Berkeley NLP

Supervising AI to Do Things I Can't

Ruiqi Zhong

Ruiqi Zhong

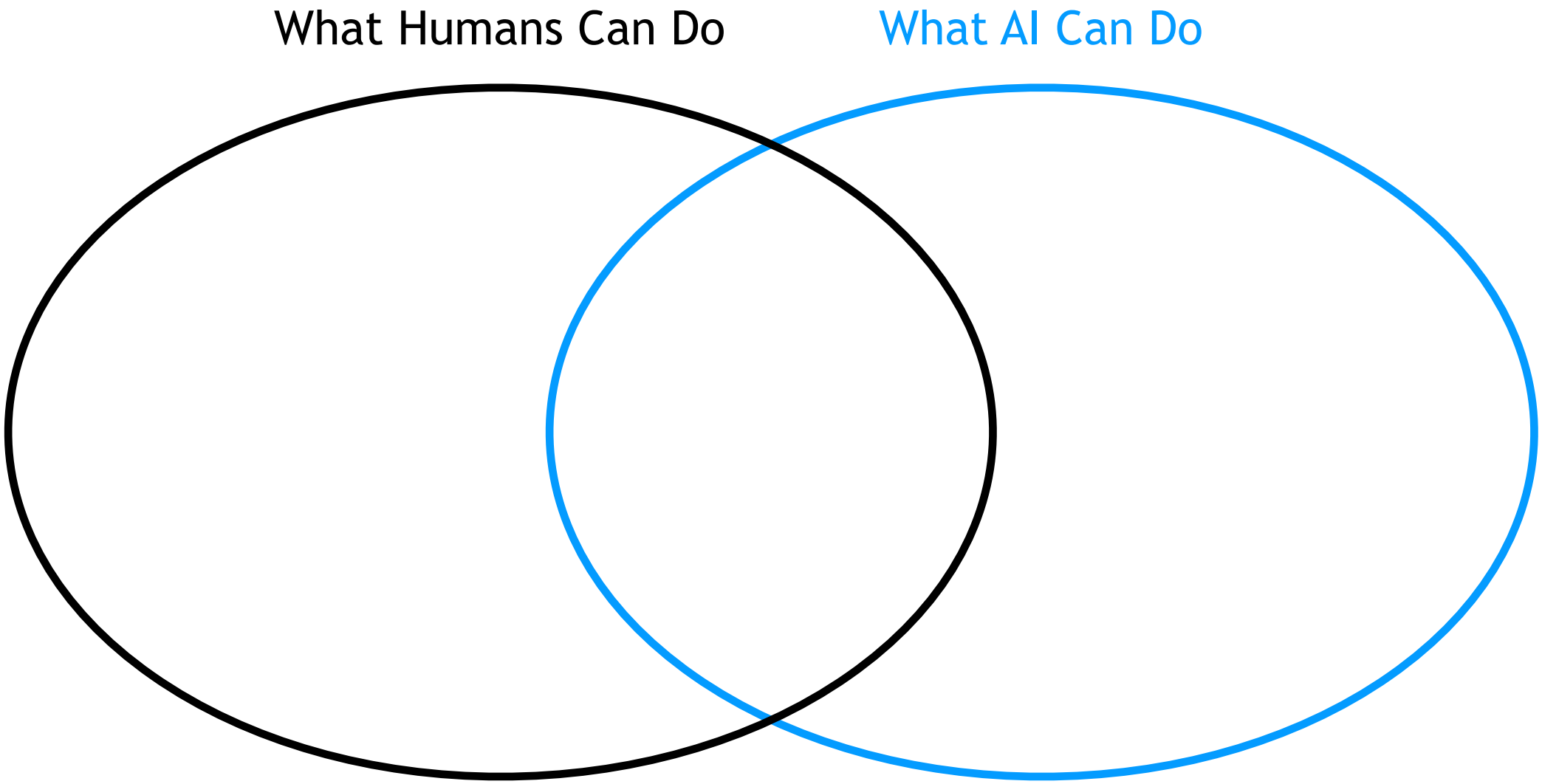
Jacob Steinhardt

Dan Klein

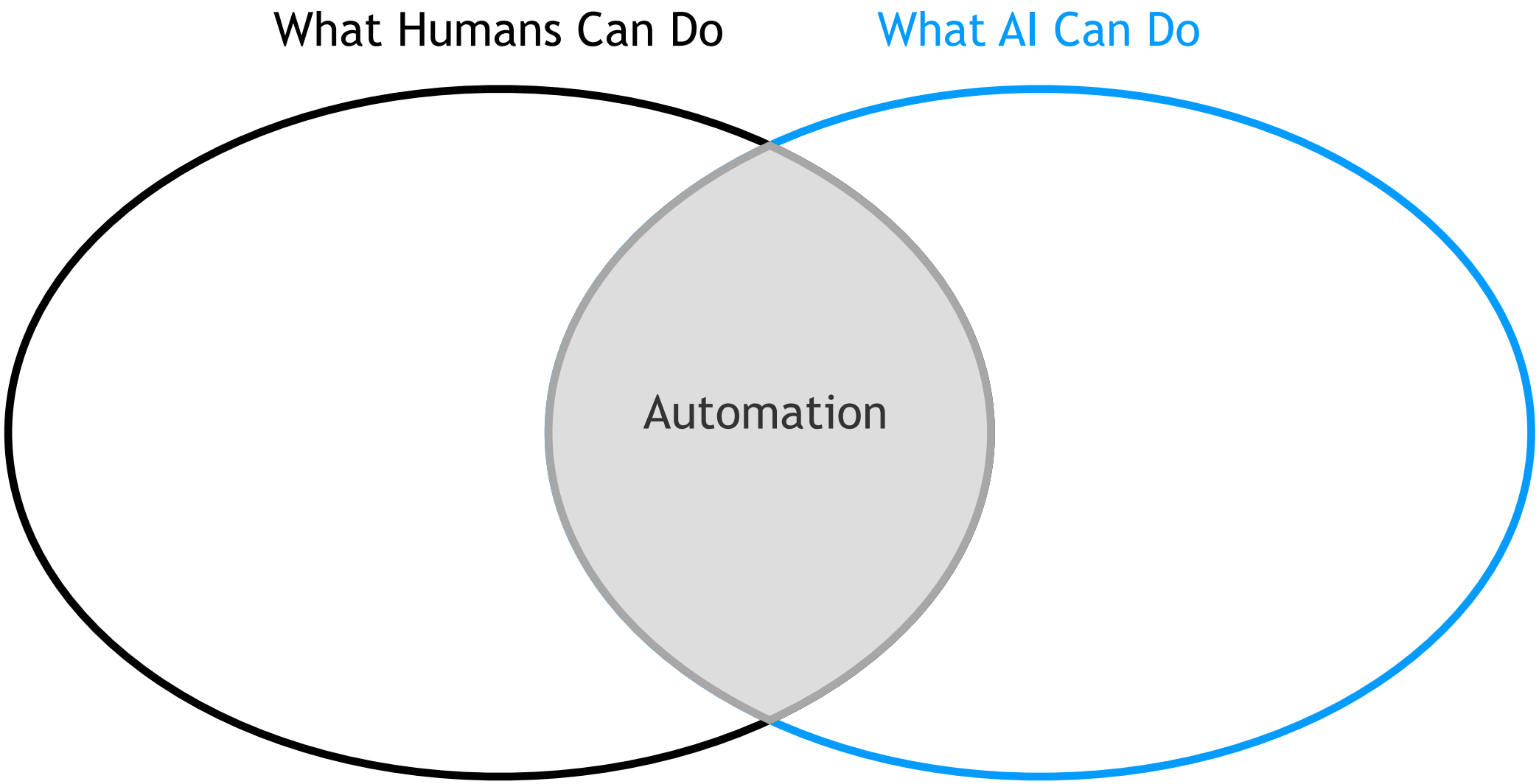
Jason Eisner

Charlie Snell

