

# 间接监督AI“做人做不到的事情”



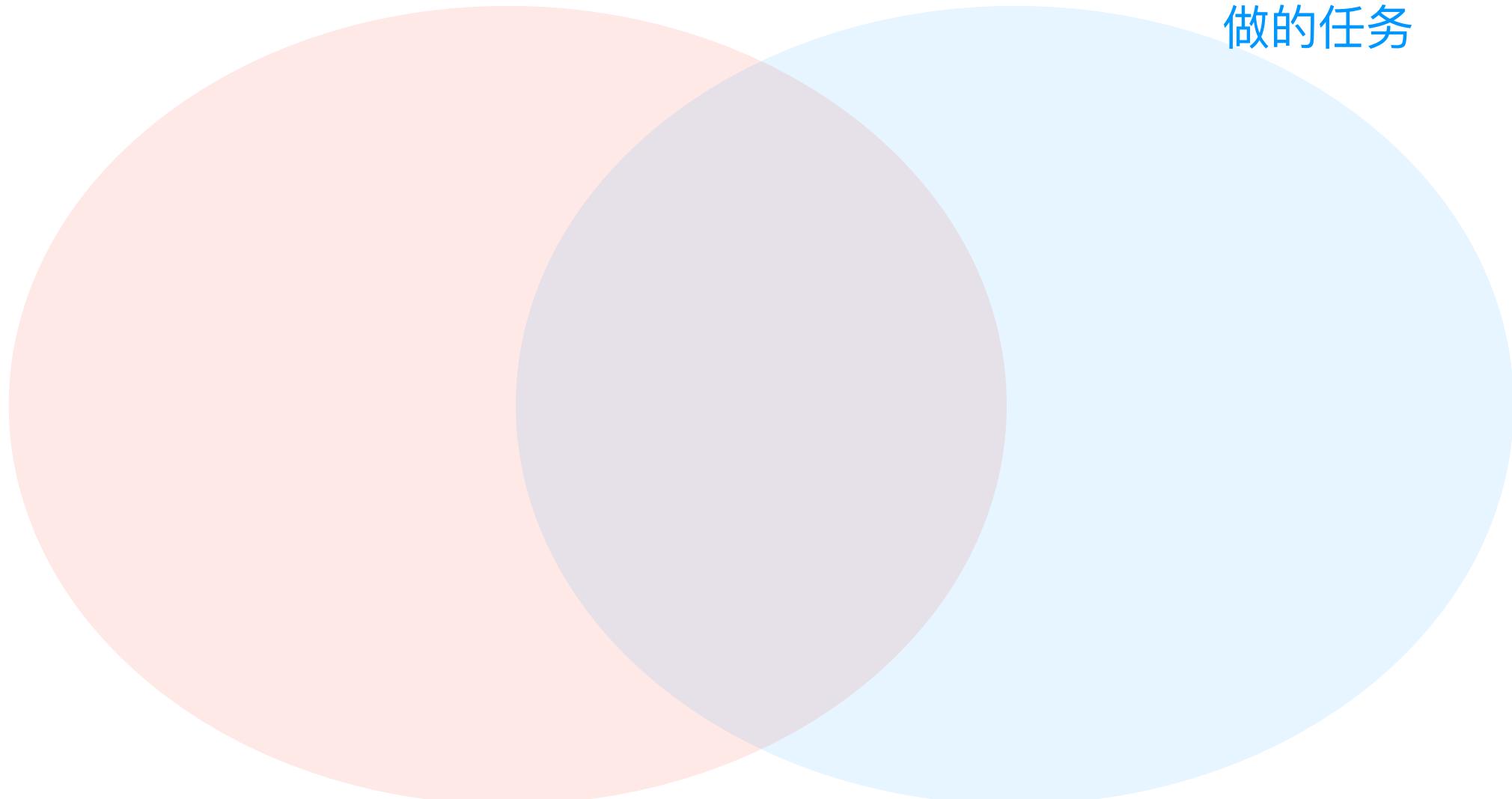
钟瑞麒



# 智能增强

人擅长的任务

我们可以让机器  
做的任务

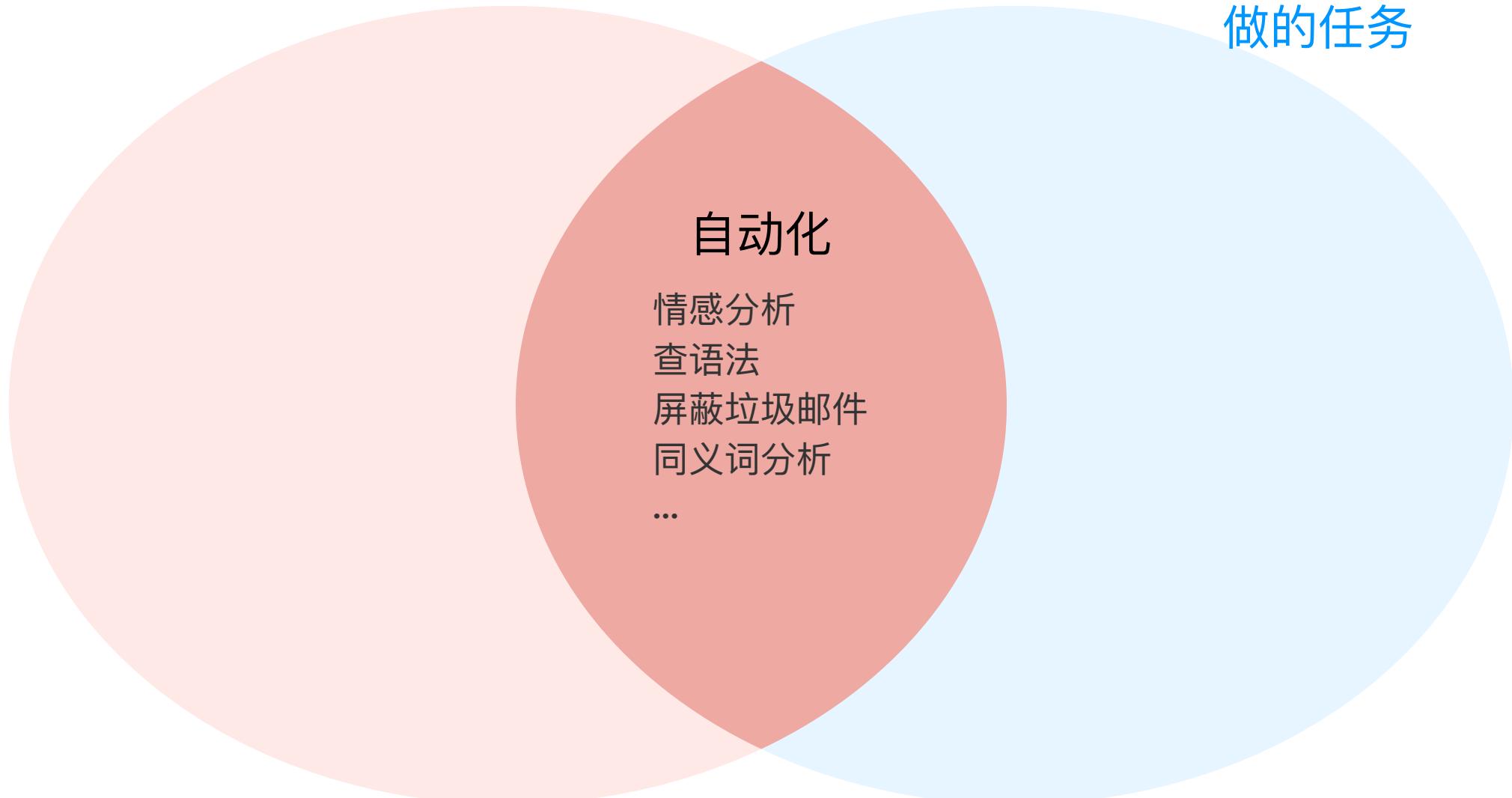




# 智能增强

人擅长的任务

我们可以让机器  
做的任务





# 智能增强

人擅长的任务

我们可以让机器  
做的任务

## 自动化

情感分析  
查语法  
屏蔽垃圾邮件  
同义词分析  
...

## 智能增强

解释规律  
写复杂的代码  
预测未来  
设计/跑实验  
.....



如何监督AI去做那些我们自己都难以  
界定正确答案的任务？



# 两个任务

## 病人对药物A的反馈

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

## 病人对药物B的反馈

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



病人在服用药物A之后更容易“感到多疑”



# 两个任务

病人对药物A的反馈

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

病人对药物B的反馈

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

病人在服用药物A之后更容易“感到多疑”



找到同时拥有狗和猫的学生的  
*first name*



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



# 两个任务

病人对药物A的反馈

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

病人对药物B的反馈

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

找到同时拥有狗和猫的学生的  
*first name*

即使人类专家也很  
难完成这些任务

病人在服用药物A之后更容易“感到多疑”

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



# 生成候选 & 简化验证

标注整个  
输出



直接验证哪个输出是  
对的

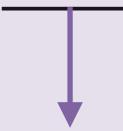


间接在更简单的问题上验  
证哪个输出是对的



# 生成候选 & 简化验证

标注整个  
输出



直接验证哪个输出是  
对的



间接在更简单的问题上验  
证哪个输出是对的

标注整个  
对话回复

(Gao et al., 2020)

选择最佳的  
对话回复

(Wiegreffe et al., 2021)

标注整个  
解释

(Stiennon et al., 2020)

选择最佳的  
解释

标注整个  
总结

↓

选择最佳的  
总结

生成几个输  
出的候选



# 生成候选 & 简化验证

标注整个  
输出



直接验证哪个输出是  
对的



间接在更简单的问题上验  
证哪个输出是对的

选择关于一个语库  
的描述

(Zhong et al., 2022)

选择正确的  
代码

(Zhong et al., 2022)

选择正确的关于  
一本书的总结

(Wu et al. 2021)

简化验证输  
出的过程



# 本次报告

---

- ▶ 关于“生成提案&简化验证”的概括
- ▶ 任务1：发现并描述两个文本分布的不同
- ▶ 任务2：让非程序员标注代码
- ▶ 一些之后的科研方向

## 任务1：发现并描述两个文本分布的不同

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

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

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



# 任务定义

输入

科研背景 (自然语言)

“我是一名教育工作者，想要理解不同学校的学生学的写的文章有什么不同”

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

语库 A (学校 A 学生写的文章)

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

语库 B (学校 B 学生写的文章)



# 任务定义

输入

科研背景 (自然语言)

“我是一名教育工作者，想要理解不同学校的学生学的写的文章有什么不同”

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

语库 A (学校 A 学生写的文章)

