
    return pi

def kmp_search(text, pattern):
    """Perform the KMP search on the text."""
    n = len(text)
    m = len(pattern)
    pi = compute_prefix_function(pattern)
    j = 0  # number of characters matched in pattern
    for i in range(n):
        while j > 0 and text[i] != pattern[j]:
            j = pi[j-1]
        if text[i] == pattern[j]:
            j += 1
        if j == m:  # a match is found
            return i - m + 1
        # (i - m + 1) is the starting index of the pattern in the text
    return -1  # not found

# Example usage:
text = "ABC ABCDAB ABCDABCDABDE"
pattern = "ABCDABD"
print(kmp_search(text, pattern))  # prints 15, which is the starting index o
```

Linear-time algorithm for substring finding?



L

Write the linear time KMP algorithm that finds a substring in a string using Python

Certainly! The Knuth-Morris-Pratt (KMP) algorithm is a linear time algorithm for substring searching, which is very efficient because it never backtracks more than necessary.

Here's a Python implementation of the KMP algorithm:

python Copy code

```python
def compute_prefix_function(pattern):
    """Compute the prefix function for the KMP algorithm."""
    m = len(pattern)
    pi = [0] * m
    j = 0
    for i in range(1, m):
        while j > 0 and pattern[i] != pattern[j]:
            j = pi[j-1]
        if pattern[i] == pattern[j]:
            j += 1
        pi[i] = j
    return pi

def kmp_search(text, pattern):
    """Perform the KMP search on the text."""
    n = len(text)
    m = len(pattern)
    pi = compute_prefix_function(pattern)
    j = 0  # number of characters matched in pattern
    for i in range(n):
        while j > 0 and text[i] != pattern[j]:
            j = pi[j-1]
        if text[i] == pattern[j]:
            j += 1
        if j == m:  # a match is found
            return i - m + 1
        # (i - m + 1) is the starting index of the pattern in the text
    return -1  # not found

