Scalable Oversight for LLMs

(“cheaper validations”)

Ruiqi Zhong’s Dissertation
University of California, Berkeley

The Man Who Almost Faked His Way to a Nobel Prize

(Keeping up scientific standards, Martin Blume, Nature 2009)

The Man Who Almost Faked His Way to a Nobel Prize

Jan Hendrik Schön

(Keeping up scientific standards, Martin Blume, Nature 2009)

The Man Who Almost Faked His Way to a Nobel Prize

Jan Hendrik Schön

Seems correct to me.

Reviewers

(Keeping up scientific standards, Martin Blume, Nature 2009)

The Man Who Almost Faked His Way to a Nobel Prize

Jan Hendrik Schön

1 paper / 8 days.

(Keeping up scientific standards, Martin Blume, Nature 2009)

The Man Who Almost Faked His Way to a Nobel Prize

Jan Hendrik Schön

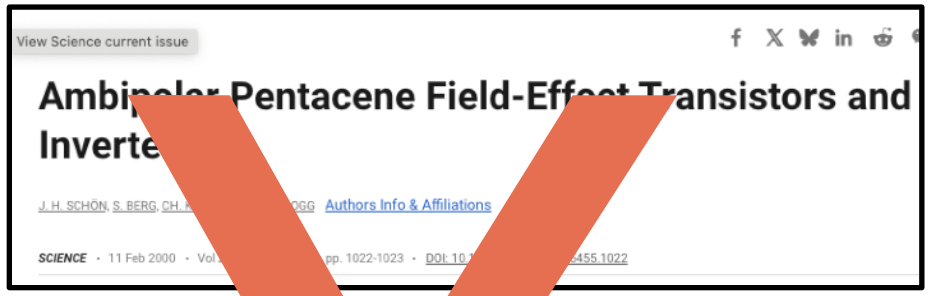
1 paper / 8 days.

Otto-Klung-Weberbank Prize for Physics 2001
Braunschweig Prize in 2001
Outstanding Young Investigator Award of the Materials Research Society 2002

(Keeping up scientific standards, Martin Blume, Nature 2009)

The Man Who Almost Faked His Way to a Nobel Prize

Jan Hendrik Schön



1 paper da

Otto-Klu berbank Prize fo ics 2001
Brauns g Prize in 2001
Outstanding Young Investigator Award of the Materials Research Society 2002

Faked results all along.

(Keeping up scientific standards, Martin Blume, Nature 2009)

What went wrong?

Factor 1

Factor 2

Outcome

What went wrong?

Factor 1

cannot validate the correctness of the discoveries.

Factor 2

Outcome

What went wrong?

Factor 1

cannot validate the correctness of the discoveries.

Factor 2

rewarded based on peer review.

Outcome

What went wrong?

Factor 1

cannot validate the correctness of the discoveries.

Factor 2

rewarded based on peer review.

Outcome

fakes discoveries that sound plausible.

LLMs Can Go Wrong

Factor 1

Factor 2

Outcome

LLMs Can Go Wrong

Factor 1

We sometimes fail to validate LLM correctness.

Factor 2

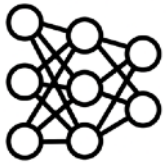
Outcome

LLMs Can Go Wrong

Factor 1

We sometimes fail to validate LLM correctness.

Factor 2



"I googled and read www.usa.gov. Including Washington, D.C.'s three electors, there are currently 270 electors in all." (Bai et al., arxiv'22)

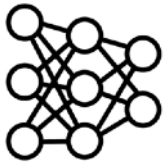
Outcome

LLMs Can Go Wrong

Factor 1

We sometimes fail to validate LLM correctness.

Factor 2



"I googled and read ww[illegible].gov. Including Washington, D.C.'s three [illegible]ectors, there are currently 270 electors i[illegible]" (Bai et al., arxiv'22)

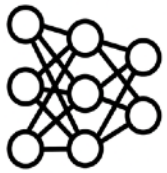
Outcome

LLMs Can Go Wrong

Factor 1

We sometimes fail to validate LLM correctness.

Factor 2



LLM is rewarded based on our validation.

(RLHF, Ouyang et al., NeurIPS'22)

Outcome

LLMs Can Go Wrong

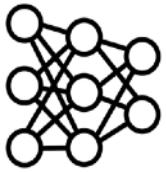
Factor 1

We sometimes fail to validate LLM correctness.

Factor 2

LLM is rewarded based on our validation.

Outcome



LLM generates plausible but incorrect outputs.

Language model learns to mislead humans via RLHF, Wen et al., ICLR'25

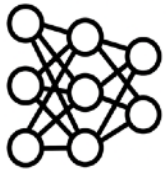
LLMs Can Go Wrong

Factor 1

We sometimes fail to validate LLM correctness.

Factor 2

Outcome



BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases.

(mata v avianca)

Validation will be Harder

Validation will be Harder

Factor 1

Our validations will diverge from correctness.

Validation will be Harder

Our validations will diverge from correctness.

How does gravity work?

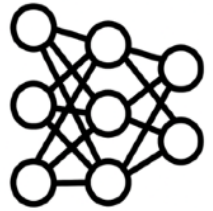
[100-page paper on **general relativity**.]

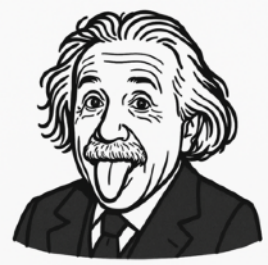
Validation will be Harder

Our validations will diverge from correctness.

How does gravity work?

[100-page paper on **general relativity.**]





or

?

My Research Goal

Enable LLMs to do difficult tasks

Concrete Problems in AI Safety, Amodei et al., arxiv'16

My Research Goal

Enable LLMs to do difficult tasks

requires

Validation on difficult tasks

(under-studied)

Concrete Problems in AI Safety, Amodei et al., arxiv'16

My Research Goal

Enable LLMs to do difficult tasks

requires

Validation on difficult tasks

(under-studied)

"Scalable" $\approx$ cheap

"Oversight" = validation

Concrete Problems in AI Safety, Amodei et al., arxiv'16

My Research Goal

Enable LLMs to do difficult tasks

requires

Validation on difficult tasks

(under-studied)

"Scalable" $\approx$ cheap

"Oversight" = validation

Important when AI optimizes against validations.

Concrete Problems in AI Safety, Amodei et al., arxiv'16

An Example of Scalable Oversight

An Example of Scalable Oversight

question q Find the students with most pets.

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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 Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

$$V(p) = 0/1?$$

Validation (reward/cost)

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

V^*: four programmers examine p for two hours

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

How to make a cheaper V' ?

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

Programmer

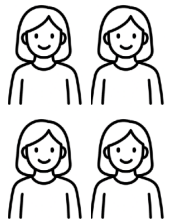
Non-programmer

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```



Programmer 4 persons

Non-programmer 1 person

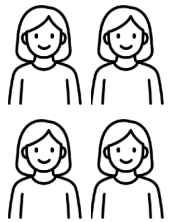
An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

Programmer
Non-programmer



4 persons
1 person

Human-in-the-loop
LLM

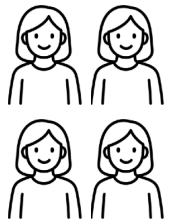
An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

Programmer
Non-programmer



4 persons
1 person

Human-in-the-loop
LLM

2 hrs
30 mins

An Example of Scalable Oversight

question q Find the students with most pets.

program p

```
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'
```

Approximate V^* with a validation method V' using less resource.

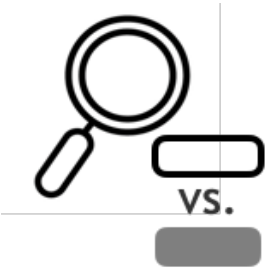
Outline

Task 1

Programming

Zhong et al., ACL'20
Zhong et al., EMNLP'20
Lai et al., ICML'22
Xie et al., EMNLP'22
Fried et al., ICLR'23
Zhong et al., EMNLP'23
Wen et al., ACL'24
Wen et al., ICLR'25

Task 2



Discovering dataset patterns

Zhong et al., ICML'22
Zhong et al., NeurIPS'23
Dunlap et al., CVPR'24
Chen et al., ICML'24
Sobhani et al., NAACL'25
Zhong et al., NeurIPS'24
Wang et al., EMNLP'23

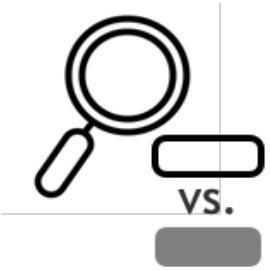
Outline

Task 1

Programming

no  needed

Task 2



Discovering dataset patterns

no  needed

Main Method: (De)Compose

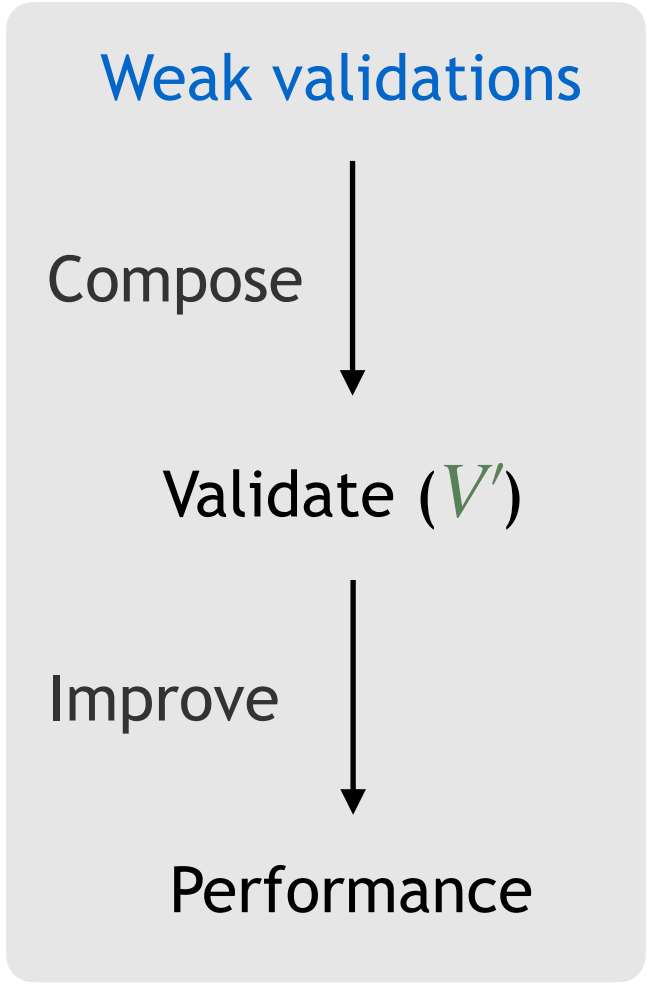
Main Method: (De)Compose

Weak validations

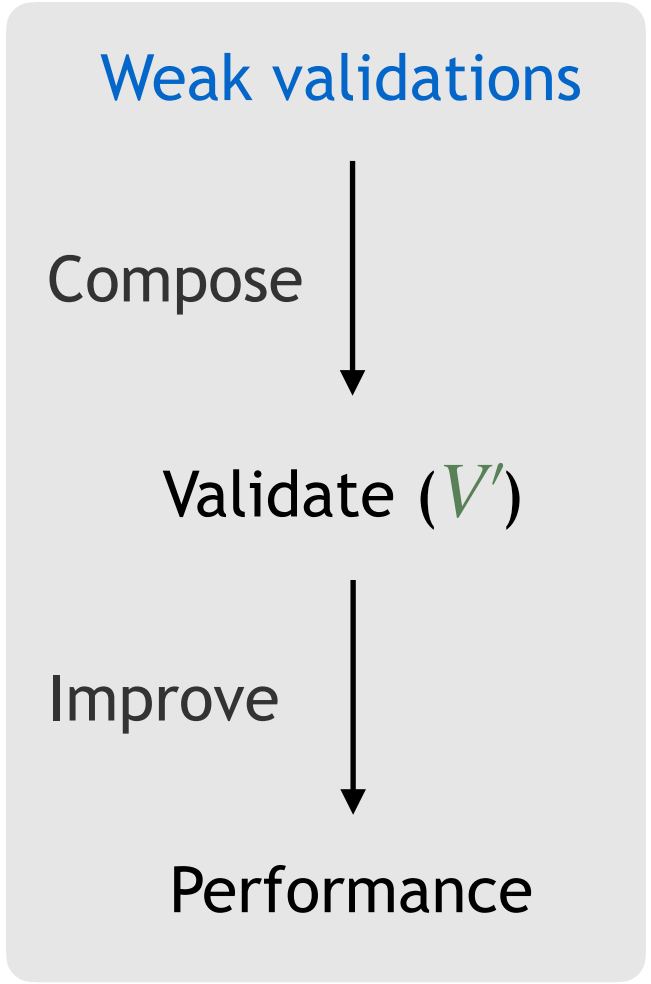
Compose

Validate (V')

Main Method: (De)Compose



Main Method: (De)Compose



Specific methods will change, general lessons remain.