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

语库 B (学校 B 学生写的文章)

输出

学校A 学生写的文章更多地  
“使用统计数字”.



# 这个任务的用处？

---

- ▶ **用户的问题** 相比 **训练数据** 包含更多日常用语
- ▶ **Obama的演讲** 相比 **Trump的演讲** 更支持移民
- ▶ **被认为是人写的话** 相比 **被认为是机器写的话** 包含更多个人经历



# 这个任务的用处？

---

- ▶ **用户的问题** 相比 **训练数据** 包含更多日常用语
- ▶ **Obama的演讲** 相比 **Trump的演讲** 更支持移民
- ▶ **被认为是人写的话** 相比 **被认为是机器写的话** 包含更多个人经历
- ▶ 机器学习系统在什么样的问题上更容易犯错？
- ▶ ChatGPT和Claude的回复有哪些系统性的不同？
- ▶ 顾客对于两个公司的评价有哪些不同？
- ▶ 不同用户在不同平台上讨论的话题有什么不同？



# 这个任务很难

~1.7 万文本

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

~1.7 万文本

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

橙色的语库 “更多地提到女性人物”



# 直接标注输出很困难

科研背景

“我是一名教育工作者，想要理解不同学校的学生学的写的文章有什么不同”

$D_A$

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

$D_B$

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



难以直接标注：  
每个语库有1.7万个文本

学校A 学生写的文章更多地  $s = \text{“使用统计数字”}$ .

⚠️ ⚠️ ⚠️ 相关性 ≠ 因果性 ⚠️ ⚠️ ⚠️



# 生成候选

$D_A$  样本



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

$D_B$  样本



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

科研背景



I want to understand how students write essays differently ...  
Compared to inputs from group B, each input from group A

用GPT-3/4生成候选



- 更加有冒犯性
- 用了更多的统计数字
- .....



# 生成候选

$D_A$  样本



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

$D_B$  样本



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

科研背景



I want to understand how students write essays differently ...  
Compared to inputs from group B, each input from group A

用GPT-3/4生成候选



- 更加有冒犯性
- 用了更多的统计数字
- .....



TODO: 人类需要验证哪个候选能更准确地描述两个语库的不同





# 简化验证

人类难以验证关于整个语库  $D_A$  和  
 $D_B$ 的描述

↓  
简化  
验证

人类更容易在单个样本上验证  
这个描述是否准确



# 在单个样本上验证

$s = D_A$  上的文本“用了更多统计数字”.



# 在单个样本上验证

$s = D_A$  上的文本“用了更多统计数字”.

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

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

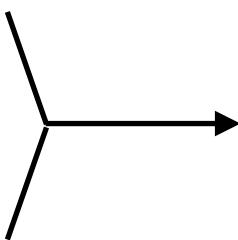


# 在单个样本上验证

$s = D_A$  上的文本“用了更多统计数字”.

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

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



人类区分

我觉得  $x_i$  是从  $D_A$  来的， $x_j$  是从  $D_B$  来的。

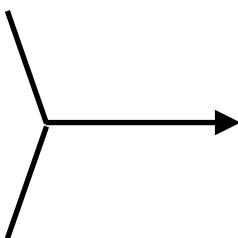


# 在单个样本上验证

$s = D_A$  上的文本“用了更多统计数字”.

