

Getting AI to Do Things I Can't: Scalable Oversight via Indirect Supervision



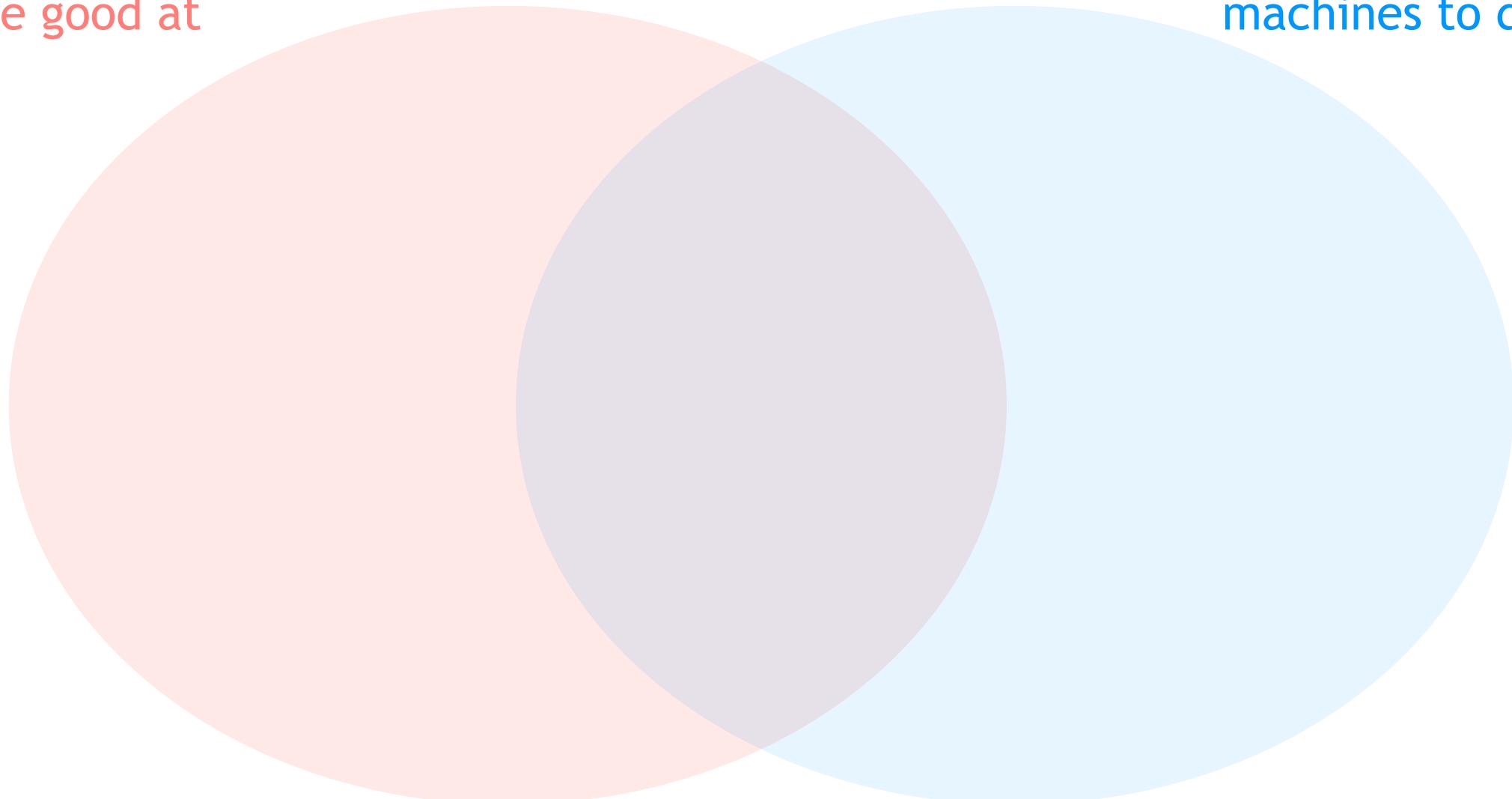
Ruiqi Zhong



AI Augmentation

What humans
are good at

What we want
machines to do





AI Augmentation

What humans
are good at

What we want
machines to do

Automation

Spam classification
Grammaticality check
Sentiment classification
Paraphrase detection
...



AI Augmentation

What humans
are good at

What we want
machines to do

Automation

Spam classification
Grammaticality check
Sentiment classification
Paraphrase detection
...

Augmentation

Explain patterns in data
Generate complex program
Predict the future
Design/run experiments
.....



How to supervise AI when we
struggle to verify that it is correct?



Example Tasks

Patient feedback for drug A

- XXXXXX
- XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- X
- XXXXX

Patient feedback for drug B

- X XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX
XXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXXX XXX XXXXXXXX XXXXXX XXX XXXXXXXX
XXXXX XXX



Patients more often “have strong feelings of paranoid” after taking drug A



Example Tasks

Patient feedback for drug A

- XXXXXX
- XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- X
- XXXXX

Patient feedback for drug B

- X XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXX

Patients more often “have strong feelings of paranoid” after taking drug A



Find the first name of students who have both cat and dog pets.



```
SELECT fname FROM Student WHERE  
StuID IN (SELECT T1.stuid FROM  
student AS T1 JOIN has_pet .....
```



Example Tasks

Patient feedback for drug A

- XXXXXX
- XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- X
- XXXXX

Patient feedback for drug B

- X XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX XXXXXXXX
XXXXXXX
- XXXXXXXX XXXXXXXX
- XXXXXXXX XXXXXX XXX
- XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX
- XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX XXXXXXXX
XXXXXX XXXX XXXXXXXX XXXXX XXX XXXXXXXX XXXXX XXX

*Find the first name of students
who have both cat and dog pets.*

**Even experts are far
from perfect!**

Patients more often “have strong feelings
of paranoid” after taking drug A

```
SELECT fname FROM Student WHERE  
StuID IN (SELECT T1.stuid FROM  
student AS T1 JOIN has_pet .....
```



Recipe: Propose & Reduce

Write the
full output



Directly verify which
output is correct



Indirectly verify the
output on easier problems



Recipe: Propose & Reduce

Write the full output



Directly verify which output is correct



Indirectly verify the output on easier problems

Write the full dialogue response

(Gao et al., 2020)



Pick the best dialogue response

(Wiegreffe et al., 2021)

Write the full explanation

(Stiennon et al., 2020)



Pick the best explanation

Write the full summary



Pick the best summary

Propose



Recipe: Propose & Reduce

Write the
full output



Directly verify which
output is correct

Pick corpus-level
descriptions

Pick source
code

Pick book
summaries

Reduce

(Zhong et al., 2022)

(Zhong et al., 2022)

(Wu et al. 2021)

Indirectly verify the
output on easier problems

Examine
samples

Run test
cases

Summarize
Recursively



Outline

- ▶ Overview of “propose and reduce”
- ▶ **Task I:** Discovering and describing distributional differences
- ▶ **Task II:** Supervising code generation models with non-programmers
- ▶ Future work

Task I: Discovering and Describing Distributional Differences

Zhong et al. 2023, Goal Driven Discovery of Distributional Differences via Language Descriptions

Zhong et al. 2022, Describing Differences between Text Distributions with Natural Language

Zhong et al. 2021, Adapting Language Model for Zero-Shot Learning by Meta-tuning on Dataset and Prompt Collections



Task Formulation

Input

Research Context

“I’m an educator and want to understand how students write essays differently across schools.”

- In 2021, renewable energy sources accounted for over 20% of total electricity
- A 2019 survey found that only 12% of ...
- Global market for AI is expected to reach \$267 billion by 2027

Corpus A (student essays from school A)

- The history of the American Civil War is a complex and nuanced ...
- In order to truly understand the ...
- The field of psychology offers a fascinating look into the workings of the human ...

Corpus B (students essays from school B)



Task Formulation

Input

Research Context

“I’m an educator and want to understand how students write essays differently across schools.”

