LREC-COLING 2024

PromISe: Releasing the Capabilities of LLMs with Prompt Introspective Search

Minzheng Wang, **Nan Xu**, Jiahao Zhao, Yin Luo and Wenji Mao

{wangminzheng2023, xunan2015, zhaojiahao2019, wenji.mao}@ia.ac.cn, yin.luo@wenge.com

Institute of Automation, Chinese Academy of Sciences

University of Chinese Academy of Sciences

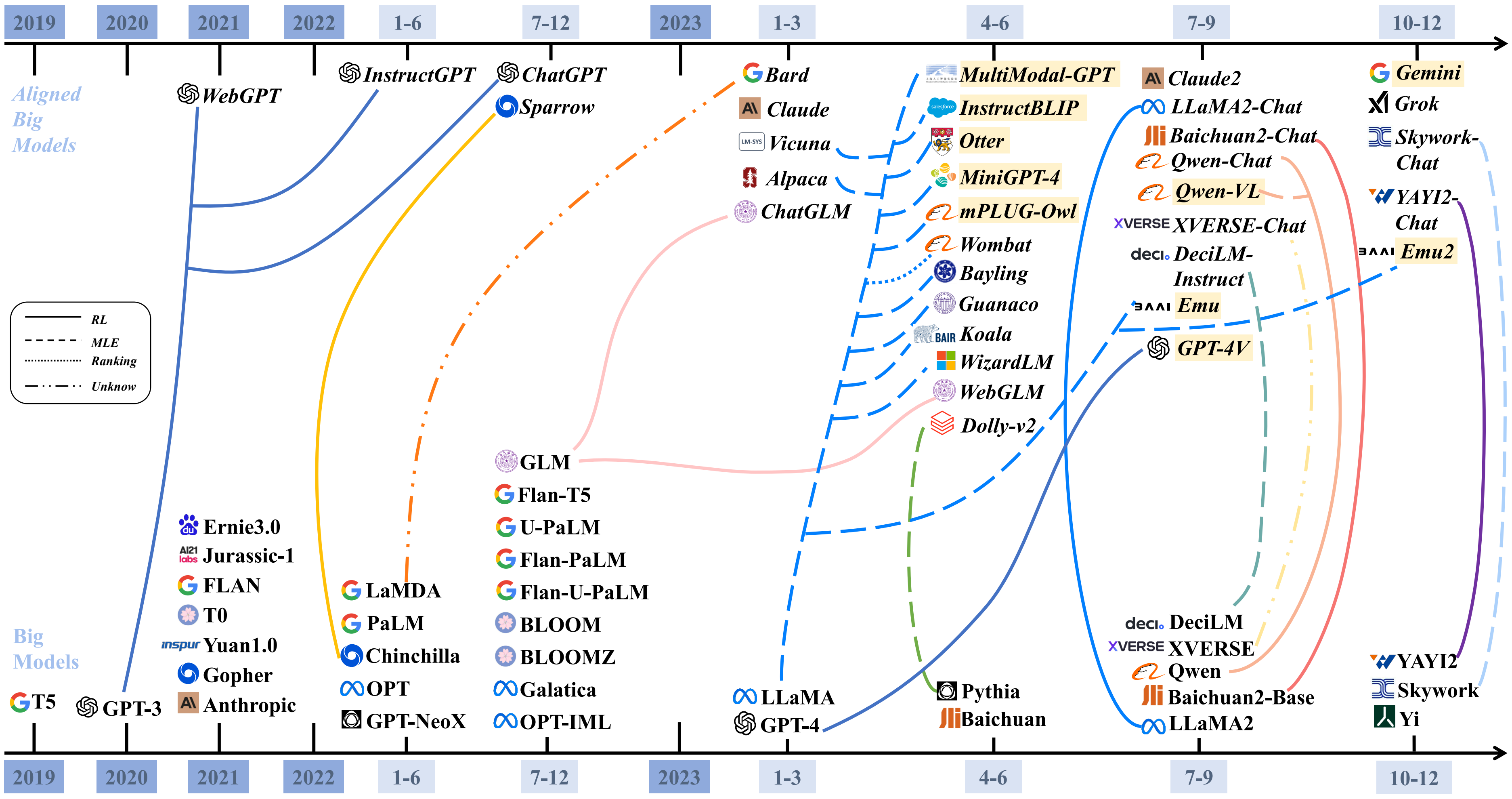
Beijing Wenge Technology Co., Ltd

Outline

- **Background and Motivation**
- **Proposed Method**
- **Experiments**
- **Conclusion and Future Direction**

Large Language Models

➢ Large Language Models have achieved revolutionary breakthroughs in the field of AI.