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

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



人类区分

我觉得  $x_i$  是从  $D_A$  来的， $x_j$  是从  $D_B$  来的。

$\text{Loss}(\underline{s})$ : 人类是否可以准确地使用  $\underline{s}$  猜测一个样本是哪个语库里来的( $D_A$  或  $D_B$ )

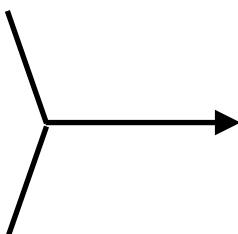


# 在单个样本上验证

$s = D_A$  上的文本“用了更多统计数字”。

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

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



人类区分

我觉得  $x_i$  是从  $D_A$  来的， $x_j$  是从  $D_B$  来的。

$\text{Loss}(\underline{s})$ : 人类是否可以准确地使用  $\underline{s}$  猜测一个样本是哪个语库里来的 ( $D_A$  或  $D_B$ )

重复 100 次，每个描述花费大约~\$10(!)来进行评估

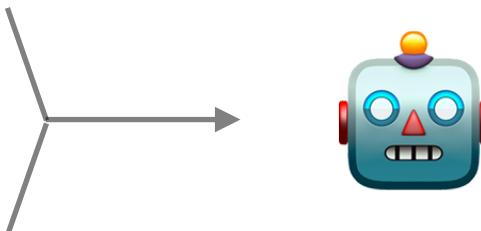


# 让验证更高效

$s = D_A$  上的文本“用了更多统计数字”.

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

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



机器区分

我觉得  $x_i$  是从  $D_A$  来的， $x_j$  是从  $D_B$  来的。

## 计算机器的准确率更加便宜

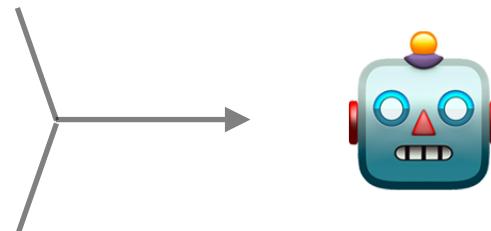


# 让验证更高效

$s = D_A$  上的文本“用了更多统计数字”。

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

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



我觉得  $x_i$  是从  $D_A$  来的， $x_j$  是从  $D_B$  来的。

机器区分

Loss( $s$ ): 机器是否可以准确地根据  $s$  区分猜测一个样本是哪个语库里来的 ( $D_A$  或  $D_B$ )

计算机器的准确率更加便宜



# 生成候选 & 简化验证



# 生成候选 & 简化验证

- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对单个样本的验证
  - ▶ 通过用机器模拟人类验证使得验证更高效



# 生成候选 & 简化验证

- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对单个样本的验证
  - ▶ 通过用机器模拟人类验证使得验证更高效
- ▶ “目标”：简化验证后 > 随机的候选



# 实验结果

---

恢复54个二分类任务上的任务介绍



# 实验结果

恢复54个二分类任务上的任务介绍

$D_A$

$D_B$

标注1的样本

标注0的样本

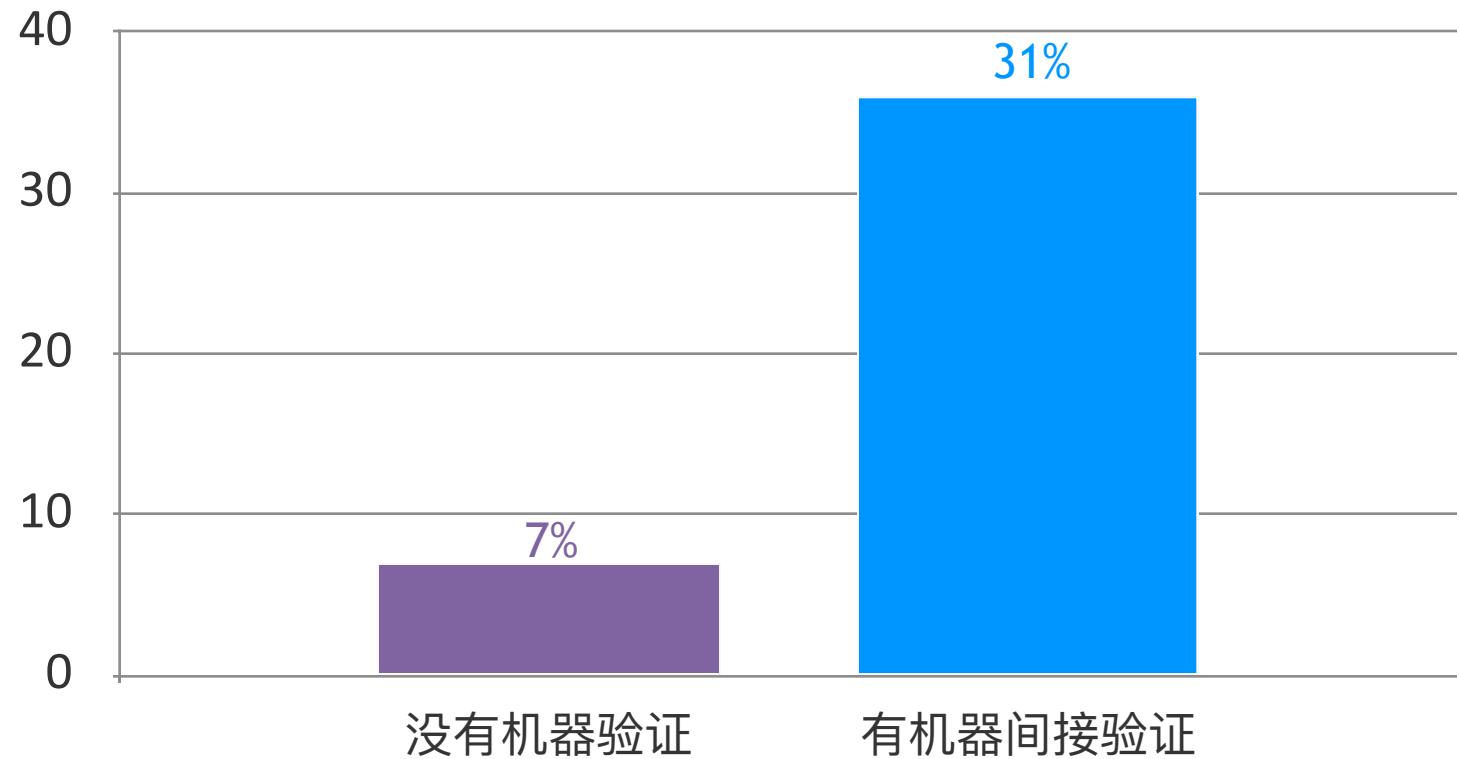
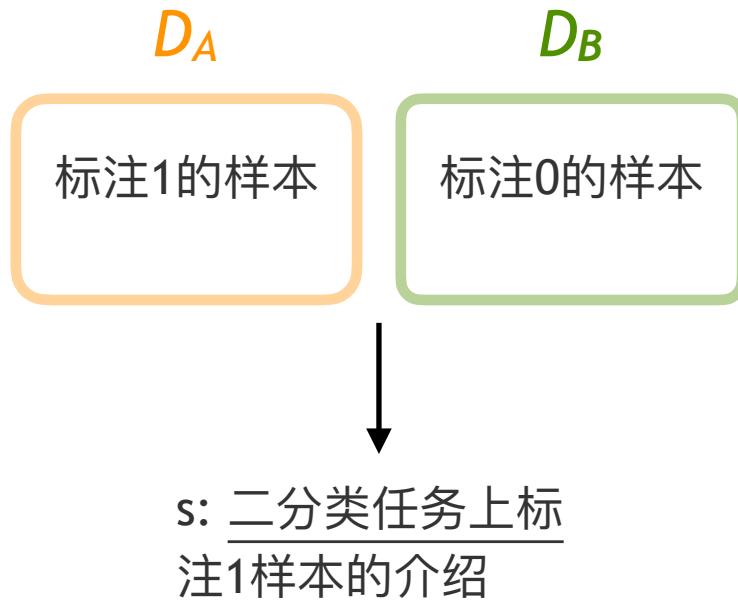