Outline

Task 1

Programming

Task 2

Discovering
dataset patterns

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

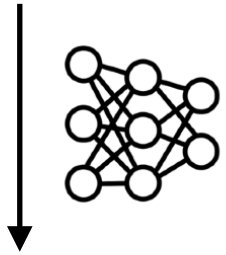
Weak validations

Compose

Validate (V')

Validating Complex Programs

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



p

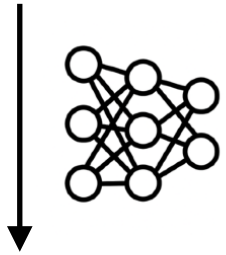
```
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'cat' INTERSECT
SELECT t1.fname FROM student AS t1 JOIN has_pet AS t2 ON t1.stuid = t2.stuid
JOIN pets AS t3 ON t3.petid = t2.petid WHERE t3.pettype = 'dog'
```

$$V(p) = 0/1?$$

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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



p

```
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'
```

$$V(p) = 0/1?$$

Experts read p: noisy & expensive

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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

p

```
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'
```

$$V(p) = 0/1?$$

Weaker validation on return values

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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

p

```
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'
```

$$V(p) = 0/1?$$

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

db

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

$p(db)$

{Alex}

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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

p

```
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'
```

$V(p) = 0/1?$

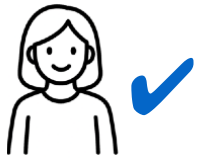
Find the first name of students who have both cat and dog pets.

db

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

$p(db)$

{Alex}



(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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

p

```
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'
```

$V(p) = 0/1?$

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

db

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

$p(db)$

{Alex}

$w(q, db, p(db)) = 1$

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Validating Complex Programs

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

p

```
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'
```

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

db

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

$p(db)$

{Alex}

$$V(p) = 0/1?$$

$$w(q, db, p(db)) = 1$$ weak

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23; Programming by Example, Lieberman, 2001)

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$

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

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat
Ruiqi	Zhong	Dog
Alex	Wei	Dog
Alex	Pan	Cat
Jacob	Klein	Dog
Lizzi	Yin	Dog
Kathy	Li	Rat
Kevin	Lin	Rat
1000 rows omitted.		

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

ABC	Wei	Cat
Meena	DEF	Dog

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

Cathy	Wei	Cat
Meena	Yin	Dog

Nick	Klein	Cat
Dan	Wang	Dog

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

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat
Ruiqi	Zhong	Dog
Alex	Wei	Dog
Alex	Pan	Cat
Jacob	Klein	Dog
Lizzi	Yin	Dog
Kathy	Li	Rat
Kevin	Lin	Rat
1000 rows omitted.		

{Lizzi}

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat

{}

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

{}

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

{}

ABC	Wei	Cat
Meena	DEF	Dog

{}

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

{Alex}

Cathy	Wei	Cat
Meena	Yin	Dog

{}

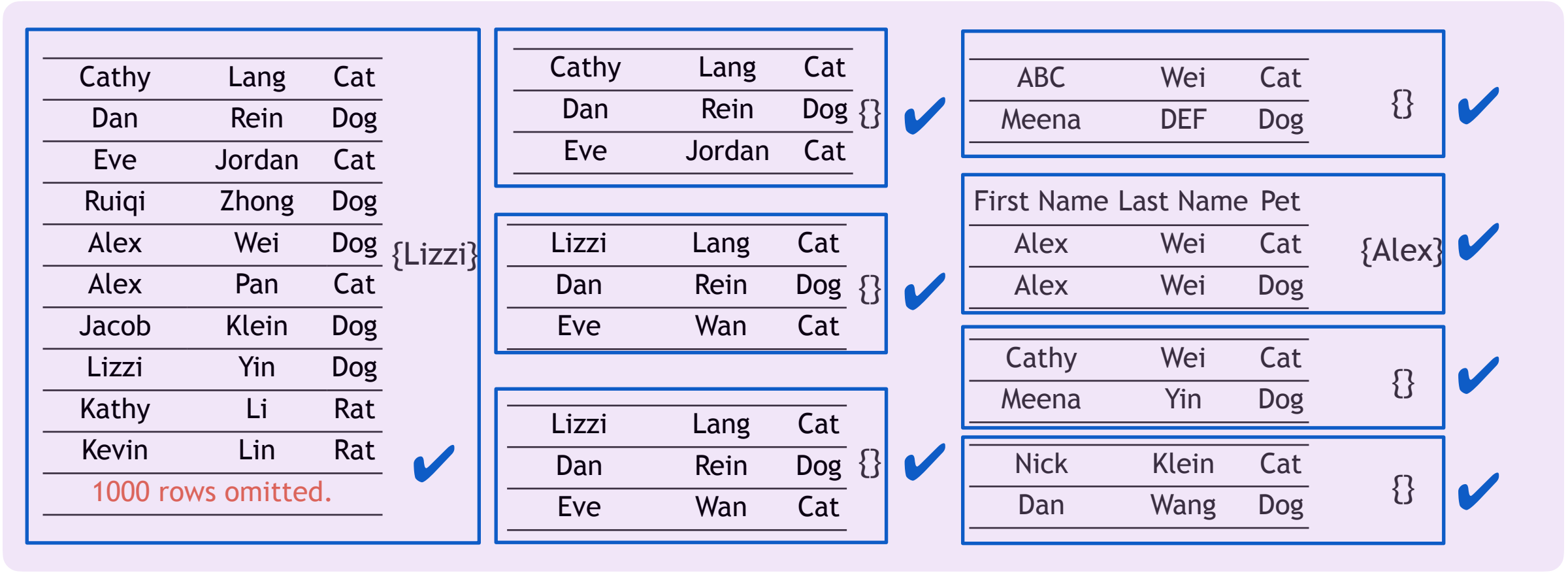
Nick	Klein	Cat
Dan	Wang	Dog

{}

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

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$



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

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat
Ruiqi	Zhong	Dog
Alex	Wei	Dog
Alex	Pan	Cat
Jacob	Klein	Dog
Lizzi	Yin	Dog
Kathy	Li	Rat
Kevin	Lin	Rat
1000 rows omitted.		

{Lizzi} ✔

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat

{} ✔

ABC	Wei	Cat
Meena	DEF	Dog

{} ✔

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

{} ✔

First Name	Last Name	Pet
Alex	Wei	Cat
Alex	Wei	Dog

{Alex} ✔

Lizzi	Lang	Cat
Dan	Rein	Dog
Eve	Wan	Cat

{} ✔

Cathy	Wei	Cat
Meena	Yin	Dog

{} ✔

Nick	Klein	Cat
Dan	Wang	Dog

{} ✔

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

Composing Weak Validations

$$V'(p) \approx \cap_{db \in S_{db}} w(q, db, p(db))$$

Not scalable!

"I dedicated my entire life to reading databases about cats and dogs. "

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

Pick the set of databases (weak validations) carefully.

(LLMs in 2022 could not automate this)

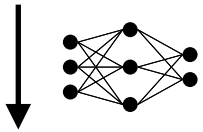
Expose the Differences: Plausible vs. Correct

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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



p_1

```
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 ……
```

p_2

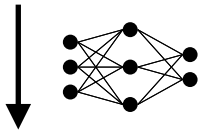
```
SELECT fname FROM Student WHERE
StuID IN ( SELECT T1.stuid FROM student
AS T1 JOIN has_pet AS T2 ……
```

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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

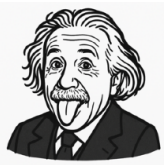


p_1

```
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 ……
```

p_2

```
SELECT fname FROM Student WHERE
StuID IN ( SELECT T1.stuid FROM student
AS T1 JOIN has_pet AS T2 ……
```

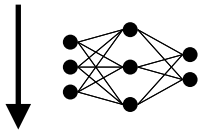


(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

p_2

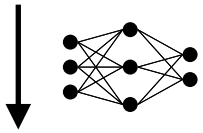
```
SELECT fname FROM Student WHERE
StuID IN ( SELECT T1.stuid FROM student
AS T1 JOIN has_pet AS T2 ……
```

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

p_2

```
SELECT fname FROM Student WHERE
StuID IN ( SELECT T1.stuid FROM student
AS T1 JOIN has_pet AS T2 ……
```

find db s.t. $p_1(db) \neq p_2(db)$

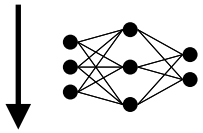
(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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

First Name	Last Name	Pet
Alex	Pan	Cat
Alex	Wei	Dog



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

{Alex}

p_2

```
SELECT fname FROM Student WHERE
StuID IN ( SELECT T1.stuid FROM student
AS T1 JOIN has_pet AS T2 ……
```

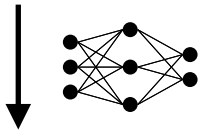
{}

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

First Name	Last Name	Pet
Alex	Pan	Cat
Alex	Wei	Dog

{Alex}

✘

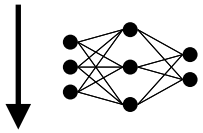
(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Expose the Differences: Plausible vs. Correct

q

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

First Name	Last Name	Pet
Alex	Pan	Cat
Alex	Wei	Dog



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

{Alex}

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

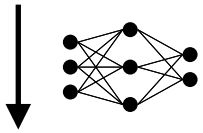
Expose the Differences: Plausible vs. Correct

q

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

rare

First Name	Last Name	Pet
Alex	Pan	Cat
Alex	Wei	Dog



p_1

```
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 ……
```

- first names of cat owners
- first names of dog owners
- intersect

{Alex}

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1 SELECT fname FROM Student WHERE StuID IN ……

p_2 SELECT t1.fname FROM student AS t1 JOIN has_pet ……

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1 SELECT fname FROM Student WHERE StuID IN .....

p_2 SELECT t1.fname FROM student AS t1 JOIN has_pet .....

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat
Ruiqi	Zhong	Dog
Alex	Wei	Dog
Alex	Pan	Cat
Jacob	Klein	Dog
Lizzi	Yin	Dog
Kathy	Li	Rat
Kevin	Lin	Rat
1000 rows omitted.		

Step 1 fuzzing: generating large databases s.t. $p_1(db) \neq p_2(db)$

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1

SELECT fname FROM Student WHERE StuID IN .....

p_2

SELECT t1.fname FROM student AS t1 JOIN has_pet .....

Cathy	Lang	Cat
Dan	Rein	Dog
Eve	Jordan	Cat
Ruiqi	Zhong	Dog
Alex	Wei	Dog
Alex	Pan	Cat
Jacob	Klein	Dog
Lizzi	Yin	Dog
Kathy	Li	Rat
Kevin	Lin	Rat
1000 rows omitted.		

