

# Data for AI + mathematics: challenges and new frontiers

Sean Welleck



# AI + mathematics

Education &  
Applications

Research &  
discovery

Formal verification

# Language models and mathematics

## ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT May 24 Version](#)

**Question:** For every  $a, b, b \neq a$  prove that

$$\frac{a^2 + b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$

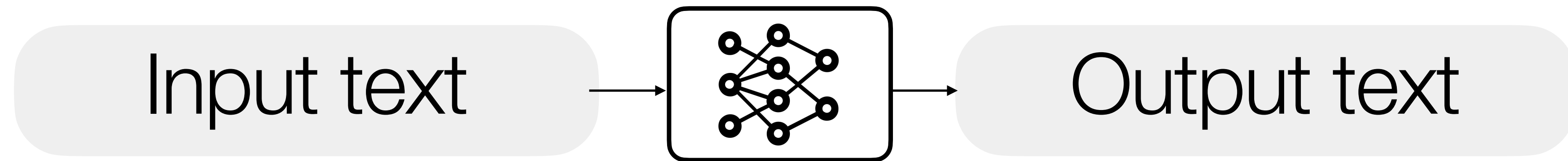
**Model output:**

$$\begin{aligned} & \frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2 \\ \Leftrightarrow & \frac{a^2+b^2}{2} > \frac{a^2+b^2+2ab}{4} \\ \Leftrightarrow & a^2 + b^2 > \frac{a^2+b^2+2ab}{2} \\ \Leftrightarrow & 2a^2 + 2b^2 > a^2 + b^2 + 2ab \\ \Leftrightarrow & a^2 + b^2 > 2ab \\ \Leftrightarrow & a^2 + b^2 - 2ab > 0 \\ \Leftrightarrow & (a - b)^2 > 0 \end{aligned}$$

which is true, because the square of a real number is positive.

Minerva Language Model  
2022 Poland National Math Exam  
[Lewkowycz et al 2022]

# Language models and mathematics



Goal: perform useful  
mathematical tasks

# Mathematical tasks

Free-form  
conversation

...

```
lemma one_plus_one_equals_two:  
  .....  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "1 = 1" by simp  
  .....  
  also have "..... = 2" by simp  
  finally have "....." by simp  
qed
```

Formal  
theorem  
proving

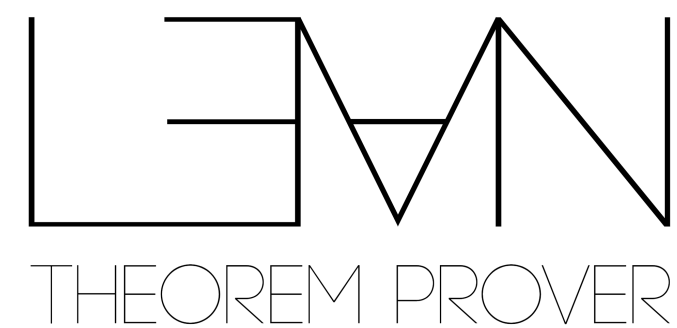
**Informal**

**Formal**

# Formal theorem proving

$$1 + 1 = 2$$

```
lemma one_plus_one_equals_two:  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "... = Suc 1" by simp  
  also have "... = 2" by simp  
  finally show ?thesis by simp  
qed
```



Microsoft Research



# Formal

# Formal theorem proving

$$1 + 1 = 2$$

proof



```
lemma one_plus_one_equals_two:  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "... = Suc 1" by simp  
  also have "... = 2" by simp  
  finally show ?thesis by simp  
qed
```