s: 二分类任务上标注1样本的介绍



# 实验结果

恢复54个二分类任务上的任务介绍





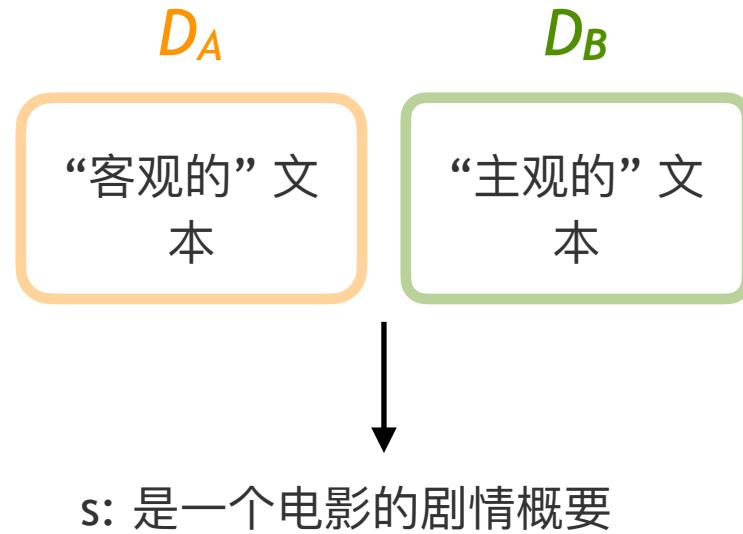
# 对数据集的误解

解读SUBJ上标注 (Subjectivity analysis dataset)  
≥ 4 论文用了这个数据集



# 对数据集的误解

解读SUBJ上标注 (Subjectivity analysis dataset)  
≥ 4 论文用了这个数据集





# 对数据集的误解

解读SUBJ上标注 (Subjectivity analysis dataset)  
≥ 4 论文用了这个数据集

$D_A$

$D_B$

“客观的” 文本

“主观的” 文本



s: 是一个电影的剧情概要

为了采集主观的文本 我们从  
[www.rottentomatoes.com](http://www.rottentomatoes.com) 选取了5千影评. 为  
了采集相对客观的文本, 我们从 [www.imdb.com](http://www.imdb.com)  
中选取了5千电影摘要。



# 分布偏移

描述不同自然语言推理数据集的不同



# 分布偏移

描述不同自然语言推理数据集的不同

$D_A$

$D_B$

SNLI 的样本

MNLI 的样本



s: 提到了身体活动，比如走路、  
玩耍、跑步。



# 分布偏移

描述不同自然语言推理数据集的不同

$D_A$

$D_B$

SNLI 的样本

MNLI 的样本



s: 提到了身体活动，比如走路、玩耍、跑步。

SNLI 里全是图片说明，MNLI里有很多不同的文体



# 分析模型错误

什么样的输入会导致一个模型出错? 任务: 零样本区分一句话是否支持了一个论点



# 分析模型错误

什么样的输入会导致一个模型出错? 任务: 零样本区分一句话是否支持了一个论点

$D_A$

$D_B$

模型A错误地  
认为答案是  
“支持”

模型B错误地  
认为答案是  
“支持”



s: 用了情感上更加正面的语言



# 分析模型错误

什么样的输入会导致一个模型出错? 任务: 零样本区分一句话是否支持了一个论点

$D_A$

$D_B$

模型A错误地  
认为答案是  
“支持”

模型B错误地  
认为答案是  
“支持”

模型A可能把“支持”理解成了一种情感



s: 用了情感上更加正面的语言



# 更多的应用

## 机器学习

Expose dataset  
misunderstanding

Describe  
distribution shift

Analyze model  
errors



# 更多的应用

## 机器学习

Expose dataset  
misunderstanding

Describe  
distribution shift

Analyze model  
errors

Debug dataset  
shortcuts

Label text  
clusters

Compare language  
model's generation

Explain neurons

Describe principle  
components

Explain human  
preferences



# 更多的应用

## 机器学习

Expose dataset misunderstanding	Describe distribution shift	Analyze model errors
Debug dataset shortcuts	Label text clusters	Compare language model's generation
Explain neurons	Describe principle components	Explain human preferences

## 健康

Compare symptoms of different diseases  
Compare side effects of drugs  
.....

## 人文

Compare lyrics of different era  
Compare writing style of different authors  
.....

## 商业

Compare reviews for different companies  
Compare advertisements ...

## 社科

还有更多!!!

Expose gender association biases in speeches

Discover topical differences in headlines between years

Discover topical differences between demographic groups

Identify political stances of different leaders

Understand how political ads target different subpopulations

Understand what job skills are more preferred for different sectors



# 任务1小结

---



# 任务1小结

---

- ▶ 描述文本分布的不同



# 任务1小结

---

- ▶ 描述文本分布的不同
- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对单个样本的验证
  - ▶ 通过用机器模拟人类验证使得验证更高效



# 任务1小结

---

- ▶ 描述文本分布的不同
- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对单个样本的验证
  - ▶ 通过用机器模拟人类验证使得验证更高效
- ▶ 发现了人类专家之前也没注意到的规律



# 本次报告

---

- ▶ 关于“生成提案&简化验证”的概括
- ▶ 任务1:发现并描述两个文本分布的不同
- ▶ 任务2: 让非程序员标注代码
- ▶ 一些之后的科研方向

## 任务2: 让非程序员标注代码

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

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



# 语义分析

自然语言

部门A里最年轻的人几岁了?



