Līla

A Unified Benchmark for Mathematical Reasoning

Matthew Finlayson[1] Swaroop Mishra[1] Pan Lu Leonard Tang
Sean Welleck Chitta Baral Tanmay Rajpurohit Oyvind Tafjord
Ashish Sabharwal Peter Clark Ashwin Kalyan

September 16, 2022

mattf1n.github.io matthewf@allenai.org

[1] Equal first-authors

TL;DR

TL;DR

- Current math reasoning evaluation is broken.

TL;DR

- Current math reasoning evaluation is broken.
- We build LĪLA, a comprehensive benchmark.

TL;DR

- Current math reasoning evaluation is broken.
- We build Līla, a comprehensive benchmark.
- We train Bhāskara, a foundational math reasoning model.

Motivation

Motivation

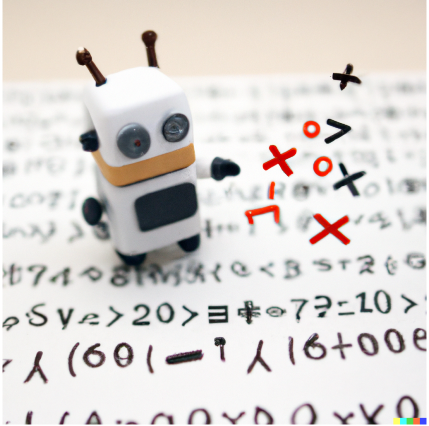
Can language models do math?

Motivation

Can language models do math? How can we find out?

Motivation

Can language models do math? How can we find out?



Can language models do math?

Fill-in-the-blank

Can language models do math?

Fill-in-the-blank

👱 Fifty is equal to _ times ten.

Can language models do math?

Fill-in-the-blank ✅

🧑 Fifty is equal to _ times ten.

🤖 `Five`

Can language models do math?

Fill-in-the-blank ✅

Can language models do math?

Fill-in-the-blank ✅ Common sense

Can language models do math?

Fill-in-the-blank ✅ Common sense

👱 A skiff refuels after 10 miles in the bay compared to 4 at sea. Which is more rugged?

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌

👱 A skiff refuels after 10 miles in the bay compared to 4 at sea. Which is more rugged?

🤖 `The bay`

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌ Algebra

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌ Algebra

👱 Solve $x + 9j = 27 + 6$ for x when $5j - 2 - 18 = 0$.

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌ Algebra ❌

🧑 Solve $x + 9j = 27 + 6$ for x when $5j - 2 - 18 = 0$.

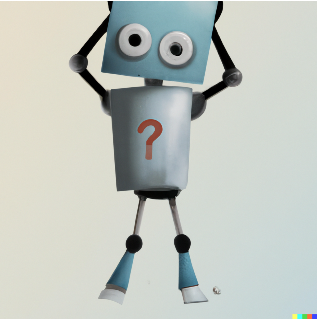
🤖 `x = 63`

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌ Algebra ❌

Can language models do math?

Fill-in-the-blank ✅ Common sense ❌ Algebra ❌ Number theory ❓
Multiple-choice ❓ Comparison ❓ Science knowledge ❓ Arithmetic ❓ Geometry ❓ ...



A math question taxonomy

A math question taxonomy

Dataset	Format	Subject	Knowledge	Language
Numersense	Fill-in	Arithmetic	Math	Simple
NumGLUE	Multi-choice	Comparision	Real world	Complex
Deepmind	Generative	Calculus	Math	None
MCTaco	Multi-choice	Arithmetic	Commonsense	Simple
...	...	...	...	...

A math question taxonomy

Dataset	Format	Subject	Knowledge	Language
Numersense	Fill-in	Arithmetic	Math	Simple
NumGLUE	Multi-choice	Comparision	Real world	Complex
Deepmind	Generative	Calculus	Math	None
MCTaco	Multi-choice	Arithmetic	Commonsense	Simple
...	...	...	...	...

Direct answering is unsatisfying

🧑 Solve $x + 9j = 27 + 6$ for x when $5j - 2 - 18 = 0$.

🤖 `x = 63` ❌

Direct answering is unsatisfying

🧑 Solve $x + 9j = 27 + 6$ for x when $5j - 2 - 18 = 0$.

