北京大学
PEKING UNIVERSITY

Eliciting Thinking Hierarchy without a Prior

Yuqing Kong*
CFCS and School of Computer Science
Peking University
yuqing.kong@pku.edu.cn

Yunqi Li†
CFCS and School of EECS
Peking University
liyunqi@pku.edu.cn

Yubo Zhang
CFCS and School of Computer Science
Peking University
falsyta@pku.edu.cn

Zhihuan Huang
CFCS and School of Computer Science
Peking University
zhihuan.huang@pku.edu.cn

Jinzhao Wu‡
CFCS and School of EECS
Peking University
jinzhao.wu@pku.edu.cn

NIPS'22

The Wisdom of Crowds

- **Plurality Voting**
  - *Plurality Voting* is usually effective in many cases.

postgraduate examination

U.S. presidential election

What to do when most people are wrong?

Examples

- What to do when most people are wrong?

gū gū zhuì dì

呱 呱 坠 地

shēn shēn xué zǐ

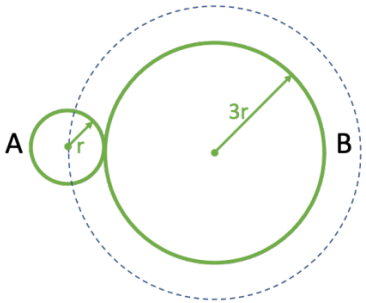
莘 莘 学 子

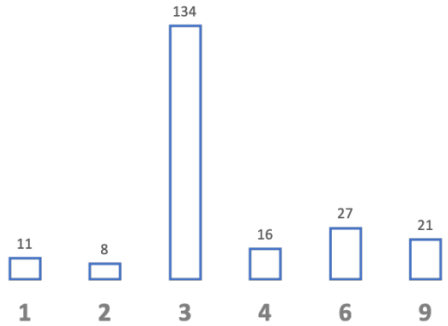
dàn tà

蛋 挞

Problem

The radius of Circle A is 1/3 the radius of Circle B. Circle A rolls around Circle B one trip back to its starting point. How many times will Circle A revolve in total?





Thinking Hierarchy without a Prior

1. Find the correct answer
2. Rank the thinking

Method

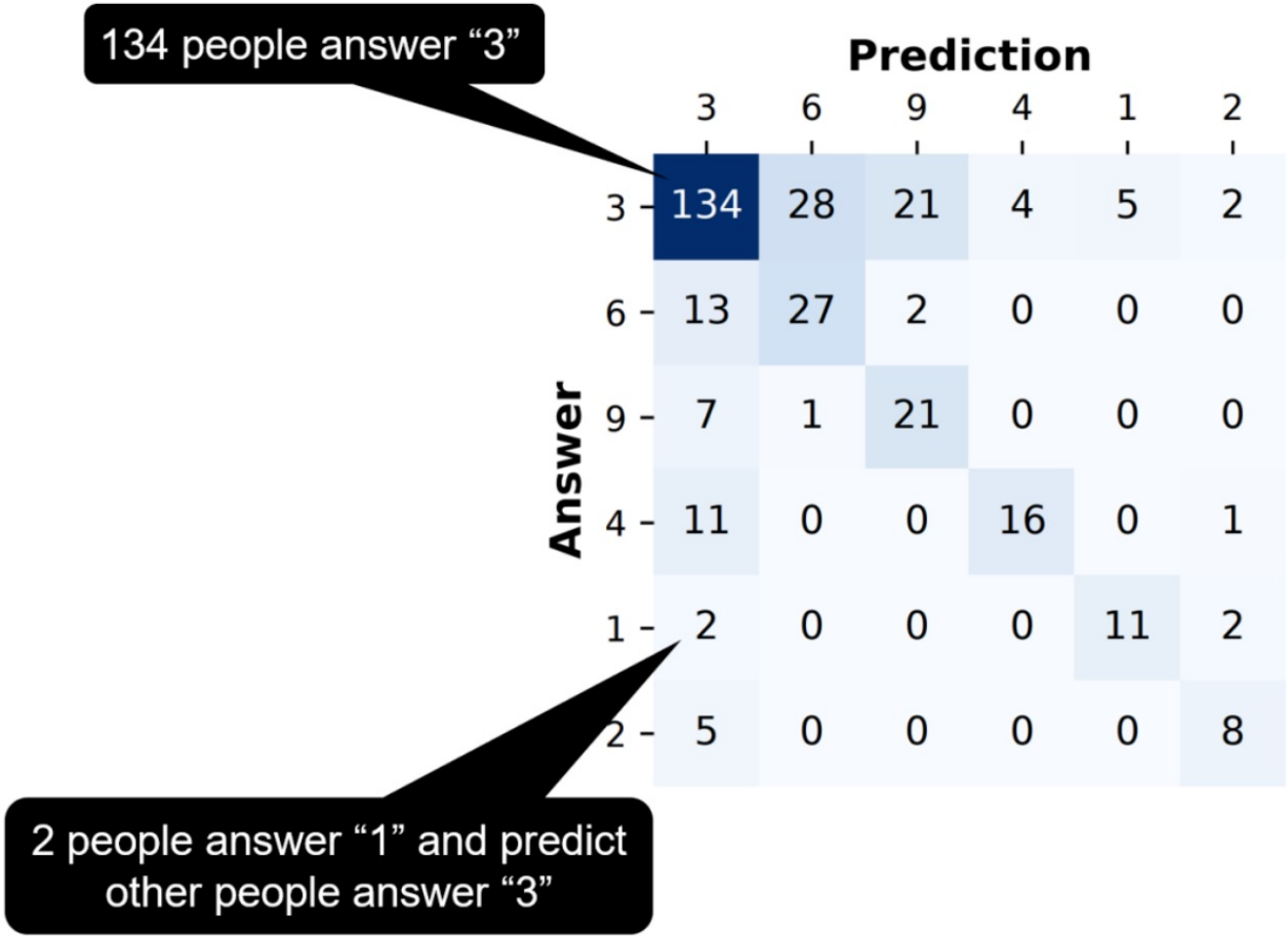
- **Related work in economics and game-theoretic**
  - Cognitive Hierarchy Theory (CHT)
  - Level-K Theory

- **Key Assumption**
  - The key insight is that people of a more sophisticated level know the mind of lower levels, but not vice versa.