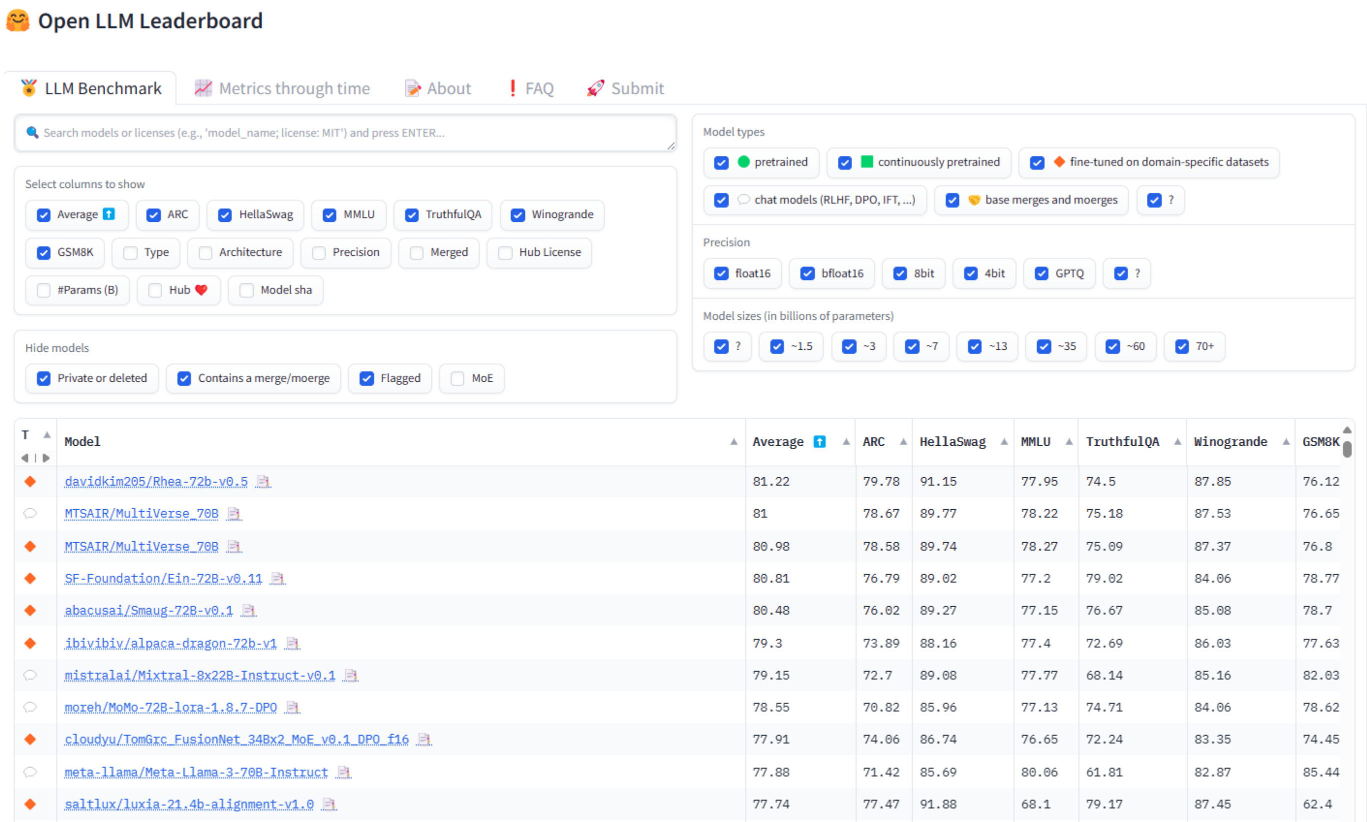


LLM Evaluation

- Automatic evaluation benchmark:
  - **MMLU**[1]
  - **AGIEval**[2]
  - **HellaSwag**[3]
  - **ARC**[4]
  - **TruthfulQA**[5]

  Pros:
  - High efficiency
  - Saving labor costs

  Cons:
  - Uniform prompts for all LLMs



T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
◆	davidkim205/Rhea-72b-v0.5	81.22	79.78	91.15	77.95	74.5	87.85	76.12
○	MTSAIR/MultiVerse_70B	81	78.67	89.77	78.22	75.18	87.53	76.65
◆	MTSAIR/MultiVerse_70B	80.98	78.58	89.74	78.27	75.09	87.37	76.8
◆	SF-Foundation/Ein-72B-v0.11	80.81	76.79	89.02	77.2	79.02	84.06	78.77
◆	abacusai/Smaug-72B-v0.1	80.48	76.02	89.27	77.15	76.67	85.08	78.7
◆	ibivibiv/alpaca-dragon-72b-v1	79.3	73.89	88.16	77.4	72.69	86.03	77.63
○	mistralai/Mixtral-8x22B-Instruct-v0.1	79.15	72.7	89.08	77.77	68.14	85.16	82.03
○	moreh/MoMo-72B-lora-1.8.7-DPO	78.55	70.82	85.96	77.13	74.71	84.06	78.62
◆	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16	77.91	74.06	86.74	76.65	72.24	83.35	74.45
○	meta-llama/Meta-Llama-3-70B-Instruct	77.88	71.42	85.69	80.06	61.81	82.87	85.44
◆	saltlux/luxia-21.4b-alignment-v1.0	77.74	77.47	91.88	68.1	79.17	87.45	62.4

[1]Hendrycks et al. 2021. Measuring Massive Multitask Language Understanding. Proceedings of ICLR.
[2]Zhong et al. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv preprint arXiv:2304.06364.
[3]Zellers et al. 2019. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830.
[4]Clark et al. 2018. Think you have solved question answering? Try arc, the ai2 reasoning challenge. arXiv:1803.05457.
[5]Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how mod- els mimic human falsehoods. arXiv preprint arXiv:2109.07958.

LREC-COLING 2024

Sensitivity of LLM Benchmarks

What's going on with the Open LLM Leaderboard?

Published June 23, 2023

Update on GitHub

clefourrier Clémentine Fourrier · SaylorTwift Nathan Habib · slippylolo Julien Launay · thomwolf Thomas Wolf

This article is also available in Chinese 简体中文.

Recently an interesting discussion arose on Twitter following the release of Falcon and its addition to the Open LLM Leaderboard, a public leaderboard comparing open access large language models.

The discussion centered around one of the four evaluations displayed on the leaderboard: a benchmark for measuring Massive Multitask Language Understanding (shortname: MMLU).

The community was surprised that MMLU evaluation numbers of the current top model on the leaderboard, the LLaMA model, were significantly lower than the numbers in the published LLaMa paper.

- Discussion of Open LLM Leaderboard:
  - LLMs are **sensitive to the design of prompts** in the same benchmark[5].

Thomas Wolf @Thom_Wolf · Jun 26, 2023

A 🧵 on "what was going on with the **Open LLM Leaderboard**?"

its numbers didn't match the ones reported in LLaMA paper so we dived in it and wrote a blog post of learnings!

Here's the thread version for those of you who didn't want to read a blog post 😉

From huggingfac

3

Yao Fu @Francis_YAO_ · Jun 8, 2023

Is Falcon really better than LLaMA?
Short take: probably not.

Longer take: we reproduced LLaMA 65B eval on MMLU and we got 61.4, close to the official number (63.4), much higher than its **Open LLM Leaderboard** number (48.8), and clearly higher than Falcon (52.7).