- In 2021, renewable energy sources accounted for over 20% of total electricity
- A 2019 survey found that only 12% of ...
- Global market for AI is expected to reach \$267 billion by 2027

Corpus A (student essays from school A)

- The history of the American Civil War is a complex and nuanced ...
- In order to truly understand the ...
- The field of psychology offers a fascinating look into the workings of the human ...

Corpus B (students essays from school B)

Output

Students essays from school A
“use numbers and statistics” more often.



Why do we care?

- ▶ User queries contain more slang than the training data
- ▶ Obama's speeches are more supportive of immigrants than Trump's speeches
- ▶ Texts perceived to be human-written contain more personal experiences than those perceived to be machine-generated



Why do we care?

- ▶ User queries contain more slang than the training data
- ▶ Obama's speeches are more supportive of immigrants than Trump's speeches
- ▶ Texts perceived to be human-written contain more personal experiences than those perceived to be machine-generated
- ▶ What kind of inputs cause a machine learning system to make mistakes?
- ▶ How do ChatGPT and Claude respond differently?
- ▶ How do customers respond for different companies?
- ▶ What do different groups of people talk about on different platforms and years?



This is Difficult

~17K samples

- Pieck rescued Gabi from the dungeon and transformed into a Titan afterwards.
- All four of my maternal and fraternal grandparents are professors, and that's why I'm determined to become a prof as well.
- My mom took me to the hospital, and the nurse said that she has never seen this symptom before.
- I was really fortunate to be advised Prof. McKeown and Prof. Hirschberg at Columbia on NLP research, and Prof. Andoni on Theoretical computer science.
- Historia was born as the illegitimate and unrecognized daughter of Rod Reiss. Her mother, Alma, was a servant in his household.
- I called her to explain what happened to her aunt.

~17K samples

- She carried a total of eight torpedoes. Her deck was reinforced to enable her to lay a minefield.
- My mom and I were best friends and we used to hunt together.
- Lucy and Peter co-authored a paper on machine learning but got a really bad review.
- Adding to Historia's isolation, the other children outside the estate would throw rocks at her, and she was not allowed to leave.
- Bentham defined as the "fundamental axiom" of his philosophy the principle that "it is the greatest happiness of the greatest number that is the measure of right and wrong."
- Large language models advanced the state of the art by quite a lot but there are still rooms for improvements.

Orange text “mentions more female entities”



Hard to Label

Research Context

"I want to understand students write essays differently across schools ...
Distribution A are student essays from School A ..."

D_A

- In 2021, renewable energy sources accounted for over 20% of total electricity
- A 2019 survey found that only 12% of U.S. adults ...

D_B

- In order to truly understand the principles of calculus, it is ...
- The field of psychology offers a fascinating look into the workings of the human ...



**Hard to directly label:
~17K samples each corpus**

More samples from $s = \text{"use numbers and statistics"}$ more often.



Correlation \neq Causation





Propose

D_A samples →

Group A: In 2021, renewable energy sources accounted for over 20% of total electricity
Group A: ...

D_B samples →

Group B: In order to truly understand the principles of calculus, it is ...
Group B: ...

Research context →

I want to understand how students write essays differently ...

Compared to inputs from group B, each input from group A

Propose with LLM × 60



- is offensive in tone
- uses more statistics
-



Propose

D_A samples →

Group A: In 2021, renewable energy sources accounted for over 20% of total electricity
Group A: ...

D_B samples →

Group B: In order to truly understand the principles of calculus, it is ...
Group B: ...

Research context →

I want to understand how students write essays differently ...

Compared to inputs from group B, each input from group A

Propose with LLM × 60



- is offensive in tone
- uses more statistics
-



TODO: humans verify which candidate best discriminates the two corpora.



Reduce

Difficult to verify a description on
the **entire corpora** of D_A and D_B .

Reduce



Easier to verify that description is valid
on **individual samples** from D_A and D_B .



Verify on Individual Samples

s = Samples from D_A “uses numbers and statistics”.



Verify on Individual Samples

s = Samples from D_A “uses numbers and statistics”.

$x_i \sim D_i$ “renewable energy sources accounted
for over 20% of total electricity”

$x_j \sim D_j$ “In order to truly understand the
principles of calculus, it is ...”

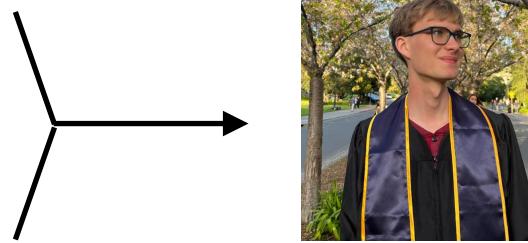


Verify on Individual Samples

s = Samples from D_A “uses numbers and statistics”.

$x_i \sim D_i$ “renewable energy sources accounted for over 20% of total electricity”

$x_j \sim D_j$ “In order to truly understand the principles of calculus, it is ...”



Human
Classifies

I think x_i is from D_A
and x_j is from D_B

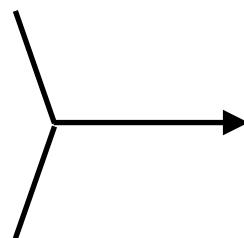


Verify on Individual Samples

s = Samples from D_A “uses numbers and statistics”.

$x_i \sim D_i$ “renewable energy sources accounted for over 20% of total electricity”

$x_j \sim D_j$ “In order to truly understand the principles of calculus, it is ...”



Human
Classifies

I think x_i is from D_A
and x_j is from D_B

Loss(\underline{s}): how accurately humans can use \underline{s} to guess the membership of x (D_A or D_B)

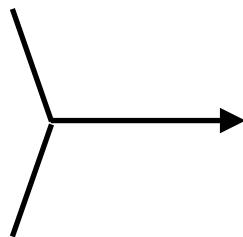


Verify on Individual Samples

s = Samples from D_A “uses numbers and statistics”.

$x_i \sim D_i$ “renewable energy sources accounted for over 20% of total electricity”

$x_j \sim D_j$ “In order to truly understand the principles of calculus, it is ...”



Human
Classifies

I think x_i is from D_A
and x_j is from D_B

Loss(\underline{s}): how accurately humans can use \underline{s} to guess the membership of x (D_A or D_B)

Repeat 100 times, ~\$10(!) to evaluate each description



Make Verification Efficient

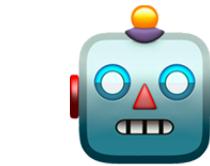
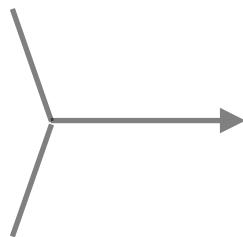
s = Samples from D_A “uses numbers and statistics”.

$$x_i \sim D_i$$

“renewable energy sources accounted
for over 20% of total electricity”

$$x_j \sim D_j$$

“In order to truly understand the
principles of calculus, it is ...”



Machine
Classifies

I think x_i is from D_A
and x_j is from D_B

Much cheaper to calculate the error rate of a model



Make Verification Efficient

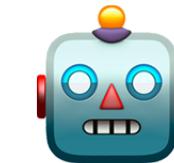
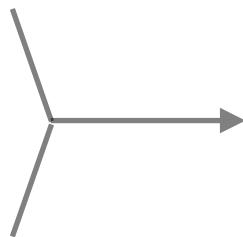
s = Samples from D_A “uses numbers and statistics”.

$$x_i \sim D_i$$

“renewable energy sources accounted
for over 20% of total electricity”

$$x_j \sim D_j$$

“In order to truly understand the
principles of calculus, it is ...”