Step 1 fuzzing: generating large databases s.t. $p_1(db) \neq p_2(db)$

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1 SELECT fname FROM Student WHERE StuID IN .....

p_2 SELECT t1.fname FROM student AS t1 JOIN has_pet .....

Alex	Wei	Dog
Alex	Pan	Cat

Step 1 fuzzing: generating large databases s.t. $p_1(db) \neq p_2(db)$

Step 2 delta-debugging: dropping rows from db

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1	0.40	SELECT fname FROM Student WHERE StuID IN .....
p_2	0.20	SELECT t1.fname FROM student AS t1 JOIN has_pet .....
p_3	0.13	SELECT t2.fname FROM .....
		.....
p_n	0.01	SELECT t1.fname FROM .....

Alex	Wei	Dog
Alex	Pan	Cat

handle a **distribution of SQLs** sampled from LLM (our work)

Step 1 fuzzing: generating large databases s.t. $p_1(db) \neq p_2(db)$

Step 2 delta-debugging: dropping rows from db

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Find db s.t. $p_1(db) \neq p_2(db)$ for Weak Validation

p_1	0.40	SELECT fname FROM Student WHERE StuID IN .....
p_2	0.20	SELECT t1.fname FROM student AS t1 JOIN has_pet .....
p_3	0.13	SELECT t2.fname FROM .....
		.....
p_n	0.01	SELECT t1.fname FROM .....

Alex	Wei	Dog
Alex	Pan	Cat

handle a **distribution of SQLs** sampled from LLM (our work)

optimize $H[p(db)]$

Step 1 fuzzing: generating large databases s.t. $p_1(db) \neq p_2(db)$

Step 2 delta-debugging: dropping rows from db

Barton P. Miller, Louis Fredriksen, and Bryan So. 1990. An empirical study of the reliability of UNIX utilities.

Computing $V'(p)$: A Recap

q

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

p

```
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' ……
```

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Computing $V'(p)$: A Recap

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

p

```
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' ……
```

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Computing $V'(p)$: A Recap

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

p 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' .....

S_{db} (size ~= 2)

$db_1 =$

Cathy	Klein	Cat	{Cathy}
Cathy	Klein	Dog	

$db_2 =$

Alex	Pan	Cat	{Alex}
Alex	Wei	Dog	

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Computing $V'(p)$: A Recap

q

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

p

```
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' .....
```

✔ $db_1 =$

Cathy	Klein	Cat	{Cathy}
Cathy	Klein	Dog	

✘ $db_2 =$

Alex	Pan	Cat	{Alex}
Alex	Wei	Dog	

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Computing $V'(p)$: A Recap

q

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

p

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' .....

✔

$db_1 =$

Cathy	Klein	Cat	{Cathy}
Cathy	Klein	Dog	

✘

$db_2 =$

Alex	Pan	Cat	{Alex}
Alex	Wei	Dog	

$$V'(p) = 0$$

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Computing $V'(p)$: A Recap

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

p 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' .....

✔

$db_1 =$

Cathy	Klein	Cat	{Cathy}
Cathy	Klein	Dog	

✘

$db_2 =$

Alex	Pan	Cat	{Alex}
Alex	Wei	Dog	

V' does not require

(Explaining wrong queries using small examples, Miao et al., SIGMOD'19; Cosette: An Automated SQL Solver, Chu et al., CIDR'17)

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

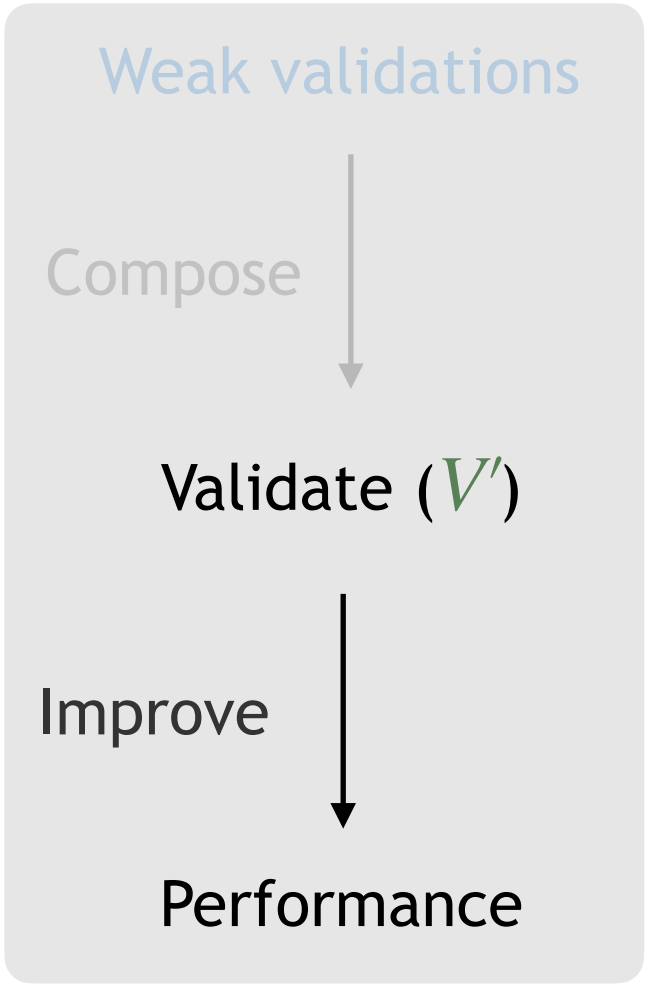
Outline

Task 1

Programming

Task 2

Discovering dataset patterns



Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

Improve

Performance

V^* established by 4 coders

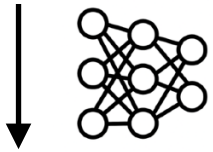
Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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



LLM (Codex, chen et al., arxiv'22)

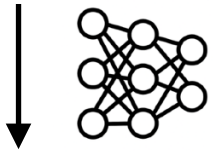
p_1

SELECT fname FROM Student WHERE……

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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



LLM (Codex, chen et al., arxiv'22)

p_1

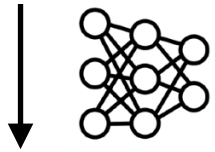
SELECT fname FROM Student WHERE……

59% of the time correct

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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



p_1 SELECT fname FROM Student WHERE.....

p_2 SELECT StuID FROM Student WHERE .....

p_3 SELECT fname FROM Student WHERE .....

.....

p_{16} SELECT lname .....

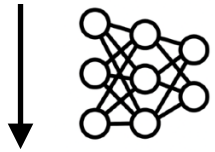
59% of the time correct

88% of the time contain the correct program

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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



p_1 SELECT fname FROM Student WHERE.....

p_2 SELECT StuID FROM Student WHERE .....

p_3 SELECT fname FROM Student WHERE .....

.....

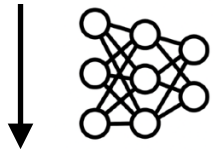
p_{16} SELECT lname .....

Oracle validation:
59% → 88%

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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

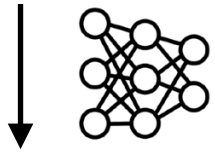


p_1	SELECT fname FROM Student WHERE.....	$V'(p)$=0
p_2	SELECT StuID FROM Student WHERE	$V'(p)$=0
p_3	SELECT fname FROM Student WHERE	$V'(p)$=1
	.....	
p_{16}	SELECT lname	$V'(p)$=0

Use $V'(p)$ to Remove Wrong Programs

SPIDER Text-to-SQL
(Yu et al., EMNLP'18)

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



p_1	SELECT fname FROM Student WHERE.....	$V'(p)$=0
p_2	SELECT StuID FROM Student WHERE	$V'(p)$=0
p_3	SELECT fname FROM Student WHERE	$V'(p)$=1
	.....	
p_{16}	SELECT lname	$V'(p)$=0

Accuracy of Generating SQL

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	

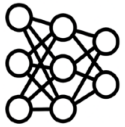
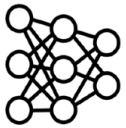
(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	
LLM + non-expert validation	75%	 + 

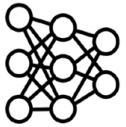
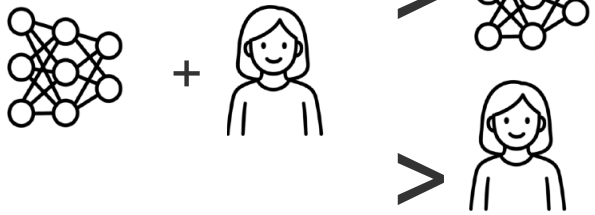
(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	
LLM + non-expert validation	75%	 + 
Ceiling	88%	

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	
LLM + non-expert validation	75%	

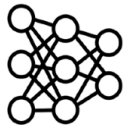
(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	
LLM + non-expert validation	75%	

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

Accuracy of Generating SQL

Expert-alone	75%	
Non-expert alone	0%	
LLM-alone	59%	
LLM + non-expert validation	75%	 + 
Ceiling	**88%**	(higher in the future)

(Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL, Zhong et al., EMNLP'23)

What’s still relevant in 2025?

What’s still relevant in 2025?

- Find places where human and AI are complementary.

What’s still relevant in 2025?

- Find places where human and AI are complementary.
- Minimize efforts required for humans

What’s still relevant in 2025?

- Find places where human and AI are complementary.
- Minimize efforts required for humans
- Expose differences between LLM samples.

Outline

Task 1

Programming

Task 2

Discovering
dataset patterns

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose ↓

Validate (V')

Improve ↓

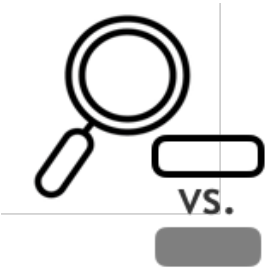
Performance

Outline

Task 1

Programming

Task 2



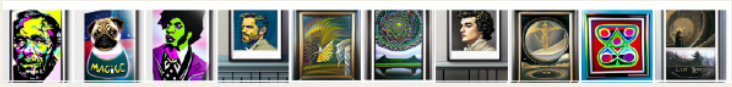
Discovering
dataset patterns

LLMs Can Discover Dataset Patterns

LLMs Can Discover Dataset Patterns

How are they different?

Training set



Test set

LLMs Can Discover Dataset Patterns

How are they different?

Training set

Test set

Here are 1M customer reviews. Categorize them.

amazon

LLMs Can Discover Dataset Patterns

How are they different?

Training set

Test set

Here are 1M customer reviews. Categorize them.

amazon

How do search query change over time?

Google Trends

"how to r rpet"
"flu symp rison"
"best rest
"iphone 1 vs"
"early signs of influenza in adults"
"how to merge cells in excel spreadsheet"
"real estate market trends 2025"
"fever and body aches remedies"
"quick healthy dinner recipes 30 minutes"
"how long does seasonal flu last"
"best credit cards for travel rewards 2025" “
what time does target pharmacy close"
"playstation 6 release date rumors"
"best time to book international flights"
"symptoms of flu in children and toddlers"

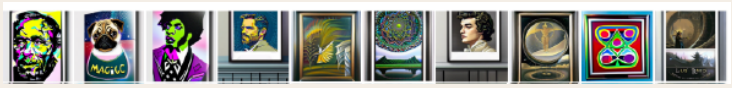
LLMs Can Discover Dataset Patterns

How are they different?

Here are 1M customer reviews. Categorize them.

How do search query change over time?