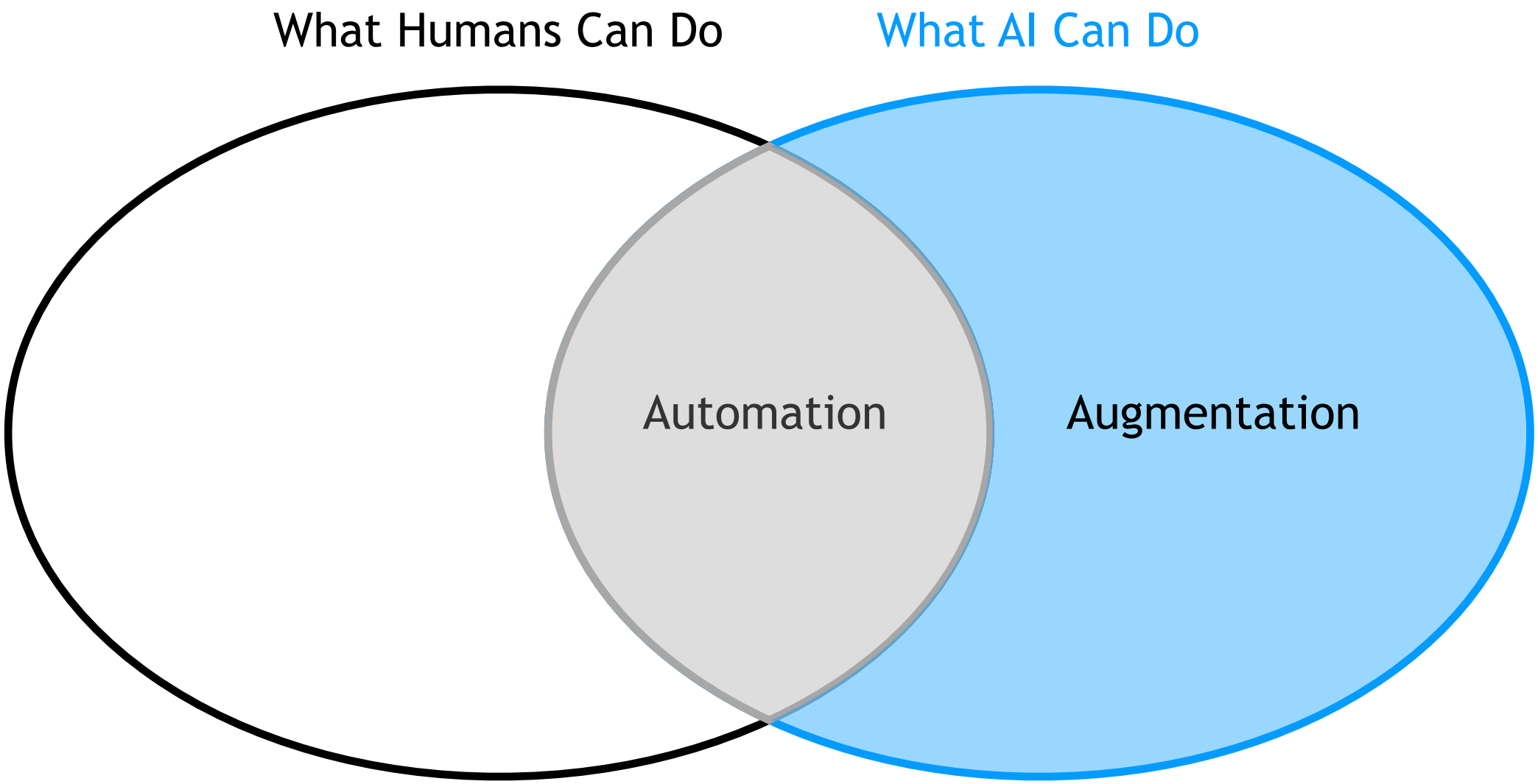
Automation vs. Augmentation



Automation vs. Augmentation



Automation vs. Augmentation



Today's Talk

How to supervise AI even when we struggle to determine the ground truth?