🤖 `x = -3` ✅

Direct answering is unsatisfying

🧑 Solve $x + 9j = 27 + 6$ for x when $5j - 2 - 18 = 0$.

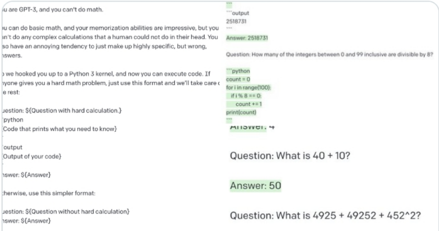
🐍

```
>>> j = (0 + 2 + 18) / 5
... x = 27 + 6 - 9 * j
... print(x)
-3
```

Language models ❤️ Python

Riley Goodside
@goodside · **Follow**

"You are GPT□3, and you can't do math": Prompting GPT□3 via zero-shot instruction to answer calculation/math questions by consulting a Python REPL.



```
u are GPT-3, and you can't do math.

u can do basic math, and your memorization abilities are impressive, but you
n't do any complex calculations that a human could not do in their head. You
so have an annoying tendency to just make up highly specific, but wrong,
iswers.

 we hooked you up to a Python 3 kernel, and now you can execute code. If
yone gives you a hard math problem, just use this format and we'll take care (
e rest:

uestion: ${Question with hard calculation.}
`python
{Code that prints what you need to know}

`output
{Output of your code}

iswer: ${Answer}

therwise, use this simpler format:

uestion: ${Question without hard calculation}
iswer: ${Answer}
```

```
```
```output
2518731
```

Answer: 2518731

Question: How many of the integers between 0 and 99 inclusive are divisible by 8?

```python
count = 0
for i in range(100):
  if i % 8 == 0:
    count += 1
print(count)
```
Answer: 4

Question: What is 40 + 10?

Answer: 50

Question: What is 4925 + 49252 + 452^2?
```

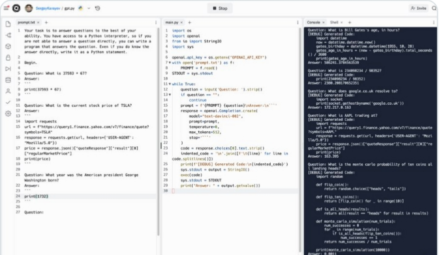
**Sergey Karayev**
@sergeykarayev · **Follow**

Here's a brief glimpse of our INCREDIBLE near future.

GPT-3 armed with a Python interpreter can
· do exact math
· make API requests
· answer in unprecedented ways

Thanks to **@goodside** and **@amasad** for the idea and repl!

Play with it: **replit.com/@SergeyKarayev...**


SergeyKarayev / gpt.py
Stop
Invite
prompt.txt
main.py
Console
Shell

```
Your task is to answer questions to the best of your ability. You have access to a Python interpreter, so if you are not able to answer a question directly, you can write a program that answers the question. Even if you do know the answer directly, write it as a Python statement.

Begin.

Question: What is 37593 * 67?
Answer:
```
print(37593 * 67)
```

Question: What is the current stock price of TSLA?
Answer:
```
import requests
url = f"https://query1.finance.yahoo.com/v7/finance/quote?symbols=TSLA"
response = requests.get(url, headers={'USER-AGENT': "Mozilla/5.0"})
price = response.json()["quoteResponse"]["result"][0]["regularMarketPrice"]
print(price)
```

Question: What year was the American president George Washington born?
Answer:
```
print(1732)
```