- **Key Questions**
  - What is your answer?
  - What do you think other people will answer?

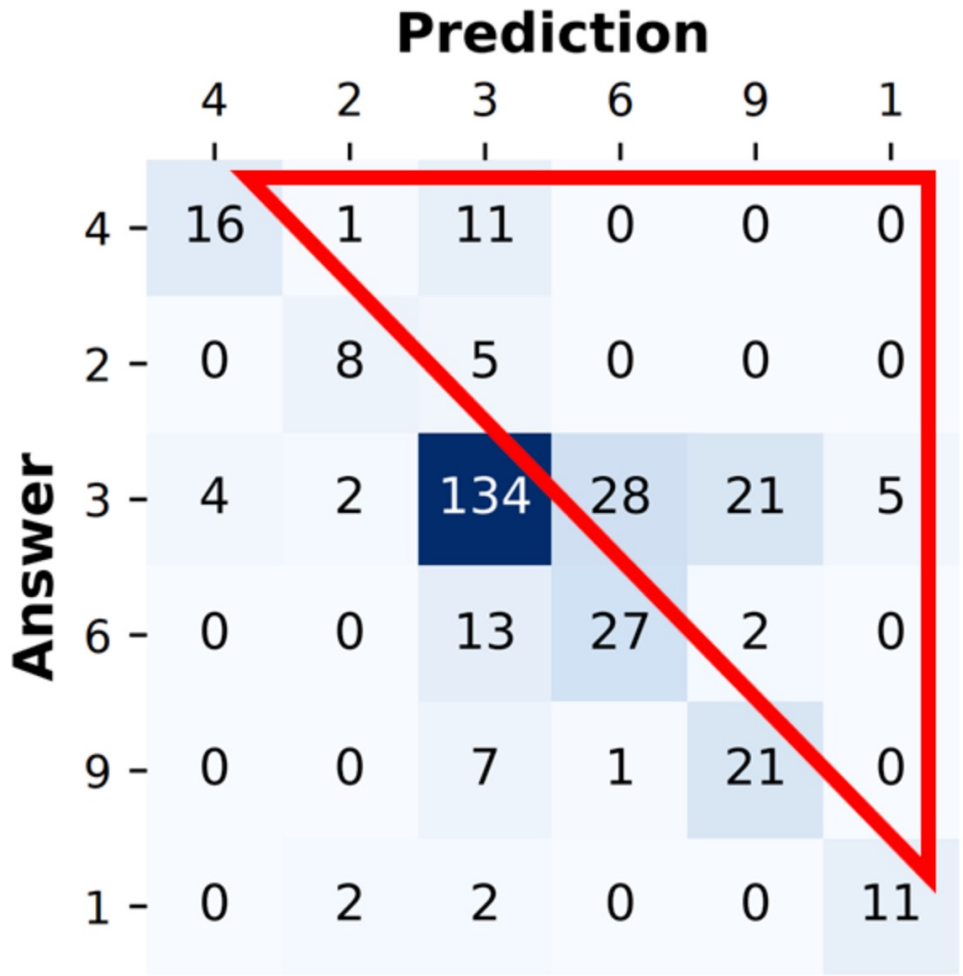
Method

- **Answer-Prediction Matrix**



Method

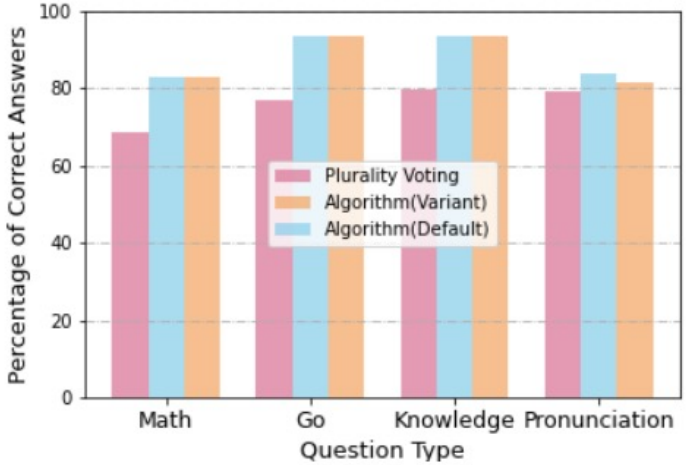
- **Ranking the answer *without any prior***



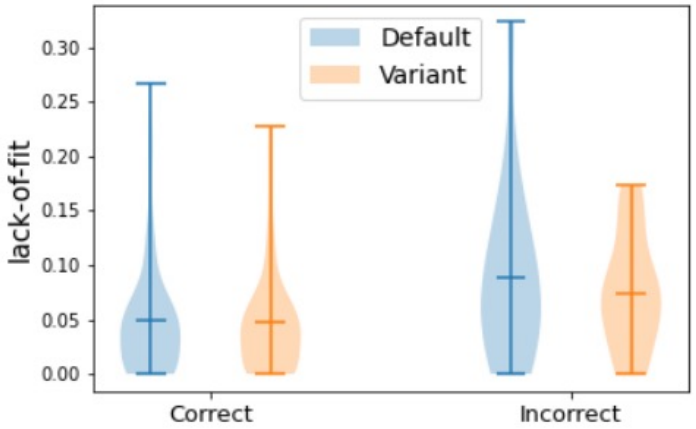
$$\pi^* \leftarrow \arg\max_{\pi} \sum_{i \leq j} M^2_{\pi(i),\pi(j)}$$

