Reasoning or Reciting?

Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

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

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

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

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









the following integral? \$\int_{-\infty}^{\infty} e^{-x^2} dx\$		
sking about is the Gaussian integral. Its value is:	0 6 9	ን

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













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 $\square \square \square \square$ searching, which is very efficient because it never backtracks more than necessary.

Here's a Python implementation of the KMP algorithm:

python	Copy code
def compute prefix function(nattern):	
""Compute the prefix function for the KMP algorithm """	
m = len(nattern)	
m = 100 + m	
$\dot{r} = 0$	
$\int - \mathbf{v}$	
while i > 0 and pattern[i] != pattern[i]:	
i = pi[i-1]	
if pattern[i] == pattern[i]:	
i += 1	
pi[i] = j	
return pi	
<pre>def kmp_search(text, pattern):</pre>	
"""Perform the KMP search on the text."""	
n = len(text)	
m = len(pattern)	
<pre>pi = compute_prefix_function(pattern)</pre>	
j = 0 # number of characters matched in pattern	
<pre>for i in range(n):</pre>	
<pre>while j > 0 and text[i] != pattern[j]:</pre>	
j = pi[j-1]	
<pre>if text[i] == pattern[j]:</pre>	
j += 1	
<pre>if j == m: # a match is found</pre>	
return i - m + 1	
# (i - m + 1) is the starting index of the pattern in $ au$	the text
return -1 # not found	
# Example usage:	
text = "ABC ABCDAB ABCDABCDABDE"	
pattern = "ABCDABD"	
print(kmp search(text nattern)) # prints 15 which is the st	arting index o
princial principal and the standard and the standard standa	LETING THUER O





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 ር ዋ searching, which is very efficient because it never backtracks more than necessary

Here's a Python implementation of the KMP algorithm:

python	🗂 Copy code
<pre>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]:</pre>	
<pre>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</pre>	e text
<pre># Example usage: text = "ABC ABCDAB ABCDABCDABDE" pattern = "ABCDABD" print(kmp_search(text, pattern)) # prints 15, which is the start</pre>	ting index o

GPT-4 can code so well!







GPT-4 can code so well!



text = "ABC ABCDAB ABCDABCDABDE" pattern = "ABCDABD" match_index = kmp_search(text, pattern) print("Pattern found at index:", match_index)

This code defines a kmp_search function that first constructs the partial match table for the given pattern. It then uses this table to efficiently search through the text. When it finds a match, it returns the starting index of that match; otherwise, it returns -1.



https://www.almabetter.com/bytes/articles/kmp-algorithm





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:

Ĵ	Copy	/ code
---	------	--------

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

text = "ABC ABCDAB ABCDABCDABDE" pattern = "ABCDABD" print(kmp_search(text, pattern))

```
for i in range(n):
       while j > 0 and text[i] != pattern[j]:
           j = pi[j-1]
       if text[i] == pattern[j]:
           j += <mark>1</mark>
       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!



Memorization





Memorization



od	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72.7, 71.5)
\bigcirc	Phi-3-mini-128k-instruct (74.1, 68.3)
\bigcirc	math-shepherd-mistral-7b-rl (74.5, 61.1)
\bigcirc	dbrx-instruct (75.0, 70.1)
\bigcirc	Meta-Llama-3-8B-Instruct (75.2. 69.0)
\bigcirc	Smaug-34B-v0.1 (75.7. 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
\bigcirc	Mixtral-8x22B-v0.1 (77.0.67.7)
	Phi-3-mini-4k-instruct (78.2.68.4)
	claude-3-haiku-20240307 (78.5, 78.5)
\bigcirc	mistral-medium-latest (79.0, 75.2)
	mistral-small-latest (79.0, 71.8)
	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
	mistral-large-latest (85.3, 85.3)
	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)



Memorization



lod	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72.7, 71.5)
\bigcirc	Phi-3-mini-128k-instruct (74.1, 68.3)
\bigcirc	math-shepherd-mistral-7b-rl (74.5, 61.1)
\bigcirc	dbrx-instruct (75.0, 70.1)
\bigcirc	Meta-Llama-3-8B-Instruct (75.2, 69.0)
\bigcirc	Smaug-34B-v0.1 (75.7, 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
	Mixtral-8x22B-v0.1
	Phi-3-mini-4k-instruct (78.2 68.4)
	claude-3-haiku-20240307 (78 5 78 5)
	mistral-medium-latest (79.0, 75.2)
	mistral-small-latest (79.0, 71.8)
	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
\bigcirc	mistral-large-latest (85.3, 85.3)
	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
\bigcirc	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)
_	





Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?

John writes 20 pages a day. How long will it take him to write 3 books that are 400 pages each?

Herbert is 10 years younger than Kris. If Kris is 24 years old now, how old will Herbert be next year?

James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?







Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?

John writes 20 pages a day. How long will it take him to write 3 books that are 400 pages each?

Herbert is 10 years younger than Kris. If Kris is 24 years old now, how old will Herbert be next year?

Cobbe et al. Training Verifiers to Solve Math Word Problems. 2021.

James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?







lod	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72.7, 71.5)
\bigcirc	Phi-3-mini-128k-instruct (74.1, 68.3)
	math-shepherd-mistral-7b-rl (74.5, 61.1)
\bigcirc	dbrx-instruct (75.0, 70.1)
\bigcirc	Meta-Llama-3-8B-Instruct (75.2, 69.0)
\bigcirc	Smaug-34B-v0.1 (75.7, 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
	Mixtral-8x22B-v0.1 (77.0, 67.7)
\bigcirc	Phi-3-mini-4k-instruct (78.2, 68.4)
\bigcirc	claude-3-haiku-20240307 (78.5, 78.5)
\bigcirc	mistral-medium-latest (79.0, 75.2)
	mistral-small-latest (79.0, 71.8)
	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
	mistral-large-latest (85.3, 85.3)
\bigcirc	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
\bigcirc	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)



7



1od	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72 7 71 5)
	Phi-3-mini-128k-instruct
	(74.1, 66.5) math-shepherd-mistral-7b-rl
	(74.5, 61.1) dbrx-instruct
\bigcirc	(75.0, 70.1) Meta-Llama-3-8B-Instruct
\bigcirc	(75.2, 69.0)
\bigcirc	(75.7, 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
\bigcirc	Mixtral-8x22B-v0.1 (77.0, 67.7)
\bigcirc	Phi-3-mini-4k-instruct (78.2, 68.4)
\bigcirc	claude-3-haiku-20240307 (78.5, 78.5)
\bigcirc	mistral-medium-latest (79.0, 75.2)
	mistral-small-latest (79.0, 71.8)
	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
	mistral-large-latest (85.3, 85.3)
	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
\bigcirc	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)



7



lod	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72.7, 71.5)
	Phi-3-mini-128k-instruct (74.1, 68.3)
	math-shepherd-mistral-7b-rl (74.5, 61.1)
\bigcirc	dbrx-instruct (75.0, 70.1)
\bigcirc	Meta-Llama-3-8B-Instruct
\bigcirc	Smaug-34B-v0.1 (75.7 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
	Mixtral-8x22B-v0.1
	Phi-3-mini-4k-instruct
	claude-3-haiku-20240307
	mistral-medium-latest
	mistral-small-latest (79.0, 71.8)
	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
	mistral-large-latest (85.3, 85.3)
	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)
_	



7



Zhang et al. A Careful Examination of Large Language Model Performance on Grade School Arithmetic. 2024.

lod	el (GSM8k, GSM1k)
\bigcirc	claude-3-sonnet-20240229 (71.9, 74.4)
\bigcirc	dbrx-base (72.7, 71.5)
\bigcirc	Phi-3-mini-128k-instruct (74.1, 68.3)
\bigcirc	math-shepherd-mistral-7b-rl (74.5, 61.1)
\bigcirc	dbrx-instruct (75.0, 70.1)
\bigcirc	Meta-Llama-3-8B-Instruct (75.2, 69.0)
\bigcirc	Smaug-34B-v0.1 (75.7, 68.8)
\bigcirc	gpt-3.5-turbo (76.0, 75.3)
	Mixtral-8x22B-v0.1 (77.0, 67.7)
\bigcirc	Phi-3-mini-4k-instruct (78.2, 68.4)
\bigcirc	claude-3-haiku-20240307 (78.5, 78.5)
\bigcirc	mistral-medium-latest (79.0, 75.2)
\bigcirc	mistral-small-latest (79.0, 71.8)
\bigcirc	gemini-pro (79.2, 78.9)
\bigcirc	Smaug-2-72B (79.9, 73.8)
\bigcirc	claude-3-opus-20240229 (80.2, 82.5)
\bigcirc	Meta-Llama-3-70B (81.1, 78.9)