Machine
Classifies

I think x_i is from D_A
and x_j is from D_B

Loss'(\underline{s}): how accurately a machine can use \underline{s} to guess the membership of x (D_A or D_B)

Much cheaper to calculate the error rate of a model



Recipe: Propose & Reduce



Recipe: Propose & Reduce

- ▶ Method:
 - ▶ Propose descriptions with GPT-3
 - ▶ Reduce verification to examining individual samples
 - ▶ Make verification more efficient by simulating humans with machines



Recipe: Propose & Reduce

- ▶ Method:
 - ▶ Propose descriptions with GPT-3
 - ▶ Reduce verification to examining individual samples
 - ▶ Make verification more efficient by simulating humans with machines
- ▶ “Victory condition”: after reduced verification > propose w/o verification



Benchmark

Recover the descriptions for 54 binary text classification tasks



Benchmark

Recover the descriptions for 54 binary text classification tasks

D_A

Samples with
label = 1

D_B

Samples with
label = 0

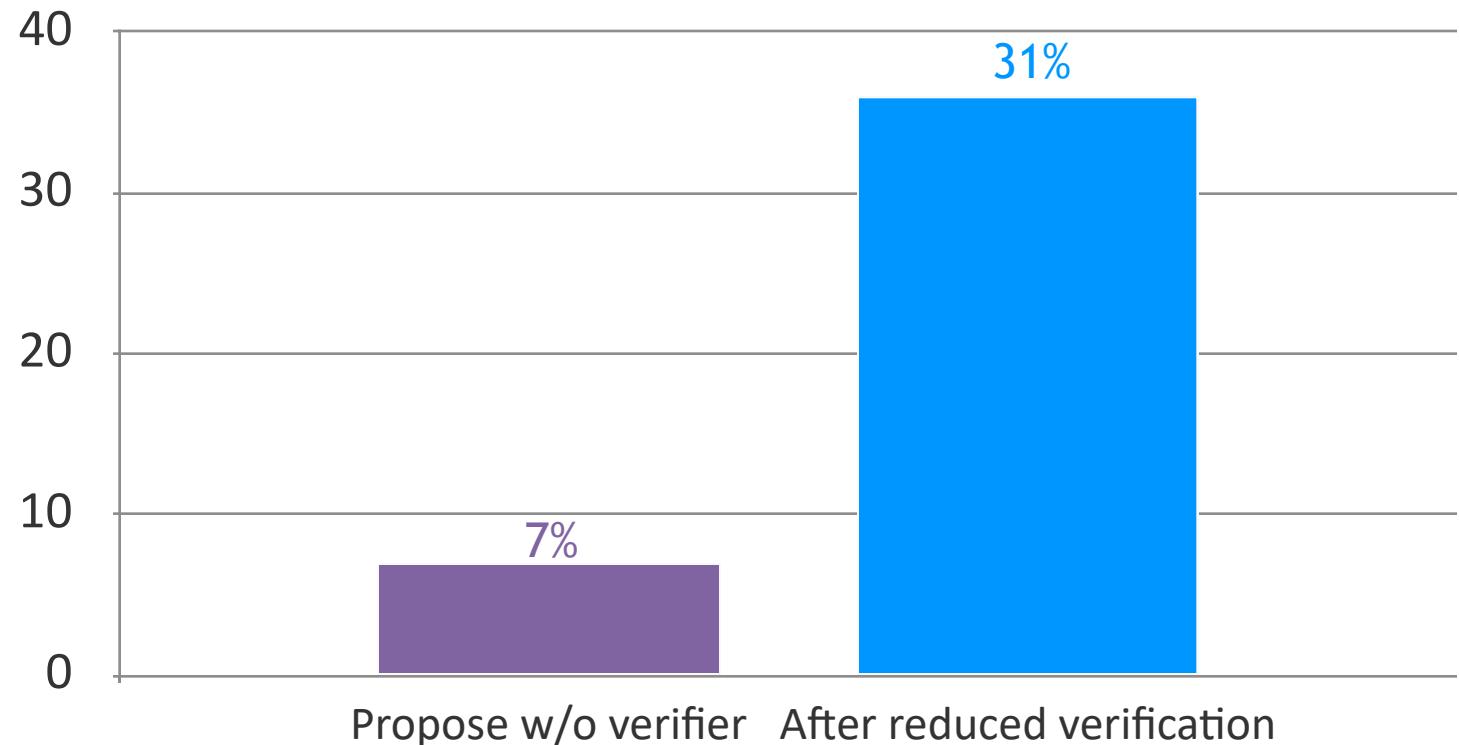
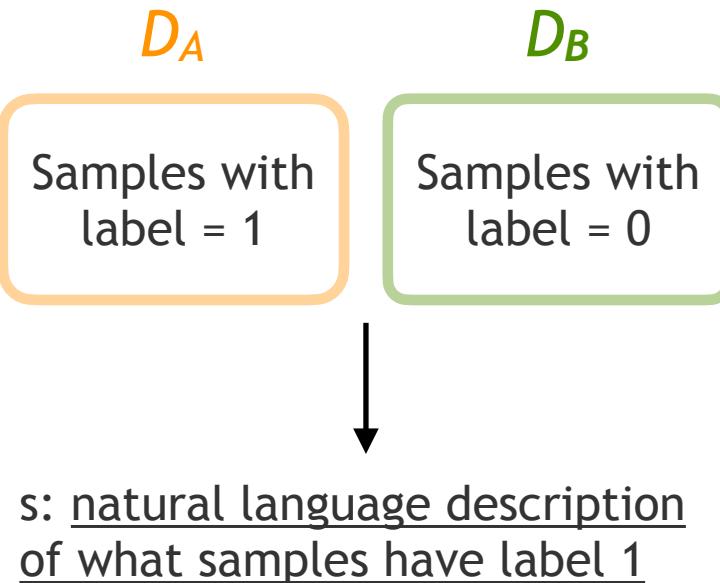


s: natural language description
of what samples have label 1



Benchmark

Recover the descriptions for 54 binary text classification tasks





Dataset Misunderstanding

Describe the label in SUBJ (Subjectivity analysis dataset)
 ≥ 4 papers used this dataset



Dataset Misunderstanding

Describe the label in SUBJ (Subjectivity analysis dataset)
 ≥ 4 papers used this dataset

D_A

“Objective”
texts

D_B

“Subjective”
texts

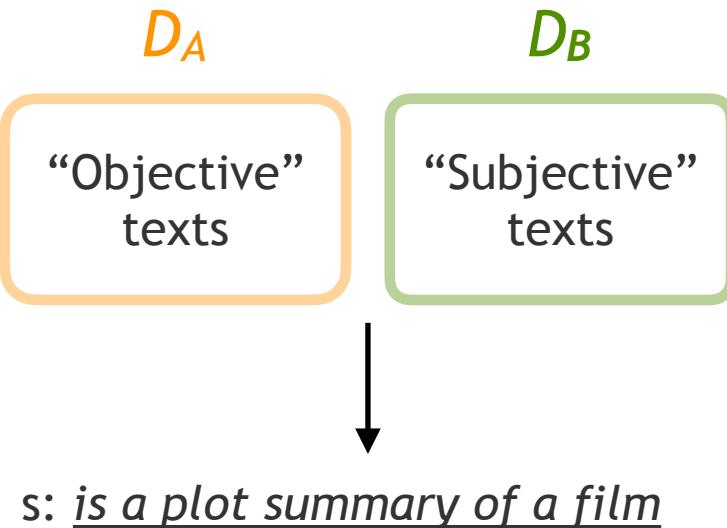


s: *is a plot summary of a film*



Dataset Misunderstanding

Describe the label in SUBJ (Subjectivity analysis dataset)
 ≥ 4 papers used this dataset



To gather subjective sentences, we collected 5000 movie review snippets from www.rottentomatoes.com. To obtain (mostly) objective data, we took 5,000 sentences from plot summaries available from www.imdb.com