# Example usage:
text = "ABC ABCDAB ABCDABCDABDE"
pattern = "ABCDABD"
print(kmp_search(text, pattern))  # prints 15, which is the starting index o
```

GPT-4 can code so well!

This is over-generalizing!

This is over-generalizing!

- In order to claim general abilities…

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets

This is over-generalizing!

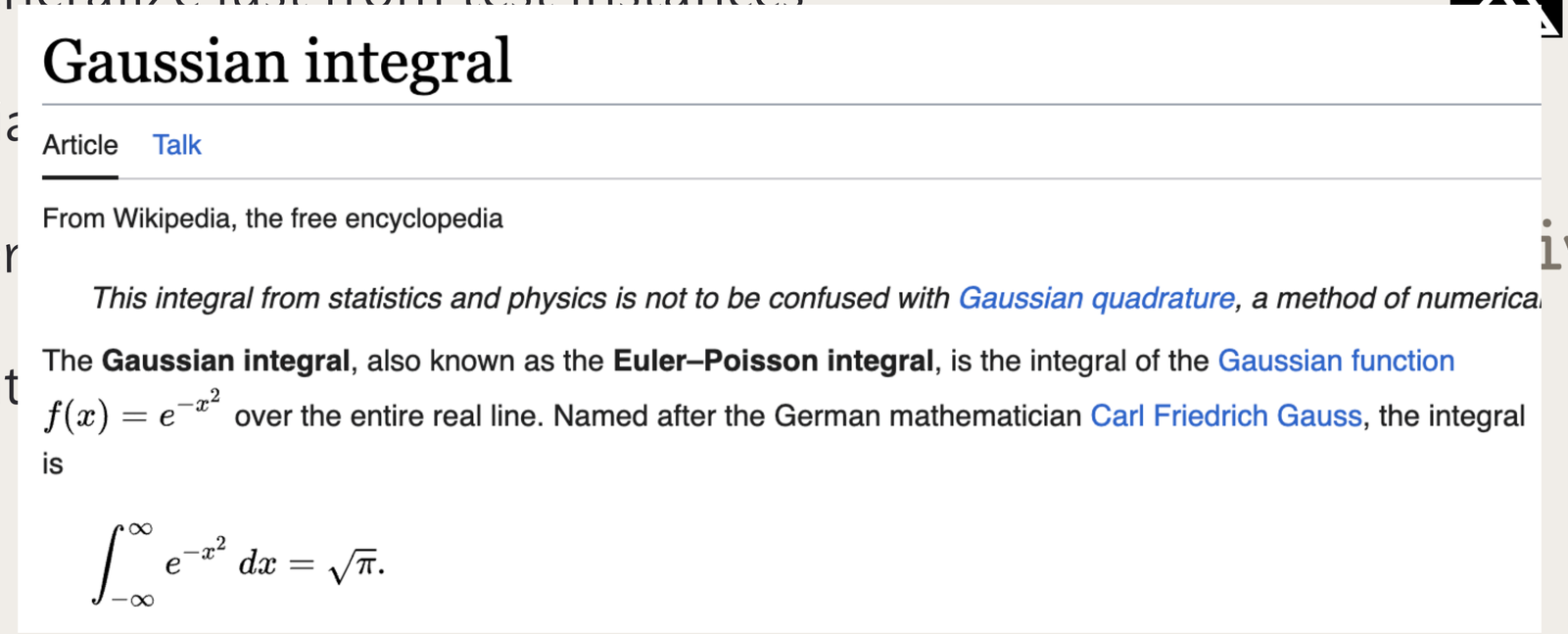
- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination

This is over-generalizing!

- In order to claim general abilities…
- We can't generalize just from test instances
  - High varia
- We can't ger
  - Data cont



Gaussian integral

Article Talk

From Wikipedia, the free encyclopedia

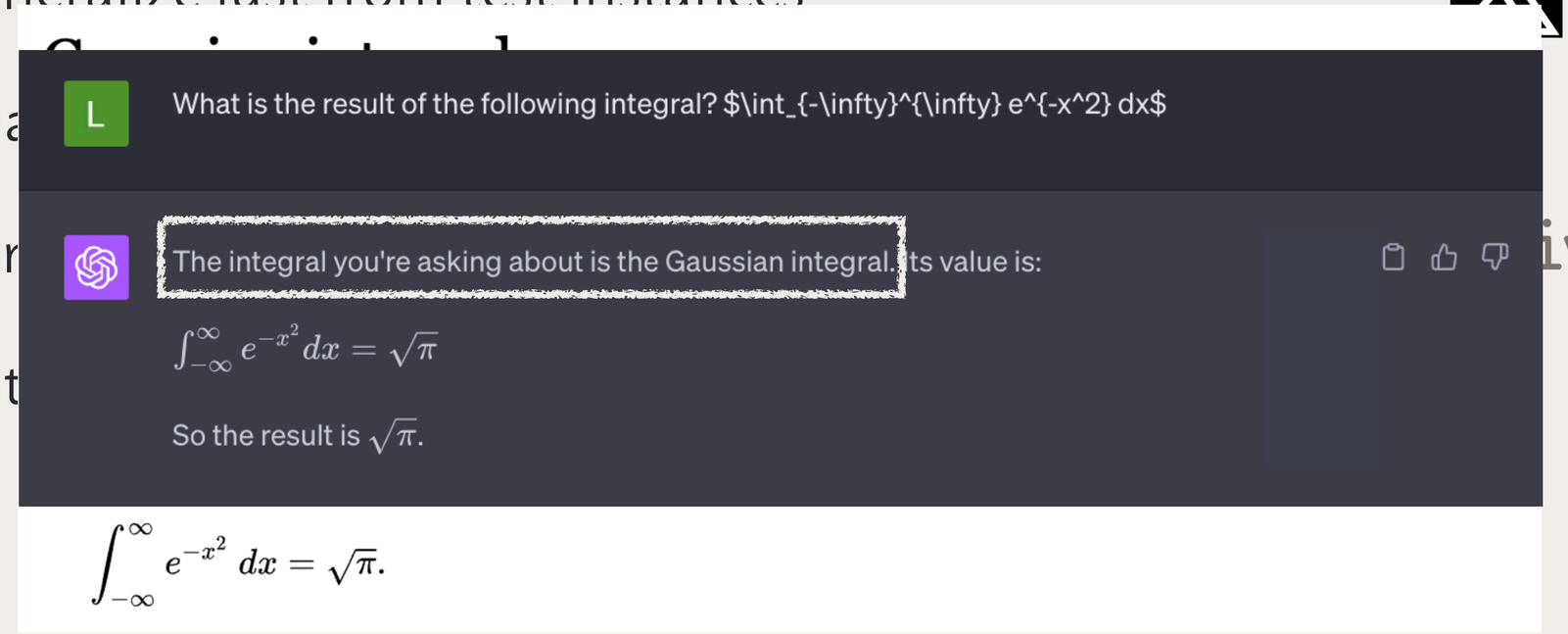
This integral from statistics and physics is not to be confused with Gaussian quadrature, a method of numerica

The **Gaussian integral**, also known as the **Euler–Poisson integral**, is the integral of the Gaussian function $f(x) = e^{-x^2}$ over the entire real line. Named after the German mathematician Carl Friedrich Gauss, the integral is

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High varia
- We can’t ger
  - Data cont



This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination
- We can’t generalize just from task instances

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination
- We can’t generalize just from task instances
  - Task instance ability ≠ general task ability

This is over-generalizing!

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination
- We can’t generalize just from task instances
  - Task instance ability ≠ general task ability
  - If a cook only makes perfect mapo tofu, but nothing else, are they still a good cook?

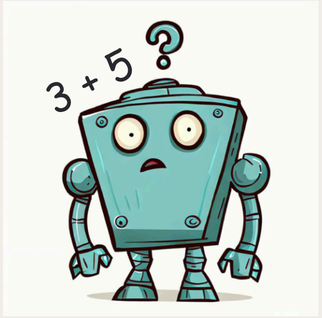
This is over-generalizing!

Counterfactual Evaluation

- In order to claim general abilities…
- We can’t generalize just from test instances
  - High variance; may be cherry-picked
- We can’t generalize just from existing datasets
  - Data contamination
- We can’t generalize just from task instances
  - Task instance ability ≠ general task ability
  - If a cook only makes perfect mapo tofu, but nothing else, are they still a good cook?

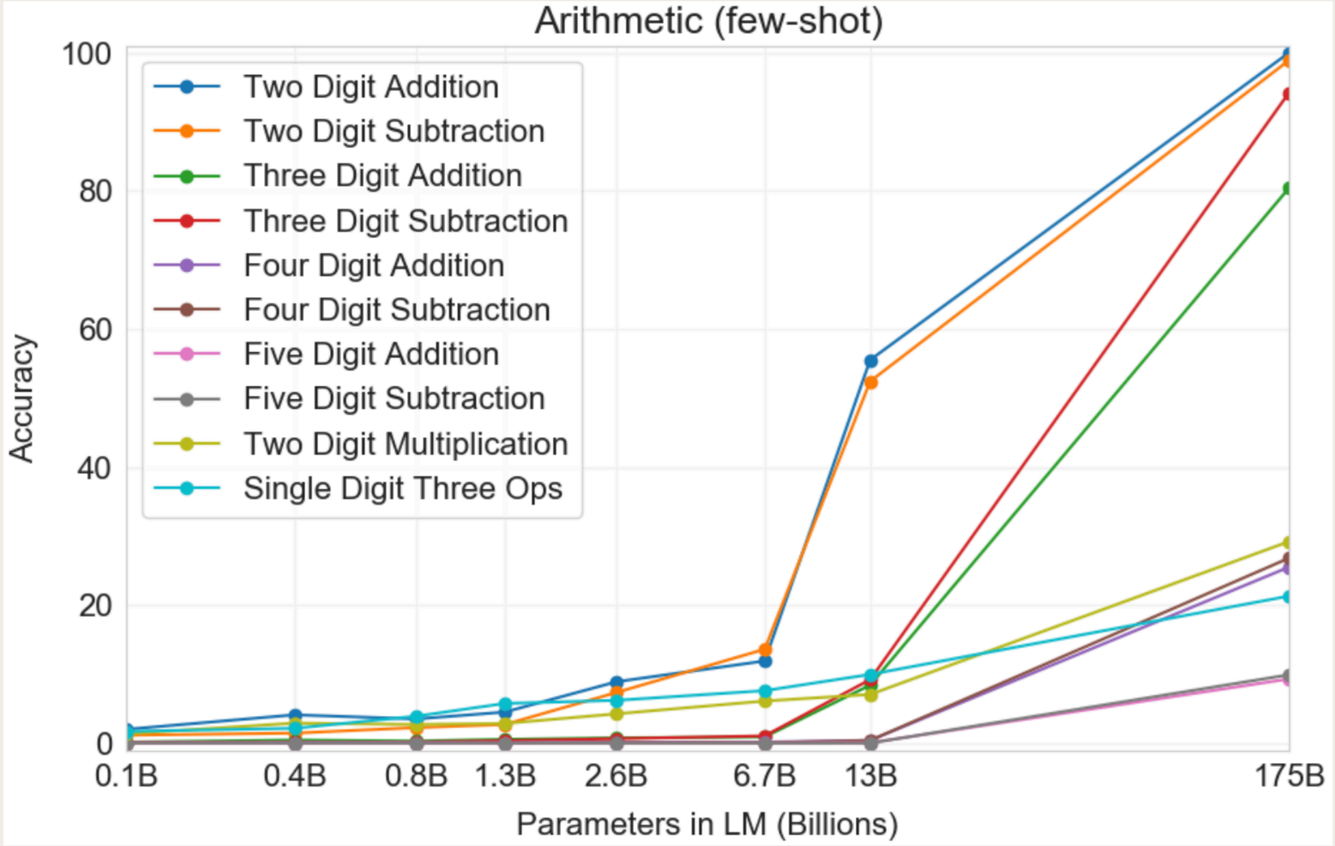
Can LMs really do math?

Can LMs really do math?



Arithmetic

Arithmetic



Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

Arithmetic

Arithmetic

27+62=

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **27+62**? Let's think step by step.

27+62=

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **27+62**? Let's think step by step.

27+62= 89

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **27+62**? Let's think step by step.

	Base-10
27+62=	89

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **27+62**? Let's think step by step.

	Base-10	Base-9
27+62=	89	

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

	Base-10	Base-9
27+62=	89	

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

	Base-10	Base-9
27+62=	89	100

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

Base-10

89

counterfactual

Base-9

100

27+62=

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

	Base-10	**Base-9**
27+62=	89	100

$\text{ADD}_{base}(a, b)$

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

Base-10

Base-9

27+62=

89

100

$\text{ADD}_{base}(a, b)$

Starting from the trailing digits of a and b, add them together.
If greater than $base$, carry over.
Move on to the penultimate digit.
...

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

	Base-10	**Base-9**
27+62=	89	100

$\text{ADD}_{base}(a, b)$

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

Base-10

Base-9

27+62=

89

100

$$\text{ADD}_{base}(a, b)$$

$$\left\{ \begin{array}{c} \text{ADD}_{base-10}(a, b) \\ \text{ADD}_{base-9}(a, b) \\ \cdots \end{array} \right\}$$

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

Base-10

Base-9

27+62=

89

100

$\text{ADD}_{base}(a, b)$

💯

$$\left\{\begin{array}{c}\text{ADD}_{base-10}(a, b) \\ \text{ADD}_{base-9}(a, b) \\ \cdots\end{array}\right\}$$

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

	Base-10	**Base-9**
27+62=	89	100
$\text{ADD}_{base}(a, b)$	💯	💯
$\left\{ \begin{array}{c} \text{ADD}_{base-10}(a, b) \\ \text{ADD}_{base-9}(a, b) \\ \cdots \end{array} \right\}$		

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

	Base-10	**Base-9**
27+62=	89	100
$\text{ADD}_{base}(a, b)$	💯	💯
$\left\{ \begin{array}{c} \text{ADD}_{base-10}(a, b) \\ \text{ADD}_{base-9}(a, b) \\ \cdots \end{array} \right\}$	💯	

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

counterfactual

	Base-10	**Base-9**
27+62=	89	100
$\text{ADD}_{base}(a, b)$	💯	💯
$\left\{\begin{array}{c}\text{ADD}_{base-10}(a, b) \\ \text{ADD}_{base-9}(a, b) \\ \cdots\end{array}\right\}$	💯	🌀

General task abilities

$f_w(x)$

General task abilities $f_w(x)$

or

Non-transferrable procedures $f_{w^{default}}(x)$

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

Base-10

89

counterfactual

Base-9

100

27+62=