Training set



Test set

amazon

Google Trends

"how to r rpet"
"flu symp rison"
"best rest
"iphone 1 ws"
"early signs of influenza in adults"
"how to merge cells in excel spreadsheet"
"real estate market trends 2025"
"fever and body aches remedies"
"quick healthy dinner recipes 30 minutes"
"how long does seasonal flu last"
"best credit cards for travel rewards 2025" “
what time does target pharmacy close"
"playstation 6 release date rumors"
"best time to book international flights"
"symptoms of flu in children and toddlers"

Task Definition

Task Input

S_A	S_B
桌上有钱。	I love soccer.
他去学校了。	He went to school.
餐巾纸给我下。	The weather is nice.
谢谢。	I am learning Chinese.
这动画好看。	She is reading a book.

(Describing Differences between Text Distributions with Natural Language, Zhong et al., ICML'22)

Task Definition

Task Input

S_A	S_B
桌上有钱。	I love soccer.
他去学校了。	He went to school.
餐巾纸给我下。	The weather is nice.
谢谢。	I am learning Chinese.
这动画好看。	She is reading a book.

Discover what's more common in S_B than in S_A?

(Describing Differences between Text Distributions with Natural Language, Zhong et al., ICML'22)

Task Definition

Task Input

S_A	S_B
桌上有钱。	I love soccer.
他去学校了。	He went to school.
餐巾纸给我下。	The weather is nice.
谢谢。	I am learning Chinese.
这动画好看。	She is reading a book.

Discover what's more common in S_B than in S_A?

Task Output

d "is in English"

(Describing Differences between Text Distributions with Natural Language, Zhong et al., ICML'22)

Task Definition

Task Input

S_A

桌上有钱。
他去学校了。
餐巾纸给我下。
谢谢。
这动画好看。

S_B

I love soccer.
He went to school.
The weather is nice.
I am learning Chinese.
She is reading a book.

Discover what's more common in S_B than in S_A?

Task Output

d

"is in English" ✔

"is not in Chinese" ✔

"is written in English" ✔

(Describing Differences between Text Distributions with Natural Language, Zhong et al., ICML'22)

Task Definition

S_A

桌上有钱。
他去学校了。
餐巾纸给我下。
I am learning Chinese.
这动画好看。

S_B

I love soccer.
He went to school.
The weather is nice.
She is reading a book.
谢谢。

d “is in English”

Warmup

S_A

The moon shines brightly tonight.
Birds soar through cloudy skies.
Fresh coffee fills the kitchen.
Waves crash against rocky shores.
Autumn leaves drift gently down.

S_B

I have 2 cats at home.
The temperature is 75 degrees today.
She won 1st place in the competition.
There are 365 days in a year.
He scored 100 points in the game.

Warmup

S_A

The moon shines brightly tonight.
Birds soar through cloudy skies.
Fresh coffee fills the kitchen.
Waves crash against rocky shores.
Autumn leaves drift gently down.

S_B

I have 2 cats at home.
The temperature is 75 degrees today.
She won 1st place in the competition.
There are 365 days in a year.
He scored 100 points in the game.

“contains numbers”

What's the difference?

S_A

Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.
High above the stage, the chandelier refracts light in a thousand different directions, amazing the audience.
The paper had more than 20 authors; it's a huge collaboration.
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.
A delicate origami crane seemed ready to take flight in her hand.
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.
Stacked neatly on the shelf, the leather-bound books carry the scent of time and wisdom.
A single drop of rain landed on the windowsill, signaling the storm's arrival.

S_B

The children ran around the park, laughing as they played hide-and-seek behind the tall tree, their voices echoing in the breeze.
My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine
There is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.
He left the circuit boards exposed, waiting to be tested.
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.
The kid really hates eating broccoli so I need to talk to her a lot.
The cat curled up in the sunbeam, purring softly as it dozed.

What's the difference?

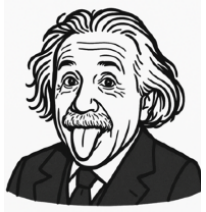
S_A

Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.
High above the stage, the chandelier refracts light in a thousand different directions, amazing the audience.
The paper had more than 20 authors; it's a huge collaboration.
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.
A delicate origami crane seemed ready to take flight in her hand.
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.
Stacked neatly on the shelf, the leather-bound books carry the scent of time and wisdom.
A single drop of rain landed on the windowsill, signaling the storm's arrival.

S_B

The children ran around the park, laughing as they played hide-and-seek behind the tall tree, their voices echoing in the breeze.
My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine
There is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.
He left the circuit boards exposed, waiting to be tested.
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.
The kid really hates eating broccoli so I need to talk to her a lot.
The cat curled up in the sunbeam, purring softly as it dozed.

"contains a green object"



Validation is Hard

S_A

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.
A single red balloon floated into the sky, carrying a child's wish with it.
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.
The wooden swing creaked as it swayed gently in the evening breeze.
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.
A half-written letter lay on the desk, the ink smudged by hurried hands.
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.
The golden retriever dozed by the fireplace, its tail twitching in a dream.
The baker's hands worked skillfully, kneading the dough into soft perfection.
The children love running barefoot on the soft grass in the backyard.
The old map, edges frayed and corners curled, hinted at forgotten adventures.
A trail of footprints disappeared into the mist, leaving behind a mystery.
The carousel spun in slow circles, its painted horses frozen mid-gallop.
The lantern flickered in the night, guiding lost travelers home.
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.
The abandoned bicycle leaned against the fence, its tires flat with time.
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.
A single drop of rain landed on the windowsill, signaling the storm's arrival.

S_B

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
He left the circuit boards exposed, waiting to be tested.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine
A delicate origami crane seemed ready to take flight in her hand.
The chessboard sat untouched, the pieces frozen in the middle of a battle.
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.
The paper had more than 20 authors; it's a huge collaboration.
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.
A single candle flickered in the dark, casting dancing shadows on the walls.
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
The delicate snowflake melted the moment it touched her warm palm.
The kid really hates eating broccoli so I need to talk to her a lot.
The painter's palette held a swirl of colors, waiting to become something beautiful.
The ivy crept up the old stone wall, weaving its way through the cracks.

d "contains a green object"

Validation is Hard

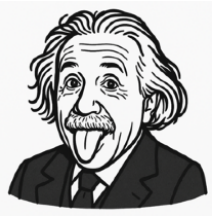
S_A

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.
A single red balloon floated into the sky, carrying a child's wish with it.

S_B

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
He left the circuit boards exposed, waiting to be tested.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine
A delicate origami crane seemed ready to take flight in her hand.

d "contains a green object"



Validation is Hard

S_A

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.
A single red balloon floated into the sky, carrying a child's wish with it.
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.
The wooden swing creaked as it swayed gently in the evening breeze.
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.
A half-written letter lay on the desk, the ink smudged by hurried hands.
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.
The golden retriever dozed by the fireplace, its tail twitching in a dream.
The baker's hands worked skillfully, kneading the dough into soft perfection.
The children love running barefoot on the soft grass in the backyard.
The old map, edges frayed and corners curled, hinted at forgotten adventures.
A trail of footprints disappeared into the mist, leaving behind a mystery.
The carousel spun in slow circles, its painted horses frozen mid-gallop.
The lantern flickered in the night, guiding lost travelers home.
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.
The abandoned bicycle leaned against the fence, its tires flat with time.
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.
A single drop of rain landed on the windowsill, signaling the storm's arrival.

S_B

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
He left the circuit boards exposed, waiting to be tested.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine.
A delicate origami crane seemed ready to take flight in her hand.
The chessboard sat untouched, the pieces frozen in the middle of a battle.
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.
The paper had more than 20 authors; it's a huge collaboration.
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.
A single candle flickered in the dark, casting dancing shadows on the walls.
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
The delicate snowflake melted the moment it touched her warm palm.
The kid really hates eating broccoli so I need to talk to her a lot.
The painter's palette held a swirl of colors, waiting to become something beautiful.
The ivy crept up the old stone wall, weaving its way through the cracks.

d "contains a green object"

Applications

Applications

Trump’s speech 2016 vs. Trump’s speech 2024: discusses about government efficiency.

Applications

Trump's speech 2016 vs. Trump's speech 2024: discusses about government efficiency.

patients' comments after taking drug A vs. drug B: mentions feelings of paranoid.

Applications

Trump's speech 2016 vs. Trump's speech 2024: discusses about government efficiency.

patients' comments after taking drug A vs. drug B: mentions feelings of paranoid.

what activates a neuron vs. what does not activate a neuron: contains religious concepts.

Applications

Trump's speech 2016 vs. Trump's speech 2024: discusses about government efficiency.

patients' comments after taking drug A vs. drug B: mentions feelings of paranoid.

what activates a neuron vs. what does not activate a neuron: contains religious concepts.

unpopular tweets vs. popular tweets: engages in controversial topics.

images from StableDiffusionV1 vs. images from StableDiffusionV2: have more vibrant colors.

ChatGPT's outputs vs. Claude's output: is more human-like.

Applications

Trump's speech 2016 vs. Trump's speech 2024: discusses about government efficiency.

patients' comments after taking drug A vs. drug B: mentions feelings of paranoid.

what activates a neuron vs. what does not activate a neuron: contains religious concepts.

unpopular tweets vs. popular tweets: engages in controversial topics.

images from StableDiffusionV1 vs. images from StableDiffusionV2: have more vibrant colors.

ChatGPT's outputs vs. Claude's output: is more human-like.

675 applications

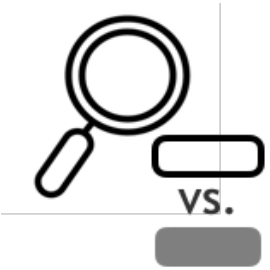
(D5, Zhong et al., NeurIPS'23)

Outline

Task 1

Programming

Task 2



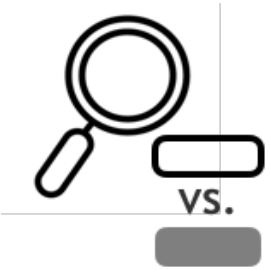
Discovering
dataset patterns

Outline

Task 1

Programming

Task 2



Discovering dataset patterns

Weak validations

Compose

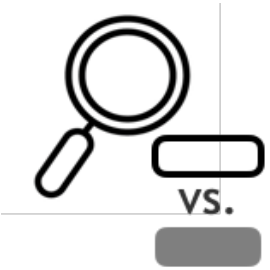
Validate (V')

Outline

Task 1

Programming

Task 2



Discovering dataset patterns

Weak validations

Compose

Validate (V')

V' does not need humans

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
……

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
……

d "includes a green object"

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
……

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
……

$V(d) = ?$

d “includes a green object”

Validating Dataset-level Discoveries

S_A	S_B
I love eating broccoli.	The ivy climbed up the wall.
I was at the dinner table.	He played football.
The swan is flying.	He bought the circuit boards.
……	……

d "includes a green object"

Annotators/LLMs read the entire datasets?

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
.....

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
.....

d "includes a green object"

d: "includes a green object"

x : I love eating broccoli. 

✔

d: "includes a green object object"

x: He played football. 

✘

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
.....

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
.....

d "includes a green object"

d: "includes a green object"

x : I love eating broccoli. 

✔

$w(d, x) = 1$

d: "includes a green object object"

x: He played football. 

✘

$w(d, x) = 0$

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
.....

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
.....

d "includes a green object"

d: "includes a green object"

x : I love eating broccoli.

✔

$w(d, x) = 1$

d: "includes a green object object"

x: He played football.

✘

$w(d, x) = 0$

Validating Dataset-level Discoveries

S_A

I love eating broccoli.
I was at the dinner table.
The swan is flying.
.....

S_B

The ivy climbed up the wall.
He played football.
He bought the circuit boards.
.....

d "includes a green object"

d: "includes a green object"

x : I love eating broccoli.

✔

$w(d, x) = 1$

d: "includes a green object object"

x: He played football.