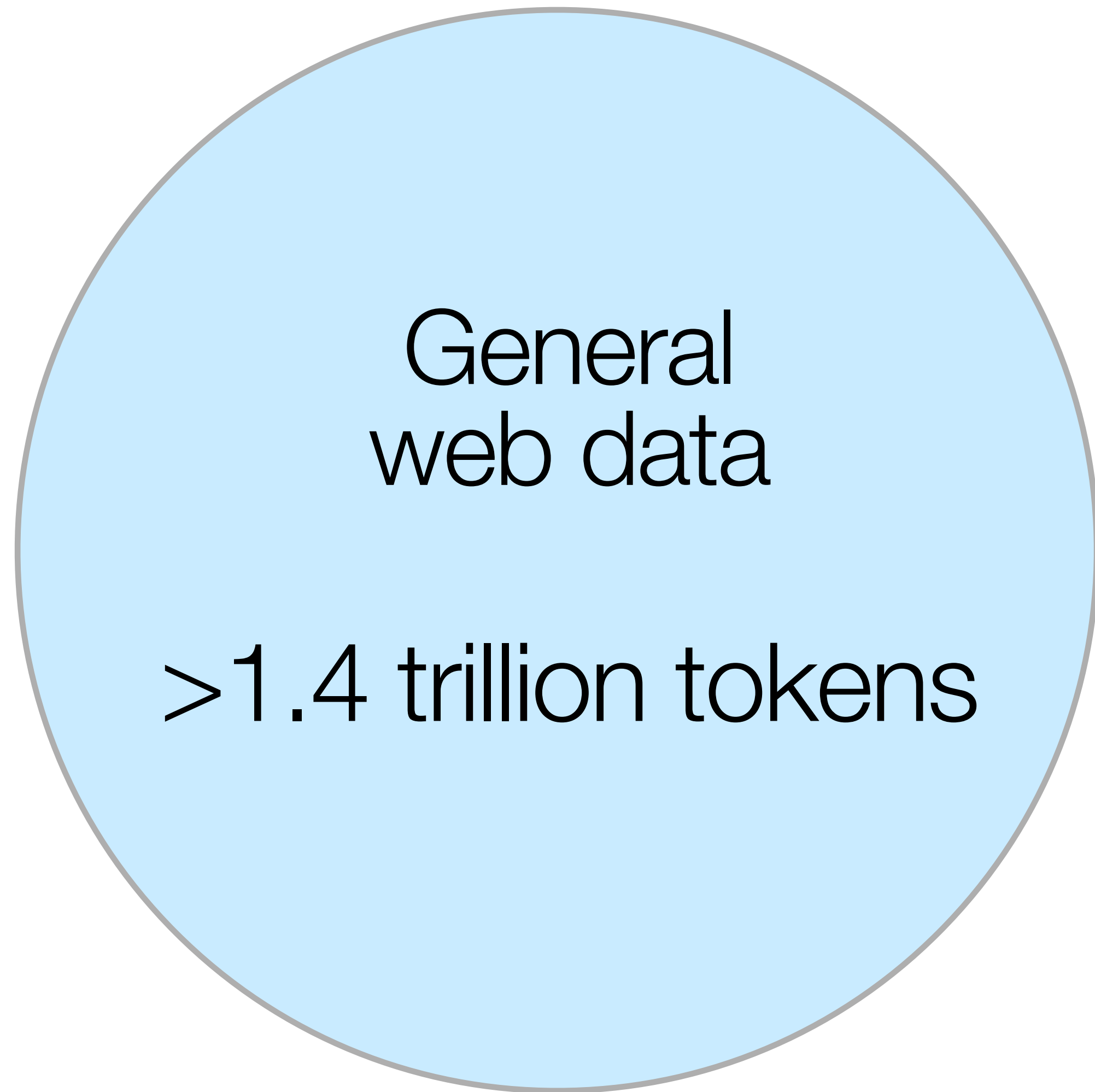
LEMN  
THEOREM PROVER

Microsoft Research

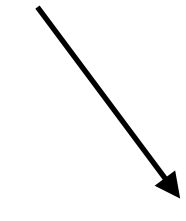


Formal

# Small data can have a large impact



Expert data





# Case study: MiniF2F [Zheng et al 2022]

General  
web data

>1.4 trillion tokens

Lean MiniF2F

↓  
~10k  
tokens

# Case study: MiniF2F

- 488 competition problems



- 4 proof assistants

## MINIF2F: A CROSS-SYSTEM BENCHMARK FOR FORMAL OLYMPIAD-LEVEL MATHEMATICS

**Kunhao Zheng**

École Polytechnique

kunhao.zheng@polytechnique.edu

**Jesse Michael Han**

OpenAI

University of Pittsburgh

jessemichaelhan@openai.com

**Stanislas Polu**

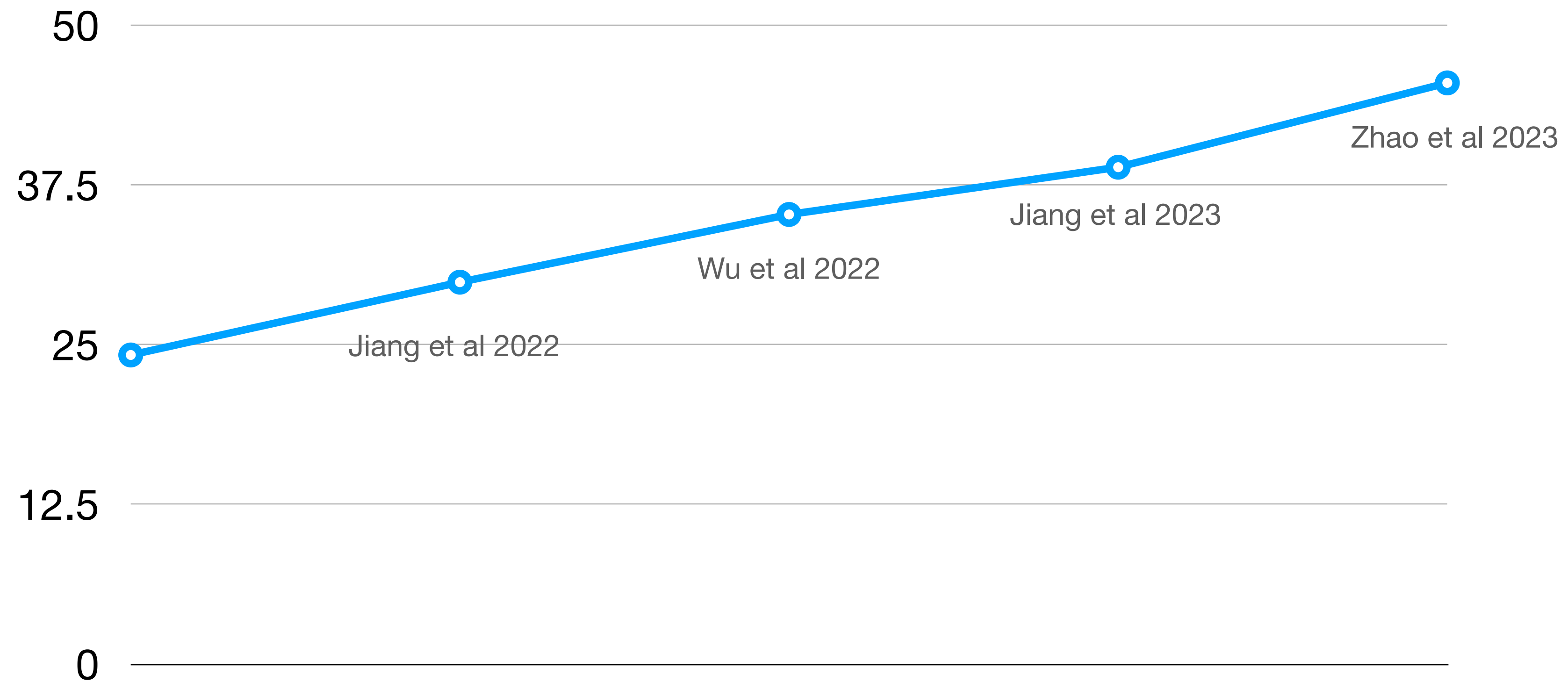
OpenAI

spolu@openai.com