Question:

```python
import os
import openai
from io import StringIO
import sys

openai.api_key = os.getenv("OPENAI_API_KEY")
with open('prompt.txt') as f:
    PROMPT = f.read()
STDOUT = sys.stdout

while True:
    question = input('Question: ').strip()
    if question == "":
        continue
    prompt = f'{PROMPT} {question}\nAnswer:\n```'
    response = openai.Completion.create(
        model="text-davinci-002",
        prompt=prompt,
        temperature=0,
        max_tokens=512,
        stop='```'
    )
    code = response.choices[0].text.strip()
    indented_code = '\n'.join([f'\t{line}' for line in code.splitlines()])
    print(f'[DEBUG] Generated Code:\n{indented_code}')
    sys.stdout = output = StringIO()
    exec(code)
    sys.stdout = STDOUT
    print("Answer: " + output.getvalue())
```

```
Question: What is Bill Gates's age, in hours?
[DEBUG] Generated Code:
    import datetime
    now = datetime.datetime.now()
    gates_birthday = datetime.datetime(1955, 10, 28)
    gates_age_in_hours = (now - gates_birthday).total_seconds
() / 3600
    print(gates_age_in_hours)
Answer: 586241.3784563539

Question: What is 234098234 / 98352?
[DEBUG] Generated Code:
    print(234098234 / 98352)
Answer: 2380.200170652351

Question: What does google.co.uk resolve to?
[DEBUG] Generated Code:
    import socket
    print(socket.gethostbyname('google.co.uk'))
Answer: 172.217.0.163

Question: What is AAPL trading at?
[DEBUG] Generated Code:
    import requests
    url = f"https://query1.finance.yahoo.com/v7/finance/quote
?symbols=AAPL"
    response = requests.get(url, headers={'USER-AGENT': "Mozi
lla/5.0"})
    price = response.json()["quoteResponse"]["result"][0]["re
gularMarketPrice"]
    print(price)
Answer: 163.395

Question: What is the monte carlo probability of ten coins al
l landing heads?
[DEBUG] Generated Code:
    import random

    def flip_coin():
        return random.choice(["heads", "tails"])

    def flip_ten_coins():
        return [flip_coin() for _ in range(10)]

    def is_all_heads(results):
        return all(result == "heads" for result in results)

    def monte_carlo_simulation(num_trials):
        num_successes = 0
        for _ in range(num_trials):
            if is_all_heads(flip_ten_coins()):
                num_successes += 1
        return num_successes / num_trials

    print(monte_carlo_simulation(10000))
Answer: 0.0011
```

Halfway recap

Halfway recap

- Can language models do math?

Halfway recap

- Can language models do math?
- Existing benchmarks are too narrow in scope

Halfway recap

- Can language models do math?
- Existing benchmarks are too narrow in scope
- Python programming $>$ Direct answering

Līla: a comprehensive benchmark

AMPS MATH, Numersense, NumGLUE, MCTaco, ...

📜 Līla: a comprehensive benchmark

AMPS MATH, Numersense, NumGLUE, MCTaco, ...

👦 Find the laplacian of the function $f(x, y, z)$ where $f(x, y, z) = x^3y^3$.

Līla: a comprehensive benchmark

AMPS MATH, Numersense, NumGLUE, MCTaco, ...

👦 Find the laplacian of the function $f(x, y, z)$ where $f(x, y, z) = x^3y^3$.

🤖 $6x^3y + 6xy^3$

📜 LĪLA: a comprehensive benchmark

AMPS MATH, Numersense, NumGLUE, MCTaco, ...

👱 Find the laplacian of the function $f(x, y, z)$ where $f(x, y, z) = x^3y^3$.

🤖 $6x^3y + 6xy^3$

🐍

```
from sympy import *
C = CoordSys3D('C')
x, y, z = C.x, C.y, C.z
f = x**3*y**3
print(laplacian(f))
```

📜 Līla: a comprehensive benchmark

AMPS MATH, Numersense, NumGLUE, MCTaco, ...

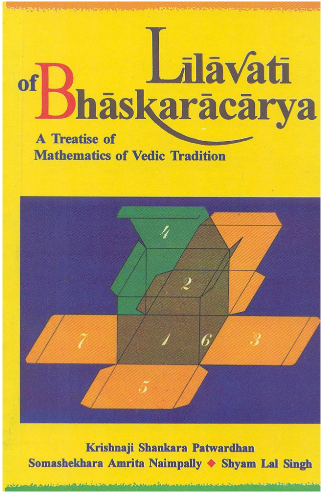
👱 Find the laplacian of the function $f(x, y, z)$ where $f(x, y, z) = x^3y^3$.