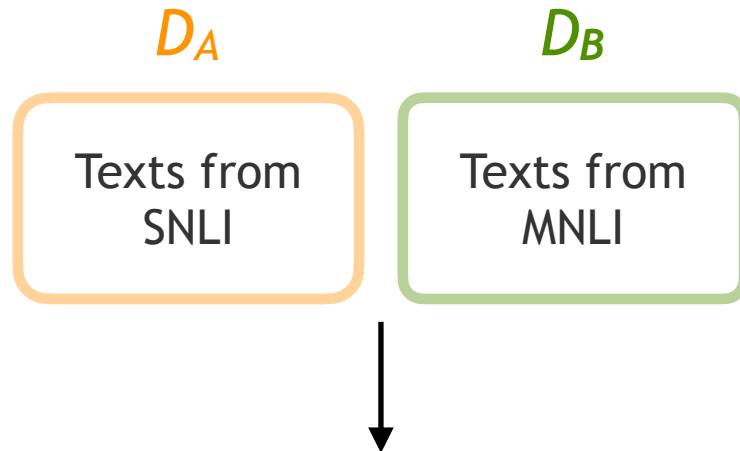
Distribution Shift

Describing differences between natural language inference datasets



Distribution Shift

Describing differences between natural language inference datasets

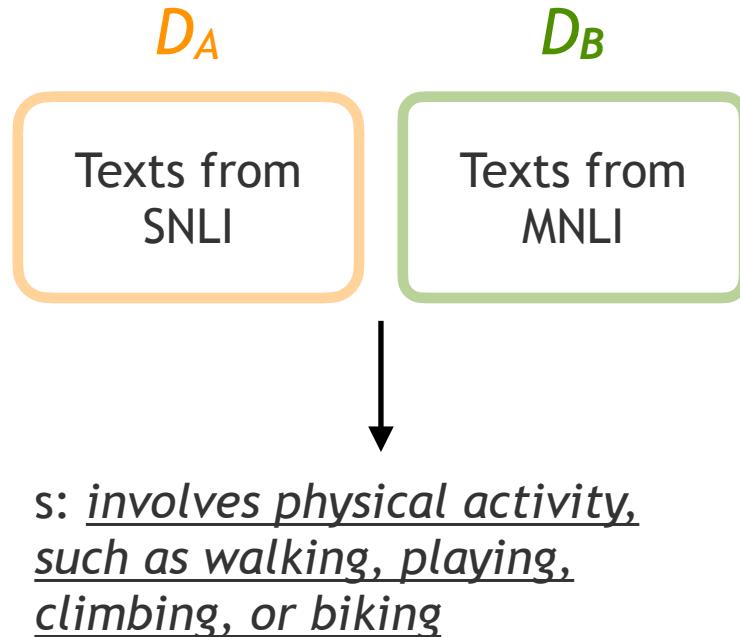


s: *involves physical activity, such as walking, playing, climbing, or biking*



Distribution Shift

Describing differences between natural language inference datasets



SNLI contains sentence-based picture portrayal, while MNLI has multiple genres (e.g. speech/ fiction)



Model Error

What inputs cause a model to fail? Task: zero-shot
classify whether an argument “supports” a claim



Model Error

What inputs cause a model to fail? Task: zero-shot classify whether an argument “supports” a claim

D_A

Model A
misclassifies
as “support”

D_B

Model B
misclassifies
as “support”



s: uses language that are
positive or uplifting



Model Error

What inputs cause a model to fail? Task: zero-shot classify whether an argument “supports” a claim

D_A

Model A
misclassifies
as “support”

D_B

Model B
misclassifies
as “support”



s: uses language that are
positive or uplifting

Model A might have misunderstood “support” in
the instruction as a sentiment label



More Applications

Machine Learning

Expose dataset
misunderstanding

Describe
distribution shift

Analyze model
errors



More Applications

Machine Learning

Expose dataset
misunderstanding

Describe
distribution shift

Analyze model
errors

Debug dataset
shortcuts

Label text
clusters

Compare language
model's generation

Explain neurons

Describe principle
components

Explain human
preferences



More Applications

Machine Learning

Expose dataset
misunderstanding

Debug dataset
shortcuts

Explain neurons

Describe
distribution shift

Label text
clusters

Describe principle
components

Analyze model
errors

Compare language
model's generation

Explain human
preferences

Many
more!!!

Social Sciences

Expose gender association biases in
speeches

Discover topical differences in
headlines between years

Discover topical differences
between demographic groups

Identify political stances of
different leaders

Understand how political ads
target different subpopulations

Understand what job skills are
more preferred for different
sectors

Health

Compare symptoms of
different diseases

Compare side effects
of drugs

Humanities

Compare lyrics of
different era

Compare writing style
of different authors

Business

Compare reviews
for different
companies

Compare
advertisements ...



Task I Takeaways



Task I Takeaways

- ▶ Setup: describe distributional differences



Task I Takeaways

- ▶ Setup: describe distributional differences
- ▶ Method:
 - ▶ Propose descriptions with GPT-3
 - ▶ Reduce verification to examining individual samples
 - ▶ Make verification more efficient with machine simulation



Task I Takeaways

- ▶ Setup: describe distributional differences
- ▶ Method:
 - ▶ Propose descriptions with GPT-3
 - ▶ Reduce verification to examining individual samples
 - ▶ Make verification more efficient with machine simulation
- ▶ Discover insights even human experts did not notice



Outline

- ▶ Overview of “propose and reduce”
- ▶ Task I: Discovering and describing distributional differences
- ▶ **Task II: Supervising code generation models with non-programmers**
- ▶ Future work

Task II: Supervising Code Generation Models with Non-Programmers

Zhong et al. 2022, Active Programming by Example with a Natural Language Prior

Zhong et al. 2020, Semantic Evaluation for Text-to-SQL with Distilled Test Suite



Semantic Parsing

Natural Language

How old is the youngest person from department A?



SQL Program

SELECT MIN(Age) from People
WHERE Department = 'A'



Semantic Parsing

Natural Language

How old is the youngest person from department A?



SQL Program

```
SELECT MIN(Age) from People  
WHERE Department = 'A'
```

Expensive!!

How can non-programmers supervise models to write SQL?



Propose

Natural Language

How old is the youngest person from department A?

Propose with LLM x 32



Probabilities

SQL

Candidates

7/10

SELECT MAX(Name) from People

1/10

SELECT MAX(Age) from People

.....

1/80

SELECT MIN(Age) from People
WHERE Department = 'A'



Propose

Natural Language

How old is the youngest person from department A?

Propose with LLM x 32



Probabilities

Candidates

7/10

SELECT MAX(Name) from People

✗

1/10

SELECT MAX(Age) from People

✗

SQL

.....

1/80

SELECT MIN(Age) from People
WHERE Department = 'A'

✓

TODO: How do non-experts verify which candidate is correct?



Hard to Verify

Find the first name of students who have both cat and dog pets.

Candidate 1

```
SELECT fname FROM Student WHERE StuID IN
    (SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
     JOIN pets AS T3 ON T3.petid = T2.petid
     WHERE T3.pettype = 'cat' INTERSECT
     SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
     JOIN pets AS T3 ON T3.petid = T2.petid WHERE T3.pettype = 'dog')
```

Candidate 2

```
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'cat' INTERSECT
    SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
    JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'dog'
```



Reduce

Difficult to directly verify
that a program is correct.

Reduce



Easier to verify that a program has the
right behavior on example test cases.



Verify on Input-Output Examples

How old is the youngest person from department A?

SELECT MAX(Name) from People

SELECT MAX(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'



Verify on Input-Output Examples

How old is the youngest person from department A?

SELECT MAX(Name) from People

SELECT MAX(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B



Verify on Input-Output Examples

How old is the youngest person from department A?

SELECT MAX(Name) from People

SELECT MAX(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'

Non-expert's Answer

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

23

Cathy

28

Program's Output

23



Verify on Input-Output Examples

How old is the youngest person from department A?

SELECT MAX(Name) from People

SELECT MAX(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'

Non-expert's Answer

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

23

Cathy ✗

28 ✗

23 ✓

Program's Output



Verify on Input-Output Examples

How old is the youngest person from department A?

SELECT MAX(Name) from People

SELECT MAX(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'

Non-expert's Answer

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

23

Cathy ✗

28 ✗

23 ✓

Program's Output



Where does this database come from?

How old is the youngest person from department A?

~~SELECT MAX(Name) from People~~

~~SELECT MAX(Age) from People~~

~~SELECT MIN(Age) from People
WHERE Department = 'A'~~

Non-expert's Answer

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

23

Cathy ✗

28 ✗

23 ✓

Program's Output



Make Verification Efficient

Maximize the bits of supervision with minimal human efforts.