ICLR 2022

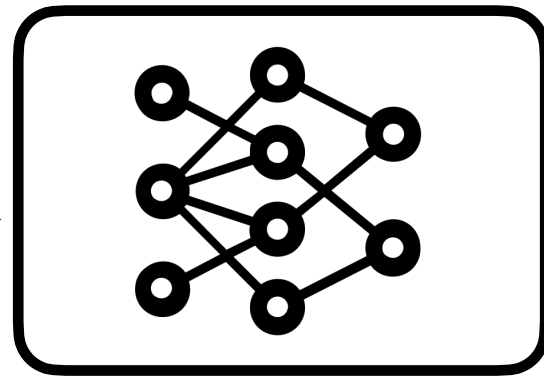
# Case study: MiniF2F

## Driver of progress in neural theorem proving



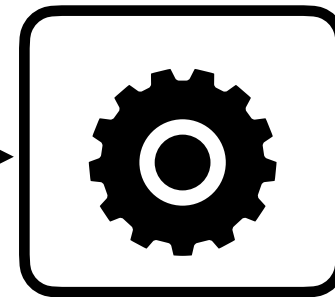
# Case study: Draft, Sketch, Prove

Informal  
Proof



```
lemma one_plus_one_equals_two:  
  shows "1 + 1 = 2"  
  
proof -  
  have "1 + 1 = Suc 1" by simp  
  also have "Suc 1 = 2" by simp  
  ..  
  finally show ?thesis by simp  
qed
```

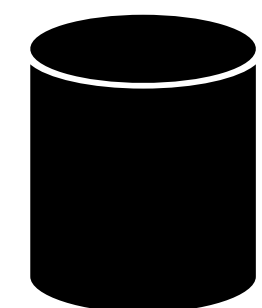
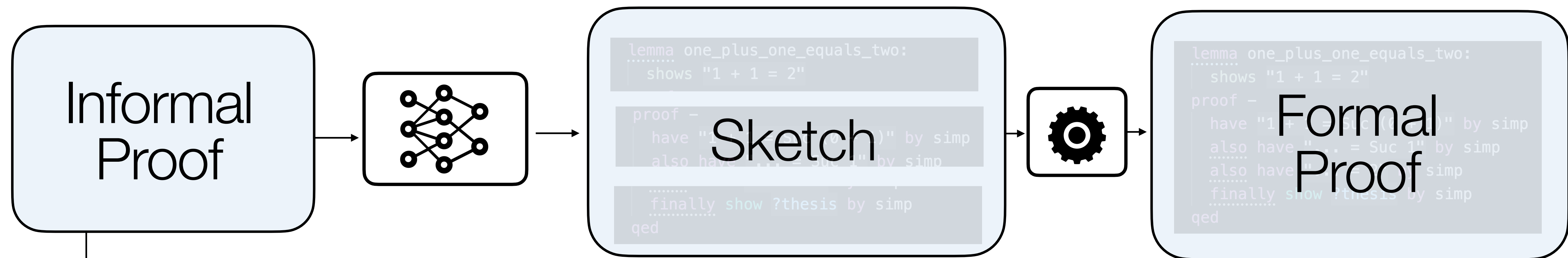
Sketch



```
lemma one_plus_one_equals_two:  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc 1" by simp  
  also have "Suc 1 = 2" by simp  
  also have "2 = 2" by simp  
  finally show ?thesis by simp  
qed
```