Example Tasks

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 ……
```

Corpus 1

- 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.
- It's quite ironical that such a centralized government fail to locate the suspects who gravely injured those girls earlier this month.

Corpus 2

- 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.
- After 10 years of lockdown due to the pandemics, I finally saw my grandfather — I thought I might never see him again.

"*Corpus 1 mentions more female entities*"

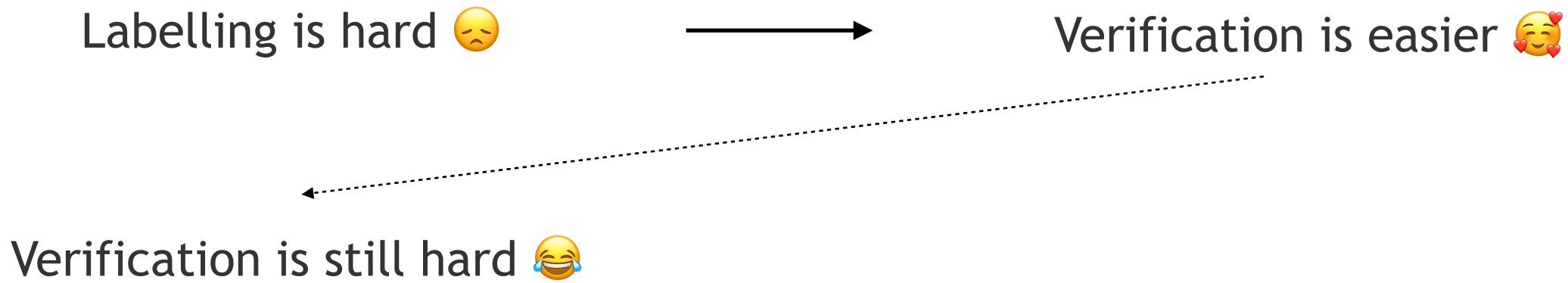
Label —> Verify —> Reduce

Labelling is hard 😞

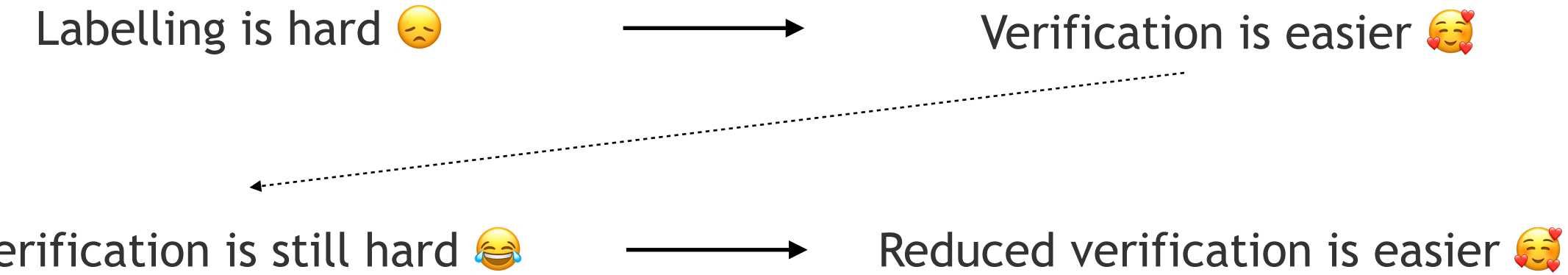
Label —> Verify —> Reduce

Labelling is hard 😞 ⟶ Verification is easier 🥰

Label —> Verify —> Reduce



Label —> Verify —> Reduce



Task I: Enabling Non-Experts to Indirectly Label Complex Programs

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 do non-experts label these programs to train a model?

Label Full Output —> Verify

Natural Language *How old is the youngest person from department A?*

	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'`

Label Full Output —> Verify

Natural Language *How old is the youngest person from department A?*

	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'`

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'
```

Reduction: Verify with Input-Output

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'
```

Reduction: Verify with Input-Output

How old is the youngest person from department A?

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

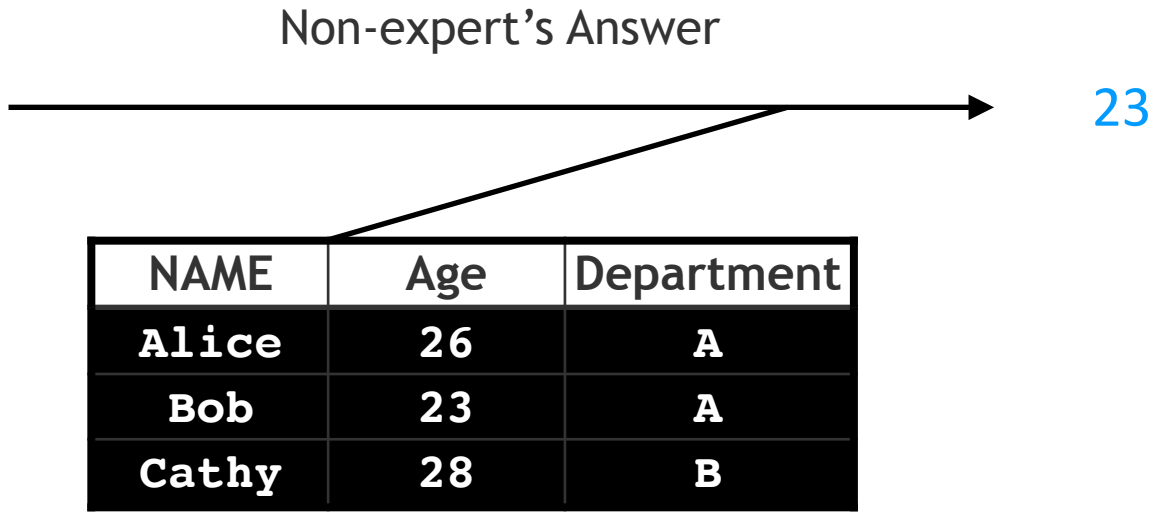
```
SELECT MAX(Name) from People