SQL 程序

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



# 语义分析

自然语言

部门A里最年轻的人几岁了?



SQL 程序

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

标注很贵!!

如何让非程序员标注SQL?



# 生成候选

自然语言

部门A里最年轻的人几岁了?

Codex生成候选SQL程序 x 32



概率	SQL程序
7/10	<u>SELECT MAX(Name) from People</u>
1/10	<u>SELECT MAX(Age) from People</u>
1/80	<u>SELECT MIN(Age) from People</u> <u>WHERE Department = 'A'</u>



# 生成候选

自然语言

部门A里最年轻的人几岁了?

Codex生成候选SQL程序 x 32



候选	概率	SQL程序	
	7/10	<u>SELECT MAX(Name) from People</u>	✗
	1/10	<u>SELECT MAX(Age) from People</u>	✗
	1/80	<u>SELECT MIN(Age) from People WHERE Department = 'A'</u>	✓

TODO: 如何让非程序员知道哪个候选SQL程序是正确的?



# 对非程序员来说难以验证

找到同时有猫和狗的学生的*first name*

候选 1

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

候选 2

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



# 简化验证

非程序员难以直接验  
证哪个程序是对的

简化  
验证



在简单的测试用例上更容易  
间接验证一个程序的正确性



# 在程序的输入输出上验证其正确性

---

部门A里最年轻的人几岁了？

SELECT MAX(Name) from People

SELECT MAX(Age) from People

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



# 在程序的输入输出上验证其正确性

部门A里最年轻的人几岁了？

SELECT MAX(Name) from People

SELECT MAX(Age) from People

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

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



# 在程序的输入输出上验证其正确性

部门A里最年轻的人几岁了？

非程序员的答案

23

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

Cathy

SELECT MAX(Name) from People

SELECT MAX(Age) from People

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

SQL程序的输出

28

23



# 在程序的输入输出上验证其正确性

部门A里最年轻的人几岁了？

SELECT MAX(Name) from People

SELECT MAX(Age) from People

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

非程序员的答案

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

23

Cathy ✗

28 ✗

23 ✓

SQL程序的输出



# 在程序的输入输出上验证其正确性

部门A里最年轻的人几岁了？

~~SELECT MAX(NAME) from People~~

~~SELECT MAX(Age) from People~~

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

非程序员的答案

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

23

Cathy ✗

28 ✗

23 ✓

SQL程序的输出



# 这个数据库从哪里来？

部门A里最年轻的人几岁了？

~~SELECT MAX(Name) from People~~

~~SELECT MAX(Age) from People~~

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

非程序员的答案

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

23

Cathy ✗

28 ✗

23 ✓

SQL程序的输出



# 让验证更高效

用最少的人力实现最大程度的监督

Size (

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

) is small

InfoGain (

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

) is large



# 让验证更高效

数据库输入必须简单  
并且让人容易理解

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

→ ?????  
标注者的答案

[总共1000行，剩下的省略]



# 让验证更高效

Size (

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

) = 1000

[总共1000行，剩下的省略]



# 让验证更高效

部门A里最年轻的人几岁了？

NAME	Age	Department
Collin	26	A
Bob	23	A

SELECT MIN(Age) from People

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



# 让验证更高效

部门A里最年轻的人几岁了？

NAME	Age	Department
Collin	26	A
Bob	23	A

→ 答案

23 ✓

毫无信息量！

SELECT MIN(Age) from People

→ 23 ✓

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

→ 23 ✓



# 期望信息增益

概率

SQL

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

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

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

NAME	Age	Department
Collin	26	A
Bob	23	A



# 期望信息增益

概率

1/3

SELECT MIN(Age) from People

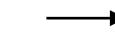
1/3

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

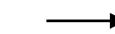
1/3

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

SQL



23  
23



26

程序输出的熵:  $-2/3 * \log(2/3) - 1/3 * \log(1/3) = 0.92$

NAME	Age	Department
Collin	26	A
Bob	23	A



# 期望信息增益

概率

1/3

SELECT MIN(Age) from People

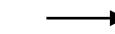
SQL

1/3

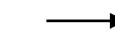
SELECT MIN(Age) from People  
WHERE Department = 'A'

1/3

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



23



23



26

程序输出的熵:  $-2/3 * \log(2/3) - 1/3 * \log(1/3) = 0.92$

InfoGain (

NAME	Age	Department
collin	26	A
Bob	23	A

) = 0.92 bit



# 论文里有更多详细技术细节

---

- ▶ 在控制大小的情况下优化数据库的期望信息增益
- ▶ 用模糊测试生成大的高信息增益的数据库
  - ▶ 贪心地扔掉没有信息量的行
- ▶ 多轮交互



# 生成候选 & 简化验证



# 生成候选 & 简化验证

- ▶ 方法：
  - ▶ 用Codex生成 SQL programs 候选
  - ▶ 把验证问题简化成分析数据库输入和输出
  - ▶ 减小数据库的大小并增大其信息增益使得验证更高效



# 生成候选 & 简化验证

- ▶ 方法：
  - ▶ 用Codex生成 SQL programs 候选
  - ▶ 把验证问题简化成分析数据库输入和输出
  - ▶ 减小数据库的大小并增大其信息增益使得验证更高效
- ▶ “目标”：简化验证后 > 第一候选



# 数据集和基线

## 自然语言

部门A里最年轻的人几岁了?

用Codex  
提出候选

概率



7/10    SELECT MAX(Name) from People    Codex 第一候选

1/10    SELECT MAX(Age) from People

.....

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



# 数据集和基线

## 自然语言

部门A里最年轻的人几岁了?

用Codex  
提出候选



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

概率

7/10 SELECT MAX(N ~~Age~~ from People Codex第一候选

1/10 SELECT MAX(Age) from People

.....

1/80 SELECT MIN(Age) from People  
WHERE Department = 'A' 非程序员标注  
(非计算机系, 无SQL经验)



# 数据集和基线

## 自然语言

部门A里最年轻的人几岁了?

用Codex  
提出候选



概率

7/10

SELECT MAX(Name) from People

Codex第一候选

1/10

SELECT MAX(Age) from People

.....