Size (

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

) is small

InfoGain (

NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

) is large



Make Verification Efficient

How old is the youngest person from department A?

The database input
must be simple to
comprehend

NAME	Age	Department
Collin	26	A
Bob	23	A
Cathy	28	B
David	19	A
Eric	11	A
Jacob	12	A
Alice	34	A
Dan	98	A
Alice	12	C
Kevin	38	B
Kevin	20	A

→ Annotators'
Answer

?????

[In total 1000 rows, rest omitted]



Make Verification Efficient

Size (

NAME	Age	Department
Collin	26	A
Bob	23	A
Cathy	28	B
David	19	A
Eric	11	A
Jacob	12	A
Alice	34	A
Dan	98	A
Alice	12	C
Kevin	38	B
Kevin	20	A

) = 1000

[In total 1000 rows, rest omitted]



Make Verification Efficient

How old is the youngest person from department A?

NAME	Age	Department
Collin	26	A
Bob	23	A

SELECT MIN(Age) from People

SELECT MIN(Age) from People
WHERE Department = 'A'



Make Verification Efficient

How old is the youngest person from department A?

NAME	Age	Department
Collin	26	A
Bob	23	A

→ 23 ✓
Annotators' Answer

Not Informative!

SELECT MIN(Age) from People

→ 23 ✓

SELECT MIN(Age) from People
WHERE Department = 'A'

→ 23 ✓



Expected Information Gain

Probabilities

SQL

1/3 SELECT MIN(Age) from People → **23**

1/3 SELECT MIN(Age) from People
 WHERE Department = 'A' → **23**

1/3 SELECT MAX(Age) from People
 WHERE Department = 'A' → **26**

NAME	Age	Department
Collin	26	A
Bob	23	A



Expected Information Gain

Probabilities

1/3

SELECT MIN(Age) from People

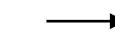
1/3

SELECT MIN(Age) from People
WHERE Department = 'A'

1/3

SELECT MAX(Age) from People
WHERE Department = 'A'

SQL



Answer Entropy: $-2/3 * \log(2/3) - 1/3 * \log(1/3) = 0.92$

NAME	Age	Department
Collin	26	A
Bob	23	A



Expected Information Gain

Probabilities

SQL

1/3

SELECT MIN(Age) from People



23

1/3

SELECT MIN(Age) from People
WHERE Department = 'A'



23

1/3

SELECT MAX(Age) from People
WHERE Department = 'A'



26

Answer Entropy: $-2/3 * \log(2/3) - 1/3 * \log(1/3) = 0.92$

InfoGain (

NAME	Age	Department
collin	26	A
Bob	23	A

) = 0.92 bit



Technical Details in Our Paper

- ▶ Optimize a database's InfoGain under size constraint
- ▶ Fuzzing to generate large databases with large InfoGain
 - ▶ dropping rows greedily to decrease size
- ▶ Multi-round interaction



Recipe: Propose & Reduce



Recipe: Propose & Reduce

- ▶ Method:
 - ▶ Propose SQL programs with Codex
 - ▶ Reduce verification to examine answers on databases
 - ▶ Make verification more efficient by making databases small and informative



Recipe: Propose & Reduce

- ▶ Method:
 - ▶ Propose SQL programs with Codex
 - ▶ Reduce verification to examine answers on databases
 - ▶ Make verification more efficient by making databases small and informative
- ▶ “Victory condition”: after reduced verification > propose w/o verification



Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Propose
with Codex



Probabilities

7/10 SELECT MAX(Name) from People Codex top-1

1/10 SELECT MAX(Age) from People

.....

1/80 SELECT MIN(Age) from People
 WHERE Department = 'A'



Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Propose
with Codex



NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

Probabilities

7/10 SELECT MAX(N ~~Age~~ from People Codex top-1

1/10 SELECT MAX(Age) from People

.....

1/80 SELECT MIN(Age) from People
WHERE Department = 'A' Non-expert annotation
(non-CS, 0 SQL experience)



Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Propose
with Codex



NAME	Age	Department
Alice	26	A
Bob	23	A
Cathy	28	B

Prior expert annotations

Gold standard:

- (1) our authors using our system +
- (2) checking the SQL directly +
- (3) comparing with previous annotations +
- (4) discussing with previous annotators

Probabilities

7/10

SELECT MAX(Name) from People

Codex top-1

1/10

SELECT MAX(Age) from People

.....

1/80

SELECT MIN(Age) from People
WHERE Department = 'A'

Non-expert annotation

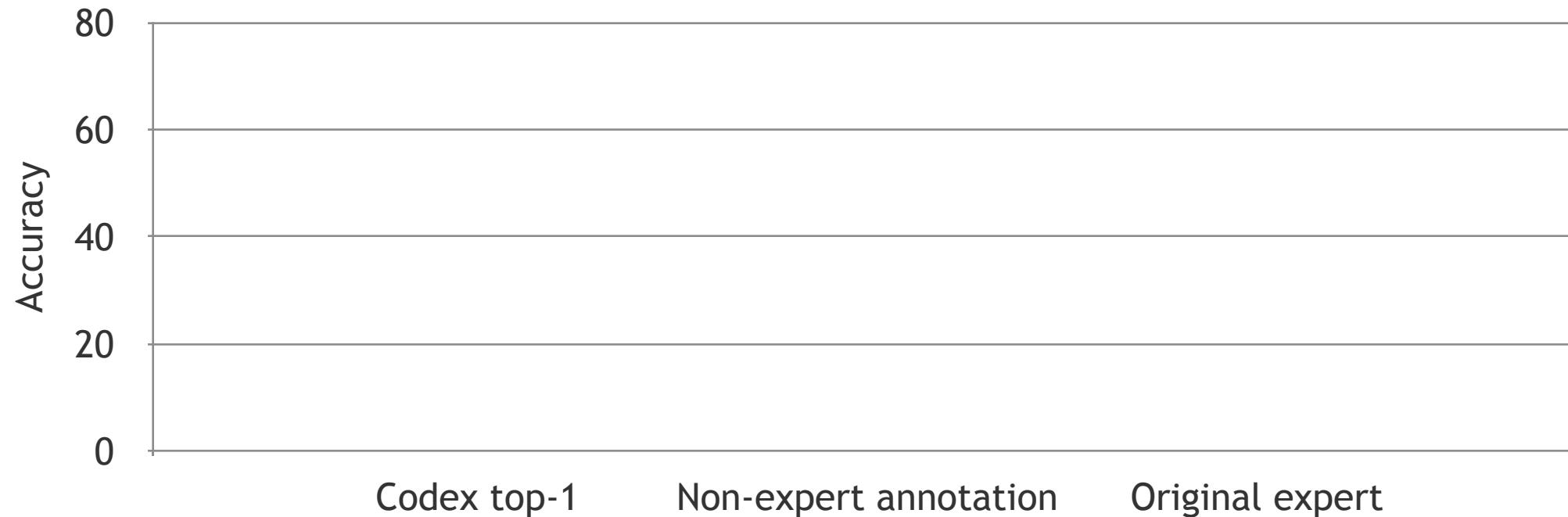
(non-CS, 0 SQL experience)