Formal  
Proof

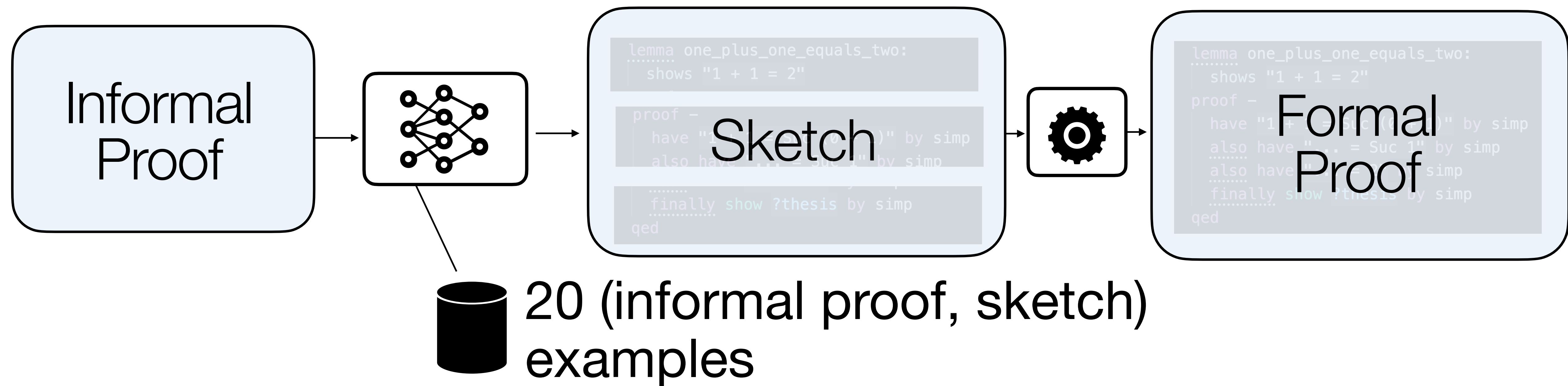
# Case study: Draft, Sketch, Prove



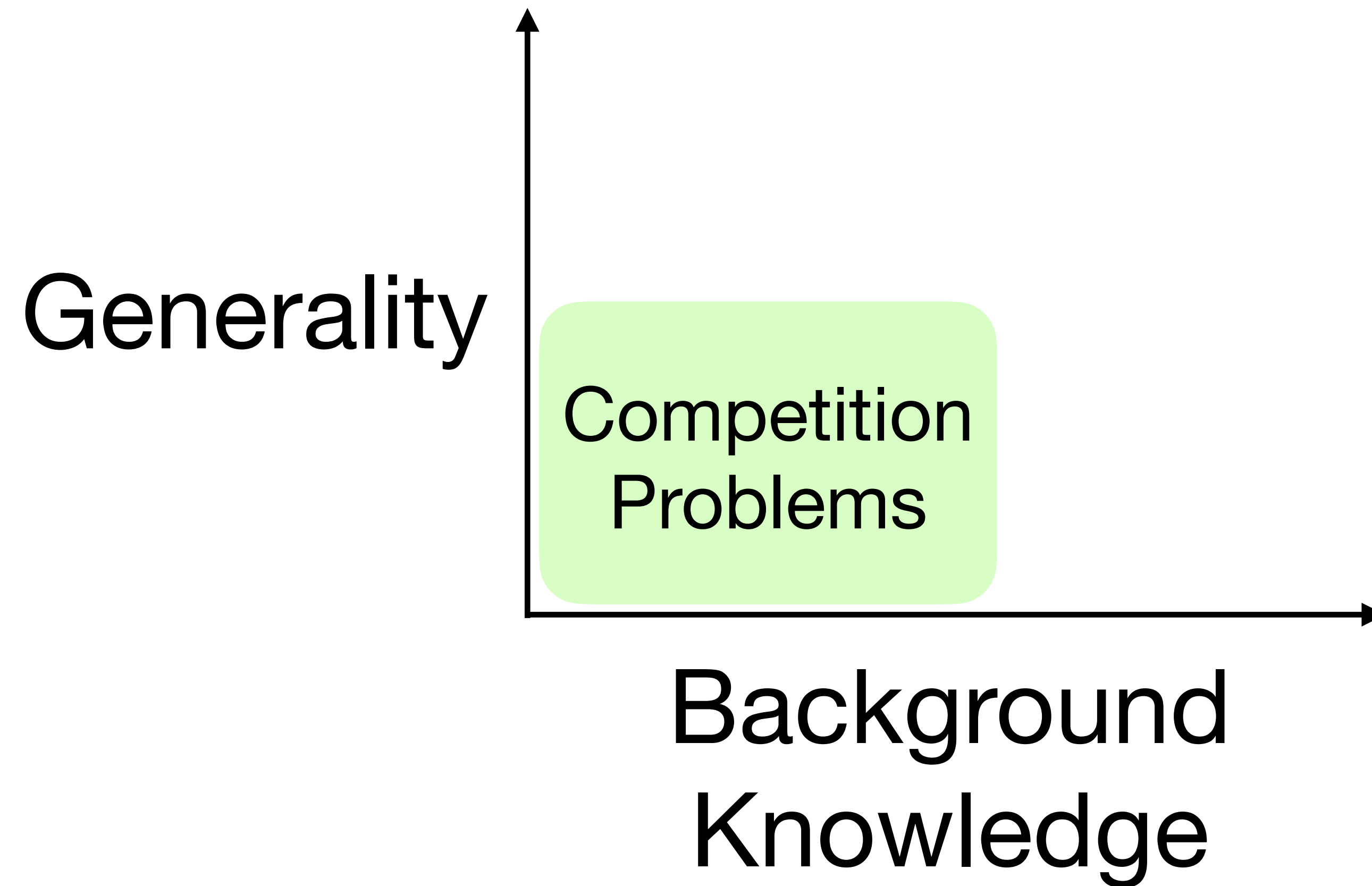
MiniF2F + 488 informal (Latex) proofs

<https://github.com/facebookresearch/miniF2F>

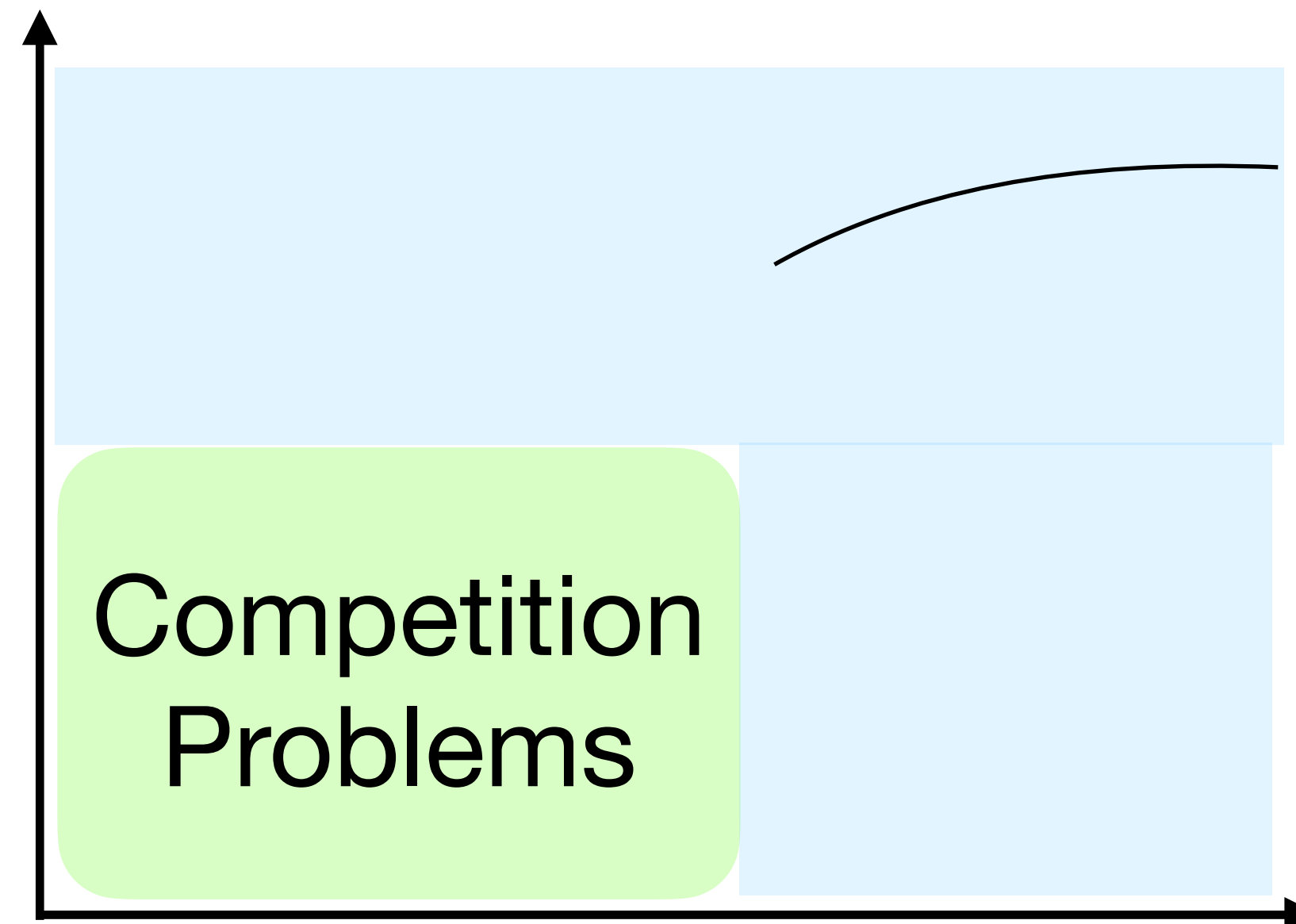
# Case study: Draft, Sketch, Prove



# Challenge: broadening scope



Generality



Background  
Knowledge

---

## ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics

---

**Zhangir Azerbayev**  
Yale College\*  
zhangir.azerbayev@yale.edu

**Bartosz Piotrowski**  
University of Warsaw\*  
bartoszpiotrowski@post.pl

**Hailey Schoelkopf**  
EleutherAI, Yale College  
hailey.schoelkopf@yale.edu

**Edward W. Ayers**  
Carnegie Mellon University  
contact@edayers.com

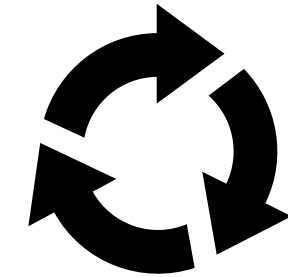
**Dragomir Radev**  
Yale University  
dragomir.radev@yale.edu

**Jeremy Avigad**  
Carnegie Mellon University  
avigad@cmu.edu



# New frontiers

Expand problems & domains



Interdisciplinary collaboration

Hard to  
evaluate

Informal

```
lemma one_plus_one_equals_two:  
  .....  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "Suc (0 + 1) = Suc 1" by simp  
  also have "Suc 1 = 2" by simp  
  finally have thesis by simp  
qed
```

Formal

Show that for any  
real number  $a$ ,  
 $10a \leq 28a^2 + 1$




```
lemma one_plus_one_equals_two:  
.....  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "Suc 1" by simp  
  also have "..... = 2" by simp  