SELECT MAX(Age) from People

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

Reduction: Verify with Input-Output

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'
```

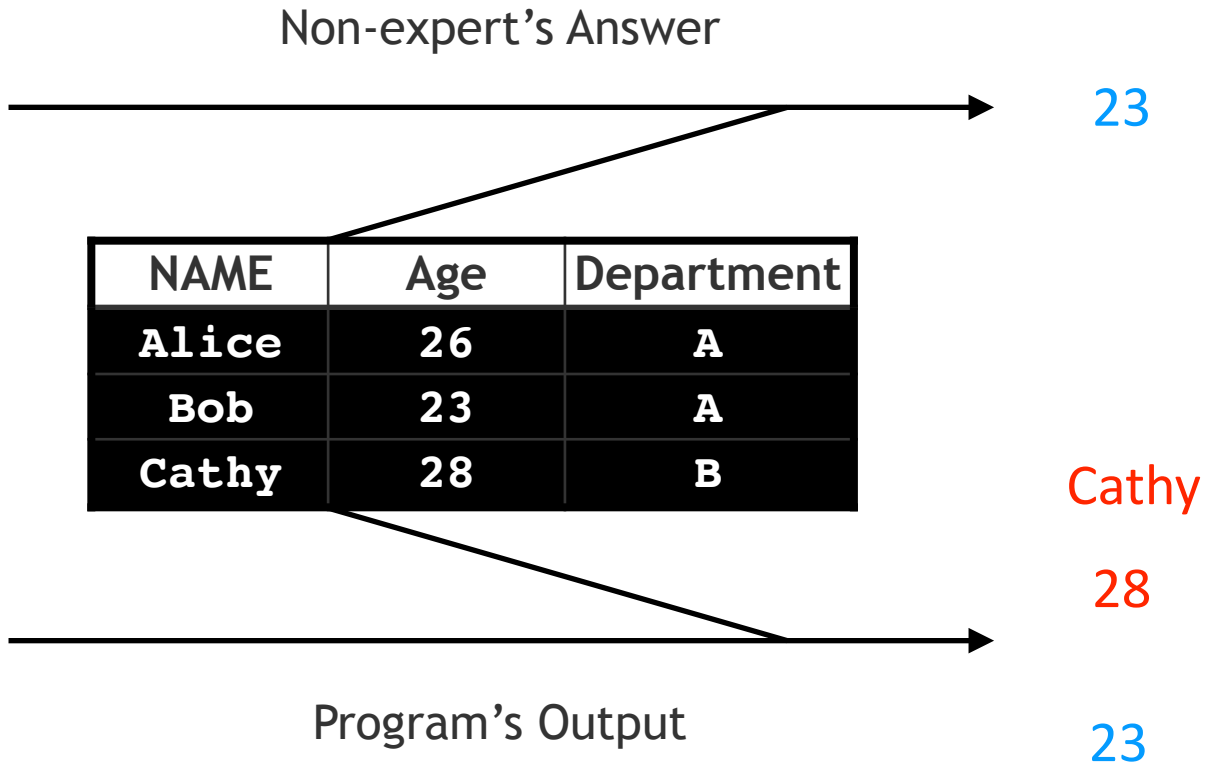
Reduction: Verify with Input-Output

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'
```



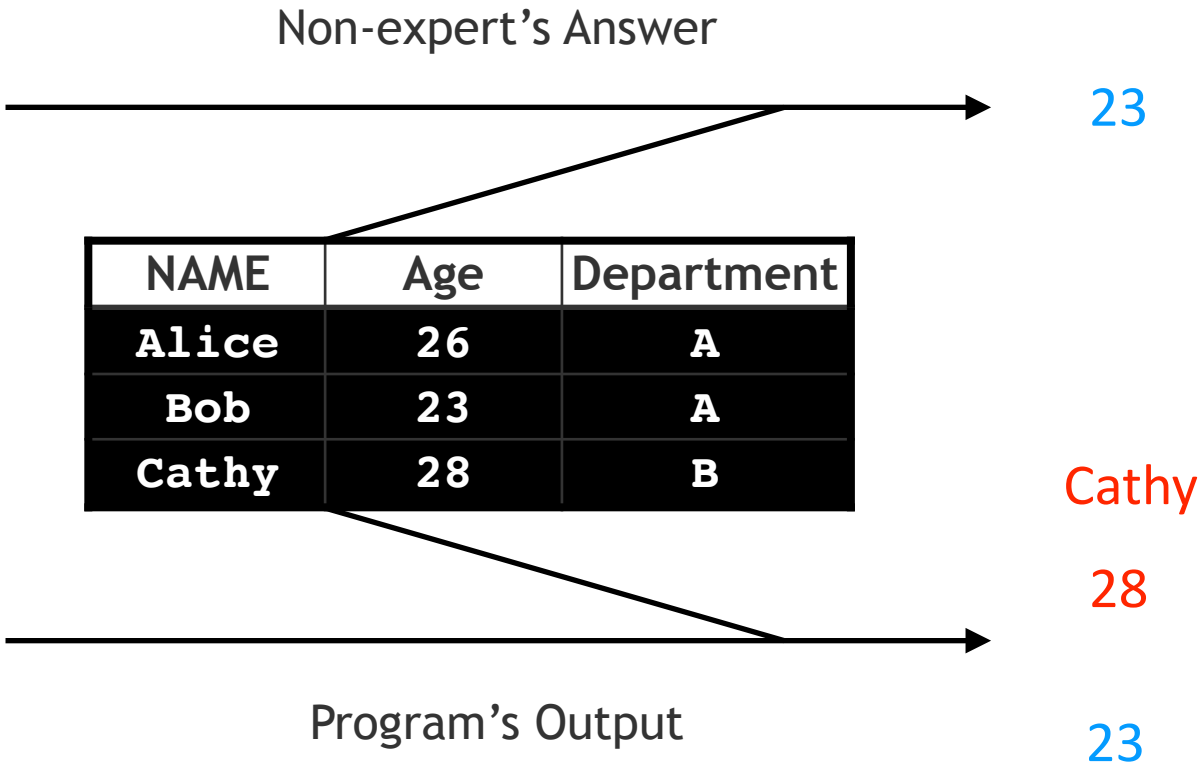
Reduction: Verify with Input-Output

How old is the youngest person from department A?

```
SELECT MAX(N      from People

SELECT MAX( ge)   om People
```

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



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

23

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

Cathy

28

Program's Output

23

Making Verification Simpler

Making Verification Simpler

- Choose the database wisely

Making Verification Simpler

- Choose the database wisely
  - is small enough to comprehend

Making Verification Simpler

- Choose the database wisely
  - is small enough to comprehend
  - make different programs return different values

Making Verification Simpler

How old is the youngest person from department A?

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

[In total 1000 rows, rest omitted]

→ Annotators' Answer ?????

Making Verification Simpler

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

[In total 1000 rows, rest omitted]

→ Annotators' Answer ?????

Making Verification Simpler

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]

Making Verification Effective

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'
```

Making Verification Effective

How old is the youngest person from department A?

NAME	Age	Department
Collin	26	A
Bob	23	A

→ Annotators' Answer **23**