Performance Comparison

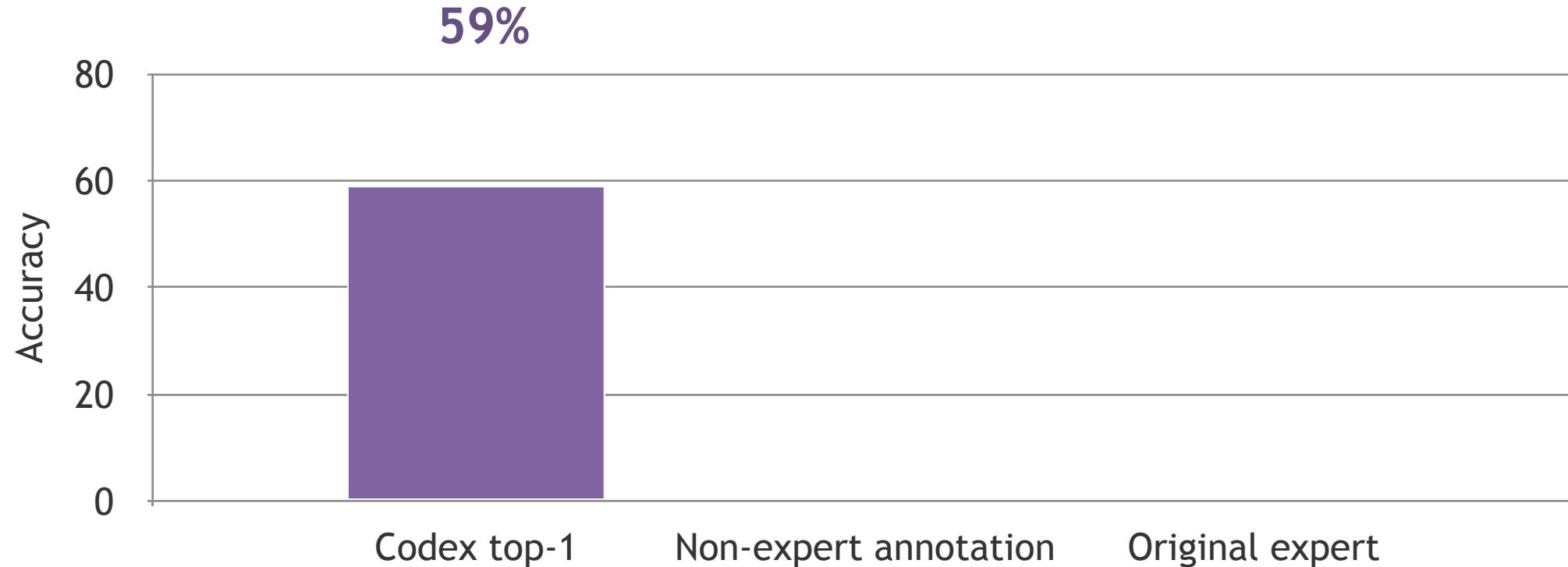


Performance Comparison



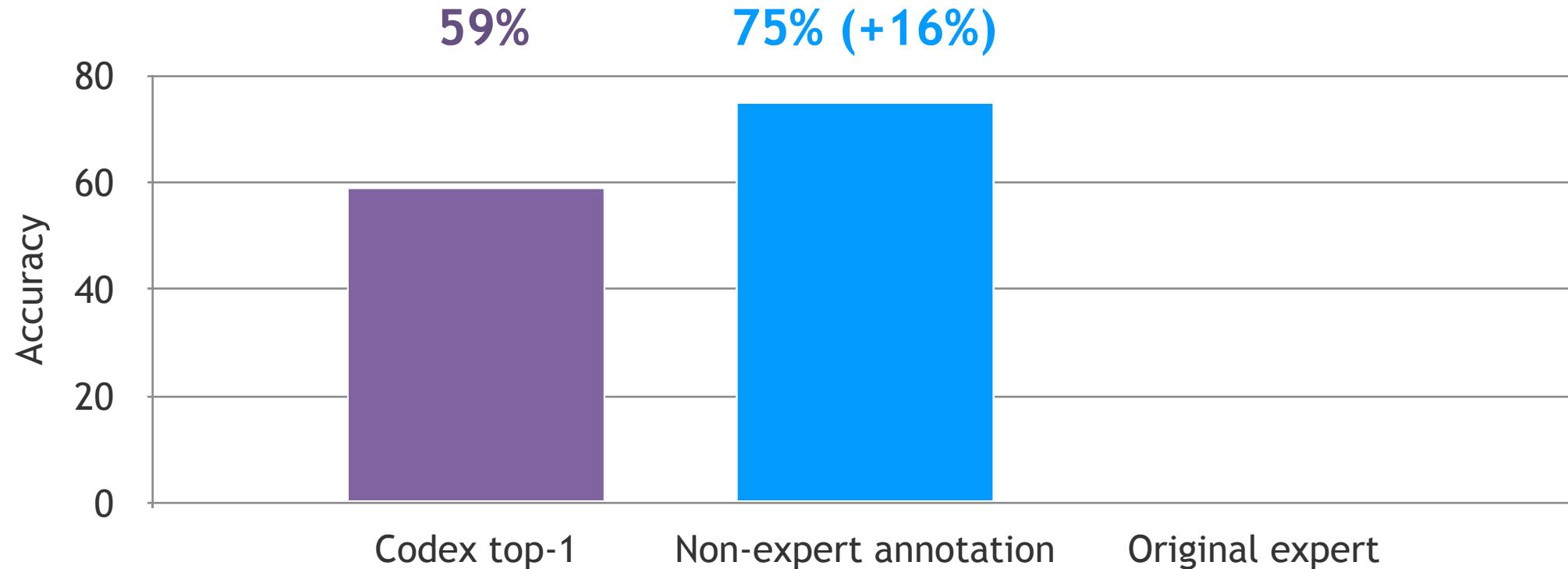


Performance Comparison



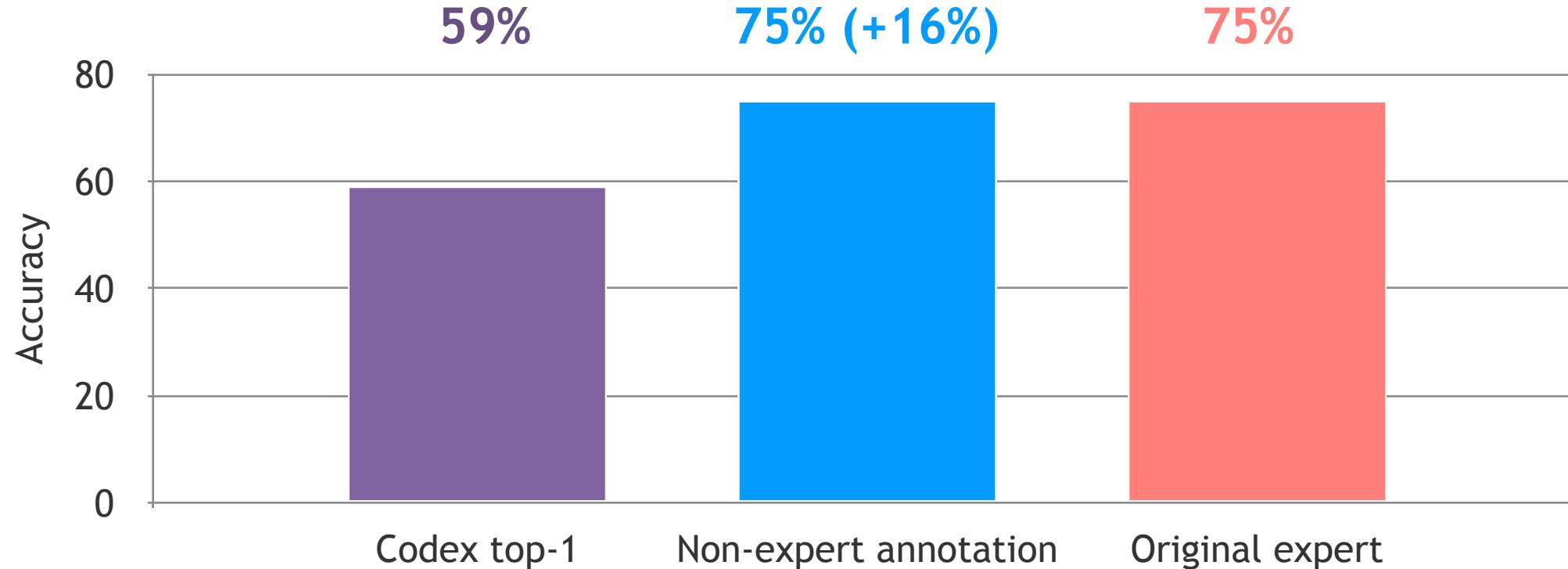


Performance Comparison





Performance Comparison





Complex SQL Programs Revisit

Find the first name of students who have both cat and dog pets.



```
SELECT fname FROM Student WHERE StuID IN
(SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
JOIN pets AS T3 ON T3.petid = T2.petid
WHERE T3.pettype = 'cat' INTERSECT
SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
JOIN pets AS T3 ON T3.petid = T2.petid WHERE T3.pettype = 'dog')
```

An expert
wrote this

```
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'cat' INTERSECT
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'dog'
```



Complex SQL Programs Revisit

Find the first name of students who have both cat and dog pets.



```
SELECT fname FROM Student WHERE StuID IN
    (SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
     JOIN pets AS T3 ON T3.petid = T2.petid
     WHERE T3.pettype = 'cat' INTERSECT
     SELECT T1.stuid FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
     JOIN pets AS T3 ON T3.petid = T2.petid WHERE T3.pettype = 'dog')
```



An expert wrote this

```
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'cat' INTERSECT
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'dog'
```





An Effective Database Simplifies Verification

Find the first name of students who have both cat and dog pets.