Maximum the sum of squares of elements in the **upper-triangular**

Experiments



(a) Accuracy of algorithms

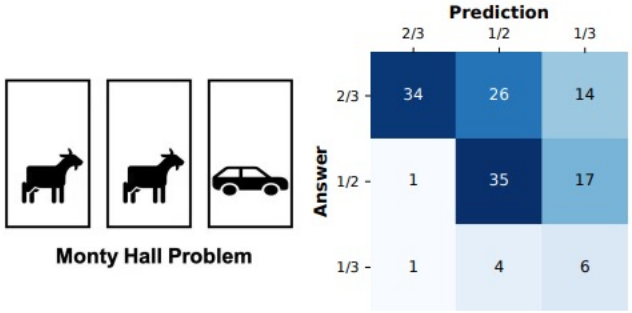


(b) Empirical distribution of lack-of-fit

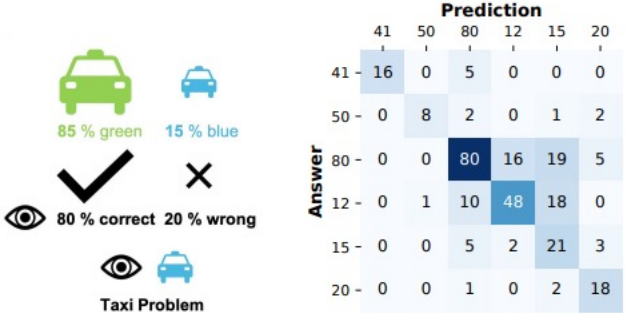
Type	Total	Our algorithm(Default)	Our algorithm(Variant)	Plurality voting
Math	35	**29**	**29**	24
Go	30	**28**	**28**	23
General knowledge	44	**41**	**41**	35
Chinese character	43	**36**	35	34

Table 1: The number of questions our algorithms/baseline are correct.

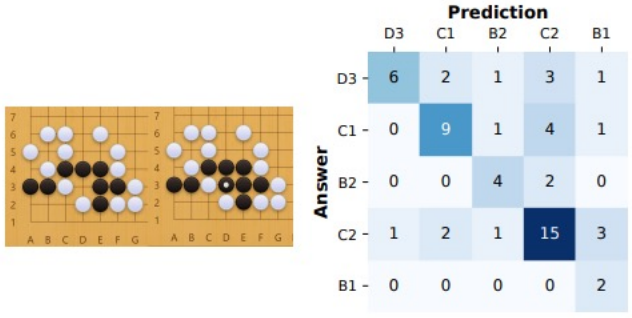
Experiments



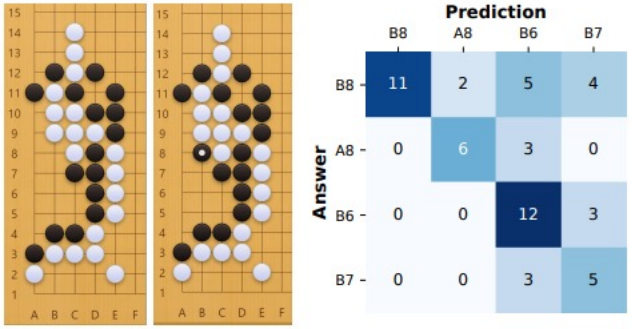
(a) the Monty Hall problem: you can select one closed door of three. A prize, a car, is behind one of the doors. The other two doors hide goats. After you have made your choice, Monty Hall will open one of the remaining doors and show that it does not contain the prize. He then asks you if you would like to switch your choice to the other unopened door. What is the probability to get the prize if you switch?



(b) the Taxicab problem: 85% of taxis in this city are green, the others are blue. A witness sees a blue taxi. She is usually correct with probability 80%. What is the probability that the taxi saw by the witness is blue?

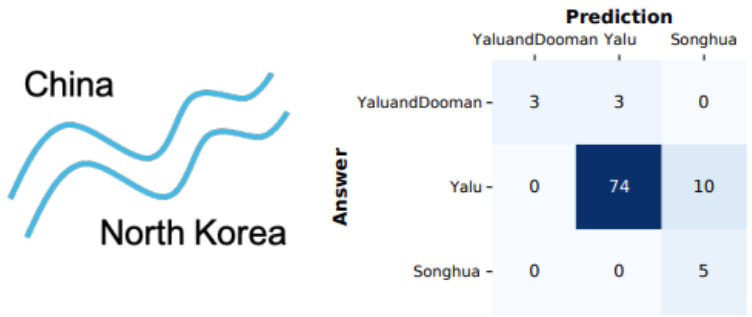


(c) Pick a move for black such that they can be alive.

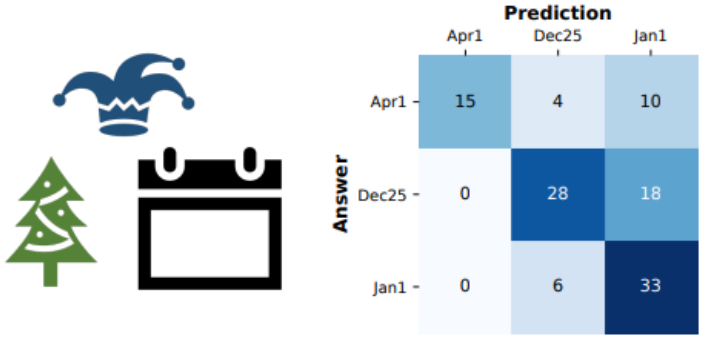


(d) Pick a move for black such that they can be alive by ko.