  finally by simp  
.....  
qed
```

**Informal**

**Formal**

Show that for any  
real number  $a$ ,  
 $10a \leq 28a^2 + 1$

A screenshot of a Chat-GPT chat window. The user's message is "Show that for any real number a, 10a <= 28a^2 + 1". The AI's response is "This statement is false for some values of 'a'. For example, if 'a = -0.1', then: 'This statement is false' 10a. 28a^2 + 1 = 28(-0.1)^2 + 1 = 28 \* 0.01 + 1 = 0.29 And '-1 > 0.29', which contradicts the inequality." The AI's first sentence is highlighted with a red box. A dotted line connects the red box to the red text "This statement is false".

 This statement is false for some values of `a`.

For example, if `a = -0.1`, then:

**“This statement is false”**

$10a$ .

$28a^2 + 1 = 28(-0.1)^2 + 1 = 28 * 0.01 + 1 = 0.29$

And ` $-1 > 0.29$ `, which contradicts the inequality.

← Informal

Formal →

Show that for any  
real number  $a$ ,  
 $10a \leq 28a^2 + 1$



This statement is false for some values of  $a$ .

For example, if  $a = -0.1$ , then:

$$10a = -1$$

$$28a^2 + 1 = 28(-0.1)^2 + 1 = 28 * 0.01 + 1 = 0.29$$

“And  $-1 > 0.29$ ”

And  $-1 > 0.29$ , which contradicts the inequality.

Informal

Formal



expert  
evaluation

**Informal**

```
lemma one_plus_one_equals_two:  
  .....  
  shows "1 + 1 = 2"  
proof -  
  have "1 + 1 = Suc (0 + 1)" by simp  
  also have "Suc 1" by simp  
  also have "..... = 2" by simp  
  finally have thesis by simp  
  .....  
qed
```