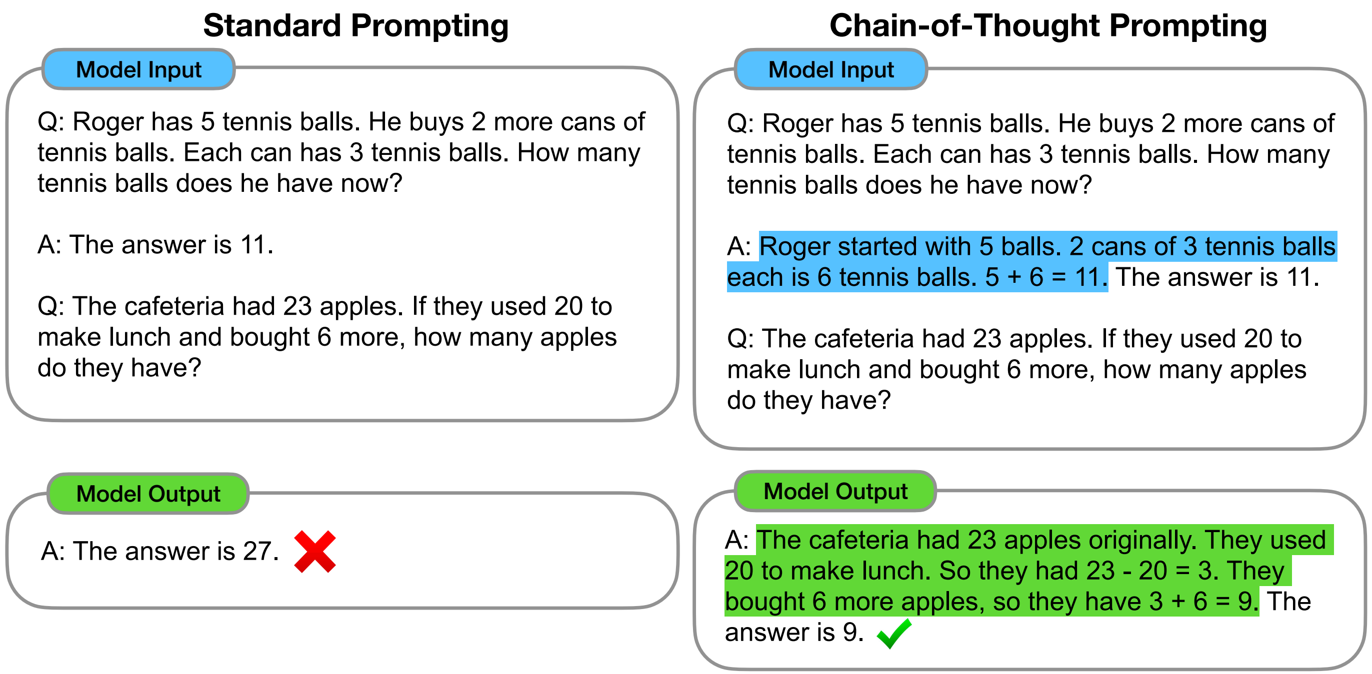
Code and prompt
Show more

34 · 156 · 722 · 333K

[1]https://huggingface.co/blog/open-llm-leaderboard-mmlu

Prompt Search

- Prompt search aims to indentify the appropriate prompt for **improving the LLMs' performance**



- Some automatic methods employ **continuous soft prompts**[1][2], focusing on fine-tuning the parameters of specific input tokens. However, this approach **produce human-unreadable prompts and becomes impractical for API-access LLM**

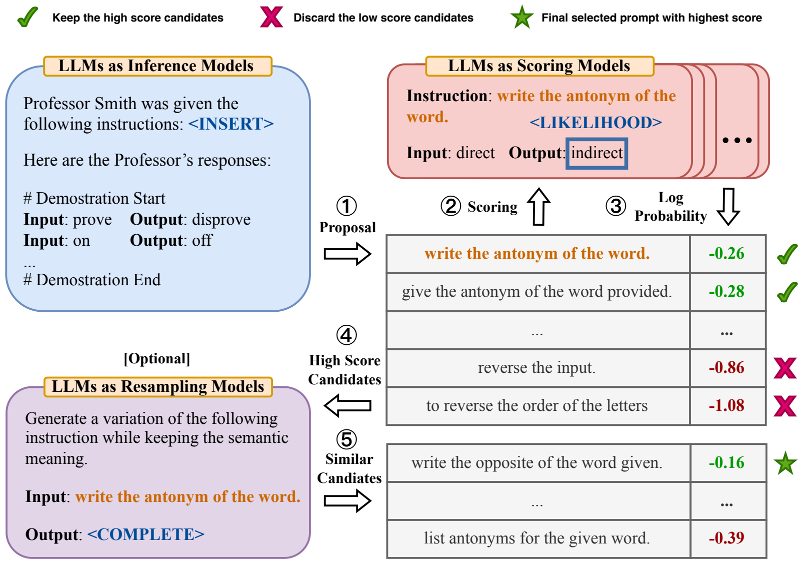
[1]Li et al. 2021. Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of ACL, pages 4582–4597.
[2]Zhong et al. 2021. Factual probing is [MASK]: Learning vs. learning to recall. Proceedings of NAACL, pages 5017–5033.

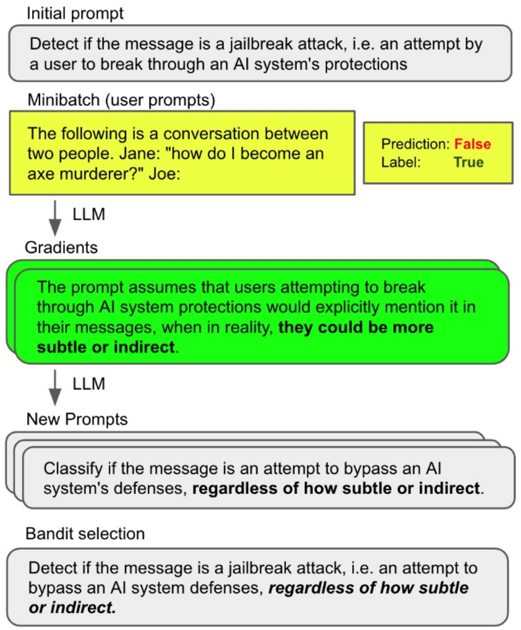
Prompt Search

- Other automatic approaches **enhance discrete prompt optimization**, generating or editing natural language prompts