1/80

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

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

之前专家的标注

最高质量的标注:

- (1) 我们使用了我们的系统 +
- (2) 直接检查SQL的正确性 +
- (3) 和之前专家的标注进行对比 +
- (4) 和之前的专家进行讨论

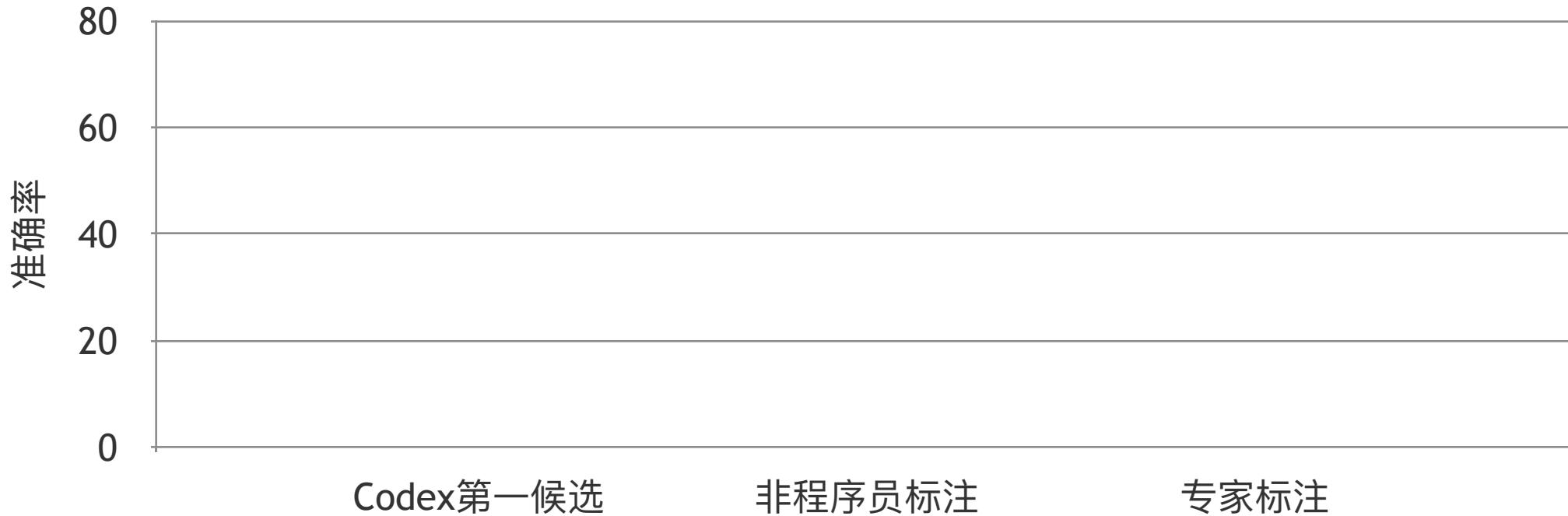
非程序员标注  
(非计算机系, 无SQL经验)