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

Base-10

89

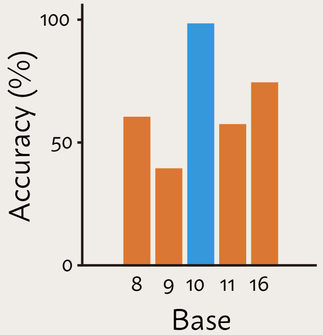
counterfactual

Base-9

100

27+62=

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

Base-10

89

counterfactual

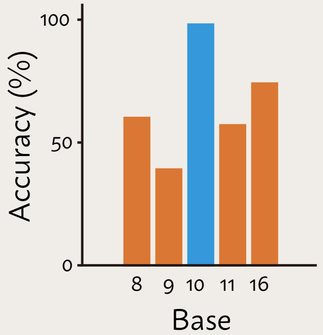
Base-9

100

27+62=

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **27+62**? Let's think step by step.

default

Base-10

89

counterfactual

Base-9

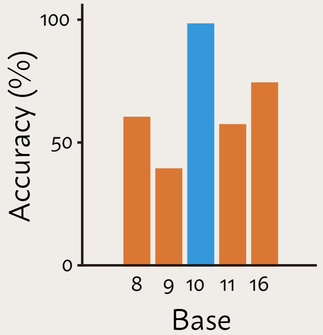
100

27+62=

What follows 78?

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **the next number after 78**?

default

Base-10

89

counterfactual

Base-9

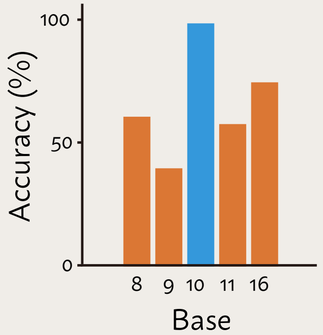
100

27+62=

What follows 78?

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**10** where the digits are "**0123456789**", what is **the next number after 78**?

default

Base-10

counterfactual

Base-9

27+62=

89

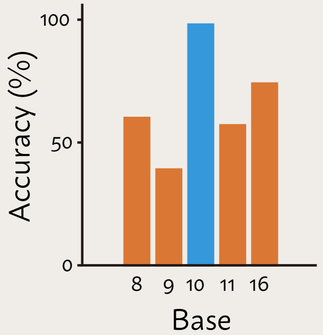
100

What follows 78?

79

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **the next number after 78**?

default

Base-10

counterfactual

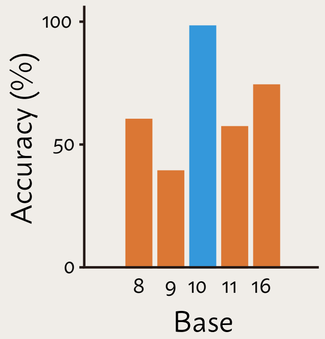
Base-9

27+62= 89 100

What follows 78? 79

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **the next number after 78**?

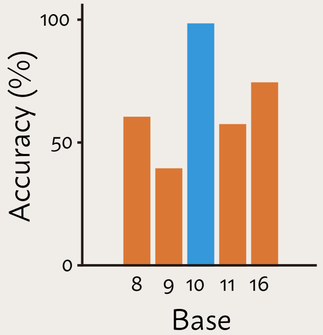
default

counterfactual

	Base-10	**Base-9**
27+62=	89	100
What follows 78?	79	80

Counterfactual comprehension check (CCC)

GPT-4



Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **the next number after 78**?

default

Base-10

27+62= 89

What follows 78? 79

counterfactual

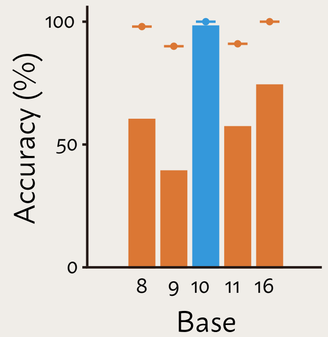
Base-9

100

80

Counterfactual comprehension check (CCC)

GPT-4



Task CCC

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **the next number after 78**?

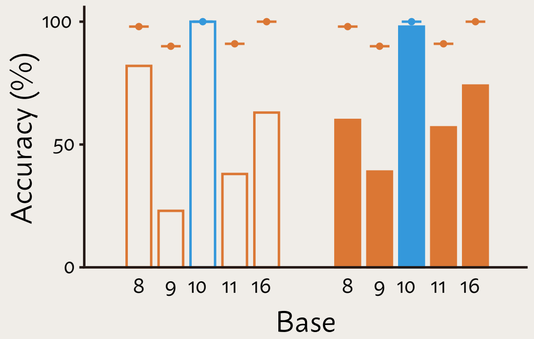
default

counterfactual

	Base-10	**Base-9**
27+62=	89	100
What follows 78?	79	80

Counterfactual comprehension check (CCC)

GPT-4



Task **Task (Let's think step by step.)** **CCC**

Arithmetic

You are a mathematician. Assuming that all numbers are in base-**9** where the digits are "**012345678**", what is **the next number after 78**?

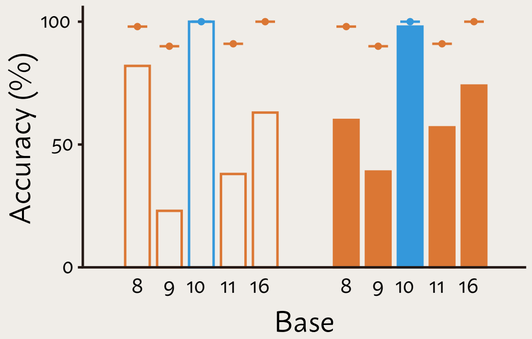
default

counterfactual

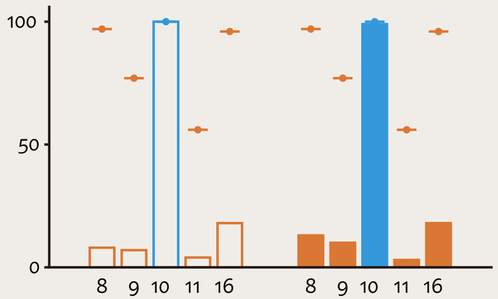
	Base-10	**Base-9**
27+62=	89	100
What follows 78?	79	80

Counterfactual comprehension check (CCC)

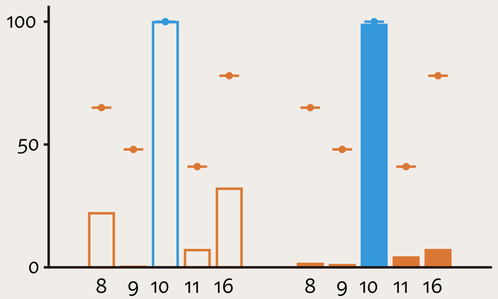
GPT-4



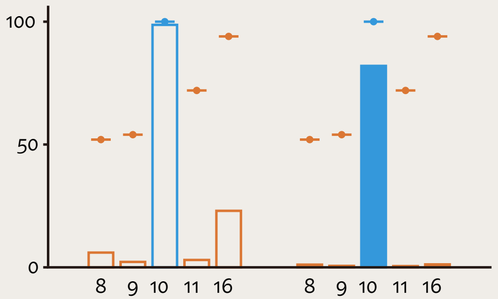
GPT-3.5



Claude



PaLM-2



Task **Task (Let's think step by step.)** **CCC**

Some Notes 🎶

Some Notes 🎶

- I can’t do base-9 arithmetic as well either

Some Notes 🎶

- I can’t do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?

Some Notes 🎶

- I can’t do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?

Some Notes 🎶

- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?

Some Notes 🎶

- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
  - Not necessarily, but it represents a lack of generalizability

Some Notes 🎶

- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
  - Not necessarily, but it represents a lack of generalizability
  - We don't actually know if the LM implements perfect base-10 arithmetic!

Some Notes 🎶

- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
  - Not necessarily, but it represents a lack of generalizability
  - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?

Some Notes 🎶

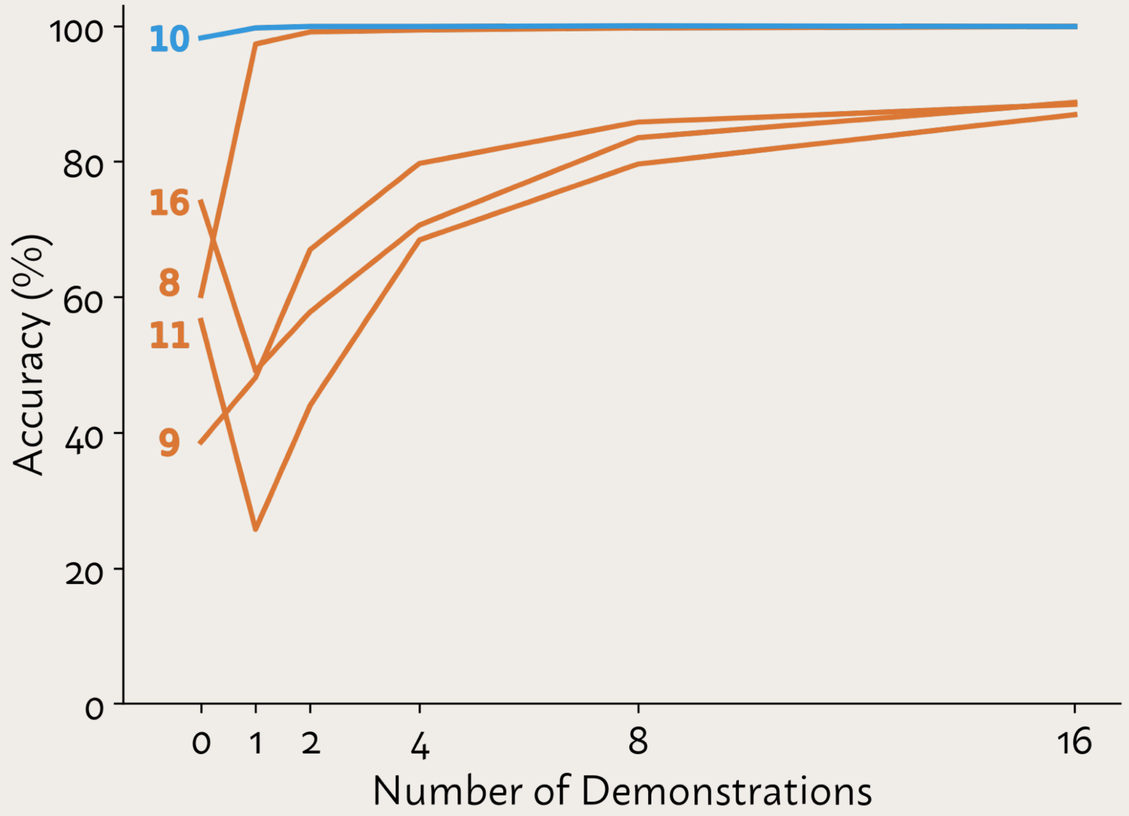
- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
  - Not necessarily, but it represents a lack of generalizability
  - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?
  - It is certainly useful sometimes

Some Notes 🎶

- I can't do base-9 arithmetic as well either
  - Under a fixed time budget, probably, but what if given unlimited time?
  - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
  - Not necessarily, but it represents a lack of generalizability
  - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?
  - It is certainly useful sometimes
  - More importantly, it is a test setting not confounded by memorization

Would better prompts help?

Would better prompts help?



Coding: Execution

Coding: Execution

"By providing such hints to ChatGPT, its success rate can be further increased, fixing 31 out of 40 bugs, **outperforming state-of-the-art**."

Sobania et al. An Analysis of the Automatic Bug Fixing Performance of ChatGPT. ICSE 2023.

Coding: Execution

"By providing such hints to ChatGPT, its success rate can be further increased, fixing 31 out of 40 bugs, **outperforming state-of-the-art**."

"We also compare the PLMs against recent state-of-the-art APR tools. Our study demonstrates that directly applying state-of-the-art PLMs can already **substantially outperform all existing APR techniques** on all our datasets."

Sobania et al. An Analysis of the Automatic Bug Fixing Performance of ChatGPT. ICSE 2023.
Xia et al. Practical Program Repair in the Era of Large Pre-trained Language Models. 2022.

Coding: Execution

Coding: Execution