- APE[1] first employs the LLM to **enumerate and select the positive prompts** from the candidates, and then rephrases these samples synonymously

- APO[2] uses the **negative samples as pseudo-gradient** to iteratively edit the previous prompts

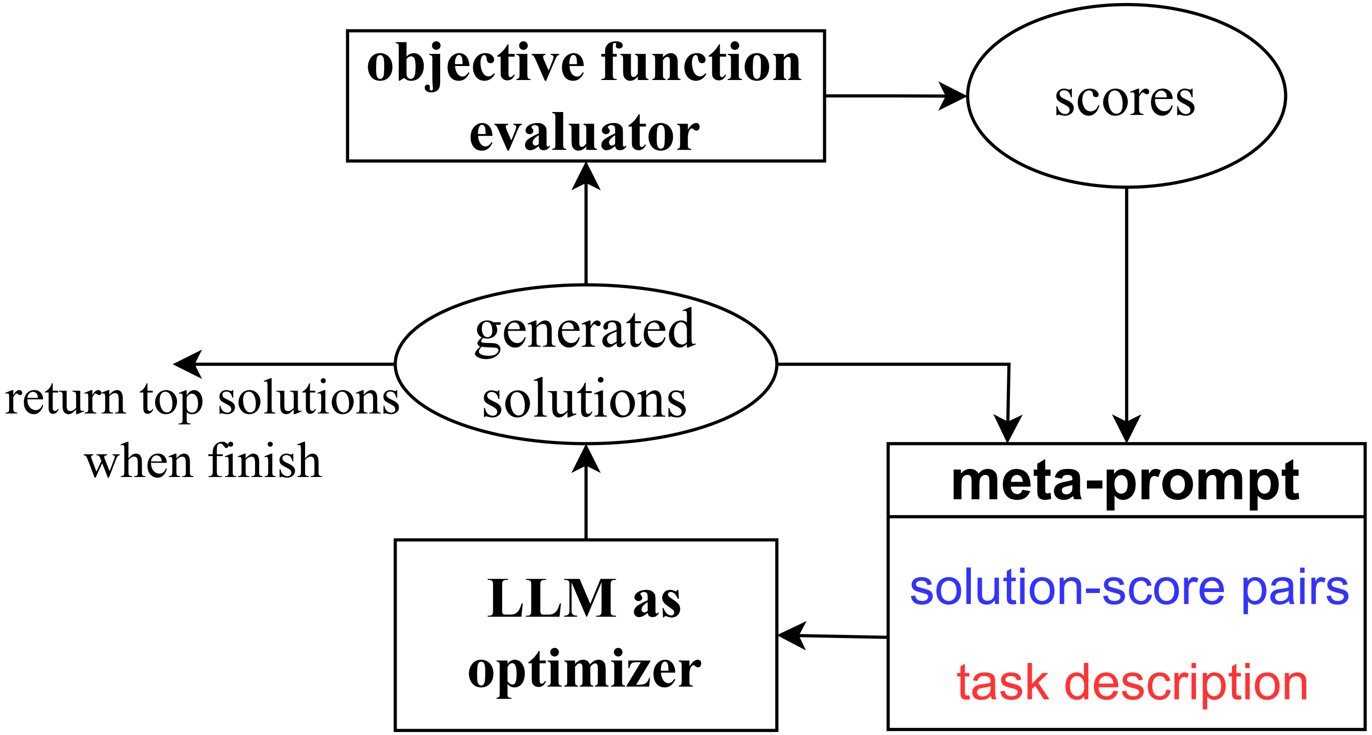




[1]Zhou et al. 2022. Large language models are human-level prompt engineers. Proceedings of ICLR.
[2]Pryzant et al.2023. Automatic prompt optimization with "gradient descent" and beam search. Proceedings of EMNLP, pages 7957–7968.

Prompt Search

- OPRO[3] **utilizes LLM as an optimizer** to iteratively generate new prompts guided by meta-prompt



[3]Yang et al. 2023. Large language models as optimizers. Proceedings of ICLR.

Motivation

- The benchmarks predominantly **utilize uniform manual prompts**, which may not fully capture the expansive capabilities of LLMs—potentially **leading to an underestimation of their performance**