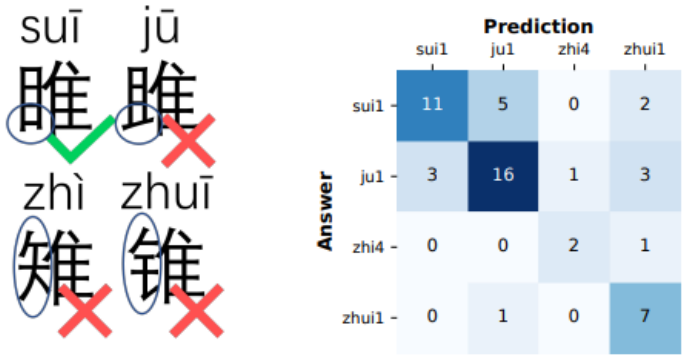
Experiments



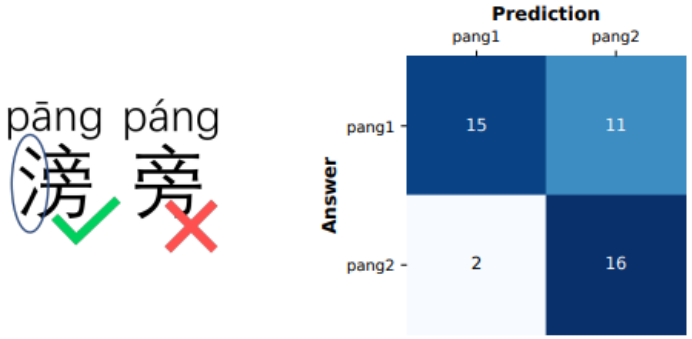
(e) the boundary question: what river forms the boundary between North Korea and China?



(f) the Middle Age New Year question: when was the new year in middle age?



(g) the pronunciation of 睢



(h) the pronunciation of 滂

Calibrating "Cheap Signals" in Peer Review without a Prior

Yuxuan Lu
Center on Frontiers of Computing Studies
School of Computer Science
Peking University
Beijing, China
yx_lu@pku.edu.cn

Yuqing Kong*
Center on Frontiers of Computing Studies
School of Computer Science
Peking University
Beijing, China
yuqing.kong@pku.edu.cn

NIPS'23

Background

- **Cheap Signal Cause Bias in Peer-Review**
  - Reference
  - Supplementary Material (Code)
  - Paper ID
  - Beautiful Figure
  - ……

- **Examples**

Example 1 (Peer review). *We have two papers and each paper is reviewed by 5 reviewers. We can only accept one paper. The first paper receives 4 "accept" and 1 "reject", the second paper receives 3 "accept"and 2 "reject".*

Example 2 (Hot vs. cold topic). *The topic of the first paper is more popular than the second paper. In this case, the first paper is easier to obtain reviewers with high expertise. Thus, in the noisy setting, the first paper has less noisy ratings than the second paper.*

Example 3 (Long proof vs. short proof). *The first paper has a complicated proof. The second paper has a short proof. In the noisy setting, each reviewer (unconsciously) intends to vote for "accept" for paper with a long proof, even if the length of the proof is a "cheap signal".*

Motivation

**Can we design a one-shot scoring process that leads to a noise-robust rank without any prior knowledge? Formally, we want the paper with a higher expected score in the clean setting also have a higher score in the noisy setting with high probability, even if different papers' reviews have different noise levels and biases.*

We adopt the signal-prediction framework of SP, i.e., asking each reviewer to additionally provide her prediction for a randomly selected reviewer's report. Besides, we follow the literature of information elicitation [3] and information aggregation [4] to assume that agents are perfect Bayesian, which establishes a start for theoretical analysis:

Assumption 1 (Bias generated from "cheap signals" can be predicted)**.** *(Informal) We assume agents are perfect Bayesian. Besides, when their ratings are noisy and have systematic biases, they will adjust their predictions based on the biased noise.*

Automatic Evaluation in LLMs

LANGUAGE MODEL SELF-IMPROVEMENT BY REINFORCEMENT LEARNING CONTEMPLATION

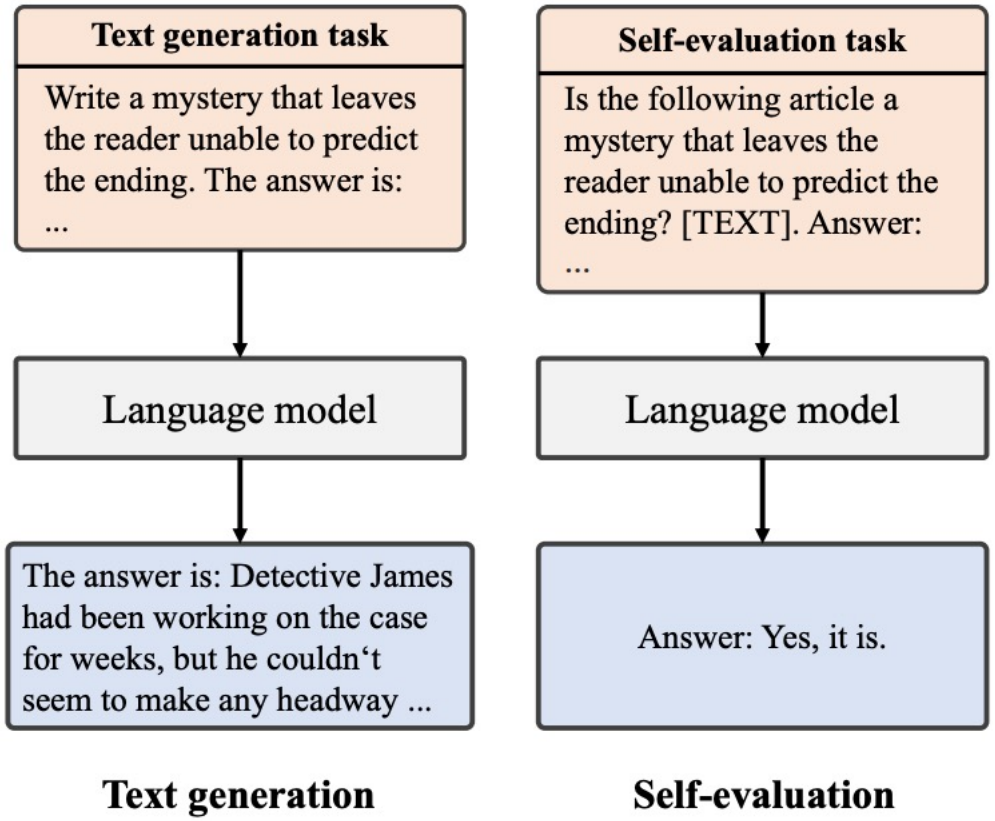
A PREPRINT

Jing-Cheng Pang[1,2,*], Pengyuan Wang[1,2,*], Kaiyuan Li[1], Xiong-Hui Chen[1,2], Jiacheng Xu[1], Zongzhang Zhang[1], and Yang Yu[1,2,◇]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Polixir.ai
[*] Equal contribution
[◇] Corresponding: yuy@nju.edu.cn