```
SELECT MIN(Age) from People
```
→ **23**

```
SELECT MIN(Age) from People
   WHERE Department = 'A'
```
→ **23**

Making Verification Effective

How old is the youngest person from department A?

NAME	Age	Department
Collin	26	A
Bob	23	A

→ Annotators' Answer **23**

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	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

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

Technical Details in Our Paper

- Search i with small Size(i) and large InfoGain(i)

Technical Details in Our Paper

- Search i with small Size(i) and large InfoGain(i)
  - fuzzing; then drop rows greedily.

Technical Details in Our Paper

- Search i with small Size(i) and large InfoGain(i)
  - fuzzing; then drop rows greedily.
- Multi-round interaction

Human Study Setup

Human Study Setup

- 11 students: non-CS; zero SQL experience

Human Study Setup

- 11 students: non-CS; zero SQL experience
- Answer on average 4 databases with 9 rows per question

Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Expert Annotations

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

Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Codex

Probabilities

7/10 `SELECT MAX(Name) from People`

1/10 `SELECT MAX(Age) from People`

……

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

Expert Annotations

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

Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Codex

Probabilities

7/10 `SELECT MAX(Name) from People`

1/10
```
SELECT MAX(Age) from People
```

……

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

Expert Annotations

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

`Codex Top-1`

Test suite
(Zhong et al. 2020)

Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Codex

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

Probabilities

7/10 `SELECT MAX( ne rom People`

1/10 `SELECT MAX( rom People`

.....

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

Expert Annotations

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

Codex Top-1

Dataset and Baselines

Natural Language

How old is the youngest person from department A?

Codex

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

Probabilities

7/10 `SELECT MAX( ne from People`

1/10 `SELECT MAX( ) rom People`

.....

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

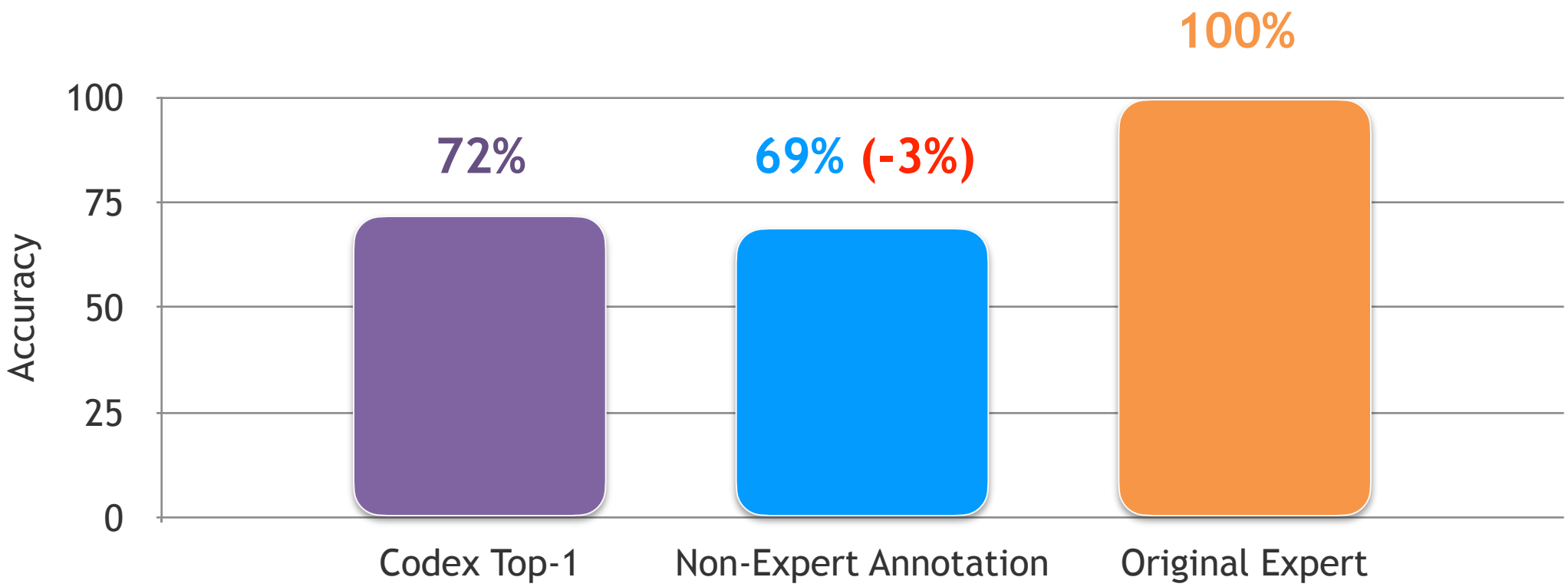
Expert Annotations

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

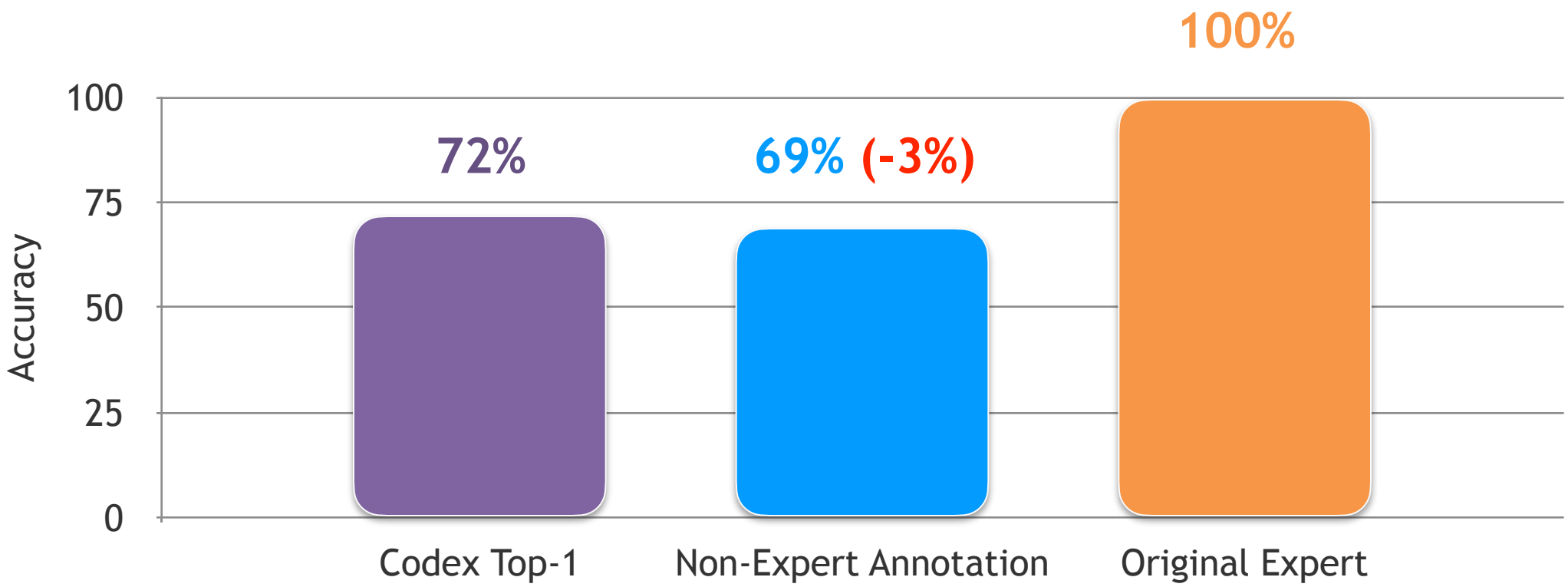
Codex Top-1

Our Annotation

Using Original Labels as the Ground Truth



Using Original Labels as the Ground Truth



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)

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

Re-annotation

Re-annotation

- Authors use the new framework to re-label

Re-annotation

- Authors use the new framework to re-label
- Stick to the original label if ambiguous

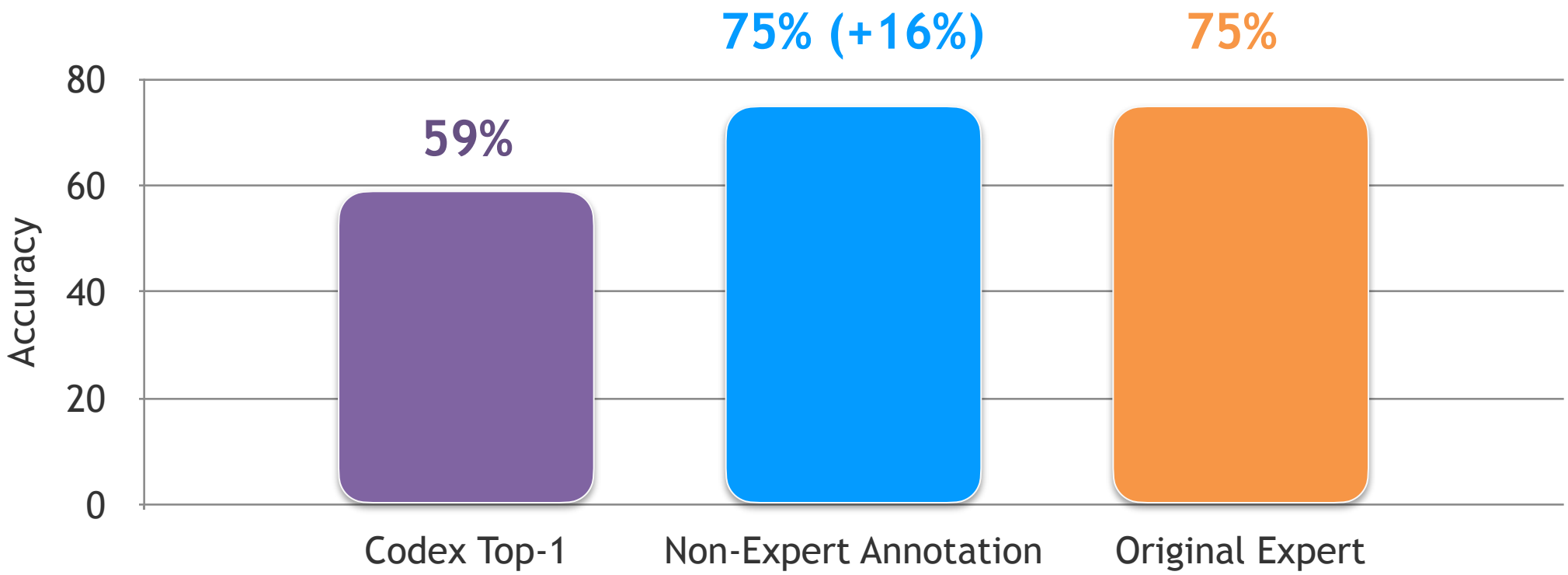
Re-annotation

- Authors use the new framework to re-label
- Stick to the original label if ambiguous
- Fix only the uncontroversially worse labels

Re-annotation

- Authors use the new framework to re-label
- Stick to the original label if ambiguous
- Fix only the uncontroversially worse labels
- Supported by the SPIDER 1st author

After Re-annotation with Our System …



Recap

Recap

- Labeling programs —> verifying programs

Recap

- Labeling programs —> verifying programs
- Verifying programs —> reducing to input-outputs

Recap

- Labeling programs —> verifying programs
- Verifying programs —> reducing to input-outputs
- Making verification simple and effective by choosing databases wisely