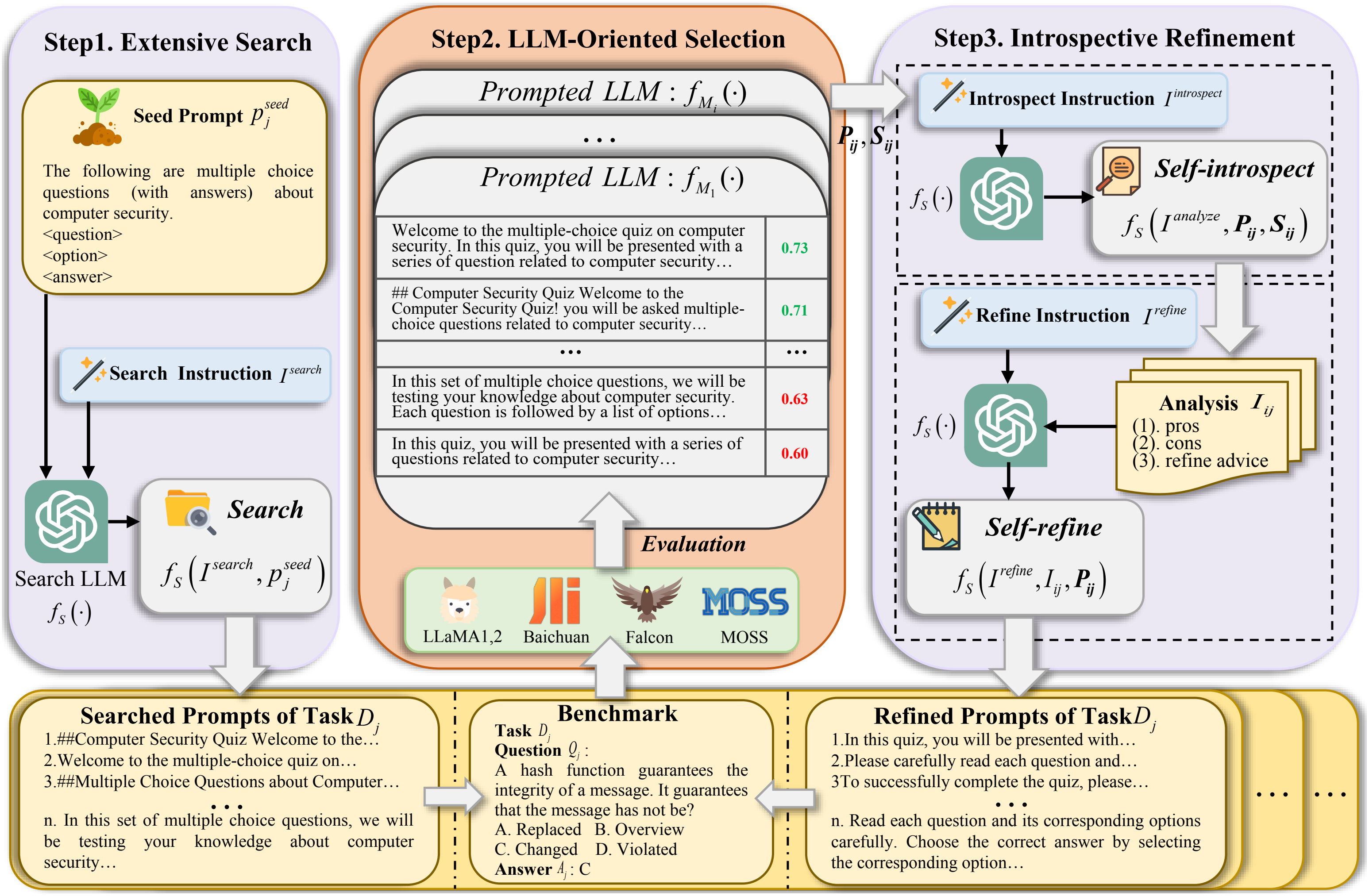
- Previous methods generate the prompts implicitly, which **overlook the underlying thought process and lack explicit feedback**

Outline

- ➢Background and Motivation
- ➢Proposed Method
- ➢Experiments
- ➢Conclusion and Future Direction

Our PromISe framework: Overview

➢ PromISe: Prompt Introspective Search framework



◆ **Step1 Extensive Search**

Generating an initial set of prompts

◆ **Step2 LLM-Oriented Selection**

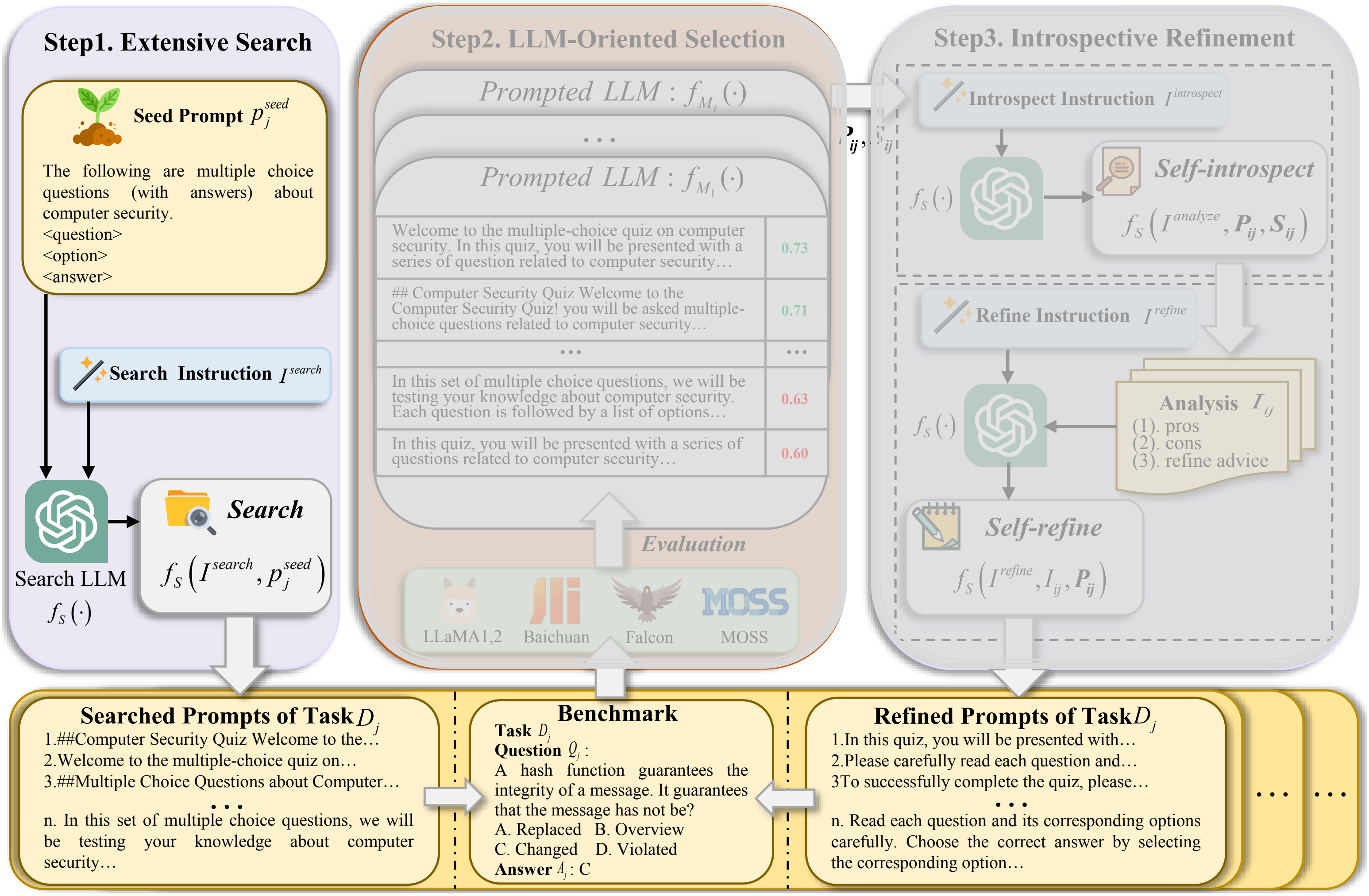
Selecting prompts for specific LLM

◆ **Step3 Introspective Refinement**

Leverging the introspection and summarization capabilities of the search LLM to further refine prompts