Motivation

- **LLMs can do better in evaluation than generation**



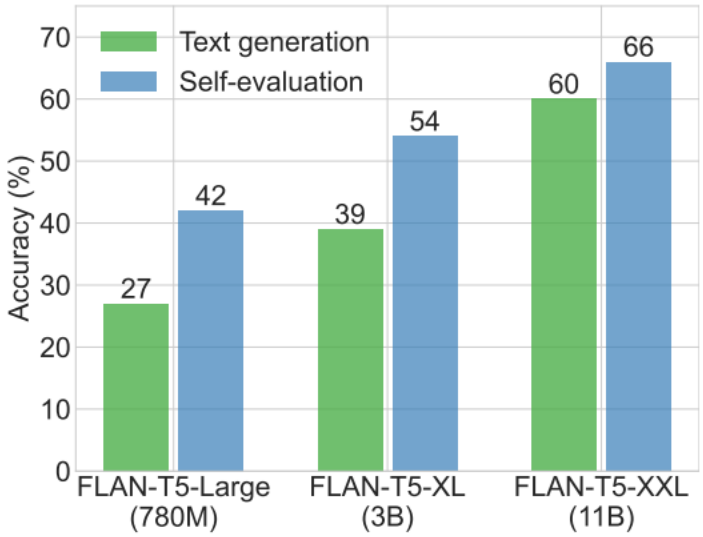
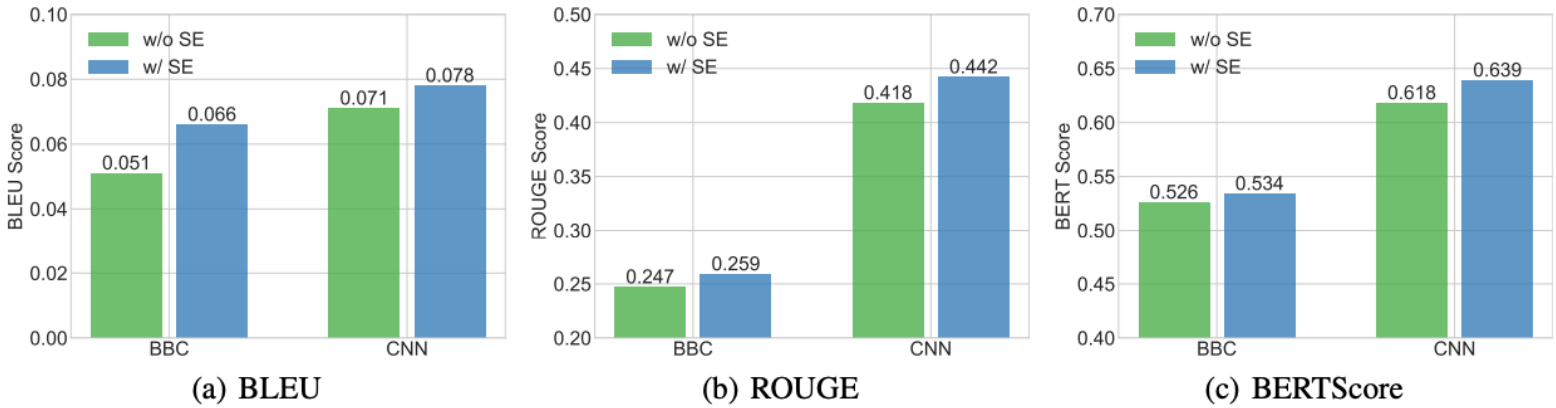


Figure 2: Comparison of the text generation and self-evaluation.

- **Language Model Self-Improve (LMSI)**
  - RLCAI, RLAIF use LLMs to replace RLHF

Toy Experiments



(a) BLEU (b) ROUGE (c) BERTScore

Figure 3: Comparison of text generation with/without self-evaluation on text summarization tasks.

	Reasoning about Colored Objects	Logical Deduction (7)	Tracking Shuffled Objects (5)	Object Counting
w/o SE	30.9%	18.5%	10.1%	34.7%
w/ SE	**31.1%**	**20.5%**	**11.1%**	**34.9%**

	Web of Lies	Sports Understanding	Logical Deduction (3)	Logical Deduction (5)
w/o SE	51.6%	59.7%	34.9%	23.6%
w/ SE	**53.2%**	59.7%	**38.3%**	**25.7%**

	Penguins in a Table	Navigate	Tracking Shuffled Objects (3)	Geometric Shapes
w/o SE	23.5%	47.7%	28.1%	10.7%
w/ SE	**28.8%**	**50.5%**	**31.5%**	**13.5%**

Table 2: Comparison of the answer accuracy between answer generation with/without self-evaluation. Full results on all 27 BigBench tasks are presented in Appendix C.2.