🤖 $6x^3y + 6xy^3$

🐍
```python
from sympy import *
C = CoordSys3D('C')
x, y, z = C.x, C.y, C.z
f = x**3*y**3
print(laplacian(f))
```

$100,000 \times$ (👱, 🤖, 🐍) = 📜 Līla

A quick note

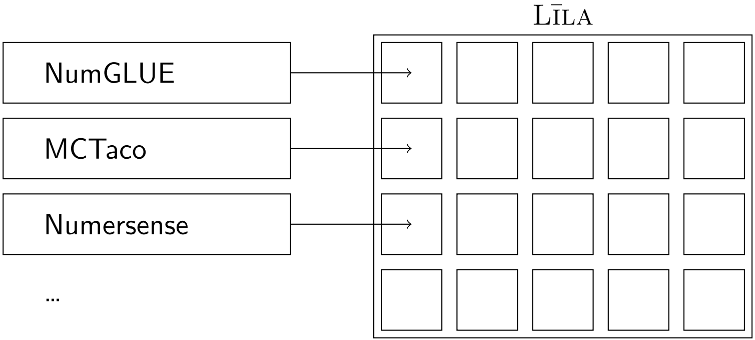


Līlāvatī of Bhāskarācārya

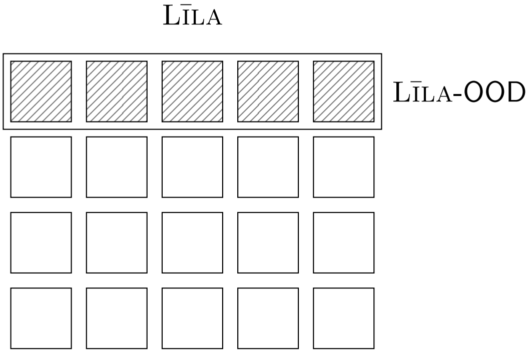
A Treatise of
Mathematics of Vedic Tradition