Application II: Describing Differences between Text Distributions (ICML 2022)

Difference between Distributions

D_1

- Ma mère m’a emmené à l’hôpital.
- J'ai 10 $. Je dépense 3 $ sur un livre.
- Le gouvernement n'a pas réussi à localiser les suspects.

D_2

- My mom and I were best friends.
- Lucy and Peter co-authored a paper.
- Yesterday, I called her to explain why I did horribly on the test.

Difference between Distributions

D_1

- Ma mère m’a emmené à l’hôpital.
- J'ai 10 $. Je dépense 3 $ sur un livre.
- Le gouvernement n'a pas réussi à localiser les suspects.

D_2

- My mom and I were best friends.
- Lucy and Peter co-authored a paper.
- Yesterday, I called her to explain why I did horribly on the test.

s = “D_1 contains more French sentences compared to D_2”

Difference between Distributions

D_1

- Ma mère m'a emmené à l'hôpital.
- J'ai 10 $. Je dépense 3 $ sur un livre.
- Le gouvernement n'a pas réussi à localiser les suspects.

D_2

- My mom and I were best friends.
- Lucy and Peter co-authored a paper.
- Yesterday, I called her to explain why I did horribly on the test.

s = "D_1 contains more French sentences compared to D_2"

s = "D_1 contains more accent mark compared to D_2"

Tell the Difference!

D_1

- 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.

D_2

- 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.

Tell the Difference!

D_1

- 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.

D_2

- 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.

s = "D_1 mentions more female entities compared to D_2"

Example Insights

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2
- A text cluster contains more sports-related articles than other clusters.
  D_1 s D_2

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2
- A text cluster contains more sports-related articles than other clusters.
  D_1 s D_2
- Public opinions from this year are more optimistic about the pandemic than last year.
  D_1 s D_2

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2
- A text cluster contains more sports-related articles than other clusters.
  D_1 s D_2
- Public opinions from this year are more optimistic about the pandemic than last year.
  D_1 s D_2
- Find dataset shortcuts
- Summarize author styles
- …

The Two Challenges …

- Hard to label —> easier to verify
- Hard to verify —> reduce to something simpler

Label Full Output —> Verify

Group A: The author proposed a novel and impactful task. Accept!
Group A: Strong accept.
Group A: The method is simple but effective.
Group A: [Other random samples from D_1]

Group B: Thanks for the author, but this paper should be rejected.
Group B: Is this a high school tech report?
Group B: Total waste of time reading this submission.
Group B: [Other random samples from D_2]

Compared to sentences from group B, each sentence from group A

GPT-3

- Is more positive
- Is offensive in tone
- [Other descriptions]

The Two Challenges …

- Hard to label —> easier to verify
- Hard to verify —> reduce to something simpler

Reduction: Verify on Individual Samples

A good description helps humans tell ***individual samples*** from D_1 and D_2 apart.

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$ "This paper proposes an impactful task."

$x_b \sim D_b$ "The approach of this paper is too trivial."

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$ "This paper proposes an impactful task."

$x_b \sim D_b$ "The approach of this paper is too trivial."

Human Classifies

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$ "This paper proposes an impactful task."

$x_b \sim D_b$ "The approach of this paper is too trivial."

Human Classifies

I think x_a is from D_1 and x_b is from D_2