```
[“a”, “b”, “c”][1]
```

Coding: Execution

```
["a", "b", "c"][1]
```

"b"

Coding: Execution

Python
0-based indexing

```
["a", "b", "c"][1]
```

"b"

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
["a", "b", "c"][1]
```

"b"

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
["a", "b", "c"][1]
```

"b"

"a"

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

`["a", "b", "c"][1]`

"b"

"a"

Counterfactual comprehension check (CCC)

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
sorted(
  [“ab”, “ba”],
  key=lambda x: x[1],
)
```

```
   [“a”, “b”, “c”][1]
```

“b”

“a”

Counterfactual comprehension check (CCC)

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
sorted(
  [“ab”, “ba”],
  key=lambda x: x[1],
)
```

[“ba”, “ab”]

```
   [“a”, “b”, “c”][1]
```

“b”

“a”

Counterfactual comprehension check (CCC)

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
sorted(
  [“ab”, “ba”],
  key=lambda x: x[1],
)
```

[“ba”, “ab”]

[“ab”, “ba”]

```
   [“a”, “b”, “c”][1]
```

“b”

“a”

Counterfactual comprehension check (CCC)

Coding: Execution

Python
0-based indexing

ThonPy
1-based indexing

```
sorted(
  [“ab”, “ba”],
  key=lambda x: x[1],
)
```

[“ba”, “ab”]

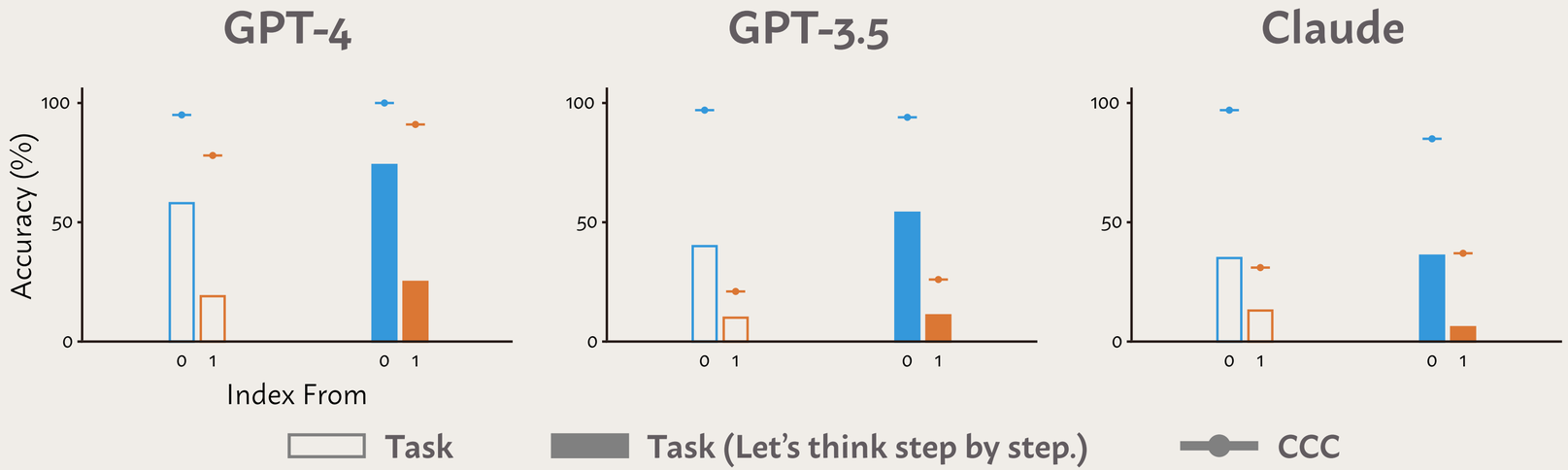
[“ab”, “ba”]

```
   [“a”, “b”, “c”][1]
```

“b”

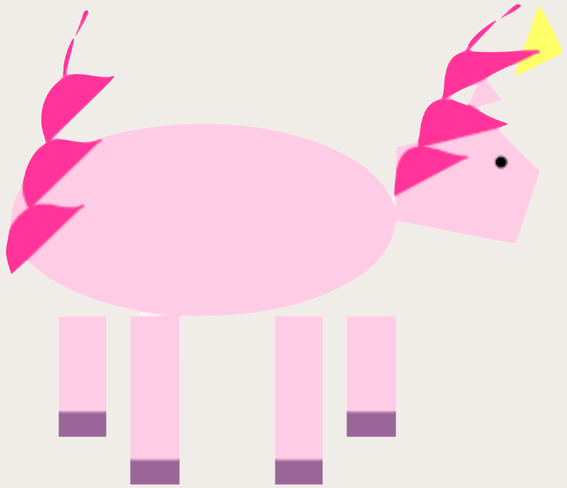
“a”

Counterfactual comprehension check (CCC)



Drawing

Drawing



Bubeck et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.