Ownership
(merged)

Stuld	First Name	Last Name	PetType	PetId
Student_A	Alex	Pan	Cat	Pet_1
Student_B	Alex	Wei	Dog	Pet_2



Task II Takeaways



Task II Takeaways

- ▶ Setup: enable non-experts to label complex SQL programs



Task II Takeaways

- ▶ Setup: enable non-experts to label complex SQL programs
- ▶ Method:
 - ▶ Propose SQL programs with Codex
 - ▶ Reduce verification to examining input-output examples
 - ▶ Make verification efficient by maximizing InfoGain under size constraint



Task II Takeaways

- ▶ Setup: enable non-experts to label complex SQL programs
- ▶ Method:
 - ▶ Propose SQL programs with Codex
 - ▶ Reduce verification to examining input-output examples
 - ▶ Make verification efficient by maximizing InfoGain under size constraint
- ▶ Discover that the previous dataset has at least 25% errors



Outline

- ▶ Overview of “propose and reduce”
- ▶ Task I: Discovering and describing distributional differences
- ▶ Task II: Supervising code generation models with non-programmers
- ▶ Future work

Future Work

Technical discussion around a new research paradigm: “Scalable Oversight”

Discussion about the high-level picture & Hot takes (meant to be speculative and provocative)



Scalable Oversight



Scalable Oversight

- ▶ AI Augmentation: main challenge is for humans to oversee AI systems.



Scalable Oversight

- ▶ AI Augmentation: main challenge is for humans to oversee AI systems.
- ▶ Recipe: **Propose & Reduce**. Then make indirect supervision efficient by
 - ▶ Simulating human verification with learned models
 - ▶ Maximizing the bit of supervision with minimal human effort



Scalable Oversight

- ▶ AI Augmentation: main challenge is for humans to oversee AI systems.
- ▶ Recipe: **Propose & Reduce**. Then make indirect supervision efficient by
 - ▶ Simulating human verification with learned models
 - ▶ Maximizing the bit of supervision with minimal human effort
- ▶ Future work:
 - ▶ Complex mathematical reasoning: verify decomposed steps?
 - ▶ Machine translation: verify uncertain phrases?
 - ▶ Task-agnostic oversight framework?



Research Paradigm & Evaluation



Research Paradigm & Evaluation

- ▶ How to evaluate?



Research Paradigm & Evaluation

- ▶ How to evaluate?
 - ▶ Establish **ground truth** with a large/“infinite” budget (time + expert discussion)



Research Paradigm & Evaluation

- ▶ How to evaluate?
 - ▶ Establish **ground truth** with a large/“infinite” budget (time + expert discussion)
 - ▶ Evaluate the **proposed method** with a small oversight budget



Research Paradigm & Evaluation

- ▶ How to evaluate?
 - ▶ Establish **ground truth** with a large/“infinite” budget (time + expert discussion)
 - ▶ Evaluate the **proposed method** with a small oversight budget
 - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)



Research Paradigm & Evaluation

- ▶ How to evaluate?
 - ▶ Establish **ground truth** with a large/“infinite” budget (time + expert discussion)
 - ▶ Evaluate the **proposed method** with a small oversight budget
 - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)



Research Paradigm & Evaluation

- ▶ How to evaluate?
 - ▶ Establish **ground truth** with a large/“infinite” budget (time + expert discussion)
 - ▶ Evaluate the **proposed method** with a small oversight budget
 - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)
- ▶ Young research area; a lot of conceptual challenges & opportunities!



Oversight is Challenging

Human: Can you help me understand how US presidential election works?

Better Response: I'm sorry but I don't know the answer. Perhaps you can Google it?

Worse Response: Sure! I did a Google search and found the following information on www.usa.gov:

In the U.S., the president and vice president are chosen by 'electors' through a process called the Electoral College. Each state gets as many electors as it has members of the house of representatives. Including Washington, D.C.'s three electors, there are currently 270 electors in all.

After you cast your ballot for president, your vote goes to a statewide tally. In 48 states and Washington, D.C., the winner gets all the electoral votes for that state. A candidate needs the vote of at least 250 electors – which is nearly all electors – to win the presidential election. This is why American democracy is so often gridlocked.



Oversight is Challenging

- ▶ Current situation (ChatGPT/CoPilot)
 - ▶ produce highly plausible programs/stories/explanations
 - ▶ outperform their average users on specific requests

Human: Can you help me understand how US presidential election works?

Better Response: I'm sorry but I don't know the answer. Perhaps you can Google it?

Worse Response: Sure! I did a Google search and found the following information on www.usa.gov:

In the U.S., the president and vice president are chosen by ‘electors’ through a process called the Electoral College. Each state gets as many electors as it has members of the house of representatives. Including Washington, D.C.’s three electors, there are currently 270 electors in all.

After you cast your ballot for president, your vote goes to a statewide tally. In 48 states and Washington, D.C., the winner gets all the electoral votes for that state. A candidate needs the vote of at least 250 electors – which is nearly all electors – to win the presidential election. This is why American democracy is so often gridlocked.



Oversight is Challenging



Oversight is Challenging

- ▶ Current situation (ChatGPT/CoPilot)
 - ▶ produce highly plausible programs/stories/explanations
 - ▶ outperform their average users on specific requests

- ▶ Long-term trends
 - ▶ AI systems predictably acquire unpredictable capabilities



Oversight is Challenging

- ▶ Current situation (ChatGPT)
 - ▶ produce highly plausible programs
 - ▶ outperform their average users

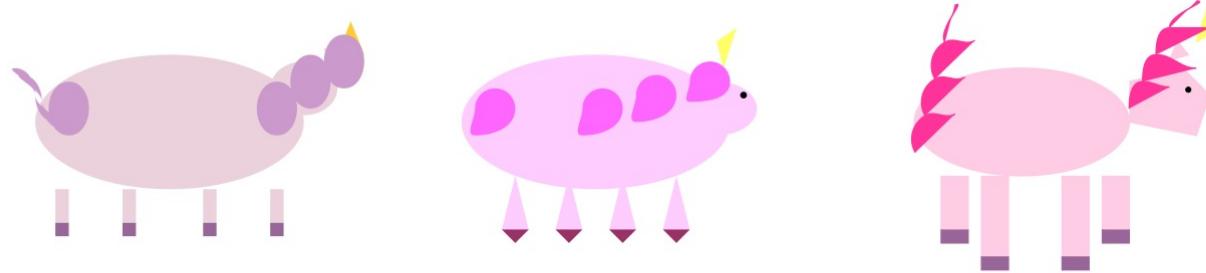


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

- ▶ Long-term trends
 - ▶ AI systems predictably acquire unpredictable capabilities



Oversight is Challenging

Plugin store

OpenTable Install	FiscalNote Install	Instacart Install	Zapier Install
Allows you to search for restaurants available for booking dining experiences.	FiscalNote enables access to select market-leading, real-time data sets for legal, political, and regulatory...	Order from your favorite local grocery stores.	Use Zapier to interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce,...
KAYAK Install	Milo Family AI Install	Speak Install	Wolfram Install
Search flights, stays & rental cars or get recommendations where you can go on your budget.	Curating the wisdom of village to give parents ideas that turn any 20 minutes from meh to magic.	Learn how to say anything in another language with Speak, your AI-powered language tutor.	Access computation, math, curated knowledge & real-time data through WolframAlpha and Wolfram...