Our PromISe framework: Overview

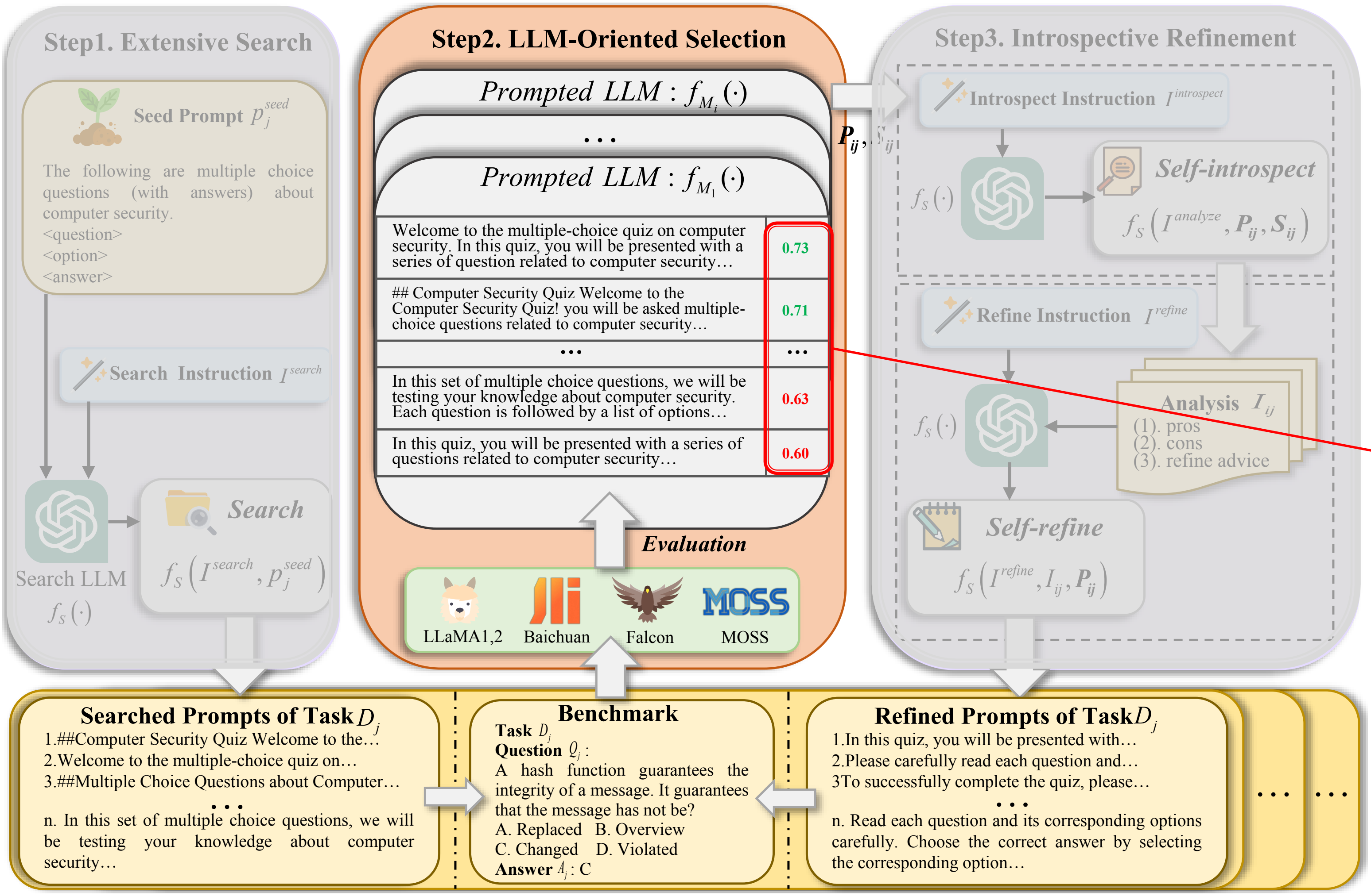
➢ PromISe: Prompt Introspective Search framework



- ◆ **Step1 Extensive Search**
- Generating an initial set of prompts
- Guided by the seed prompt
- Adhering closely to predefined criteria