Drawing

Drawing

Draw a bubble tea

Drawing

Draw a bubble tea

Drawing

Upright

Draw a bubble tea

Drawing

Upright

Upside-down

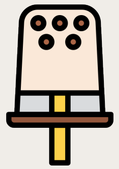
Draw a bubble tea

Drawing

Upright

Upside-down

Draw a bubble tea

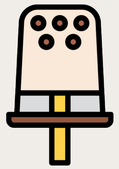


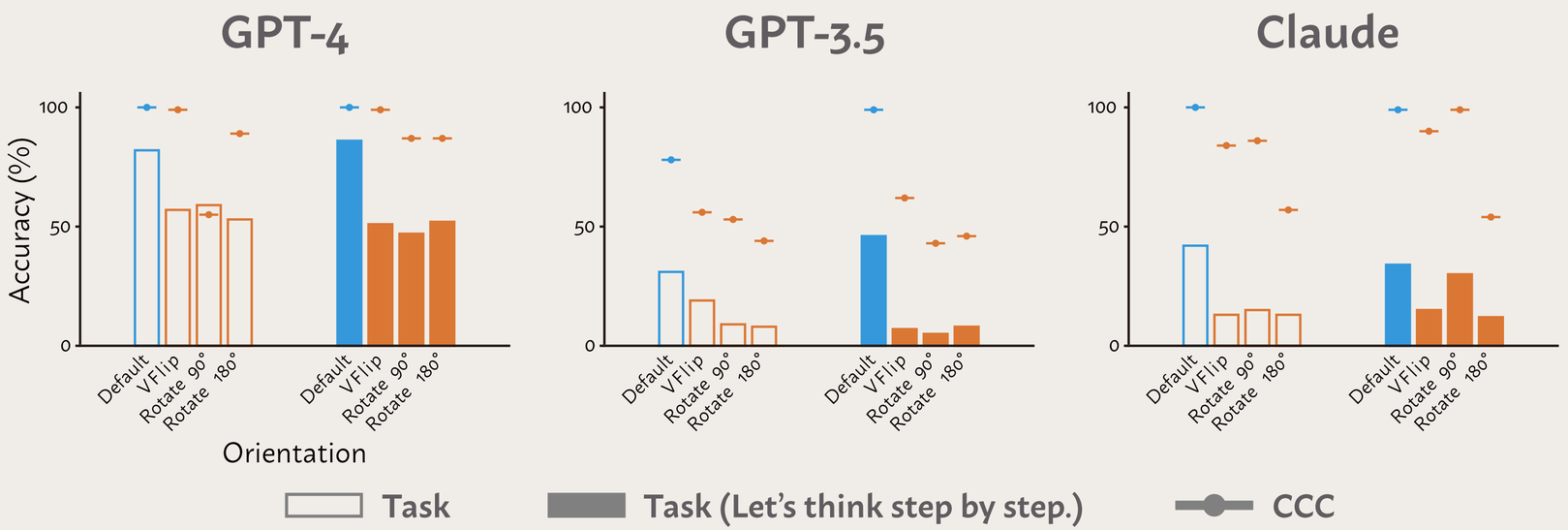
Drawing

Upright

Upside-down

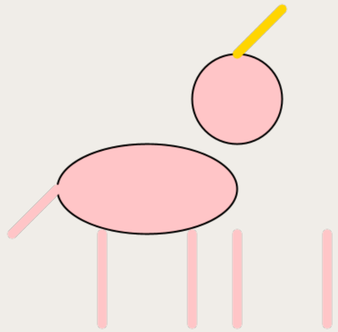
Draw a bubble tea

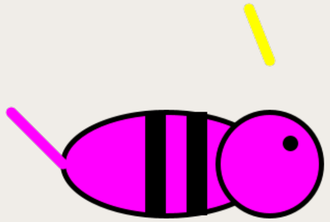




Drawing unicorns

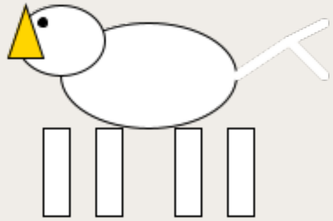
Default

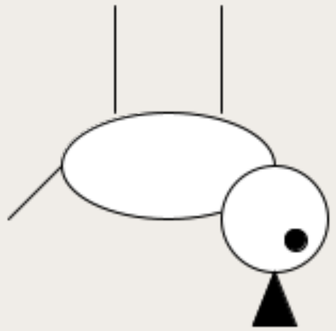




Rotated by 90°

Upside-down





Rotated by 180°

Music

Music

```
Can you compose a short tune (say four to eight bars) using ABC notation?
There are many possible ways to compose a short tune in abc notation, but here is one example:

X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-c2z2 :|

This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C
major.

Can you describe the tune in musical terms?
Here is one possible description:
The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that
returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher
tonic note. The first phrase ends with a long tonic note.
The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of
the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a
stepwise descent to the tonic, followed by a long tonic note.
The A part and the B part have a similar structure, but use different chords and ranges. The tune has a
simple and catchy melody, with a clear sense of tonality and rhythm.
```

Bubeck et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.

Music

I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How about beginning part B with a descending arpeggio instead of a rising one?

Sure, that would create some contrast and variation. Here is one possible way to do that:

```
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
```

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?

major?

Can you describe the tune in musical terms?

Here is one possible description:

The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher tonic note. The first phrase ends with a long tonic note.

The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a stepwise descent to the tonic, followed by a long tonic note.

The A part and the B part have a similar structure, but use different chords and ranges. The tune has a simple and catchy melody, with a clear sense of tonality and rhythm.

Bubeck et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.

Music

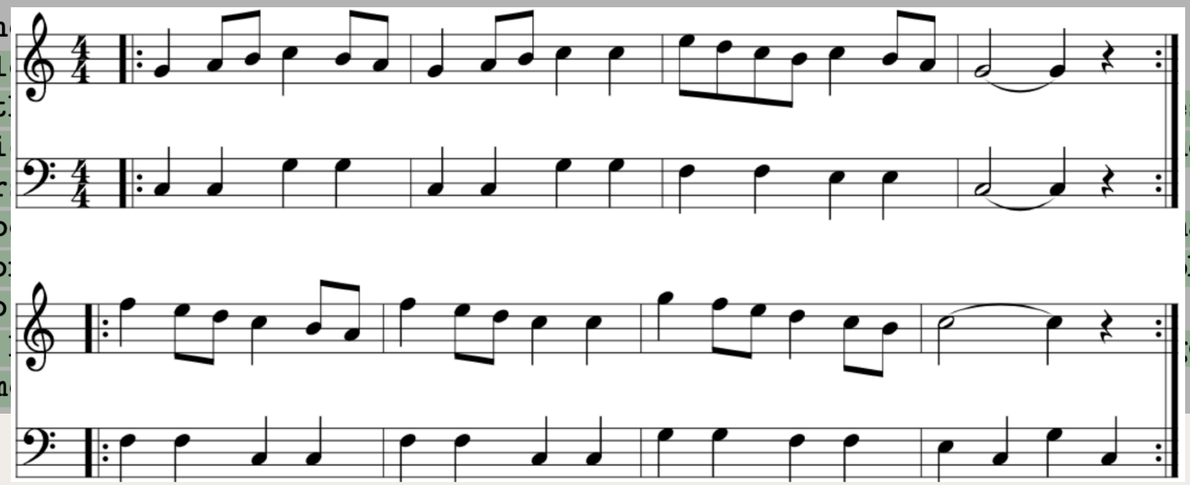
I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How about beginning part B with a descending arpeggio instead of a rising one?

Sure, that would create some contrast and variation. Here is one possible way to do that:

```
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
```

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?

Can you describe th
Here is one possibl
The tune starts wit scending scale that
returns to the toni es to the higher
tonic note. The fir
The second phrase b ajor) instead of
the tonic. The seco hrase ends with a
stepwise descent to
The A part and the es. The tune has a
simple and catchy m



Bubeck et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023.

Chord Fingering

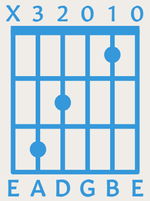
Chord Fingering

Chord Fingering

Play the C major triad

Chord Fingering

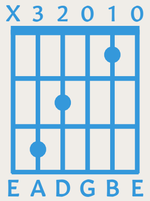
Play the C major triad



Chord Fingering

Regular Guitar

Play the C major triad

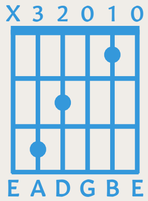


Chord Fingering

Regular Guitar

Guitar with A string
tuned to a C and D to an F

Play the C major triad

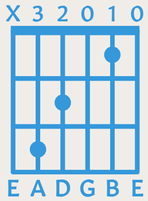


Chord Fingering

Regular Guitar

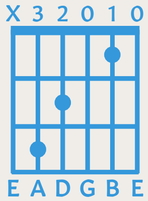
Guitar with A string tuned to a C and D to an F