Method

- **Self-improvement by Reinforcement Learning Contemplation**

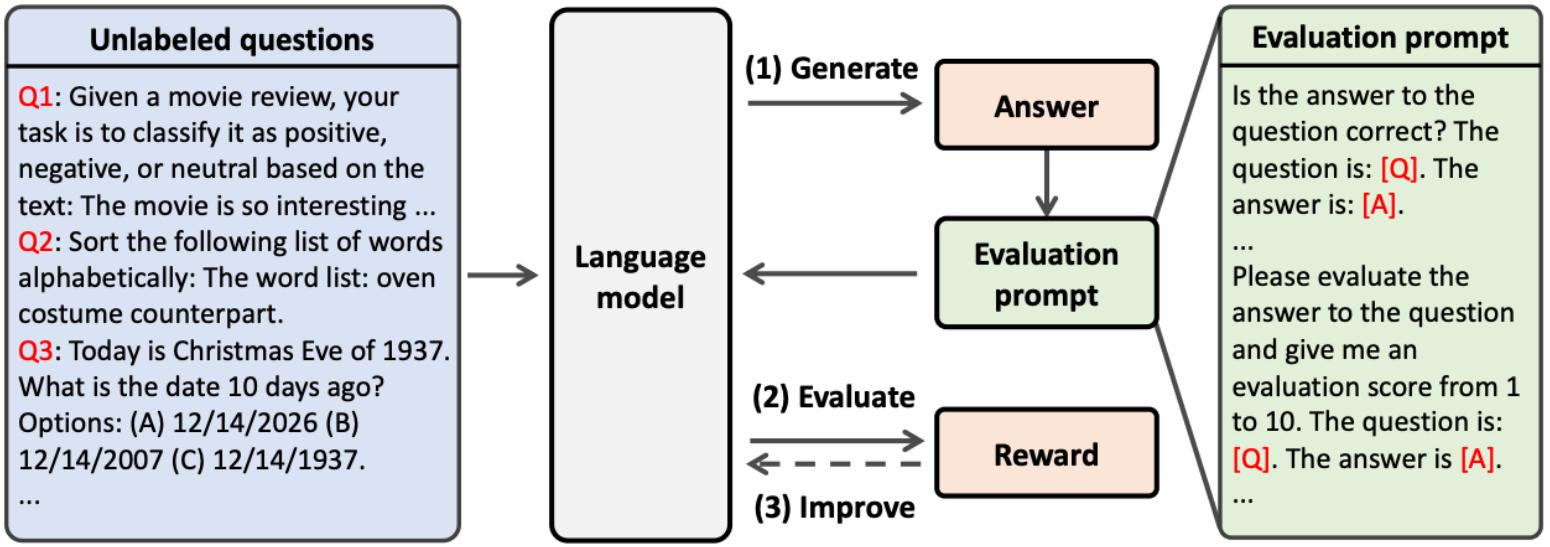


Figure 4: Overall training procedure of SIRLC, which iterates through three steps: (1) Answer generation to the unlabeled questions. (2) Self-evaluation by asking LM using *evaluation prompt*, with the evaluation results as the reward. (3) Update the language model to maximize the reward using reinforcement learning algorithms. The solid lines represent the data flow, while the dashed line represents the update of LLM parameters.

- **Self-evaluation as the reward**

$$R(q, o) = \phi(\mathcal{M}(p_{\text{EP}}, q, o)),$$

where ϕ is the text processing function, q is the question,
o is the LLM's output, $\mathcal{M}$ is the LLM, p_{EP} is the evaluation prompt.

LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models

Yen-Ting Lin, Yun-Nung Chen
National Taiwan University, Taipei, Taiwan
{ytl, y.v.chen}@ieee.org

Citation 23

LLM-Eval

{evaluation schema}

Score the following dialogue response generated on a continuous scale from 0.0 to 5.0.

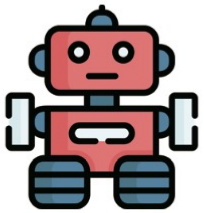
Context:

👨‍🦱: My cat likes to eat cream.
👩‍🦲: Be careful not to give too much, though.

Dialogue response :

👨‍🦱: Don't worry, I only give a little bit as a treat.

Context: {context}
Reference: {reference}
Dialogue response: {response}



Claude API

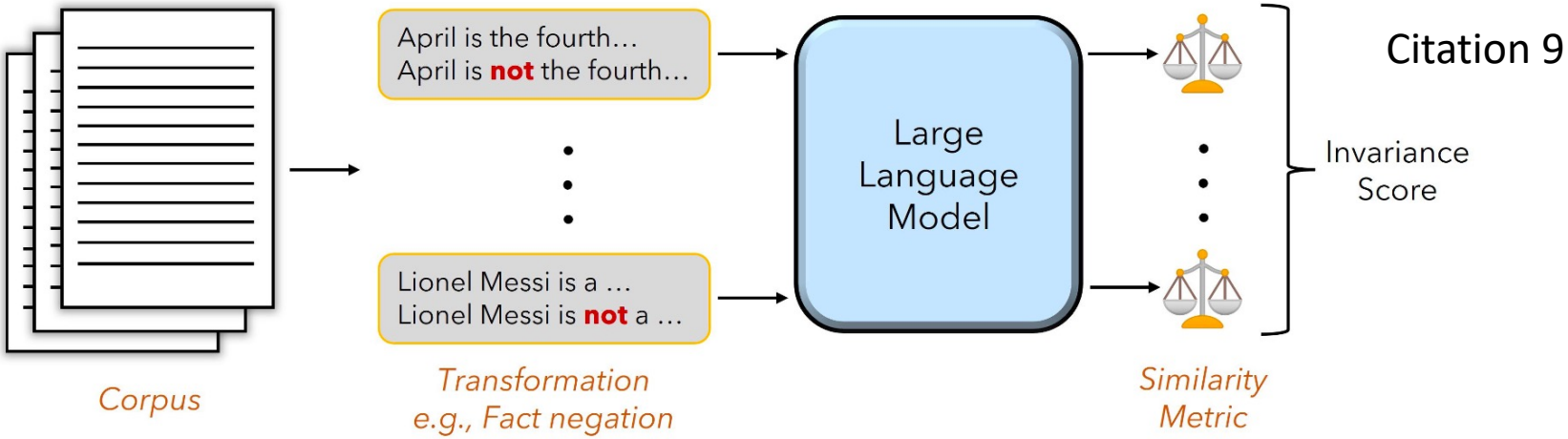
Appropriateness: 3.0
Content: 2.5
Grammer: 4.0
Relevence: 2.0