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$ "This paper proposes an impactful task."

$x_b \sim D_b$ "The approach of this paper is too trivial."

Human Classifies

I think x_a is from D_1 and x_b is from D_2

Loss(s): Repeat 100 times and calculate human classification error rate.

Learning a Natural Language Hypothesis

Learning a Natural Language Hypothesis

- Classical: find the most discriminative linear weight to separate input vectors.

Learning a Natural Language Hypothesis

- Classical: find the most discriminative linear weight to separate input vectors.
- Ours : find the most discriminative natural language string to separate input text.

Reduction: Verify on Individual Samples

s = "Samples from D_1 are more positive than those from D_2"

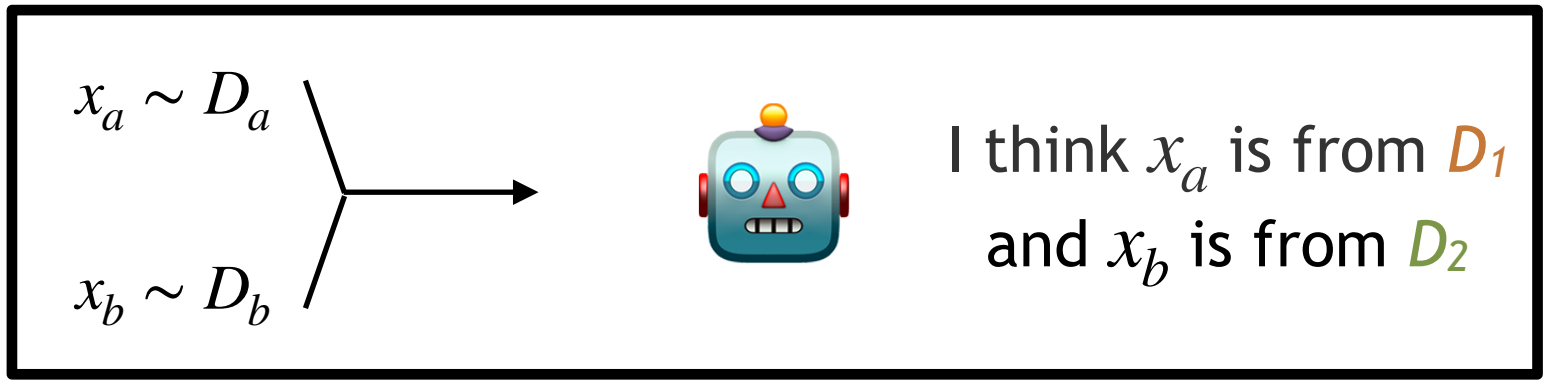
Loss(*s*): Calculate a human's error rate.

~$10 to evaluate a single description.

Making Verification Cheaper

s = "Samples from D_1 are more positive than those from D_2"

Loss(*s*): Calculate a model's error rate.



Much cheaper to calculate the error rate of a model

Method Overview

Method Overview

- GPT-3 proposes 60 descriptions

Method Overview

- GPT-3 proposes 60 descriptions
- Verify each with a learned model

Method Overview

- GPT-3 proposes 60 descriptions
- Verify each with a learned model
- Create fine-tuning data for GPT-3 (see our paper)

Benchmark

Benchmark

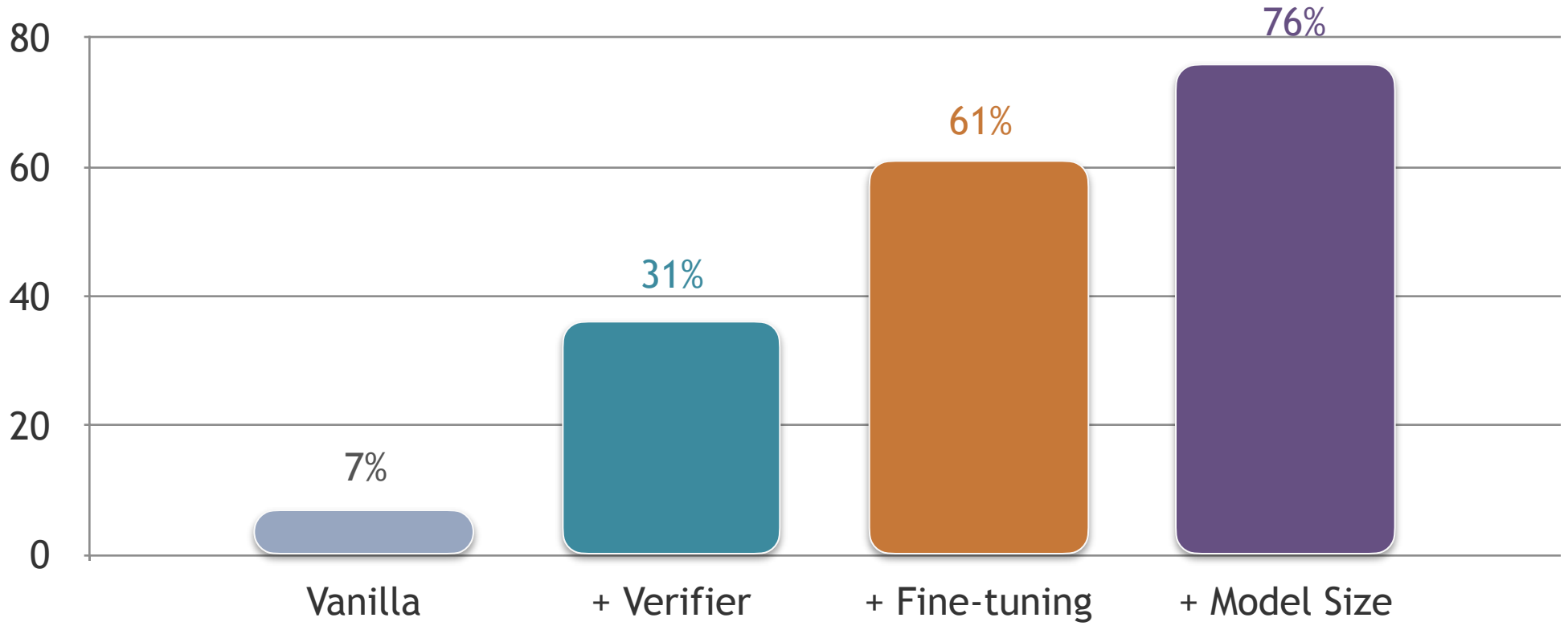
- Zhong et al 2021: 54 text classification task, with language descriptions

Benchmark

- Zhong et al 2021: 54 text classification task, with language descriptions
- Recover the descriptions for the positive class from the samples

Benchmark

- Zhong et al 2021: 54 text classification task, with language descriptions
- Recover the descriptions for the positive class from the samples



Expert Annotations Are Noisy

Expert Annotations Are Noisy

- SUBJ: Subjectivity analysis dataset

Expert Annotations Are Noisy

- SUBJ: Subjectivity analysis dataset
- >= 4 previous papers used it this way

Expert Annotations Are Noisy

- SUBJ: Subjectivity analysis dataset
- >= 4 previous papers used it this way
- Our system generates "*is a plot summary of a film*" ??

Expert Annotations Are Noisy

- SUBJ: Subjectivity analysis dataset
- >= 4 previous papers used it this way
- Our system generates "*is a plot summary of a film*" ??

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 ww.imdb.com

Exposing Dataset Flaws

Exposing Dataset Flaws

- ML models might pick up shallow correlations.

Exposing Dataset Flaws

- ML models might pick up shallow correlations.
- Binary classification: Spam vs. Non-Spam

D_1 D_2

Exposing Dataset Flaws

- ML models might pick up shallow correlations.
- Binary classification: Spam vs. Non-Spam

  D_1 D_2

- "D_1 contains more spam" / "D_1 contains more hyperlinks"

Exposing Dataset Flaws

- ML models might pick up shallow correlations.
- Binary classification: Spam vs. Non-Spam

  D_1 D_2
- "D_1 contains more spam" / "D_1 contains more hyperlinks"
- Fine-tuned RoBERTa predicts spam whenever it sees a hyperlink!!!

Recap

Recap

- Labeling descriptions —> verifying descriptions

Recap

- Labeling descriptions —> verifying descriptions
- Verifying on the entire distribution —> reducing to individual samples

Recap

- Labeling descriptions —> verifying descriptions
- Verifying on the entire distribution —> reducing to individual samples
- Making verification cheaper by simulating human verification

Future Work & Discussion

What can AI do?

What can AI do?

what one human can do

What can AI do?

what one human can verify

What can AI do?

what a large group of humans can verify with the help of AI and other tools

Augmentative AI: Opportunities

Augmentative AI: Opportunities