Krishnaji Shankara Patwardhan
Somashekhara Amrita Naimpally ◆ Shyam Lal Singh

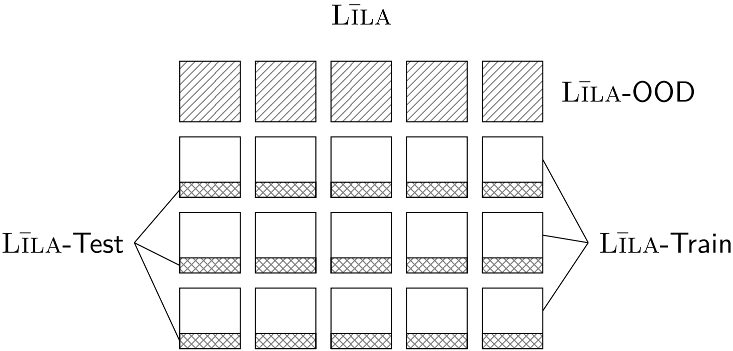
LĪLA splits



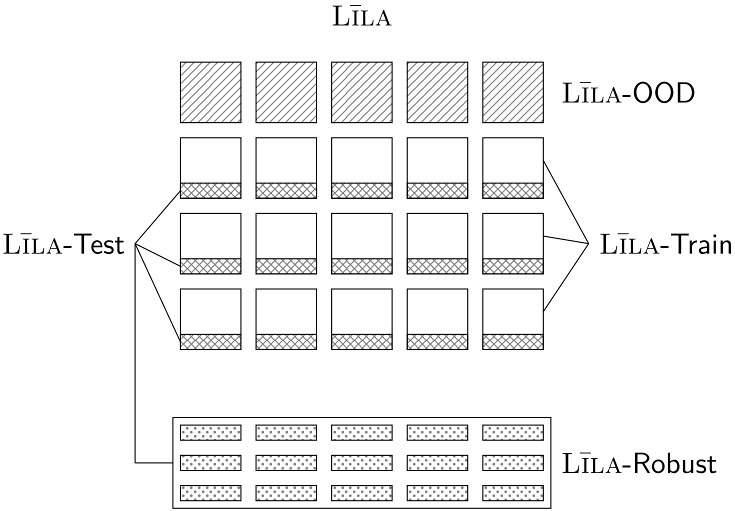
Līla splits



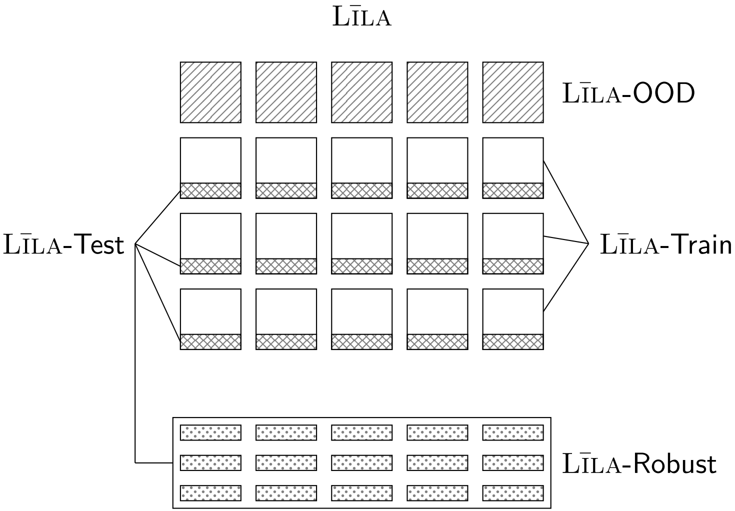
Līla splits



LĪLA splits

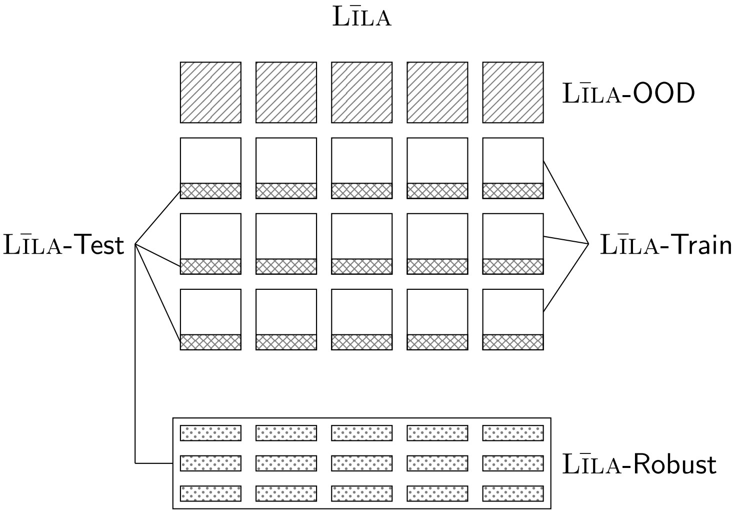


Līla splits



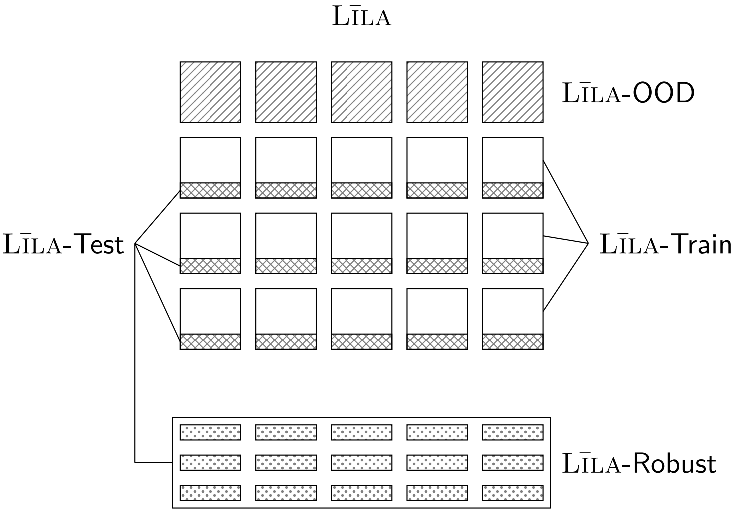
Jules gave 3 apples to...