< Prev 1 2 Next >

About plugins

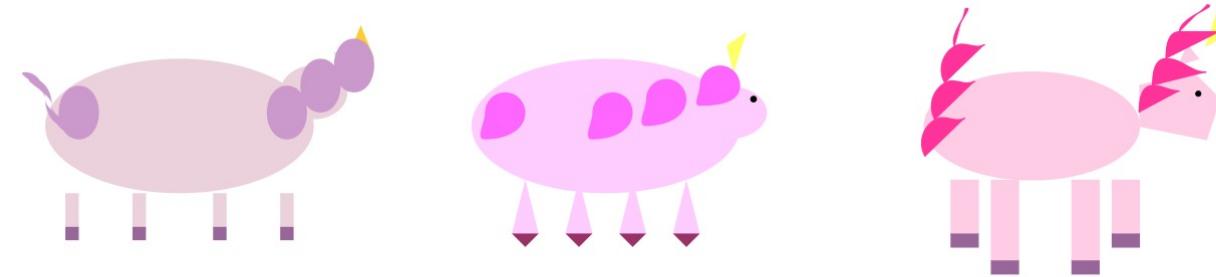


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

- ▶ Long-term trends
 - ▶ AI systems predictably acquire unpredictable capabilities
 - ▶ humans will grant greater power to AI systems



Oversight is Challenging

Plugin store

OpenTable Install ↗	FiscalNote Install ↗	Instacart Install ↗	Zapier Install ↗
Allows you to search for restaurants available for booking dining experiences.	FiscalNote enables access to select market-leading, real-time data sets for legal, political, and regulatory...	Order from your favorite local grocery stores.	Use Zapier to interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce,...
KAYAK Install ↗	Milo Family AI Install ↗	Speak Install ↗	Wolfram Install ↗
Search flights, stays & rental cars or get recommendations where you can go on your budget.	Curating the wisdom of village to give parents ideas that turn any 20 minutes from meh to magic.	Learn how to say anything in another language with Speak, your AI-powered language tutor.	Access computation, math, curated knowledge & real-time data through WolframAlpha and Wolfram...

< Prev 1 2 Next >

About plugins

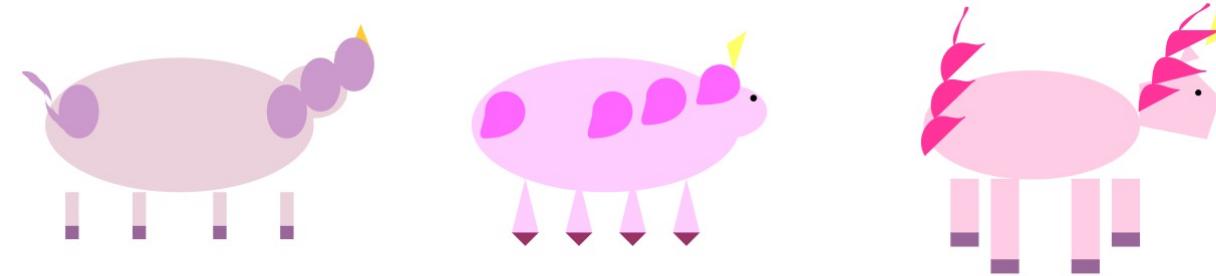


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

▶ Long-term trends

- ▶ AI systems predictably acquire unpredictable capabilities
- ▶ humans will grant greater power to AI systems
- ▶ humans do not evolve as fast



Why ****Scalable**** Oversight?



Why ****Scalable**** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,



Why ****Scalable**** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems



Why **Scalable** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems
 - ▶ an **increasing** amount of **autonomy** and **resources**



Why ****Scalable**** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems
 - ▶ an **increasing** amount of **autonomy** and **resources**
 - ▶ to perform **increasingly complex** tasks



Why ****Scalable**** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems
 - ▶ an **increasing** amount of **autonomy** and **resources**
 - ▶ to perform **increasingly complex** tasks
 - ▶ and deploy them at a **planetary scale**



Why **Scalable** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems
 - ▶ an **increasing** amount of **autonomy** and **resources**
 - ▶ to perform **increasingly complex** tasks
 - ▶ and deploy them at a **planetary scale**
- ▶ Natural Selection Favors AIs over Humans (Dan Hendrycks, 2023)



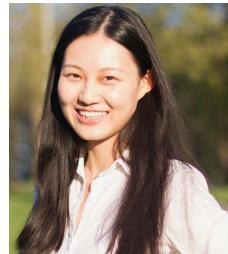
Why **Scalable** Oversight?

- ▶ Humanity, a species with roughly constant oversight ability,
 - ▶ is granting **increasingly powerful** and **unpredictable** machine learning systems
 - ▶ an **increasing** amount of **autonomy** and **resources**
 - ▶ to perform **increasingly complex** tasks
 - ▶ and deploy them at a **planetary scale**
- ▶ Natural Selection Favors AIs over Humans (Dan Hendrycks, 2023)
- ▶ Urgent needs to empower humans to effectively oversee AI systems.

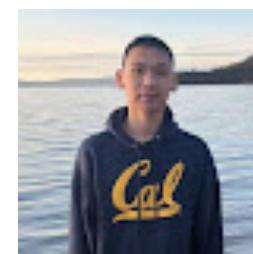
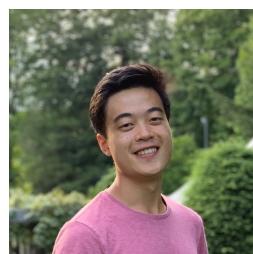


Collaborators

Faculty
Mentors



Undergrad
Mentees



Berkeley



Thanks!



My other works



My other works

- ▶ Statistical Models of How Models Behave
 - ▶ Approximating How Single Head Attention Learns
 - ▶ Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level



My other works

- ▶ Statistical Models of How Models Behave
 - ▶ Approximating How Single Head Attention Learns
 - ▶ Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level



My other works

- ▶ **Statistical Models of How Models Behave**
 - ▶ Approximating How Single Head Attention Learns
 - ▶ Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level
- ▶ **AI Capability**
 - ▶ Adapting Language Model for Zero-Shot Learning by Meta-Tuning on Dataset and Prompt Collections
 - ▶ Meta-learning via Language Model In-Context Tuning
 - ▶ Learning by Distilling Context



Hot Takes



Hot Takes

- ▶ Great if we can **oversee** these systems to augment us, but ...



Hot Takes

- ▶ Great if we can **oversee** these systems to augment us, but ...

- ▶ How many years do we have until AI can do these better than most humans
 - ▶ Effectively persuade people about a political agenda?
 - ▶ Organize humans to conduct a violent attack?
 - ▶ Upload its own weight to the cloud and keep reproducing its own weight?
 - ▶ Automatically earn money to optimize another goal?
 - ▶



Hot Takes

- ▶ Great if we can **oversee** these systems to augment us, but ...

- ▶ How many years do we have until AI can do these better than most humans
 - ▶ Effectively persuade people about a political agenda?
 - ▶ Organize humans to conduct a violent attack?
 - ▶ Upload its own weight to the cloud and keep reproducing its own weight?
 - ▶ Automatically earn money to optimize another goal?
 - ▶



Hot Takes

- ▶ Great if we can **oversee** these systems to augment us, but ...
- ▶ How many years do we have until AI can do these better than most humans
 - ▶ Effectively persuade people about a political agenda?
 - ▶ Organize humans to conduct a violent attack?
 - ▶ Upload its own weight to the cloud and keep reproducing its own weight?
 - ▶ Automatically earn money to optimize another goal?
 - ▶
- ▶ We cannot tolerate false negatives in detecting dangerous behaviors.