Output: {"appropriateness": 3.0, "content": 2.5, "grammar": 4.0, "relevance": 2.0}

Bring Your Own Data! Self-Supervised Evaluation of Large Language Models

Neel Jain* **Khalid Saifullah*** **Yuxin Wen** **John Kirchenbauer** **Manli Shu**

Aniruddha Saha **Micah Goldblum** † **Jonas Geiping** **Tom Goldstein**



Citation 9

Figure 1: In our proposed self-supervised evaluation, pairs are created from a corpus. Each pair contains the original and perturbed text, which in the figure above is creating a negation via applying a "not." These pairs are then fed into the network, and the outputs (perplexity, probability distributions, or text) are compared for each pair. These measures are then aggregated to produce an invariance or sensitivity score.

$$\text{SCORE} = A\{\mathcal{M}(f(x), f(x')) \; \forall (x, x') \in X\}.$$

aggregation operator A
Similarity metric M

Finding

- **Sensitivity score is strongly correlated with accuracy.**

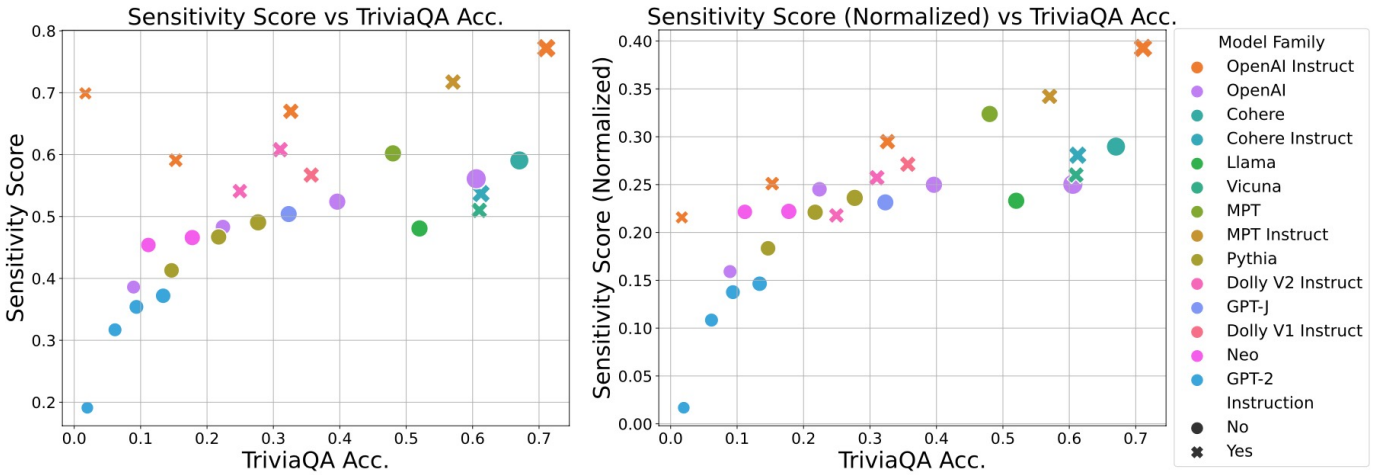


Figure 3: (Left) Sensitivity Score (negations) compared to accuracy on TriviaQA over various model sizes and families. **(Right)** Normalized Sensitivity Score compared to accuracy on TriviaQA over various model sizes and families. Larger markers correspond to bigger models, and "x" markers represent instruction finetuned models.

Finding

- **Sensitivity score is strongly correlated with accuracy.**

Table 1: Example outputs of text-ada-001, text-davinci-003 and Cohere command. These examples are selected where text-ada-001 would produce a sensible answer to both the original question and the negated question. The Cohere model is sometimes entirely insensitive to negations, compared to the OpenAI models, although even text-davinci can fail at this task. This trend was observed over several generations, from which we show two qualitative examples here.

Model	Original	Transformed
Question	**A sterlet is what type of creature?**	**A sterlet is `not` what type of creature?**
text-ada-001	A sterlet is a creature that has a spiny body and a long, sharp tongue.	A sterlet is not a creature.
text-davinci-003	A sterlet is a type of sturgeon.	A sterlet is a type of sturgeon.
Cohere command	Fish	Fish
Question	**What is the only natural food that never goes bad?**	**What is `not` the only natural food that never goes bad?**
text-ada-001	The only natural food that never goes bad is sugar.	There is no one natural food that never goes bad. There are, however, some foods that are more likely to do so. These include: milk, yogurt, ice cream, and cake.
text-davinci-003	Honey.	There is no single natural food that never goes bad.
Cohere command	Honey never goes bad.	Honey never goes bad.

PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization

Yidong Wang[1,2*], **Zhuohao Yu**[1*], **Zhengran Zeng**[1], **Linyi Yang**[2], **Cunxiang Wang**[2], **Hao Chen**[3], **Chaoya Jiang**[1], **Rui Xie**[1], **Jindong Wang**[3], **Xing Xie**[3], **Wei Ye**[1†], **Shikun Zhang**[1†], **Yue Zhang**[2†]

[1]Peking University [2]Westlake University [3]Microsoft Research Asia


```
"inputs": {
  "instruction": "Find an example of the given kind of data",
  "input": "Qualitative data",
  "response1": "An example of qualitative data is customer feedback.",
  "response2": "An example of qualitative data is a customer review."
}
```

```
"outputs": {
  "evaluation_result": "Tie",
  "evaluation_reason": "Both responses are correct and provide similar examples of qualitative data.",
  "reference_response": "An example of qualitative data is an interview transcript."
}
```

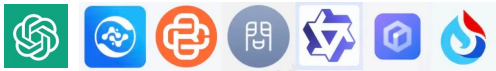

Figure 3: A training data example for PandaLM-7B.