LĪLA splits

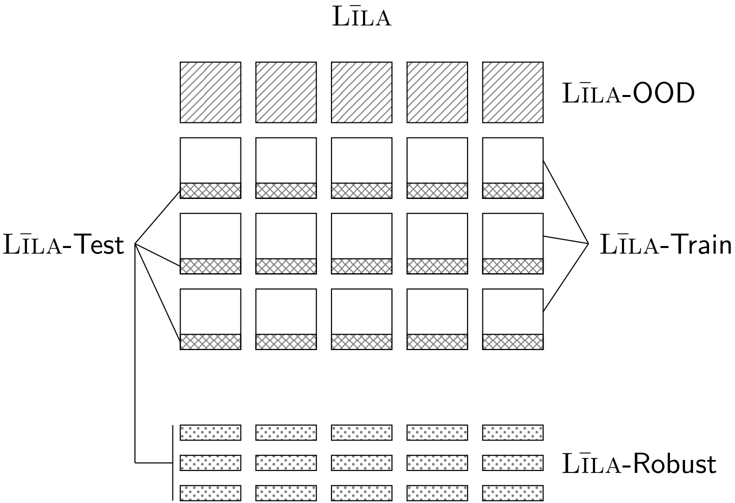


Riley is 7 years old. Jules gave 3 apples to...