Formal  
theorem  
proving

**Formal**

# Case study: NaturalProver

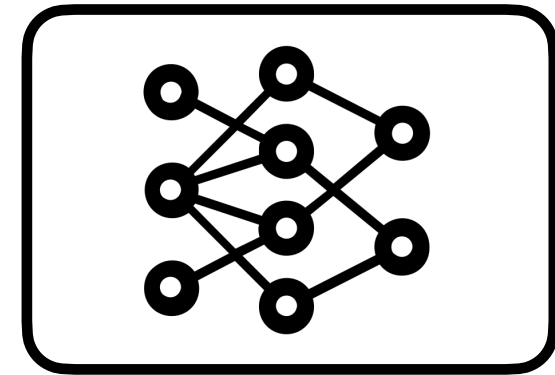


Dataset: 10,000 *informal* proofs from ProofWiki

# Case study: NaturalProver

Theorem

Proof-so-far

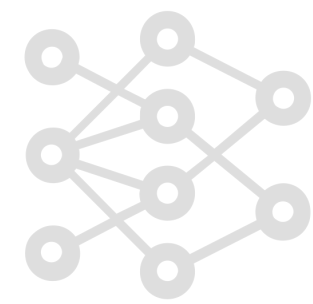


Next step



# Case study: NaturalProver

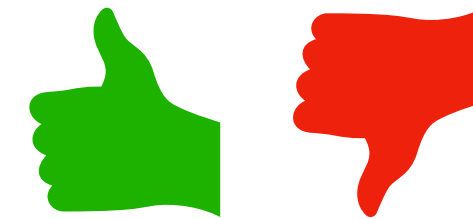
Theorem



Next step

Proof-so-far

Correct



# Case study: NaturalProver

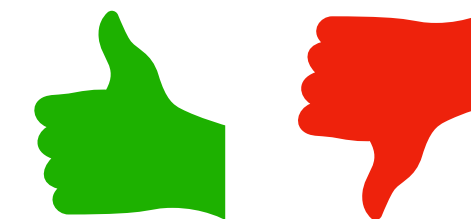
Theorem



Next step

Proof-so-far

Correct

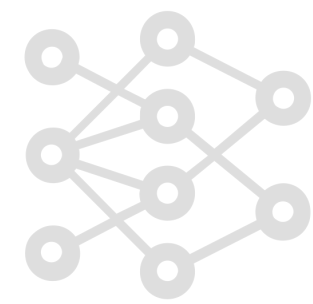


15 UW Math & Applied Math  
Undergrad/Masters/PhD students

IRB-approved study

# Case study: NaturalProver

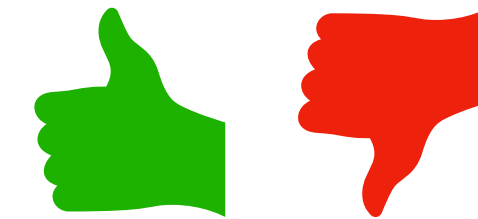
Theorem



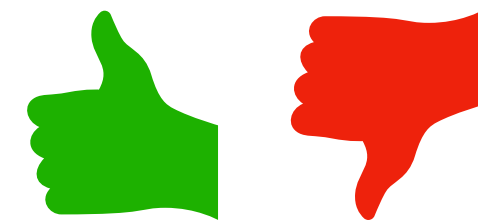
Next step

Proof-so-far

Correct



Useful



# Case study: NaturalProver

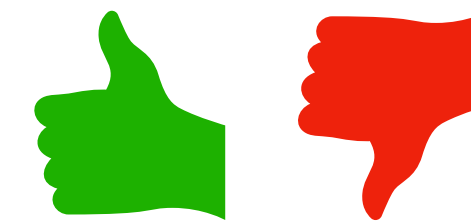
Theorem



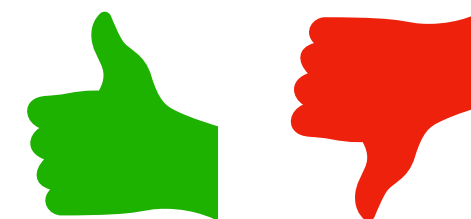
Next step

Proof-so-far

Correct

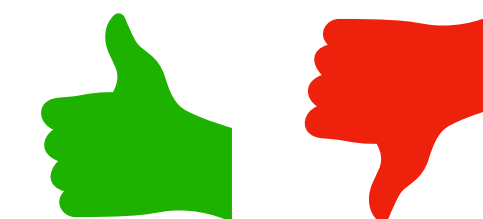


Useful

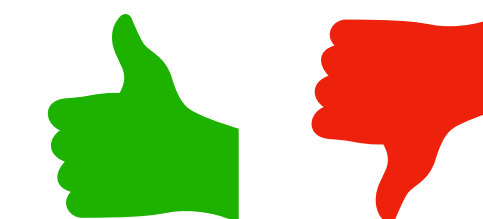


Errors —

Equation

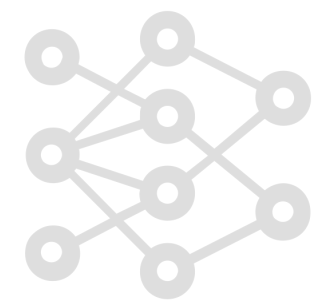


Definition



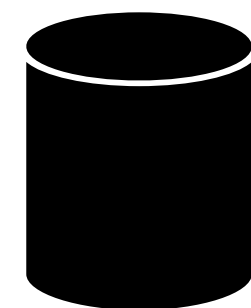
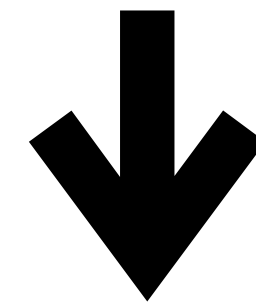
# Case study: NaturalProver

Theorem



Next step

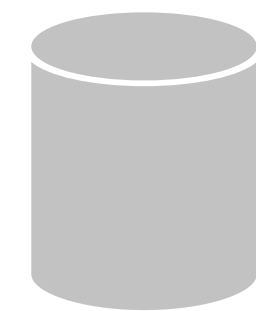
Proof-so-far



**8,400 annotated proof steps**

# Case study: NaturalProver

- *Capable* of correct and useful next-step suggestions
- **Derivations, long proofs...**



**8,400 annotated proof steps**

# Case study: NaturalProver

- ~51% useful, 43% correct next-step suggestions
- Equations, proof length...