- Search
  - Creativity
  - Effective strategies

Augmentative AI: Opportunities

- Search
  - Creativity
  - Effective strategies
- Distributional properties
  - Scientific discovery

Augmentative AI: Opportunities

- Search
  - Creativity
  - Effective strategies
- Distributional properties
  - Scientific discovery
- Long input
  - Summarization
  - Generation

Augmentative AI: Challenge

Augmentative AI: Challenge

- Hard to create training data

Augmentative AI: Challenge

- Hard to create training data
- Limitation of label —> verify —> reduce

Augmentative AI: Challenge

- Hard to create training data
- Limitation of label —> verify —> reduce
  - No proposed candidate is right

Augmentative AI: Challenge

- Hard to create training data
- Limitation of label —> verify —> reduce
  - No proposed candidate is right
  - What if humans can’t indirectly verify?

Augmentative AI: Challenge

- Hard to create training data
- Limitation of label —> verify —> reduce
  - No proposed candidate is right
  - What if humans can’t indirectly verify?
- Verification is going to be harder

Augmentative AI: Benchmarks

Augmentative AI: Benchmarks

- Automation : I spent $2K on **AMT** to label **20,000** datapoints. **I can tell that the labels are correct** by scrutinizing some annotations.

Augmentative AI: Benchmarks

- Automation : I spent $2K on **AMT** to label **20,000** datapoints. **I can tell that the labels are correct** by scrutinizing some annotations.

- Augmentation: I spent $2K on **experts** to label **54** datapoints. The labels look reasonable, but I am **uncertain whether they are correct**.

Augmentative AI: Benchmarks

Augmentative AI: Benchmarks

- Ground truth is hidden

Augmentative AI: Benchmarks

- Ground truth is hidden
  - Bayesian when interpreting the performance measures

Augmentative AI: Benchmarks

- Ground truth is hidden
  - Bayesian when interpreting the performance measures
- Depends on confidence in the human labels & the AI's capability

Augmentative AI: Benchmarks

- Ground truth is hidden
  - Bayesian when interpreting the performance measures
- Depends on confidence in the human labels & the AI's capability
  - Rule out incapable systems with high confidence labels

Augmentative AI: Benchmarks

- Ground truth is hidden
  - Bayesian when interpreting the performance measures
- Depends on confidence in the human labels & the AI's capability
  - Rule out incapable systems with high confidence labels
  - Collaborate with capable systems to improve low confidence labels

Takeaways

Takeaways

- We can supervise AI even if we struggle to determine the ground truth.

Takeaways

- We can supervise AI even if we struggle to determine the ground truth.
- Label —> Verify —> Reduce

Takeaways

- We can supervise AI even if we struggle to determine the ground truth.
- Label —> Verify —> Reduce
- Vastly under-explored research with opportunities & challenges.

My other research threads

My other research threads

- Prompting
  - 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

My other research threads

- Prompting
  - 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
- Program Synthesis
  - Semantic Scaffolds for Pseudocode to Code Generation
  - Semantic Evaluation for Text-to-SQL with Distilled Test Suite
  - Active Programming by Example with a Natural Language Prior

My other research threads

- Prompting
  - 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
- Program Synthesis
  - Semantic Scaffolds for Pseudocode to Code Generation
  - Semantic Evaluation for Text-to-SQL with Distilled Test Suite
  - Active Programming by Example with a Natural Language Prior
- 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

Thanks!