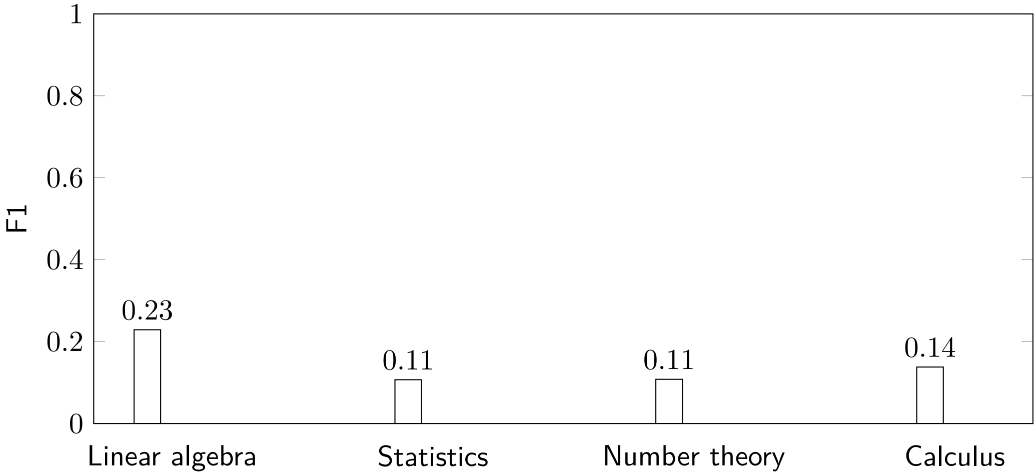
LĪLA splits



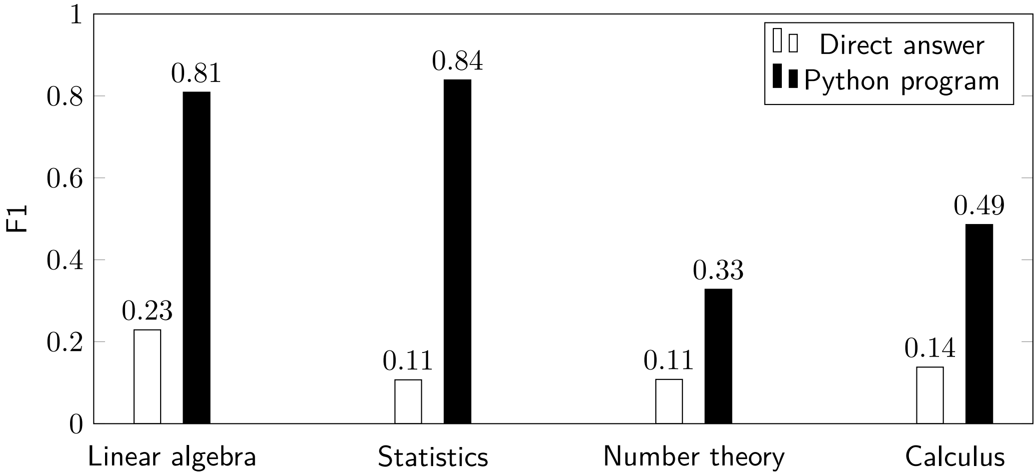
Līla splits



Python program > direct answer



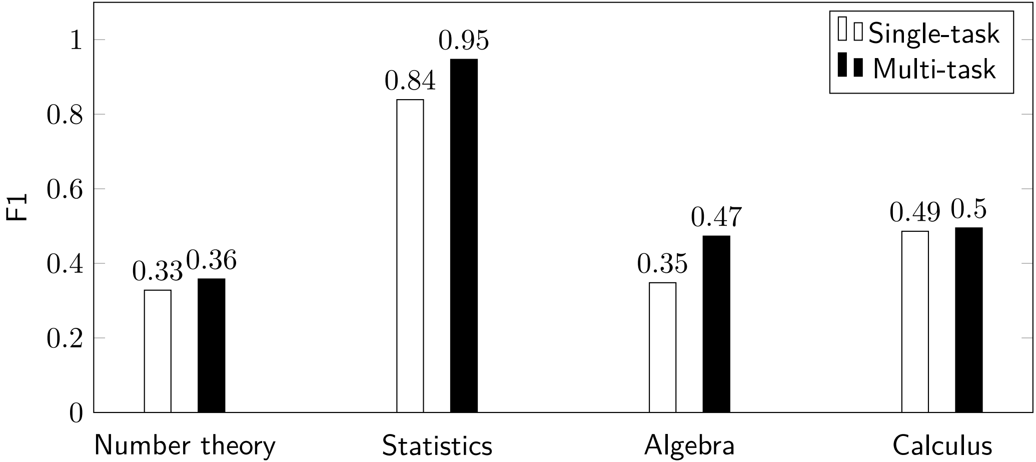
Python program > direct answer



Program answering 23 points better on average

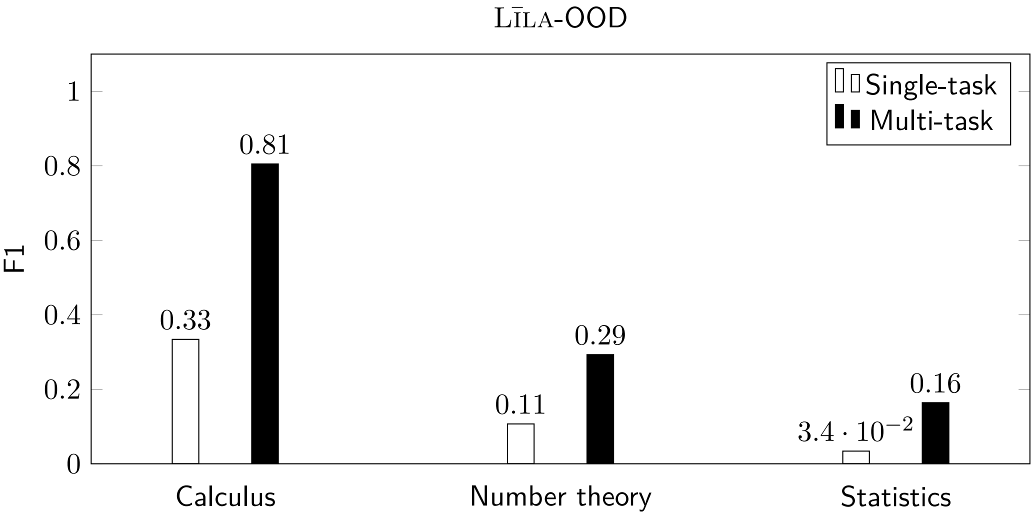
Multi-task > single-task

Multi-task > single-task



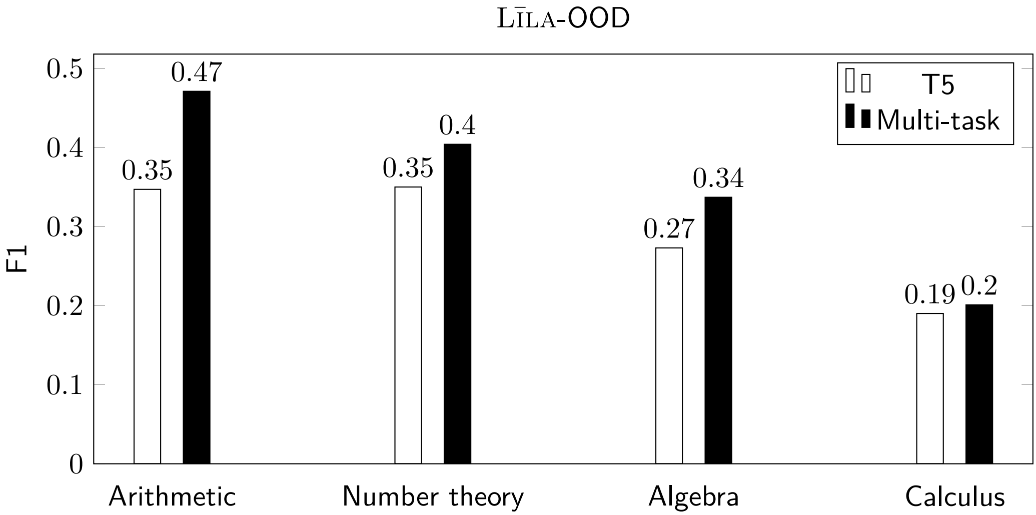
Multi-task model is 9 points better on average