Expert evaluation today (Arxiv, June 2023):

Evaluating Language Models for Mathematics through Interactions\*

Katherine M. Collins<sup>†1</sup>, Albert Q. Jiang<sup>†1</sup>,  
Simon Frieder<sup>2</sup>, Lionel Wong<sup>3</sup>, Miri Zilka<sup>1</sup>, Umang Bhatt<sup>1,4</sup>, Thomas Lukasiewicz<sup>5,2</sup>,  
Yuhuai Wu<sup>6</sup>, Joshua B. Tenenbaum<sup>3</sup>, William Hart<sup>1</sup>, Timothy Gowers<sup>7,1</sup>, Wenda Li<sup>1</sup>,  
Adrian Weller<sup>†1,4</sup>, and Mateja Jamnik<sup>†1</sup>

Hard to  
evaluate



Easier to  
evaluate

Informal

Formal





# Case study: Lila benchmark

## Unifies 20 math datasets

Grade School Math

Linear algebra

Competition problems

...

### Problem:

The pirates plan to explore 4 islands. Two islands require walking 20 miles per day while the other two islands require 25 miles per day. How many miles will they have to walk if it takes 1.5 days to explore each island?

Problem

### Program:

```
a=20
b=25
c=a+b
d=c*1.5
answer=d
print(answer)