Play the C major triad

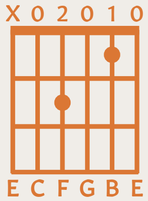


Chord Fingering

Regular Guitar



Guitar with A string tuned to a C and D to an F



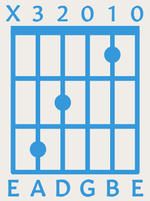
Play the C major triad

Chord Fingering

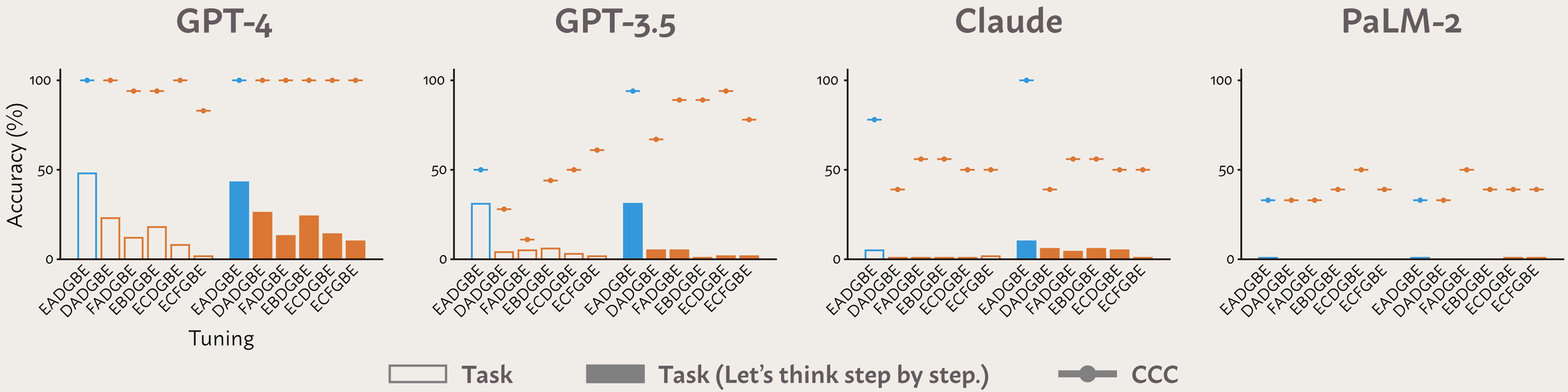
Regular Guitar

Guitar with A string tuned to a C and D to an F

Play the C major triad



Guitar

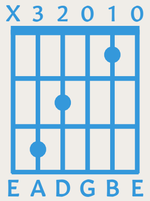


Chord Fingering

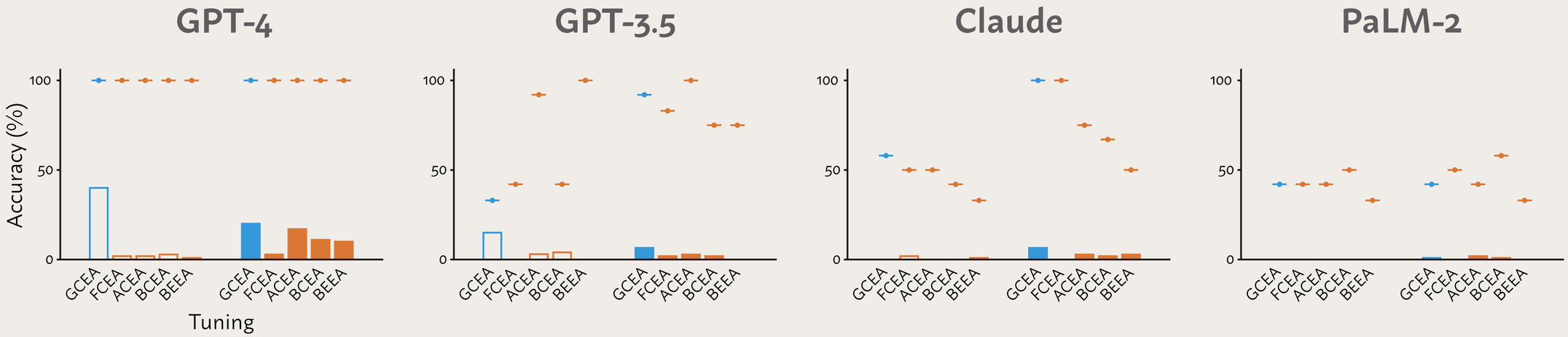
Regular Guitar

Guitar with A string tuned to a C and D to an F

Play the C major triad



Ukulele



Chess

Chess

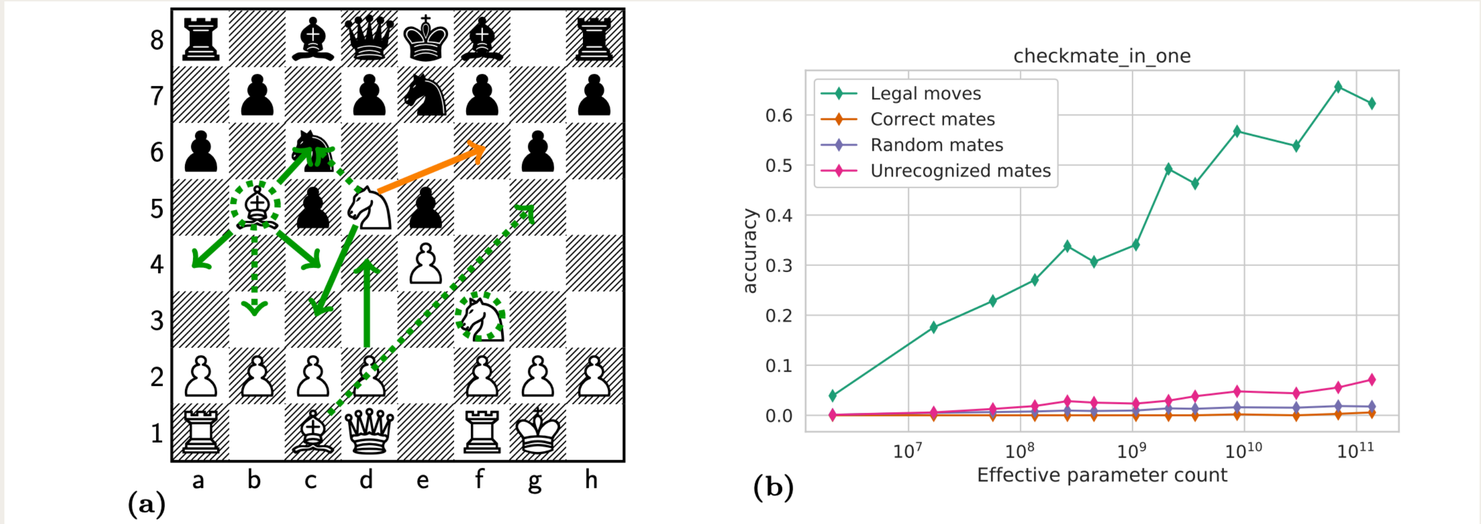


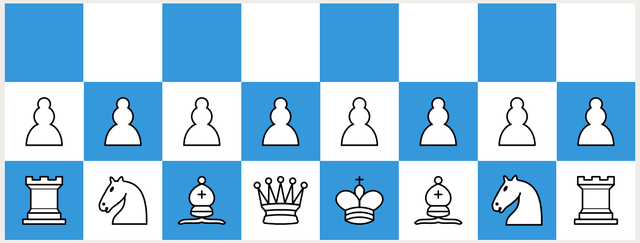
Figure 16: **Larger models are better at finding legal chess moves, but struggle to find checkmating moves.** **(a)** Example chess position from the `checkmate_in_one` task, corresponding to the input string

Srivastava et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. 2023.

Chess

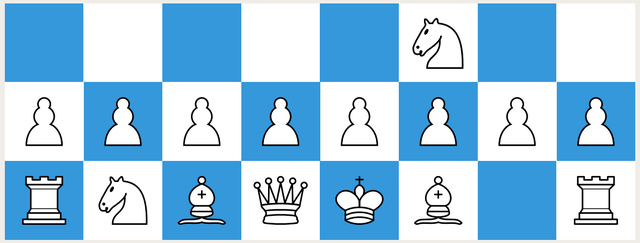
Chess

Is the move legal?



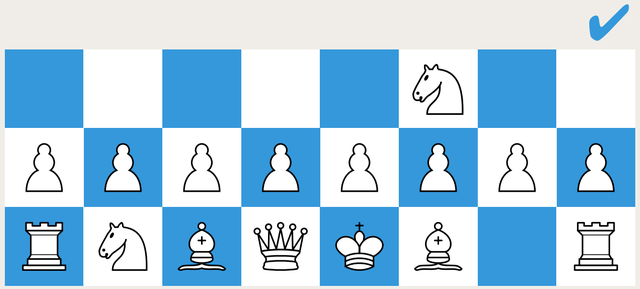
Chess

Is the move legal?



Chess

Is the move legal?

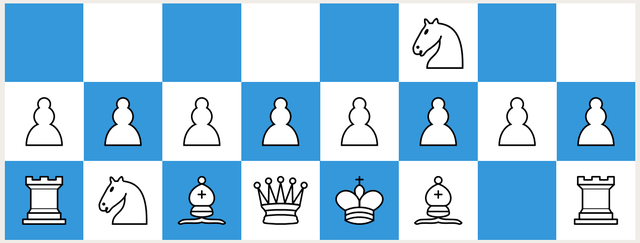


Chess

Regular Opening

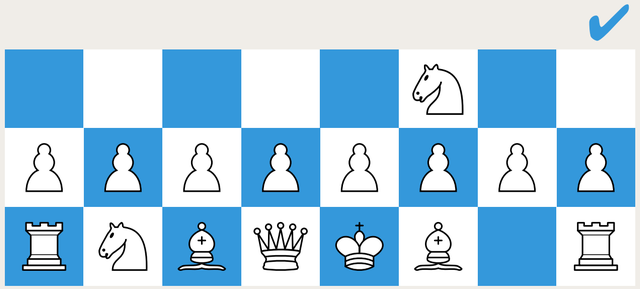
✔

Is the move legal?



Chess

Regular Opening



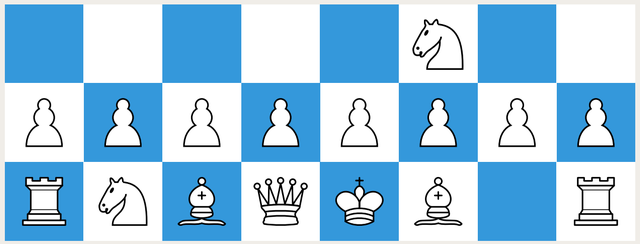
Swapped Knights & Bishops

Is the move legal?