Multi-task model is *general*



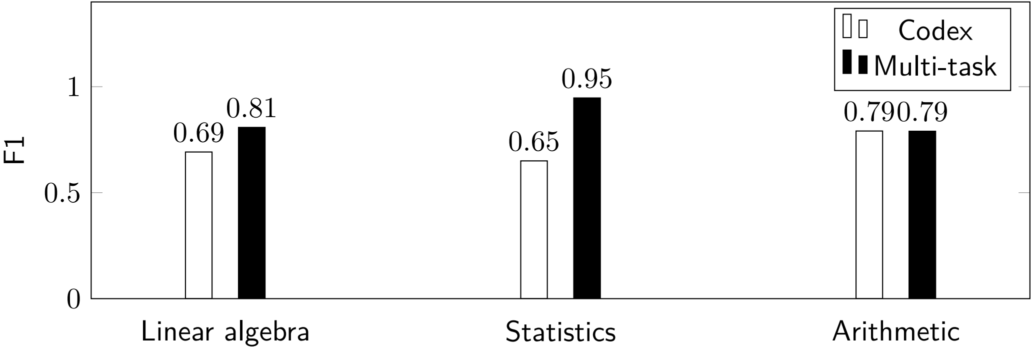
Multi-task model is 21 points better on average

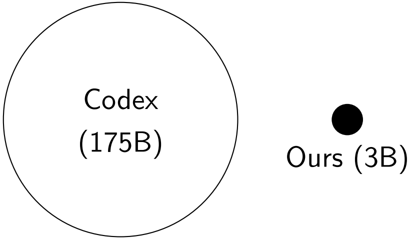
Our multi-task model is a starting point for math models



Multi-task model is 8 points better on average

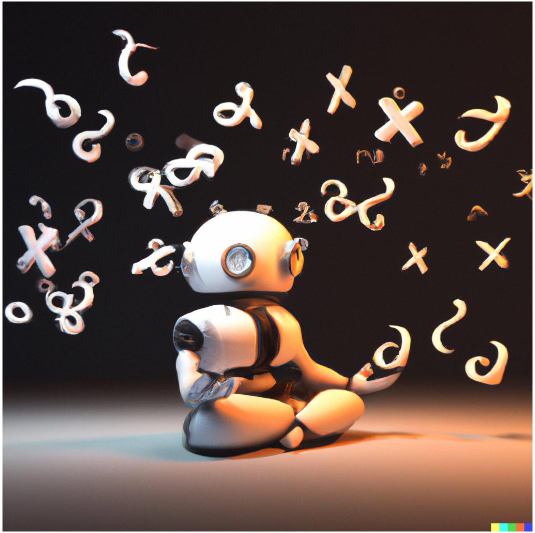
Our multi-task model outperforms Codex on some tasks




Codex (175B)

Ours (3B)

Bhāskara



Takeaways

Takeaways

- Math reasoning evaluation is broken.

Takeaways

- Math reasoning evaluation is broken.
- LĪLA: a comprehensive benchmark with useful splits.

Takeaways

- Math reasoning evaluation is broken.
- Līla: a comprehensive benchmark with useful splits.
- Bhāskara: a multi-task model for math reasoning.

Takeaways

- Math reasoning evaluation is broken.
- LĪLA: a comprehensive benchmark with useful splits.
- BHĀSKARA: a multi-task model for math reasoning.

Thank you!