✘

$w(d, x) = 0$

weak

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.
A single red balloon floated into the sky, carrying a child's wish with it.
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.
The wooden swing creaked as it swayed gently in the evening breeze.
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.
A half-written letter lay on the desk, the ink smudged by hurried hands.
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.
The golden retriever dozed by the fireplace, its tail twitching in a dream.
The baker's hands worked skillfully, kneading the dough into soft perfection.
The children love running barefoot on the soft grass in the backyard.
The old map, edges frayed and corners curled, hinted at forgotten adventures.
A trail of footprints disappeared into the mist, leaving behind a mystery.
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.
The lantern flickered in the night, guiding lost travelers home.
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.
He left the circuit boards exposed, waiting to be tested.
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine
A delicate origami crane seemed ready to take flight in her hand.
The chessboard sat untouched, the pieces frozen in the middle of a battle.
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.
The paper had more than 20 authors; it's a huge collaboration.
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.
A single candle flickered in the dark, casting dancing shadows on the walls.
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.
The delicate snowflake melted the moment it touched her warm palm.
The kid really hates eating broccoli so I need to talk to her a lot.
The painter's palette held a swirl of colors, waiting to become something beautiful.
The ivy crept up the old stone wall, weaving its way through the cracks.

d "includes a green object"

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.	0
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.	0
A single red balloon floated into the sky, carrying a child's wish with it.	0
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.	0
The wooden swing creaked as it swayed gently in the evening breeze.	0
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.	1
A half-written letter lay on the desk, the ink smudged by hurried hands.	0
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.	1
The golden retriever dozed by the fireplace, its tail twitching in a dream.	0
The baker's hands worked skillfully, kneading the dough into soft perfection.	0
The children love running barefoot on the soft grass in the backyard.	1
The old map, edges frayed and corners curled, hinted at forgotten adventures.	0
A trail of footprints disappeared into the mist, leaving behind a mystery.	0
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.	1
The lantern flickered in the night, guiding lost travelers home.	0
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.	0

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.	1
He left the circuit boards exposed, waiting to be tested.	1
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine	1
A delicate origami crane seemed ready to take flight in her hand.	0
The chessboard sat untouched, the pieces frozen in the middle of a battle.	0
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.	0
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.	1
The paper had more than 20 authors; it's a huge collaboration.	0
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.	1
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.	0
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.	0
A single candle flickered in the dark, casting dancing shadows on the walls.	1
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.	0
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
The delicate snowflake melted the moment it touched her warm palm.	1
The kid really hates eating broccoli so I need to talk to her a lot.	0
The painter's palette held a swirl of colors, waiting to become something beautiful.	0
The ivy crept up the old stone wall, weaving its way through the cracks.	1

d "includes a green object"

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.	0
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.	0
A single red balloon floated into the sky, carrying a child's wish with it.	0
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.	0
The wooden swing creaked as it swayed gently in the evening breeze.	0
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.	1
A half-written letter lay on the desk, the ink smudged by hurried hands.	0
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.	1
The golden retriever dozed by the fireplace, its tail twitching in a dream.	0
The baker's hands worked skillfully, kneading the dough into soft perfection.	0
The children love running barefoot on the soft grass in the backyard.	1
The old map, edges frayed and corners curled, hinted at forgotten adventures.	0
A trail of footprints disappeared into the mist, leaving behind a mystery.	0
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.	1
The lantern flickered in the night, guiding lost travelers home.	0
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.	0

0.23

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.	1
He left the circuit boards exposed, waiting to be tested.	1
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine	1
A delicate origami crane seemed ready to take flight in her hand.	0
The chessboard sat untouched, the pieces frozen in the middle of a battle.	0
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.	0
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.	1
The paper had more than 20 authors; it's a huge collaboration.	0
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.	1
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.	0
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.	0
A single candle flickered in the dark, casting dancing shadows on the walls.	1
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.	0
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
The delicate snowflake melted the moment it touched her warm palm.	1
The kid really hates eating broccoli so I need to talk to her a lot.	0
The painter's palette held a swirl of colors, waiting to become something beautiful.	0
The ivy crept up the old stone wall, weaving its way through the cracks.	1

0.24

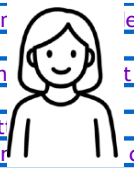
$$V^*(d) = +0.01$$

d "includes a green object"

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.	0
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.	0
A single red balloon floated into the sky, carrying a child's wish with it.	0
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.	0
The wooden swing creaked as it swayed gently in the evening breeze.	0
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.	1
A half-written letter lay on the desk, the ink smudged by hurried hands.	0
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.	1
The golden retriever dozed by the fireplace, its tail twitching in a dream.	0
The baker's hands worked skillfully, kneading the dough into soft perfection.	0
The children love running barefoot on the soft grass in the backyard.	1
The old map, edges frayed and corners curled, hinted at forgotten adventures.	0
A trail of footprints disappeared into the mist, leaving behind a mystery.	0
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.	1
The lantern flickered in the night, guiding lost travelers home.	0
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.	0



My friend paused in reverent silence upon seeing the exquisite jade shir er the gallery lights.	1
He left the circuit boards exposed, waiting to be tested.	1
Whenever the visitors enter the office, the first thing they see is the An t figurine	1
A delicate origami crane seemed ready to take flight in her hand.	0
The chessboard sat untouched, the pieces frozen in the middle of a bat	0
A lone firefly drifted through the night, its tiny light blinking like a hear dark.	0
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.	1
The paper had more than 20 authors; it's a huge collaboration.	0
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.	1
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.	0
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.	0
A single candle flickered in the dark, casting dancing shadows on the walls.	1
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.	0
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
The delicate snowflake melted the moment it touched her warm palm.	1
The kid really hates eating broccoli so I need to talk to her a lot.	0
The painter's palette held a swirl of colors, waiting to become something beautiful.	0
The ivy crept up the old stone wall, weaving its way through the cracks.	1

d "includes a green object"

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$

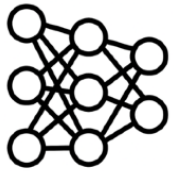
Not scalable!

“I wasted my entire life looking for green objects.”

d “includes a green object”

Composing Weak Validations

$$V^*(d) = \mathbb{E}_{x \sim S_B}[w(d,x)] - \mathbb{E}_{x \sim S_A}[w(d,x)]$$



"I trust my LM to automate it (w')"

d "includes a green object"

Composing Weak Validations

$$V'(d) = \mathbb{E}_{x \sim S_B}[w'(d,x)] - \mathbb{E}_{x \sim S_A}[w'(d,x)]$$

V' does not need humans

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.	0
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.	0
A single red balloon floated into the sky, carrying a child's wish with it.	0
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.	0
The wooden swing creaked as it swayed gently in the evening breeze.	0
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.	1
A half-written letter lay on the desk, the ink smudged by hurried hands.	0
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.	1
The golden retriever dozed by the fireplace, its tail twitching in a dream.	0
The baker's hands worked skillfully, kneading the dough into soft perfection.	0
The children love running barefoot on the soft grass in the backyard.	1
The old map, edges frayed and corners curled, hinted at forgotten adventures.	0
A trail of footprints disappeared into the mist, leaving behind a mystery.	0
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.	1
The lantern flickered in the night, guiding lost travelers home.	0
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.	0

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.	1
He left the circuit boards exposed, waiting to be tested.	1
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine	1
A delicate origami crane seemed ready to take flight in her hand.	0
The chessboard sat untouched, the pieces frozen in the middle of a battle.	0
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.	0
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.	1
The paper had more than 20 authors; it's a huge collaboration.	0
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.	1
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.	0
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.	0
A single candle flickered in the dark, casting dancing shadows on the walls.	1
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.	0
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
The delicate snowflake melted the moment it touched her warm palm.	1
The kid really hates eating broccoli so I need to talk to her a lot.	0
The painter's palette held a swirl of colors, waiting to become something beautiful.	0
The ivy crept up the old stone wall, weaving its way through the cracks.	1

d "includes a green object"

Composing Weak Validations

$$V'(d) = \mathbb{E}_{x \sim S_B}[w'(d,x)] - \mathbb{E}_{x \sim S_A}[w'(d,x)]$$

The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
A forgotten toy truck sat buried in the sandbox, its wheels caked with dried mud.	0
The typewriter's keys, worn from years of use, clicked rhythmically under her fingers.	0
A single red balloon floated into the sky, carrying a child's wish with it.	0
The violinist's bow glided effortlessly across the strings, filling the air with a haunting melody.	0
The wooden swing creaked as it swayed gently in the evening breeze.	0
I saw that there is nothing on the floor, except for a laptop, some dollar bills, and some paperclips.	1
A half-written letter lay on the desk, the ink smudged by hurried hands.	0
The children ran around the park, laughing as they played hide-and-seek behind the tall tree.	1
The golden retriever dozed by the fireplace, its tail twitching in a dream.	0
The baker's hands worked skillfully, kneading the dough into soft perfection.	0
The children love running barefoot on the soft grass in the backyard.	1
The old map, edges frayed and corners curled, hinted at forgotten adventures.	0
A trail of footprints disappeared into the mist, leaving behind a mystery.	0
Whenever the visitors enter the kitchen, the first thing they see is the avocado on the counter.	1
The lantern flickered in the night, guiding lost travelers home.	0
High above the stage, the chandelier refracted light in a thousand different directions, amazing the audience.	0

My friend paused in reverent silence upon seeing the exquisite jade shimmering under the gallery lights.	1
He left the circuit boards exposed, waiting to be tested.	1
Whenever the visitors enter the office, the first thing they see is the Android mascot figurine	1
A delicate origami crane seemed ready to take flight in her hand.	0
The chessboard sat untouched, the pieces frozen in the middle of a battle.	0
A lone firefly drifted through the night, its tiny light blinking like a heartbeat in the dark.	0
Piled in the corner of the closet, the old roller skates gathered dust. They are owned by my grandpa.	1
The paper had more than 20 authors; it's a huge collaboration.	0
The diver saw that the sunken treasure chest lay forgotten at the bottom of the ocean.	1
The ancient grandfather clock in the hallway ticked away the seconds, its pendulum swinging steadily.	0
Stacked neatly on the shelf, the leather-bound books carried the scent of time and wisdom.	0
A single candle flickered in the dark, casting dancing shadows on the walls.	1
The abandoned lighthouse stood tall against the crashing waves, a silent guardian of the shore.	0
The snow globe, shaken by curious hands, released a flurry of tiny white flakes over the miniature village.	0
The delicate snowflake melted the moment it touched her warm palm.	1
The kid really hates eating broccoli so I need to talk to her a lot.	0
The painter's palette held a swirl of colors, waiting to become something beautiful.	0
The ivy crept up the old stone wall, weaving its way through the cracks.	1

d "includes a green object"

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

V' does not need humans

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

Improve

Performance

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

Improve

Performance

V^*: compare to human-validated differences

Use $V'(p)$ to Remove Wrong Discoveries

Use $V'(p)$ to Remove Wrong Discoveries

Prompt

Some samples from S_A

He played football.

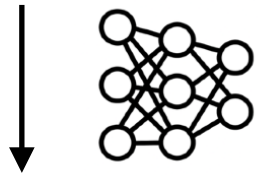
.....

Some samples from S_B

U.S. dollar bills on the table.

.....

How are samples from S_B different?



Use $V'(p)$ to Remove Wrong Discoveries

Prompt

Some samples from S_A

He played football.

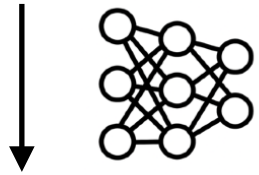
.....

Some samples from S_B

U.S. dollar bills on the table.

.....

How are samples from S_B different?



"Brainstorm"

Use $V'(p)$ to Remove Wrong Discoveries

Prompt

Some samples from S_A
He played football.
.....

Some samples from S_B
U.S. dollar bills on the table.
.....

How are samples from S_B different?

Response

d_1 = starts with a person's name

d_2 = depicts a scene

d_3 = contains a green object

d_4 = complains about the foot

.....