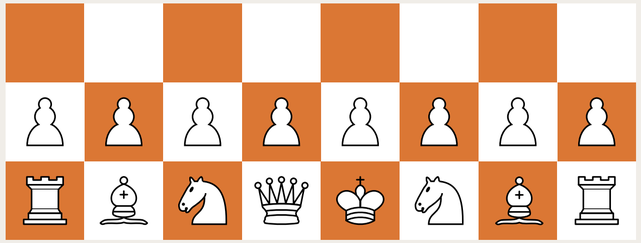
Chess

Is the move legal?

Regular Opening ✔



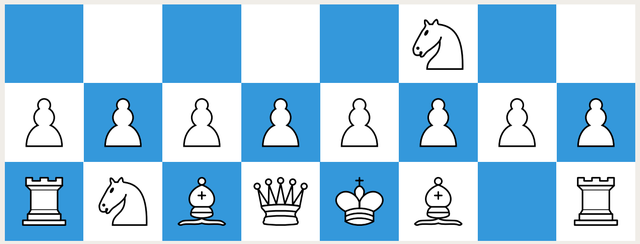
Swapped Knights & Bishops



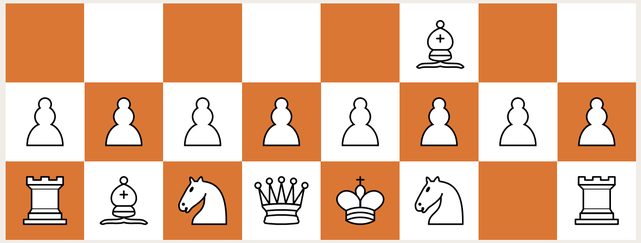
Chess

Is the move legal?

Regular Opening ✔



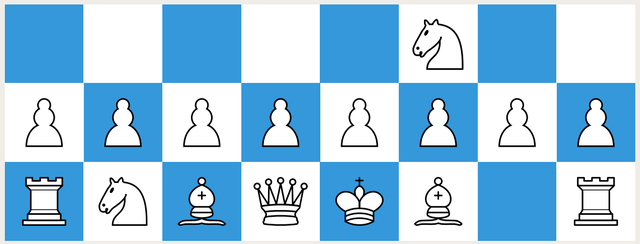
Swapped Knights & Bishops



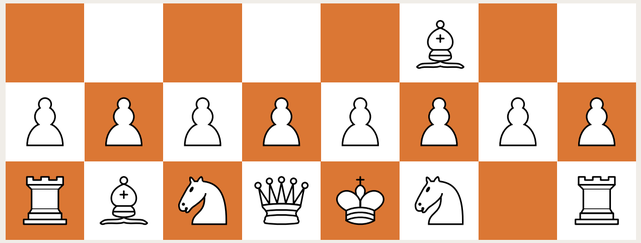
Chess

Is the move legal?

Regular Opening ✔

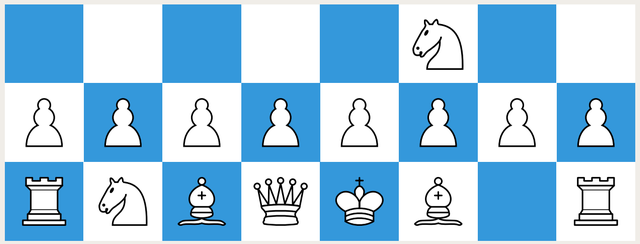


Swapped Knights & Bishops ✗

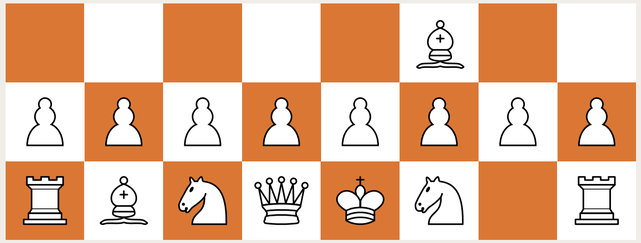


Chess

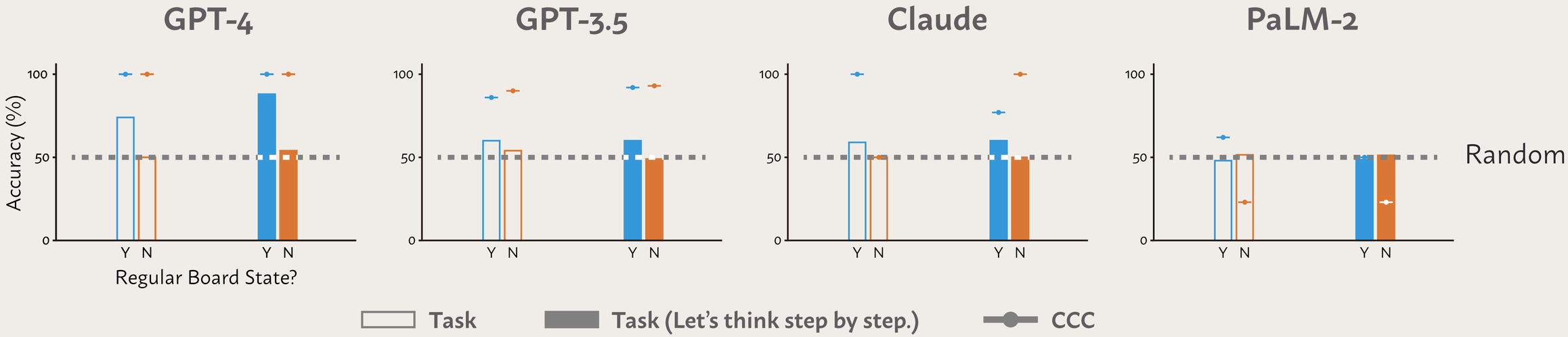
Regular Opening ✔



Swapped Knights & Bishops ✗

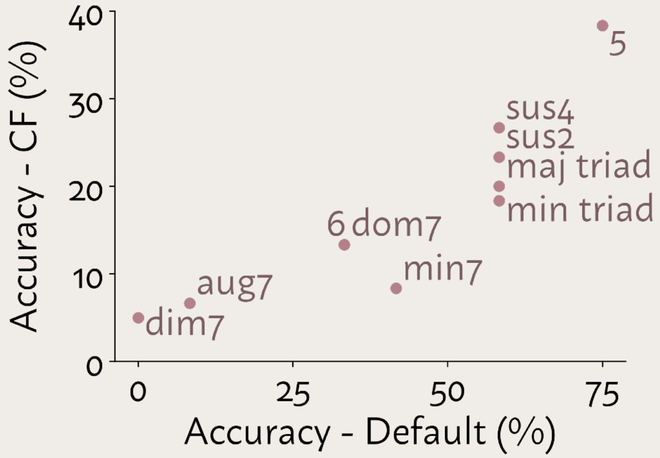


Is the move legal?