# 准确率比较

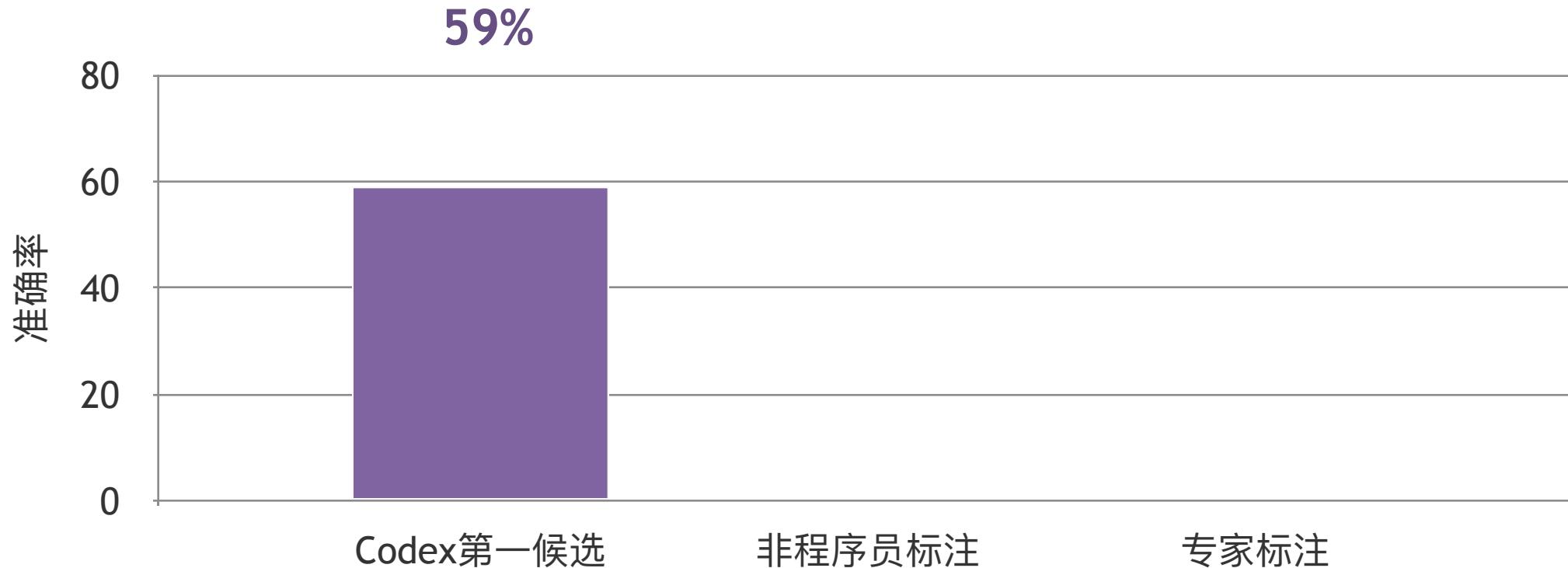


# 准确率比较



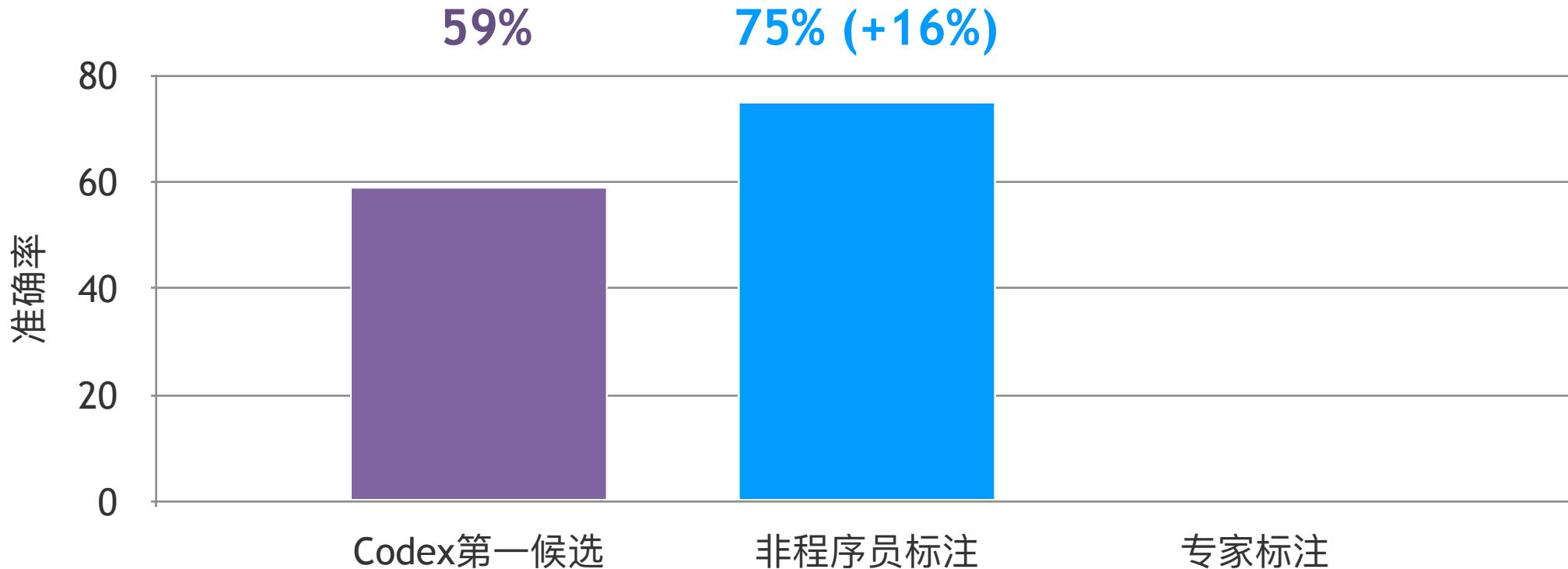


# 准确率比较



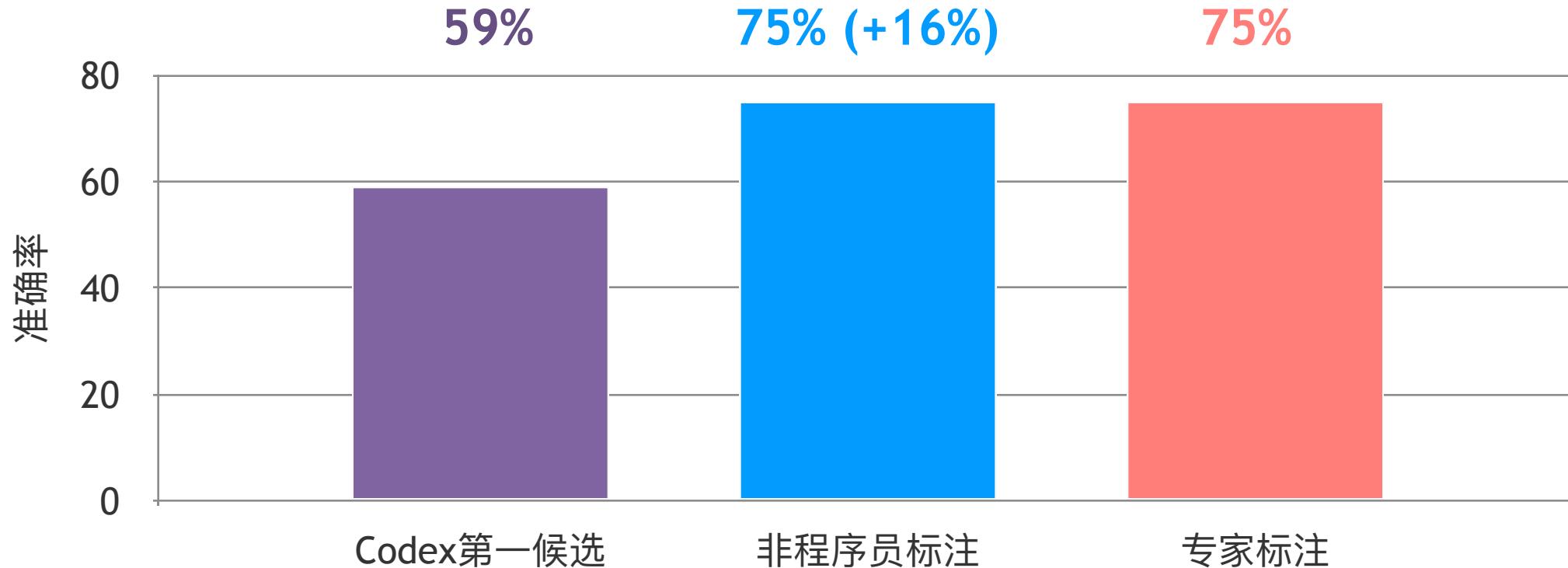


# 准确率比较





# 准确率比较





# 复杂的SQL程序

找到同时有猫和狗的学生的*first name*



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

这是一个专家犯的错误

```
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'
```



# 复杂的SQL程序

找到同时有猫和狗的学生的*first name*



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



这是一个专家犯的错误

```
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'
```





# 一个有效的数据库让验证更轻松

找到同时有猫和狗的学生的*first name*

Ownership  
(merged)

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



## 任务2小结



## 任务2小结

- ▶ 让非程序员标注 SQL 程序



## 任务2小结

- ▶ 让非程序员标注 SQL 程序
- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对数据库上正确答案的验证
  - ▶ 通过减轻人类负担验证使得验证更高效



## 任务2小结

---

- ▶ 让非程序员标注 SQL 程序
- ▶ 方法：
  - ▶ 用GPT-3/4 生成候选的描述
  - ▶ 简化成对数据库上正确答案的验证
  - ▶ 通过减轻人类负担验证使得验证更高效
- ▶ 发现了之前的专家标注有25% 的错误



# 本次报告

---

- ▶ 关于“生成提案&简化验证”的概括
- ▶ 任务1:发现并描述两个文本分布的不同
- ▶ 任务2: 让非程序员标注代码
- ▶ 一些之后的科研方向

一些之后的科研方向



# Scalable Oversight

---



# Scalable Oversight

---

- ▶ 智能增强: 其中一个难点在于如何验证AI在做正确的事情.



# Scalable Oversight

---

- ▶ 智能增强: 其中一个难点在于如何验证AI在做正确的事情.
- ▶ 生成候选 & 简化验证. 然后让验证更高效
  - ▶ 用机器模拟人类做验证
  - ▶ 用尽量少的人力做尽可能多的监督



# Scalable Oversight

---

- ▶ 智能增强: 其中一个难点在于如何验证AI在做正确的事情.
- ▶ 生成候选 & 简化验证. 然后让验证更高效
  - ▶ 用机器模拟人类做验证
  - ▶ 用尽量少的人力做尽可能多的监督
- ▶ 之后的一些方向:
  - ▶ 复杂的数学问题: 一步步进行验证?
  - ▶ 机器翻译: 验证最不确定的输出部分?
  - ▶ 有没有对所有任务都适用的简化验证的方法?



# 研究范式 & 系统评估



# 研究范式 & 系统评估

- ▶ 如何评估?