\bigcirc	mistral-large-latest (85.3, 85.3)
\bigcirc	Mixtral-8x22B-Instruct-v0.1 (85.9, 76.0)
\bigcirc	claude-2.1 (88.7, 89.4)
\bigcirc	Meta-Llama-3-70B-Instruct (89.6, 87.6)
	gemini-1.5-pro-preview-0409 (89.7, 87.9)
\bigcirc	gpt-4-turbo (89.8, 89.8)
\bigcirc	gpt-4 (91.1, 91.8)

Counterfactual **Evaluation**





Can LMs really do math?







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



27+62=

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=



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.





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=



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=





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

89

27+62=





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

89

27+62=

Base-9



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

89

27+62=

counterfactual





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

27+62=

89

 $ADD_{base}(a, b)$

counterfactual





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

89

• • •

27+62=

 $ADD_{base}(a, b)$

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





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

27+62=

89

 $ADD_{base}(a, b)$

counterfactual




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

27+62=

89

 $ADD_{base}(a, b)$

 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual





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

27+62=





 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual





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

27+62=





 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual







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

89

27+62=

 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

100

counterfactual







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

89

27+62=

 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

100











General task abilities

 $f_w(x)$

General task abilities Or Non-transferrable procedures

 $f_w(x)$

 $f_{W^{default}}(x)$



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

89

27+62=

counterfactual



100



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

89

27+62=







100







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

89

27+62=







100

Counterfactual comprehension check (CCC)







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

89

27+62=

What follows 78?







100

Counterfactual comprehension check (CCC)







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



Base-10

89

27+62=

What follows 78?







100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

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





100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

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





100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

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













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











GPT-4

GPT-3.5



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



Task (Let's think step by step.)

- CCC









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





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

"b"







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

"b"







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

"b"



ThonPy 1-based indexing







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

"b"



ThonPy 1-based indexing









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

"b"



ThonPy 1-based indexing











```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

"b"



ThonPy 1-based indexing











```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

["ba", "ab"]

"b"



ThonPy 1-based indexing











```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

["ba", "ab"]

"b"



ThonPy 1-based indexing

["ab", "ba"]

"a"









```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

["ba", "ab"]

"b"





ThonPy 1-based indexing

["ab", "ba"]

"a"













• Under a fixed time budget, probably, but what if given unlimited time?



- Under a fixed time budget, probably, but what if given unlimited time?
 - Competence vs. performance



- Under a fixed time budget, probably, but what if given unlimited time?
 - Competence vs. performance



Lewis and Mitchell. Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models. 2024.



- Under a fixed time budget, probably, but what if given unlimited time?
 - Competence vs. performance


Wait, I can't do base-9 arithmetic easily either

- Under a fixed time budget, probably, but what if given unlimited time?
 - Competence vs. performance
- Why calibrate using human performance?



15

Few-Shot Results



Few-Shot Results



16







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





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





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct instance memorization





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct instance memorization
 - Over-estimates generalizability





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct instance memorization
 - Over-estimates generalizability
 - ICL doesn't fully bridge the gap





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct instance memorization
 - Over-estimates generalizability
 - ICL doesn't fully bridge the gap