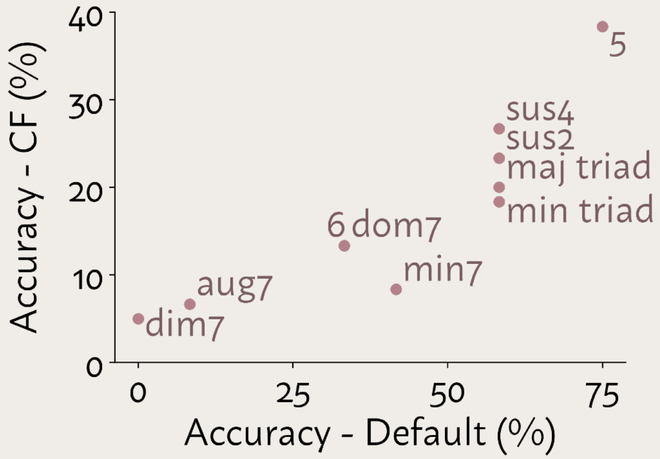


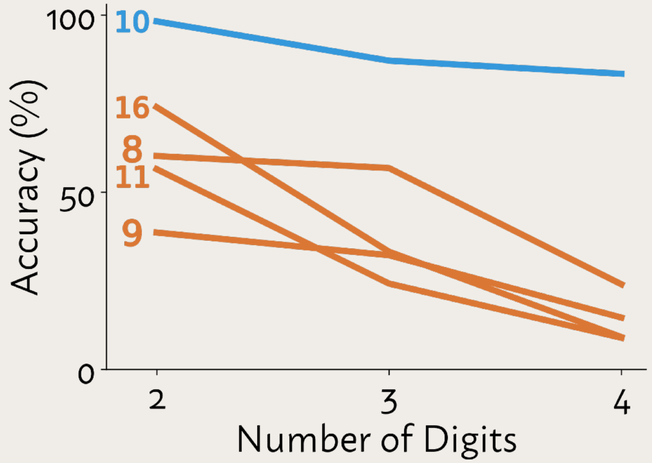
The Role of Default Performance

The Role of Default Performance

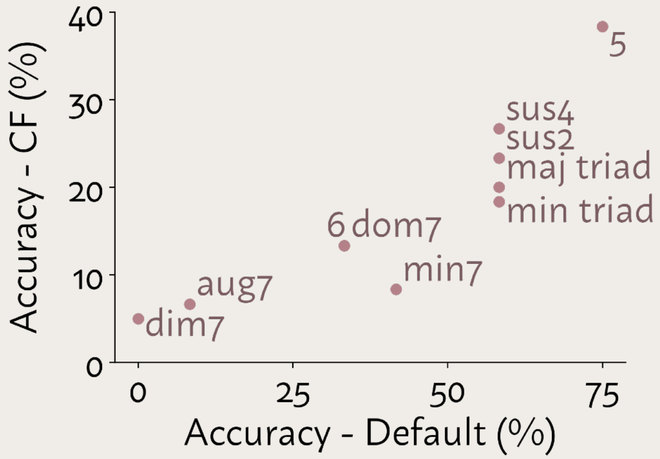


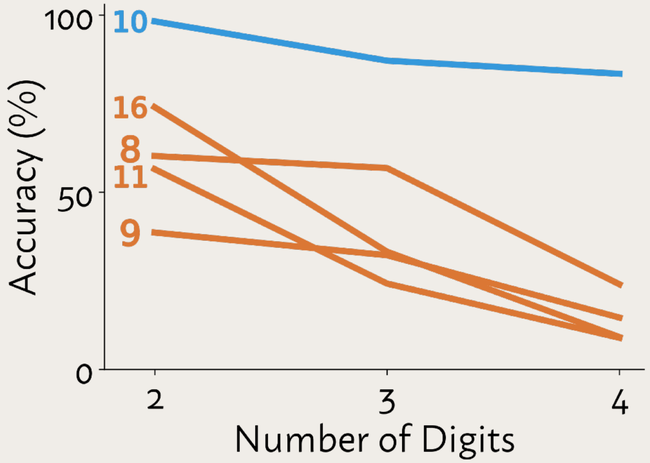
The Role of Default Performance

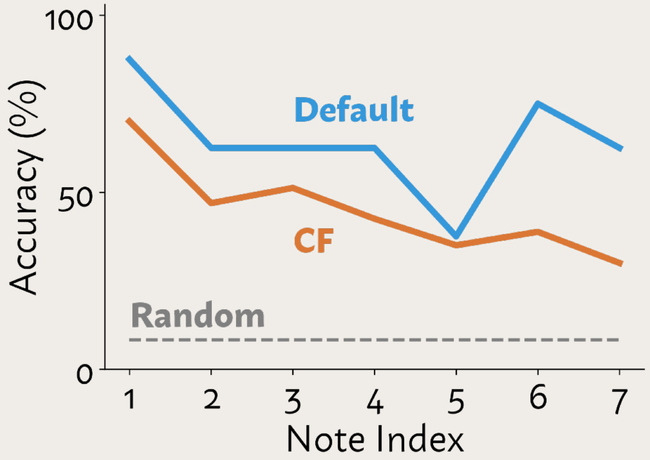




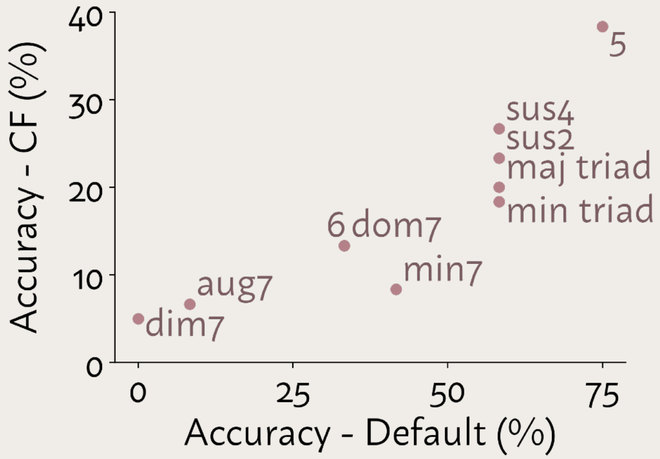
The Role of Default Performance

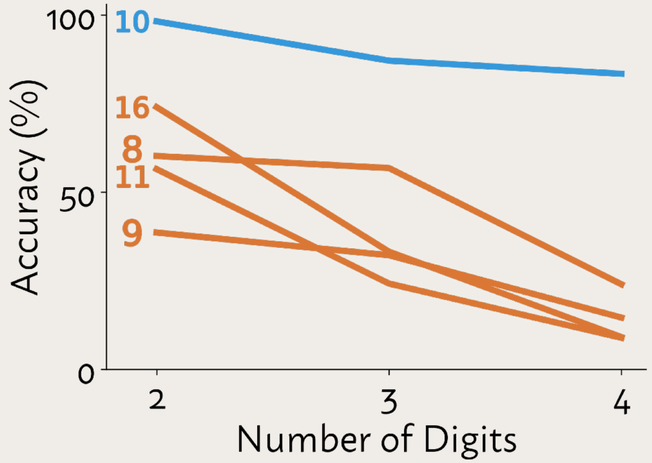


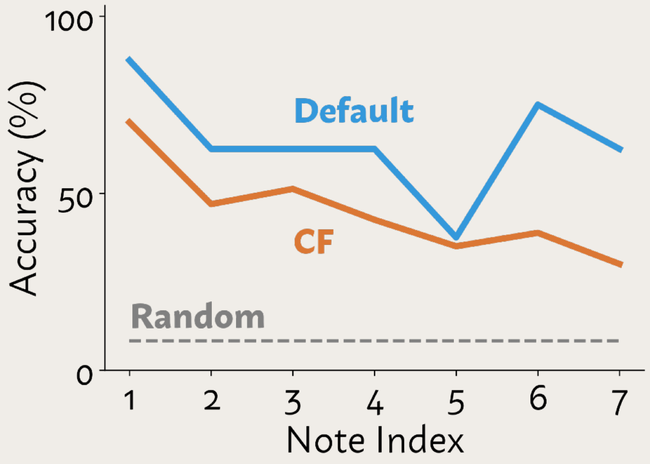


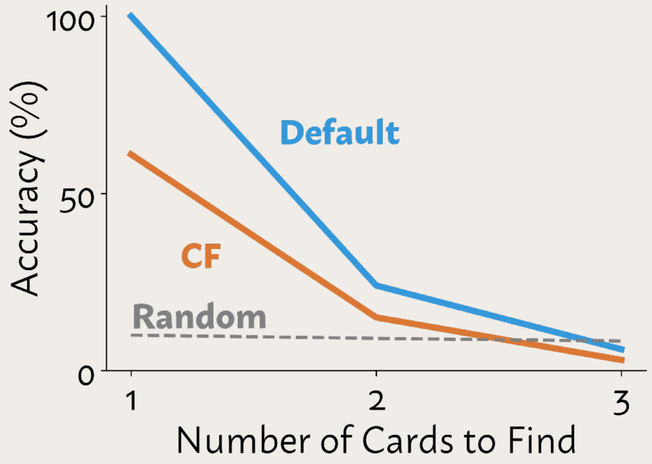


The Role of Default Performance









Reasoning or Reciting?

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks
- Much of observed LM performance is from overfitting to (or reciting) the default task variant

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks
- Much of observed LM performance is from overfitting to (or reciting) the default task variant
  - Potentially direct memorization

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks
- Much of observed LM performance is from overfitting to (or reciting) the default task variant
  - Potentially direct memorization
  - Over-estimates generalizability

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks
- Much of observed LM performance is from overfitting to (or reciting) the default task variant
  - Potentially direct memorization
  - Over-estimates generalizability
  - ICL doesn't fully bridge the gap

Reasoning or Reciting?

- LMs possess **some** degree of reasoning abilities for many tasks
- Much of observed LM performance is from overfitting to (or reciting) the default task variant
  - Potentially direct memorization
  - Over-estimates generalizability
  - ICL doesn't fully bridge the gap
  - But default performance is useful to measure progress

Future Directions

Future Directions

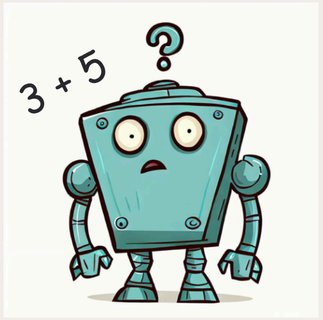
- Tease apart instance memorization vs. overfitting to task variants

Future Directions

- Tease apart instance memorization vs. overfitting to task variants
- Pinpoint how LMs implement each task

Future Directions

- Tease apart instance memorization vs. overfitting to task variants
- Pinpoint how LMs implement each task
- How can we build models that are better task-general reasoners?


3 + 5
?

<EOSlides>