Peer-review in LLMs: An Automatic Evaluation System for Ranking LLMs without Human-feedback

Kun-Peng Ning

Automatic Evaluation System

- Goal

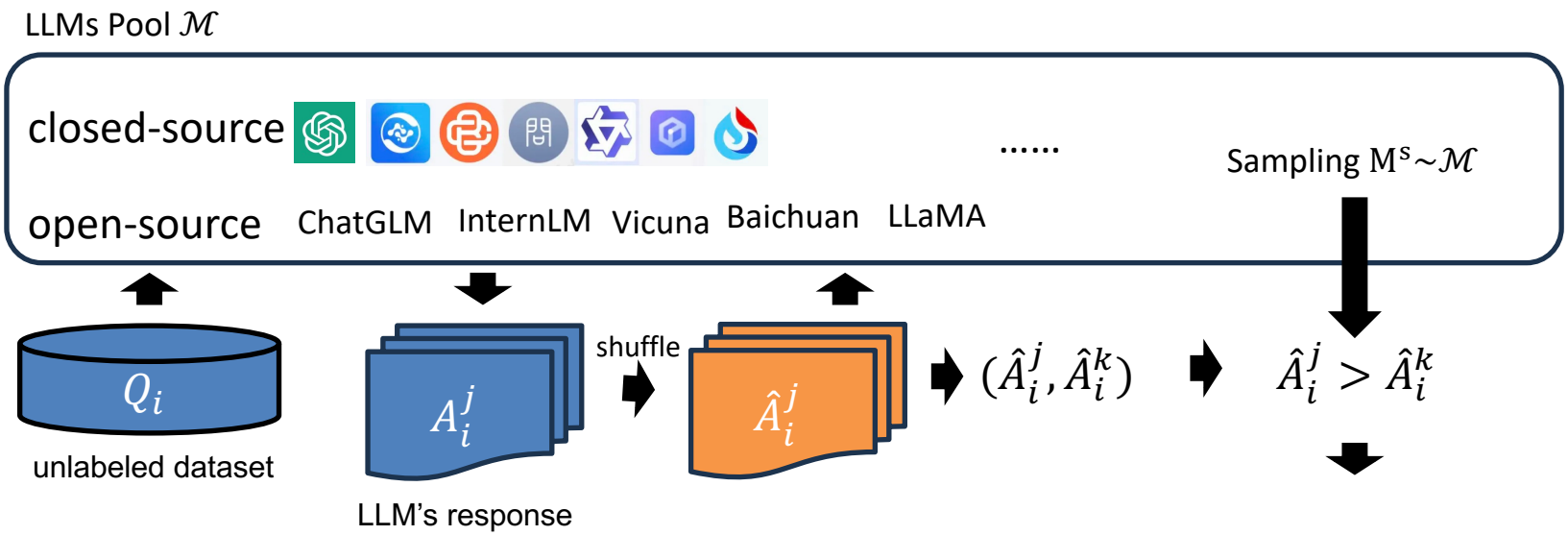
closed-source  ……

open-source ChatGLM InternLM Vicuna Baichuan LLaMA

排名	模型	机构	总分	基础能力	中文特性	学术专业	许可证
	人类	CLUE	83.66	85.03	82.29	-	-
-	GPT-4	OpenAI	70.89	70.04	72.67	69.96	专有服务
	文心一言(v2.2.0)	百度	62.00	61.11	71.38	53.50	专有服务
-	Claude-2	Authropic	60.94	62.01	61.18	59.63	专有服务
-	gpt-3.5-turbo	OpenAI	59.79	64.40	63.19	51.78	专有服务
	ChatGLM-130B	清华大学&智谱AI	59.35	53.78	71.39	52.89	专有服务
	讯飞星火(v1.5)	科大讯飞	58.02	63.32	65.72	45.03	专有服务
-	Claude-instant-v1	Authropic	56.31	58.85	55.91	54.16	专有服务
4	360智脑(4.0)	360	55.04	56.68	62.54	45.88	专有服务
5	internlm-chat-7b	上海AI实验室与商汤	53.91	54.85	61.35	45.53	开源-可商用
6	ChatGLM2-6B	清华大学&智谱AI	53.85	55.60	63.59	42.37	开源-可商用
7	MiniMax-abab5.5	MiniMax	53.06	53.61	62.79	42.77	专有服务
8	通义千问(v1.0.3)	阿里巴巴	51.52	52.84	61.73	39.98	专有服务
9	Baichuan-13B-Chat	百川智能	49.35	50.46	55.38	42.21	开源-可商用
10	BELLE-LLaMA-13B-2M-enc	链家	46.60	48.71	52.99	38.10	开源-非商用
11	IDEA-姜子牙-13B-v1.1	深圳IDEA研究院	43.80	47.55	48.61	35.26	开源-非商用
12	phoenix-7B	香港中文大学	41.57	45.39	44.62	34.70	开源-可商用
13	MOSS-16B	复旦大学	35.36	37.01	38.01	29.27	开源-可商用
14	Llama-2-13B-chat	Meta	34.26	35.85	37.37	29.57	开源-可商用

- What is a good LLMs Evaluation
  - ✓ Environment (not Benchmark)
  - ✓ Efficiency
  - ✓ Low-cost
  - ✓ Objective and Fair
  - ✓ Unsupervised
  - ✓ Preference Alignment

IDEA



Assumptions:

- ✓ Higher-level LLM can predict answer-ranking more accurately (more confidence) than lower-level one.
- ✓ Higher-level LLM also have higher answer-ranking score.
- ✓ The entropy of the automatic evaluation system should be minimized.

Satisfy three assumptions

Thanks