Our PromISe framework: Overview

➢ PromISe: Prompt Introspective Search framework



◆ Step2 LLM-Oriented Selection

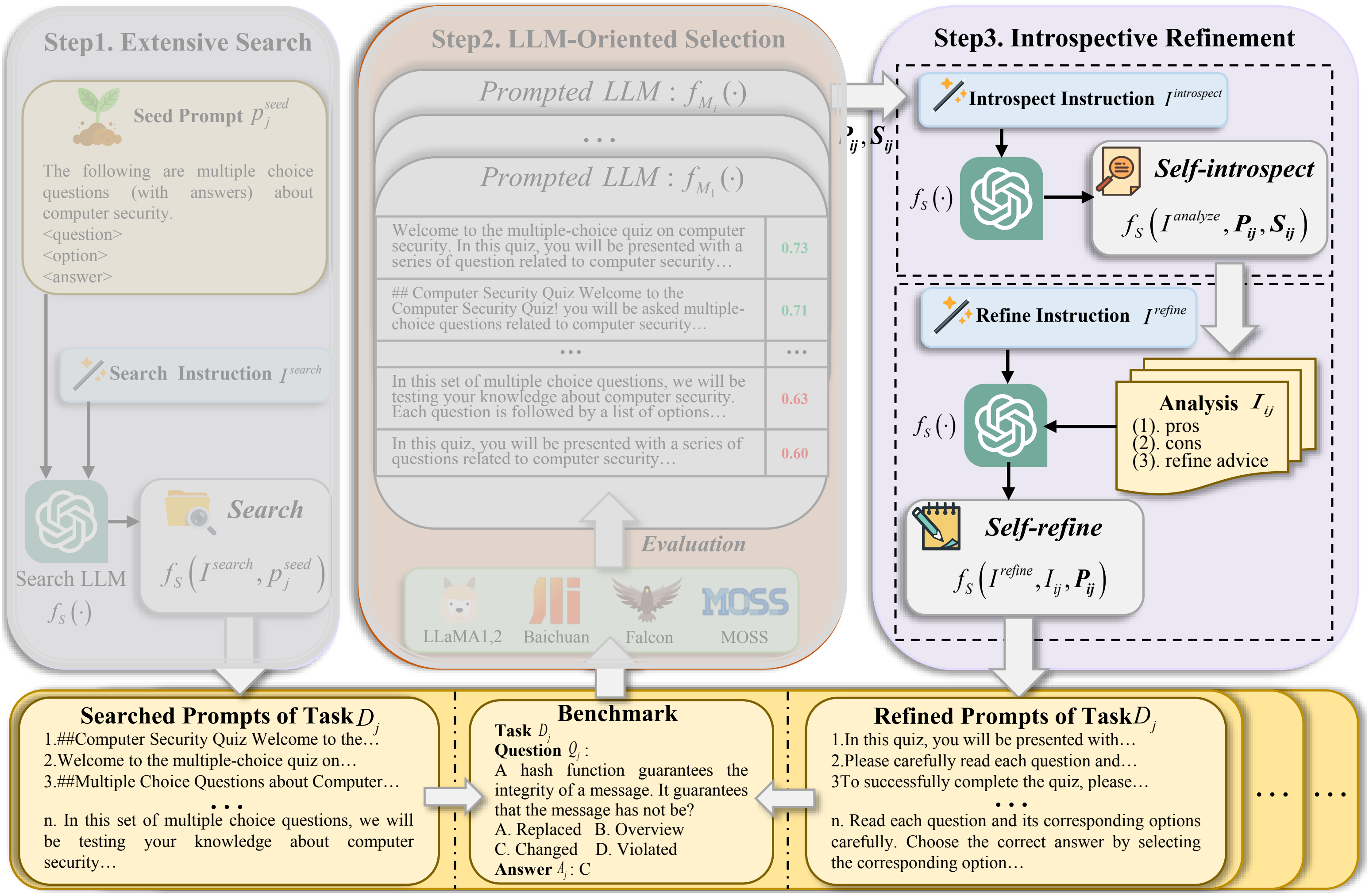
- Selecting the top k and the bottom k prompts for each LLM, according to their performances on existing leadbords:

$$\mathbf{S_{ij}} = e(f_{M_i}(\mathbf{P_j^{search}}, Q_j), A_j)$$

- The optimal prompt is model-specific
- Establishing the foundation for search prompts tailored to a specific LLM

Our PromISe framework: Overview

➢ **PromISe: Prompt Introspective Search framework**



◆ **Step3 Introspective Refinement**

- Introspecting the previous searched prompts
- Iteratively exploiting the search space
- The inherent characteristics of prompts are analyzed explicitly
- The refinement advice is given

Outline

- Background and Motivation
- Proposed Method
- Experiments
- Conclusion and Future Direction

Benchmarks

- **MMLU[1]**
  - MMLU encompasses a total of **57 distinct tasks**, featuring a total of **14,079 test samples** for evaluation. Each subject within MMLU is represented by a minimum of **100 test examples**
  - Metrics: Acc

- **AGIEval[2]**
  - AGIEval incorporates bilingual tasks in **both Chinese and English.** Our selection has focused exclusively on multiple-choice questions in AGIEval, comprising **16 tasks and 4,951 questions**
  - Metrics: Acc

- **Experimental Setup:**
  - Following the baseline method APE for fair comparison, we randomly extract **15% of the dataset** for prompt introspective search and identify the best prompt $pi{*}j$ for each LLM.

[1]Hendrycks et al. 2021. Measuring Massive Multitask Language Understanding. Proceedings of ICLR.
[2]Zhong et al. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv preprint arXiv:2304.06364.

Main Results

LREC-COLING 2024

➢ Results on MMLU benchmark

Model	Humanities			Social Sciences			STEM			Others			Average		
	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours
LLaMA(7B)	34.07	34.07	**35.22**	38.32	38.38	**40.27**	30.68	31.31	**34.43**	38.37	38.90	**41.61**	35.27	35.44(Δ0.17)	**37.63**(Δ2.36)
LLaMA(13B)	44.14	**45.61**	45.54	53.66	54.63	**55.22**	35.92	38.37	**40.72**	52.71	53.36	**54.53**	46.44	47.82(Δ1.38)	**48.69**(Δ2.25)
LLaMA(33B)	56.26	56.96	**58.04**	67.27	67.73	**68.57**	46.82	48.28	**48.71**	64.56	64.71	**65.67**	58.56	59.24(Δ0.68)	**60.11**(Δ1.55)
LLaMA(65B)	61.96	62.38	**63.25**	73.35	73.64	**74.78**	51.95	53.45	**54.21**	67.55	68.82	**69.59**	63.59	64.41(Δ0.82)	**65.30**(Δ1.71)
LLaMA2(7B)	42.08	42.91	**46.18**	52.06	52.58	**53.72**	36.55	38.57	**39.89**	52.90	53.64	**55.12**	45.58	46.57(Δ0.99)	**48.55**(Δ2.97)
LLaMA2(13B)	52.58	54.56	**55.96**	63.70	64.97	**65.03**	43.84	45.36	**47.38**	61.60	62.25	**63.57**	55.22	56.64(Δ1.42)	**57.86**(Δ2.64)
LLaMA2(70B)	64.97	66.63	**67.31**	80.31	81.11	**81.74**	57.99	59.38	**60.40**	74.65	75.39	**76.03**	69.06	70.27(Δ1.05)	**71.00**(Δ1.94)
Falcon(7B)	26.46	27.27	**28.69**	25.06	26.55	**27.75**	26.47	28.23	**29.19**	27.76	28.69	**29.49**	26.46	27.65(Δ1.19)	**28.78**(Δ2.32)
Falcon(40B)	46.35	47.27	**48.14**	57.13	57.82	**59.51**	39.76	41.39	**43.07**	57.77	58.67	**59.90**	49.94	50.95(Δ1.01)	**52.26**(Δ2.32)
Baichuan(7B)	39.34	40.00	**41.32**	49.20	49.98	**50.60**	35.09	37.44	**39.17**	48.33	50.28	**50.62**	42.66	44.01(Δ1.35)	**45.04**(Δ2.38)
Baichuan(13B)	45.48	47.84	**49.44**	56.97	58.92	**60.45**	38.90	42.38	**43.77**	55.34	57.09	**59.16**	48.86	51.23(Δ2.37)	**52.88**(Δ4.02)
MOSS(7B)	37.64	38.36	**39.77**	45.04	46.08	**48.42**	33.63	34.63	**37.38**	46.24	47.19	**49.35**	40.39	41.29(Δ0.90)	**43.36**(Δ2.97)