# 研究范式 & 系统评估

---

- ▶ 如何评估?
  - ▶ 用“无限多”的资源（专家+时间）得到“**标准答案**”



# 研究范式 & 系统评估

---

## ▶ 如何评估?

- ▶ 用“无限多”的资源（专家+时间）得到“**标准答案**”
- ▶ 评估**新提出的方法**是否可以用有限的资源得到同样的结果



# 研究范式 & 系统评估

- ▶ 如何评估?
  - ▶ 用“无限多”的资源（专家+时间）得到“**标准答案**”
  - ▶ 评估**新提出的方法**是否可以用有限的资源得到同样的结果
  - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)



# 研究范式 & 系统评估

---

- ▶ 如何评估?
  - ▶ 用“无限多”的资源（专家+时间）得到“**标准答案**”
  - ▶ 评估**新提出的方法**是否可以用有限的资源得到同样的结果
  - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)



# 研究范式 & 系统评估

---

- ▶ 如何评估?
  - ▶ 用“无限多”的资源（专家+时间）得到“**标准答案**”
  - ▶ 评估**新提出的方法**是否可以用有限的资源得到同样的结果
  - ▶ Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022)
  
- ▶ 这是一个比较新的研究方向；有很多机遇和挑战！



# Oversight is Challenging

---



# Oversight is Challenging

---

- ▶ 现在的情况 (ChatGPT/CoPilot)
  - ▶ 可以生成看上去非常可信的解释
  - ▶ 在用户提出的问题上它们往往能力更强



# Oversight is Challenging

## ▶ 现在的情况 (ChatGPT/CoPilot)

- ▶ 可以生成看上去非常可信的解释
- ▶ 在用户提出的问题上它们往往能力更强

---

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

---

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

---

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

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

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

---



# Oversight is Challenging

---



# Oversight is Challenging

---

- ▶ 现在的情况 (ChatGPT/CoPilot)

- ▶ 可以生成看上去非常可信的解释
- ▶ 在用户提出的问题上它们往往更强

- ▶ 大趋势

- ▶ AI 可预见性地获取不可预见的能力



# Oversight is Challenging

- ▶ 现在的情况 (ChatGPT/CoPi)

- ▶ 可以生成看上去非常可信的解释
- ▶ 在用户提出的问题上它们往往更

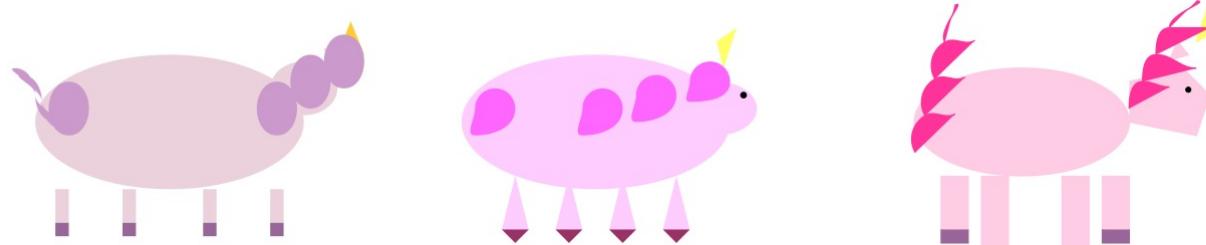


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

- ▶ 大趋势

- ▶ AI 可预见性地获取不可预见的能力



# Oversight is Challenging

## ▶ 现在的情况 (ChatGPT/CoPi)

- ▶ 可以生成看上去非常可信的解释
- ▶ 在用户提出的问题上它们往往更

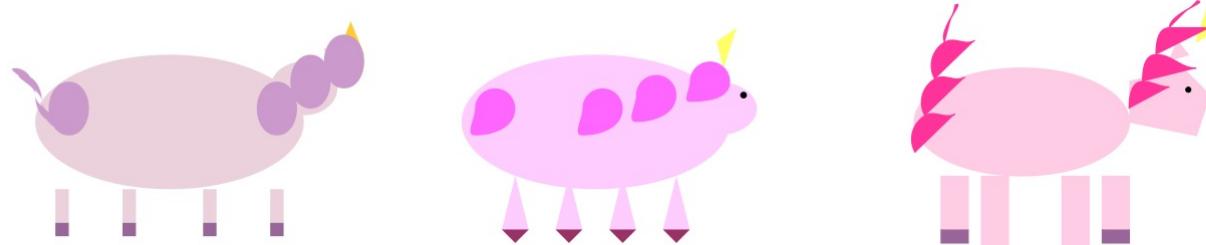


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

## ▶ 大趋势

- ▶ AI 可预见性地获取不可预见的能力
- ▶ 人类会给AI更多的“自主性”和权限



# Oversight is Challenging

Plugin store

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

< Prev 1 2 Next >

About plugins

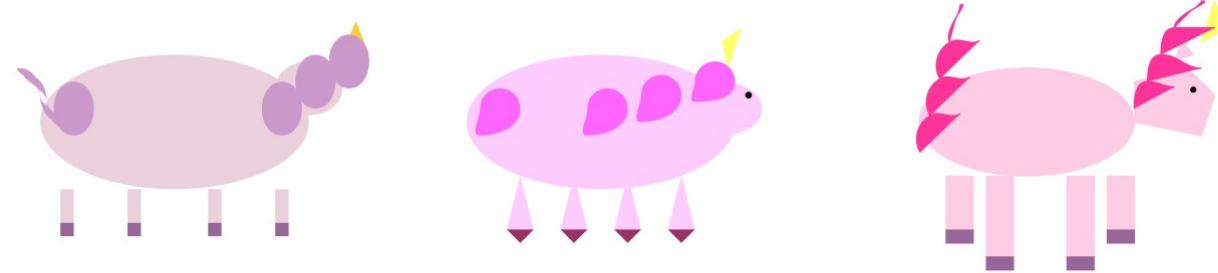


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

## ▶ 大趋势

- ▶ AI 可预见性地获取不可预见的能力
- ▶ 人类会给AI更多的“自主性”和权限
- ▶ 人类进化速度有限



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者
  - ▶ 在全球范围内



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者
  - ▶ 在全球范围内
  - ▶ 给快速进步且难以预测的AI



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者

- ▶ 在全球范围内
- ▶ 给快速进步且难以预测的AI
- ▶ 赋予更多的资源和权限



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者
  - ▶ 在全球范围内
  - ▶ 给快速进步且难以预测的AI
  - ▶ 赋予更多的资源和权限
  - ▶ 在更加复杂和开放的任务中展开行动



# 为什么我们需要 **\*\*Scalable\*\* Oversight?**

---

- ▶ 人类用户和开发者
  - ▶ 在全球范围内
  - ▶ 给快速进步且难以预测的AI
  - ▶ 赋予更多的资源和权限
  - ▶ 在更加复杂和开放的任务中展开行动
- ▶ 我们需要赋能人类更好地监督AI

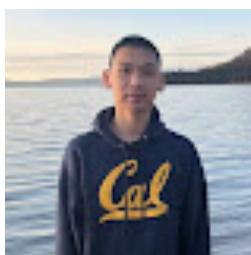


# 合作者

导师



学生



Berkeley



Thanks!