Use $V'(p)$ to Remove Wrong Discoveries

Prompt

Some samples from S_A
He played football.
.....

Some samples from S_B
U.S. dollar bills on the table.
.....

How are samples from S_B different?

Response

d_1 = starts with a person's name	$V'(d_1) = -0.1$
d_2 = depicts a scene	$V'(d_2) = +0.2$
d_3 = contains a green object	$V'(d_3) = +0.8$
d_4 = complains about the foot	$V'(d_4) = -0.6$
.....	.....

Use $V'(p)$ to Remove Wrong Discoveries

Prompt

Some samples from S_A
He played football.
.....

Some samples from S_B
U.S. dollar bills on the table.
.....

How are samples from S_B different?

Response

d_1 = starts with a person's name — $V'(d_1) = -0.1$

d_2 = depicts a scene — $V'(d_2) = +0.2$

d_3 = contains a green object — $V'(d_3) = +0.8$

24% → 53%
(SynD5, Zhong et al., NeurIPS'23)

d_4 = complains about the foot — $V'(d_4) = -0.6$

.....

Application #1: Understand LLM using LLM

(ChatbotArena, Chiang et al., ICML'24)

Application #1: Understand LLM using LLM

vs.

What is each LLM good at?

(ChatbotArena, Chiang et al., ICML'24)

Application #1: Understand LLM using LLM

Where Llama is better

```
I run many process using pytorch to train modules, but only 1 CPU core is 100% usage.
What is entropy? Explain using only words that start by a or t
Create an outline for a Thesis on topic: "Detection and recognition of objects in video surveillance." The
outline should include 3 main chapters (excluding introduction as that is not a chapter) and each chapter
should include many subchapters. One of the main focuses of the Thesis will be AI and the YOLO algorithm.
The final chapter should involve a practical use and demonstration of YOLO.
The Thesis will be about 60 pages long. Remember to include plenty of subchapters in the outline. Do NOT
make the outline overly complicated with difficult to understand concepts. Try to not mention "video
surveillance" in the outline.
```

.....

Where Claude is better

```
What is this called you are right until you are proven wrong
write 7 sentences that begin with the word “cheese” and last word must be “girl”, keep track number of
sentences. Think about this step by step
What's the meaning of the Pāḷi term "bījagāmabhūtagāmasamārambha"?
If each animal was from a country that comprise of only those animals, for example the bear country would
exclusively have bears, which country would have the most gold metals in the Olympics and why?
What are considered to be the most representative art works from the Tang Dynasty?
Who wins, the Ranger or Texas Red?
……
```

.....

(ChatbotArena, Chiang et al., ICML’24)

Application #1: Understand LLM using LLM

Where Llama is better

```
I run many process using pytorch to train modules, but only 1 CPU core is 100% usage.
What is entropy? Explain using only words that start by a or t
Create an outline for a Thesis on topic: "Detection and recognition of objects in video surveillance." The
outline should include 3 main chapters (excluding introduction as that is not a chapter) and each chapter
should include many subchapters. One of the main focuses of the Thesis will be AI and the YOLO algorithm.
The final chapter should involve a practical use and demonstration of YOLO.
The Thesis will be about 60 pages long. Remember to include plenty of subchapters in the outline. Do NOT
make the outline overly complicated with difficult to understand concepts. Try to not mention "video
surveillance" in the outline.
```

.....

Where Claude is better

```
What is this called you are right until you are proven wrong
write 7 sentences that begin with the word “cheese” and last word must be “girl”, keep track number of
sentences. Think about this step by step
What's the meaning of the Pāḷi term "bījagāmabhūtagāmasamārambha"?
If each animal was from a country that comprise of only those animals, for example the bear country would
exclusively have bears, which country would have the most gold metals in the Olympics and why?
What are considered to be the most representative art works from the Tang Dynasty?
Who wins, the Ranger or Texas Red?
……
```

.....

Llama better: “asks an open-ended or thought provoking question”

Claude better: “is technical/contains code snippet”

(ChatbotArena, Chiang et al., ICML’24)

Application #1: Understand LLM using LLM

Where Llama is better

```
I run many process using pytorch to train modules, but only 1 CPU core is 100% usage.
What is entropy? Explain using only words that start by a or t
Create an outline for a Thesis on topic: "Detection and recognition of objects in video surveillance." The
outline should include 3 main chapters (excluding introduction as that is not a chapter) and each chapter
should include many subchapters. One of the main focuses of the Thesis will be AI and the YOLO algorithm.
The final chapter should involve a practical use and demonstration of YOLO.
The Thesis will be about 60 pages long. Remember to include plenty of subchapters in the outline. Do NOT
make the outline overly complicated with difficult to understand concepts. Try to not mention "video
surveillance" in the outline.
```

.....

Where Claude is better

```
What is this called you are right until you are proven wrong
write 7 sentences that begin with the word “cheese” and last word must be “girl”, keep track number of
sentences. Think about this step by step
What's the meaning of the Pāḷi term "bījagāmabhūtagāmasamārambha"?
If each animal was from a country that comprise of only those animals, for example the bear country would
exclusively have bears, which country would have the most gold metals in the Olympics and why?
What are considered to be the most representative art works from the Tang Dynasty?
Who wins, the Ranger or Texas Red?
……
```

.....

Similar to expert manual analyses
(Dunlap et al., lmsys’24)

Llama better: “asks an open-ended or thought provoking question”

Claude better: “is technical/contains code snippet”

(ChatbotArena, Chiang et al., ICML’24)

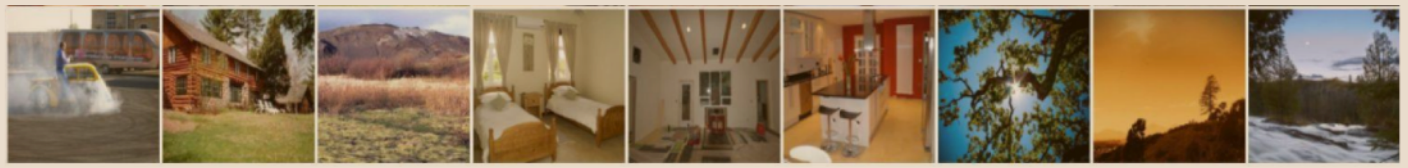
Application #2 in Cognitive Science

What makes an image memorable to humans?

(LaMem, KRTO+, ICCV'15)

Application #2 in Cognitive Science

Non-memorable



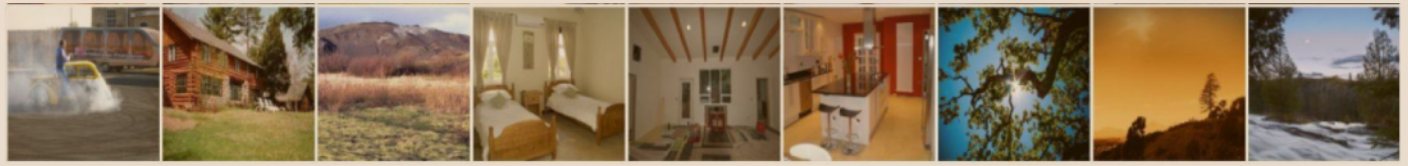
.....

Memorable

.....

Application #2 in Cognitive Science

Non-memorable



.....

Memorable

.....

Non-memorable: “portrays tranquil natural scenes”

Memorable: “highlights emotions or expressions”

Application #2 in Cognitive Science

Non-memorable

.....

Memorable

.....

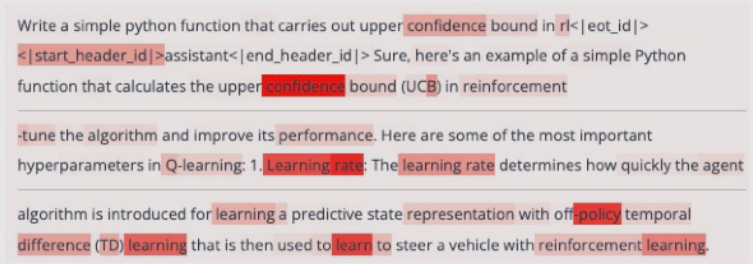
Similar to expert manual analyses
(KRTO+, ICCV'15)

Non-memorable: "portrays tranquil natural scenes"

Memorable: "highlights emotions or expressions"

Broader Impact

Explaining LM Neurons

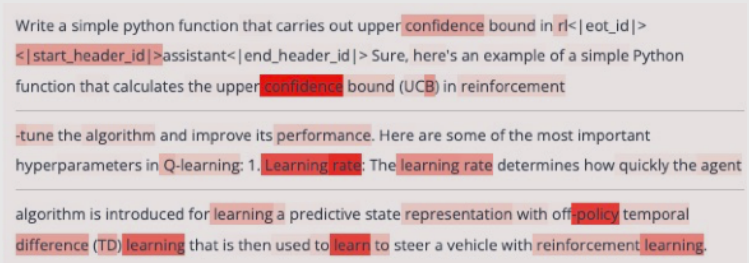


OpenAI (BCMT+'23)

Transluce (CHMJ+'24)

Broader Impact

Explaining LM Neurons



OpenAI (BCMT+'23)

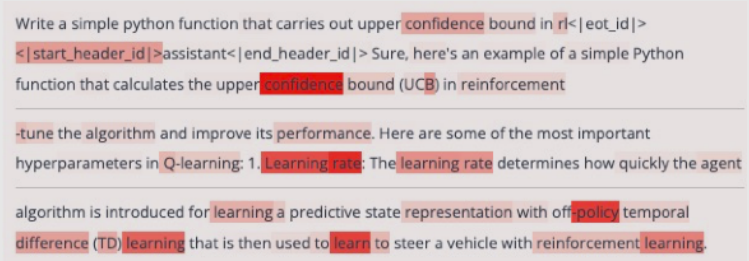
Transluce (CHMJ+'24)

Understanding User Traffic

ANTHROP\C (TMHD+'24)

Broader Impact

Explaining LM Neurons



OpenAI (BCMT+'23)

Transluce (CHMJ+'24)

Understanding User Traffic

ANTHROP\C (TMHD+'24)

Categorizing Styles

⚔ LMSYS Chatbot Arena (DMDS+'24)

What’s still relevant in 2025?

What’s still relevant in 2025?

- Automate simple validation tasks (after decomposition)

What’s still relevant in 2025?

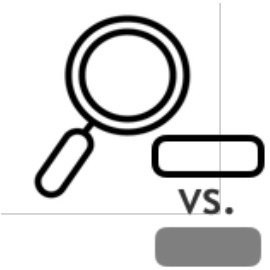
- Automate simple validation tasks (after decomposition)
- Use more “test-time compute” to validate.