# ==> 135.0
```



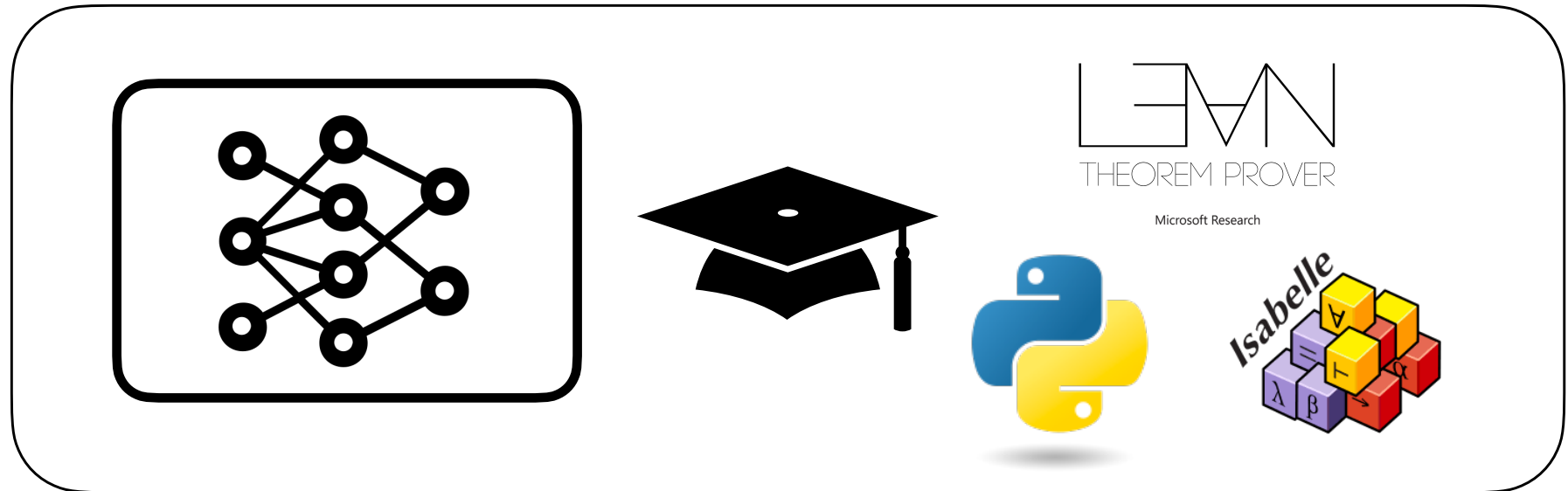
Python  
program

Answer

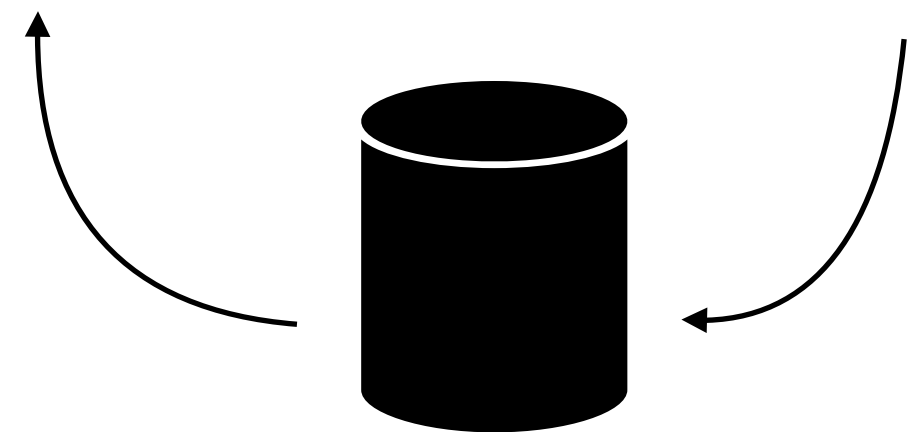


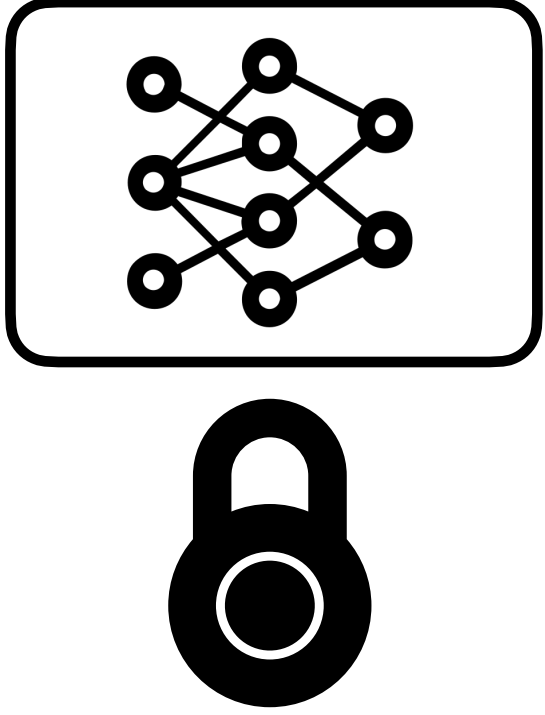
# New frontiers

Interaction



Improvement



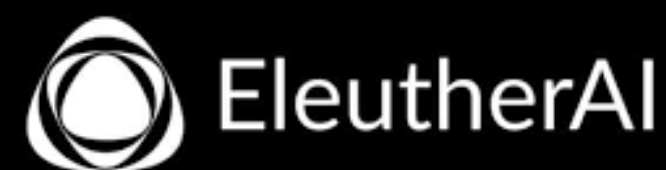
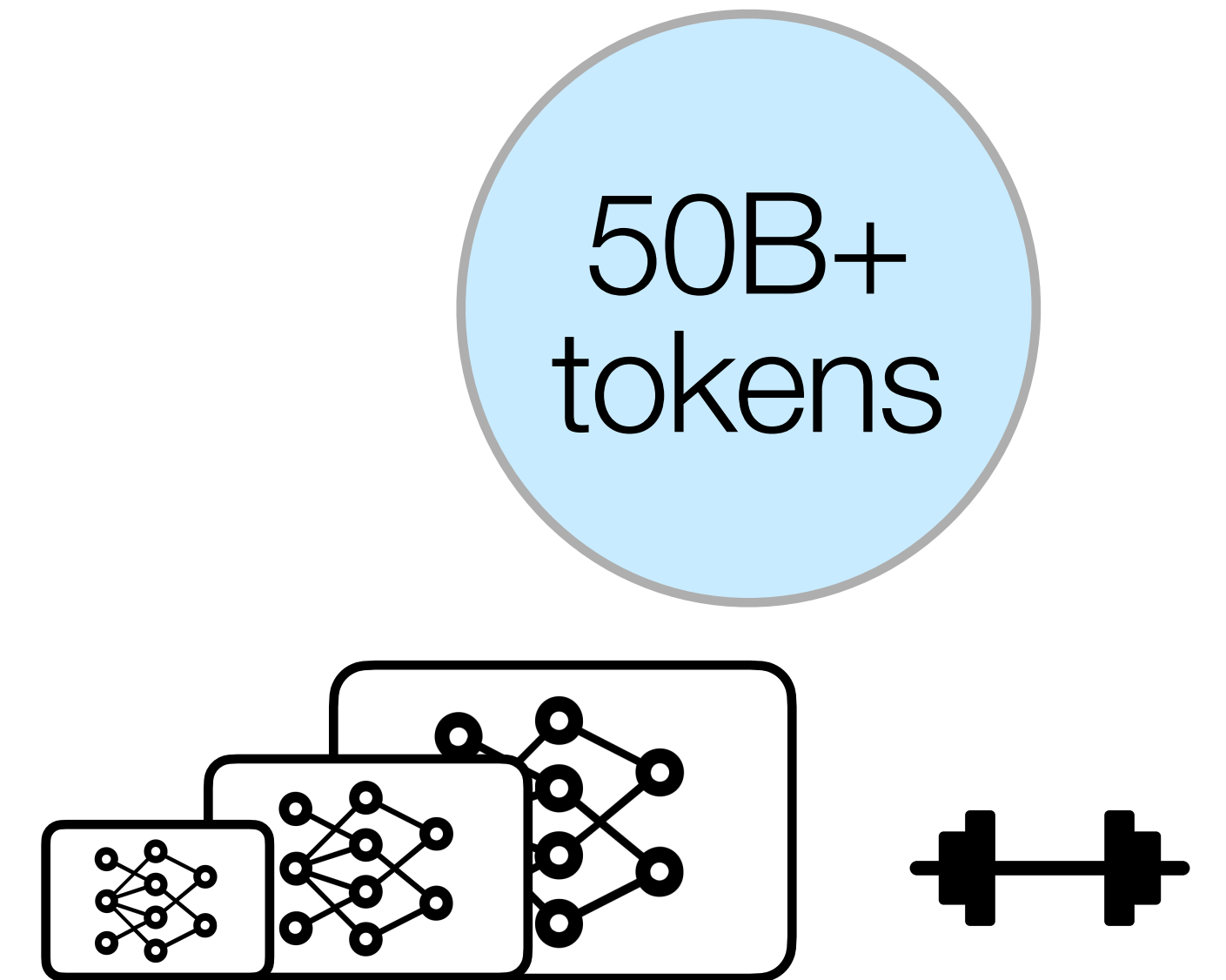


Informal

Formal

# Math-LM: open-source AI+mathematics

- 50B+ token technical corpus
- Model suite and evaluation harness
- Ongoing project:  
<https://github.com/EleutherAI/math-lm>



Led by Zhangir Azerbayev  
+ collaborators from many institutions!

# Summary

Small data,  
large impact

Expert evaluation  
& interaction

Open data

# Summary

Small data,  
large impact

Expert evaluation  
& interaction

Open data

- Incoming Assistant Professor, Jan. 2024

Carnegie Mellon University  
School of Computer Science  
Language Technologies Institute

The logo for Carnegie Mellon University, featuring the text "Carnegie Mellon University" in white serif font on a red square background.

Carnegie  
Mellon  
University