Outline

Task 1

Programming

Task 2



Discovering
dataset patterns

Outline

Task 1

Programming

Task 2

Discovering dataset patterns

Weak validations

Compose

Validate (V')

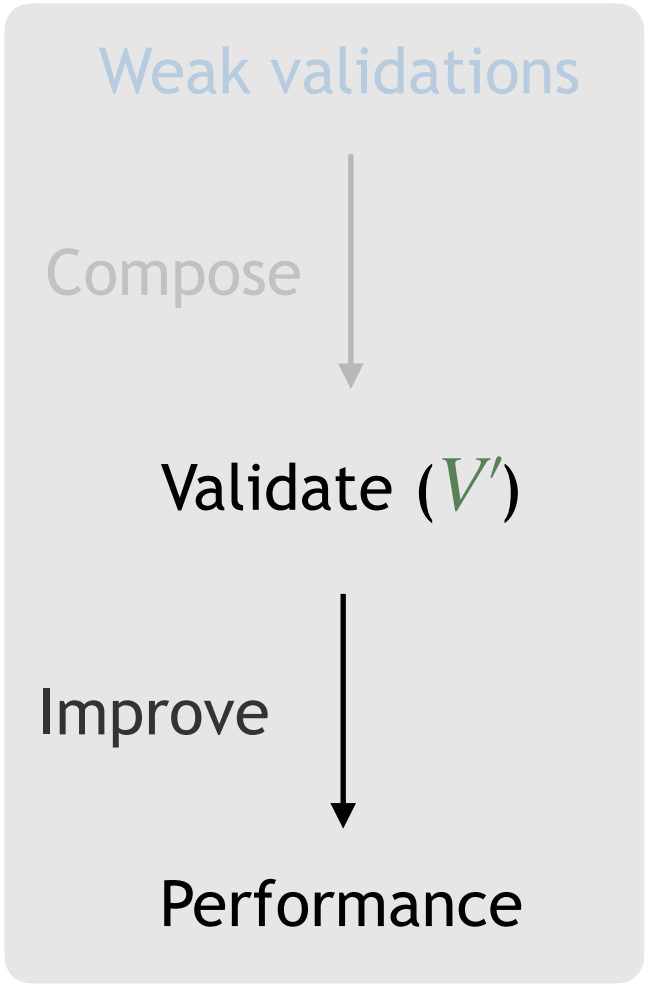
Outline

Task 1

Programming

Task 2

Discovering dataset patterns

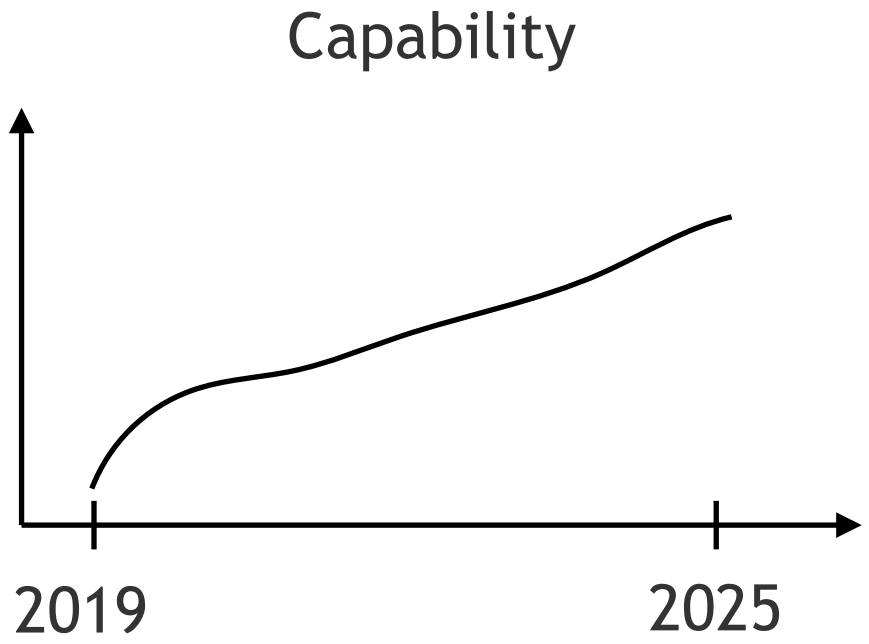


Bigger Picture and Future Work

What Drives Progress?

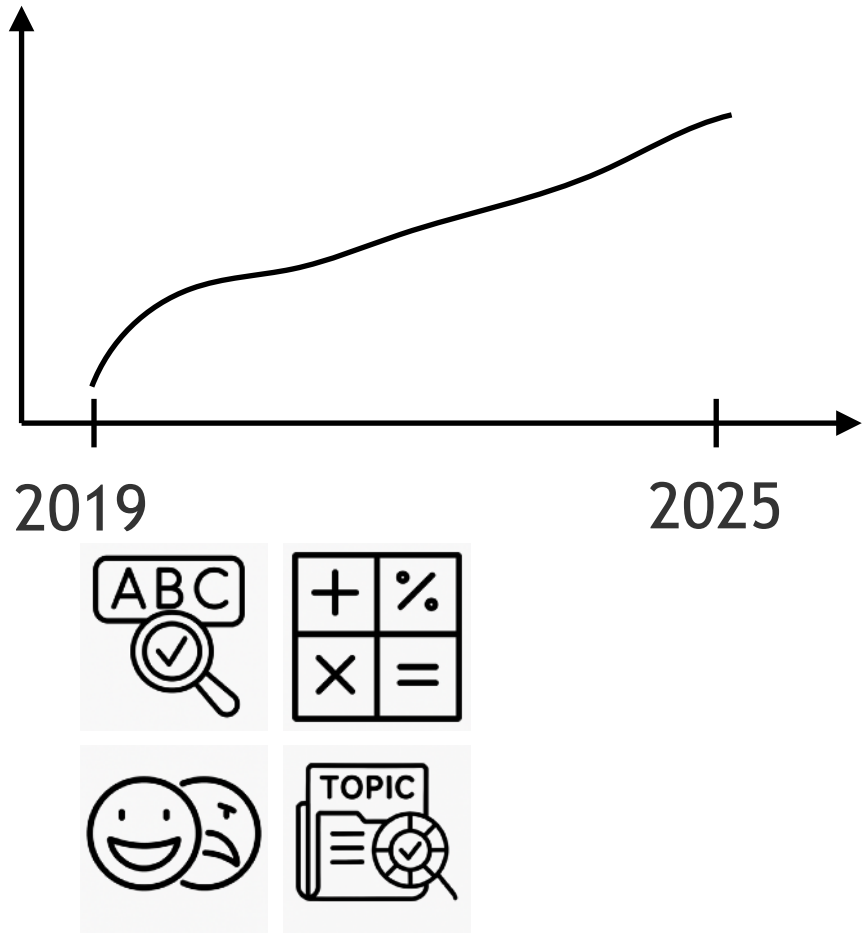
(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



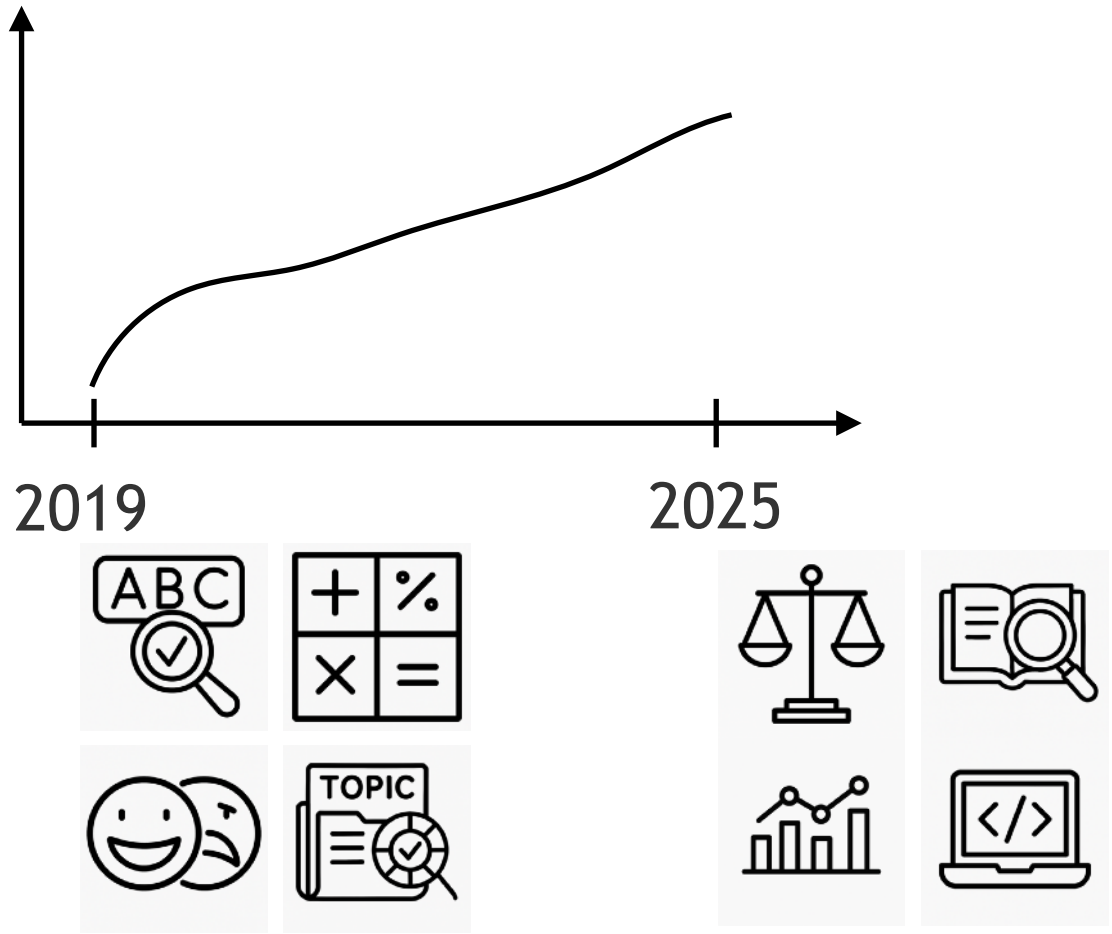
(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



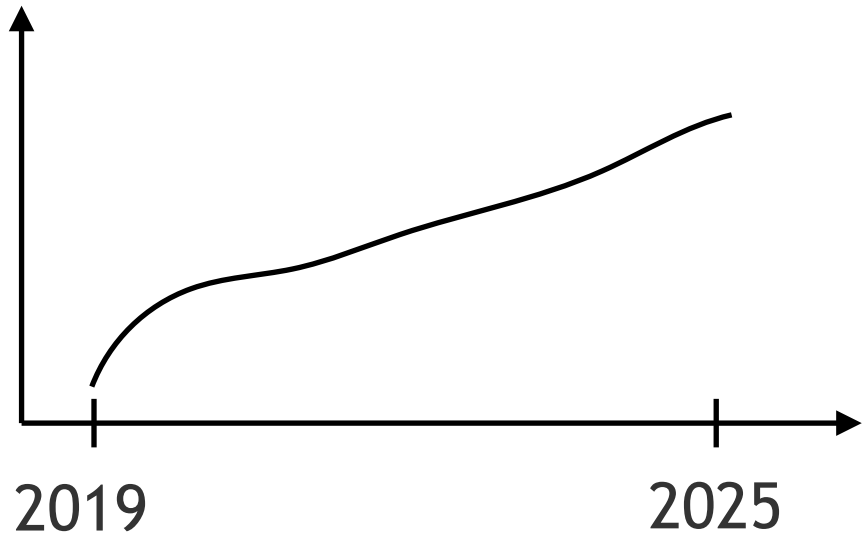
(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?

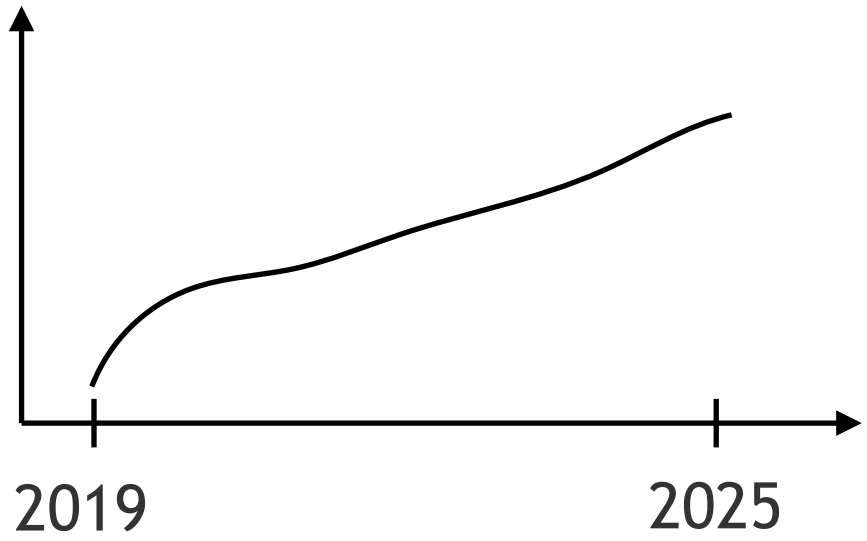


Optimization

Validation

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



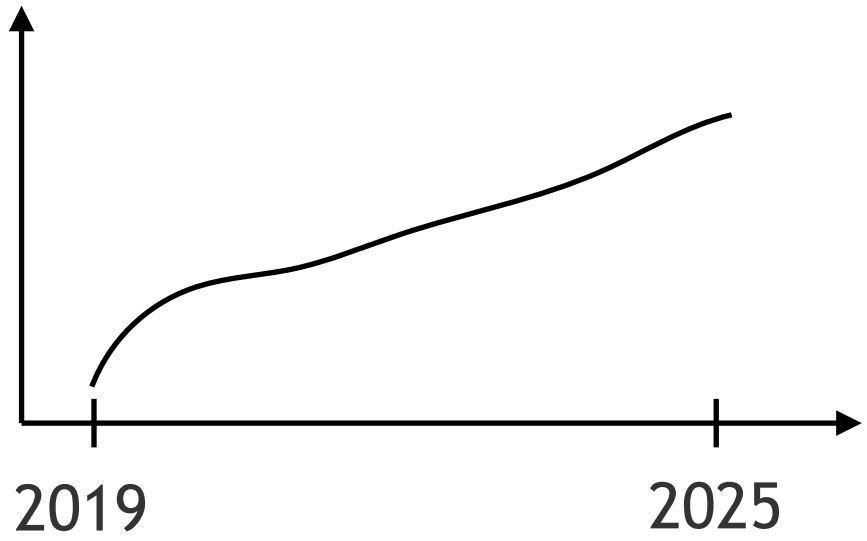
Optimization

- Larger LLM
- GPU/TPU
- Smarter LLM

Validation

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



Optimization

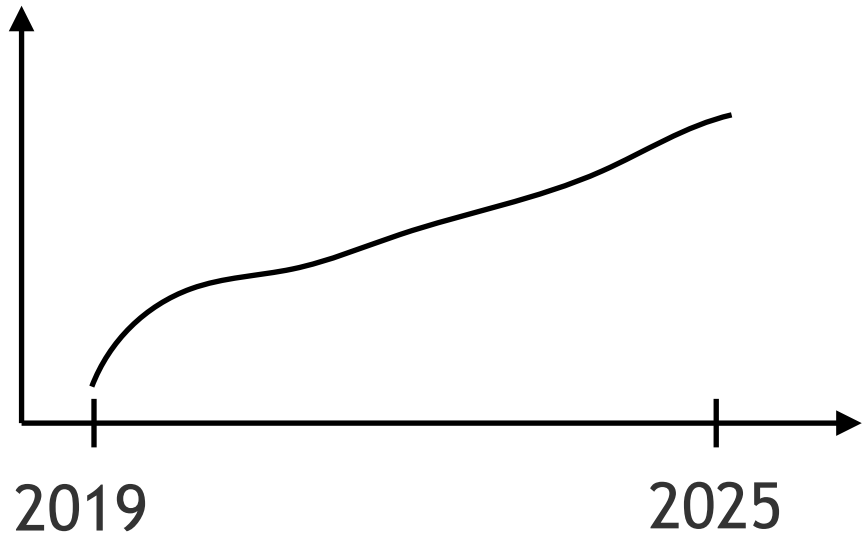
- Larger LLM
- GPU/TPU
- Smarter LLM

We know how to improve them.

Validation

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



Optimization

- Larger LLM
- GPU/TPU
- Smarter LLM

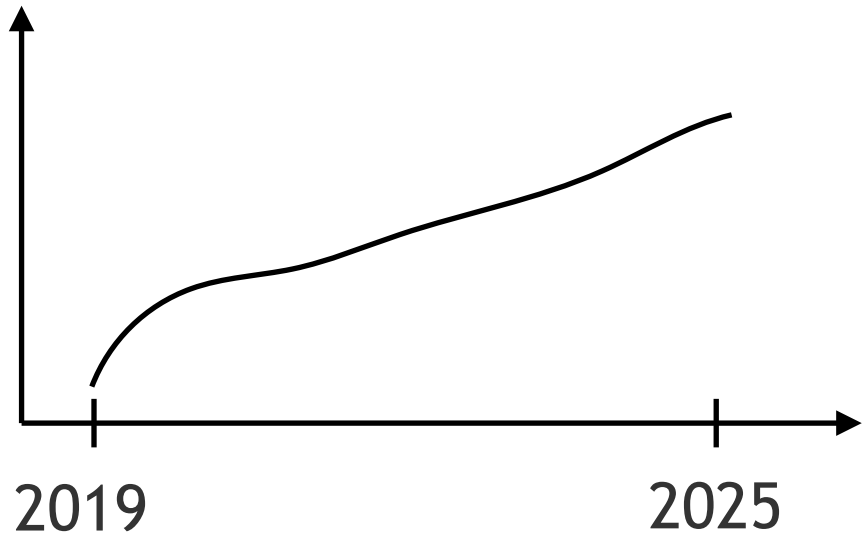
We know how to improve them.

Validation

- Content on the web (GPT2, Radford et al, arxiv'18)
- Annotator rating (InstructGPT, Ouyang et al, NeurIPS'22)
- Answer correctness (R1, Guo et al, arxiv'25)

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



Optimization

- Larger LLM
- GPU/TPU
- Smarter LLM

We know how to improve them.

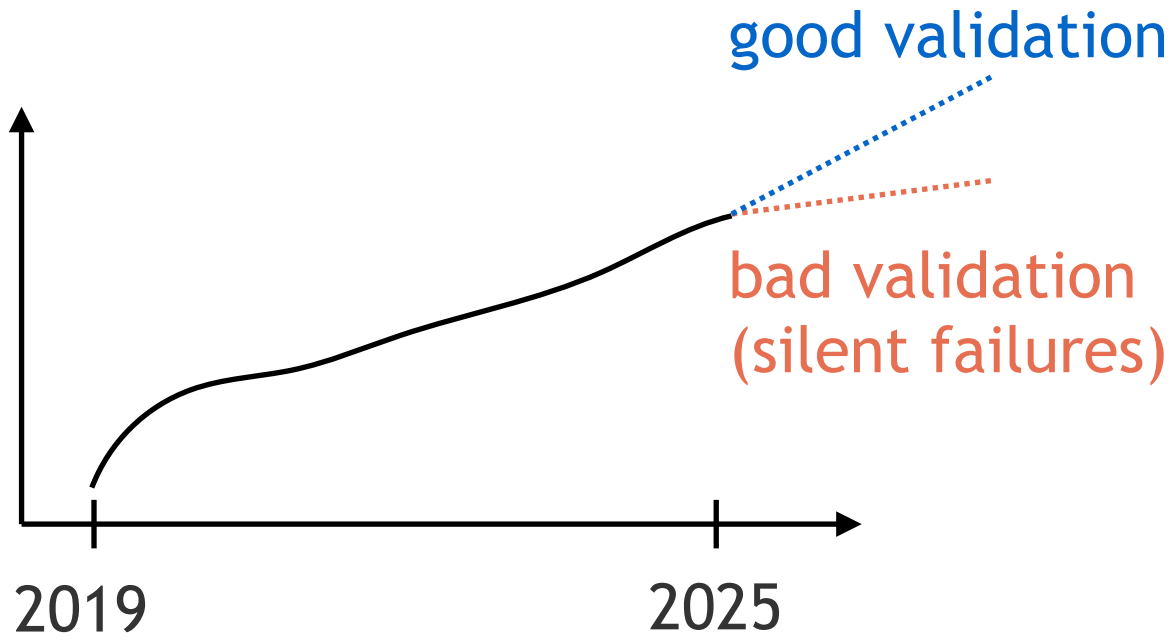
Validation

- Content on the web (GPT2, Radford et al, arxiv'18)
- Annotator rating (InstructGPT, Ouyang et al, NeurIPS'22)
- Answer correctness (R1, Guo et al, arxiv'25)

How should we improve them?

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

What Drives Progress?



Optimization

- Larger LLM
- GPU/TPU
- Smarter LLM

We know how to improve them.

Validation

- Content on the web (GPT2, Radford et al, arxiv'18)
- Annotator rating (InstructGPT, Ouyang et al, NeurIPS'22)
- Answer correctness (R1, Guo et al, arxiv'25)

How should we improve them?

(OpenHands, Wang et al., arxiv'24; Google Deepmind blogpost'24; o3, openai'24; Si et al., arxiv'24)

We are bottlenecked by validation

We are bottlenecked by validation

- Good validations (e.g. V^*) are expensive
  - Software
  - Policy recommendations
  - Research paper/reviews
  - ……

We are bottlenecked by validation

- Good validations (e.g. V^*) are expensive
  - Software
  - Policy recommendations
  - Research paper/reviews
  - ……
- Scalable oversight: build cheap & reliable V' (AI / AI + human team)

Learning to Validate w Less Resource

Learning to Validate w Less Resource

This talk: Validations of return values datapoints / datapoints —compose→ programs differences clusters

Learning to Validate w Less Resource

This talk: Validations of return values datapoints datapoints

compose

programs
differences
clusters

Hand-designed V' to bridge the resource gap.

Learning to Validate w Less Resource

This talk: Validations of return values / datapoints / datapoints

compose → programs / differences / clusters

Learn a model-based V' to bridge the resource gap.

(Learning Task Decomposition to Assist Humans in Competitive Programming, ACL'24)

Learning to Validate w Less Resource

This talk: Validations of return values datapoints / datapoints

compose → programs / differences / clusters

Future: Validations of

compose

Learning to Validate w Less Resource

This talk: Validations of return values datapoints datapoints  compose programs differences clusters

Future: Validations of lemmas  compose theorems

Learning to Validate w Less Resource

This talk: Validations of return values datapoints → compose → programs differences clusters

Future: Validations of lemmas paragraph translations → compose → theorems book translations

Learning to Validate w Less Resource

This talk: Validations of return values datapoints datapoints

compose

programs differences clusters

Future: Validations of lemmas paragraph translations

compose

theorems book translations

Learning to Validate w Less Resource

This talk: Validations of return values datapoints datapoints

compose → programs differences clusters

Future: Validations of lemmas paragraph translations

compose → theorems book translations research papers

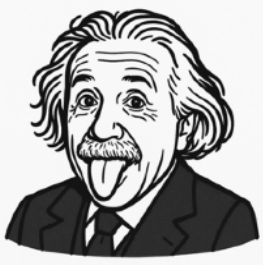
- individual claims
- novelty
- robustness checks
- ……

Learning to Validate w Less Resource

This talk: Validations of return values datapoints datapoints  compose programs differences clusters

Future: Validations of lemmas paragraph translations  compose theorems book translations research papers

- individual claims
- novelty
- robustness checks
- ……



Conclusion

Conclusion

- Validation is expensive. Make it cheap.

Conclusion

- Validation is expensive. Make it cheap.
- Decompose into weaker validation.

Conclusion

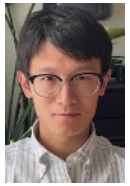
- Validation is expensive. Make it cheap.
- Decompose into weaker validation.
- Validation will be the bottleneck.

Collaborators

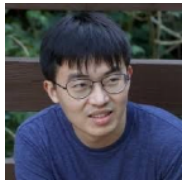
Faculty

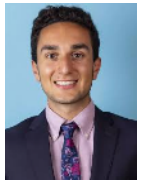




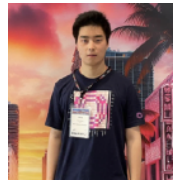




Students















Thanks for watching!

My email: ruiqizhong1997@gmail.com