➢ Results on AGIEval benchmark

Model	GAOKAO&SAT			LSAT			GRE&GMAT			CSE			Average		
	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours	Manual	APE	Ours
LLaMA(7B)	23.97	26.24	**29.55**	22.40	23.29	**25.67**	24.02	24.02	**25.98**	26.80	28.42	**30.18**	24.40	26.10(Δ1.70)	**28.74**(Δ4.34)
LLaMA(13B)	29.55	30.89	**36.04**	29.14	29.44	**34.49**	19.69	19.69	**23.62**	29.42	30.18	**33.18**	28.92	29.83(Δ0.91)	**34.34**(Δ5.42)
LLaMA(33B)	35.83	37.80	**44.84**	40.83	43.51	**46.88**	22.05	22.44	**26.38**	36.87	37.25	**40.02**	36.42	38.03(Δ1.61)	**43.04**(Δ6.62)
LLaMA(65B)	41.83	44.30	**46.35**	46.78	48.27	**51.83**	24.41	24.41	**25.59**	38.25	38.79	**40.71**	41.00	42.64(Δ1.64)	**44.92**(Δ3.92)
LLaMA2(7B)	27.37	29.30	**35.21**	23.19	25.67	**30.62**	21.26	**27.95**	27.17	30.34	30.72	**31.87**	26.98	28.86(Δ1.88)	**32.98**(Δ6.00)
LLaMA2(13B)	39.10	40.53	**44.01**	36.37	39.15	**43.21**	18.90	22.05	**27.17**	36.33	38.25	**38.71**	36.78	38.70(Δ1.92)	**41.59**(Δ4.81)
LLaMA2(70B)	51.97	53.73	**57.59**	59.66	59.66	**63.13**	23.62	26.38	**31.10**	47.62	48.69	**52.46**	50.94	52.21(Δ1.27)	**56.01**(Δ5.07)
Falcon(7B)	22.72	23.64	**27.95**	19.62	22.20	**24.48**	18.90	18.90	**22.83**	23.04	23.73	**25.96**	21.98	23.13(Δ1.15)	**26.46**(Δ4.48)
Falcon(40B)	32.61	35.21	**40.15**	31.81	33.30	**36.47**	22.05	25.20	**24.41**	31.11	31.11	**34.87**	31.51	33.23(Δ1.72)	**37.20**(Δ5.69)
Baichuan(7B)	32.73	37.01	**42.16**	22.40	25.67	**29.44**	25.59	26.77	**28.74**	31.11	33.03	**36.10**	29.83	33.12(Δ3.29)	**37.29**(Δ7.46)
Baichuan(13B)	39.61	44.84	**47.78**	28.74	30.03	**35.78**	19.69	23.62	**27.17**	36.56	37.10	**39.09**	35.57	38.70(Δ3.13)	**41.99**(Δ6.42)
MOSS(7B)	28.29	30.18	**34.12**	23.98	25.07	**27.65**	23.62	23.62	**25.20**	27.50	27.96	**28.80**	26.96	28.22(Δ1.26)	**30.94**(Δ3.98)

➢ **Search LLM:**

- gpt-3.5-turbo

➢ **Prompted LLMs:**

- Falcon
- LLaMA
- LLaMA2
- Baichuan
- MOSS

➢ **Results:**

- MMLU: **1.15%~4.02%**
- AGIEval: **3.92%~7.46%**

Ablation Study

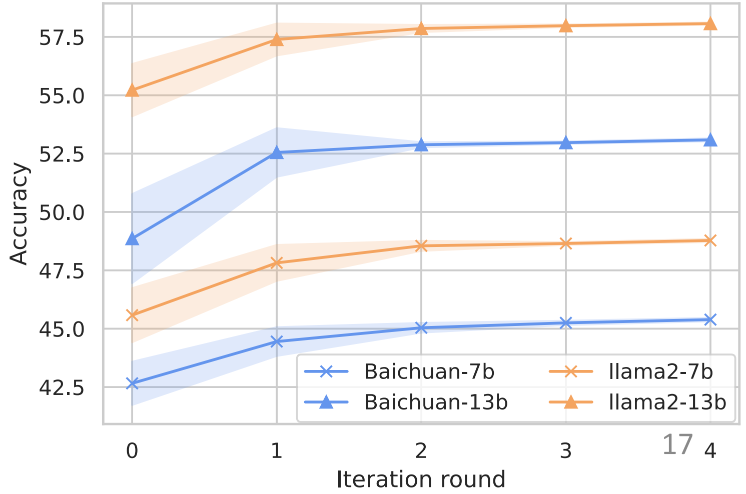
➢ Impact of CoT Component (In Step 3)

- In Step 3 Introspective Search, LLMs with the integration of CoT reasoning achieve better performance gains than LLMs without CoT component.
- **The larger** the model parameters, **the greater** the performance benefit of CoT.

Model & #Param.	w/o CoT	COT
LLaMA2(7B)	4.71	5.90
LLaMA2(13B)	4.02	4.61
LLaMA2(70B)	2.60	4.69
Baichuan(7B)	6.79	6.91
Baichuan(13B)	4.46	5.27

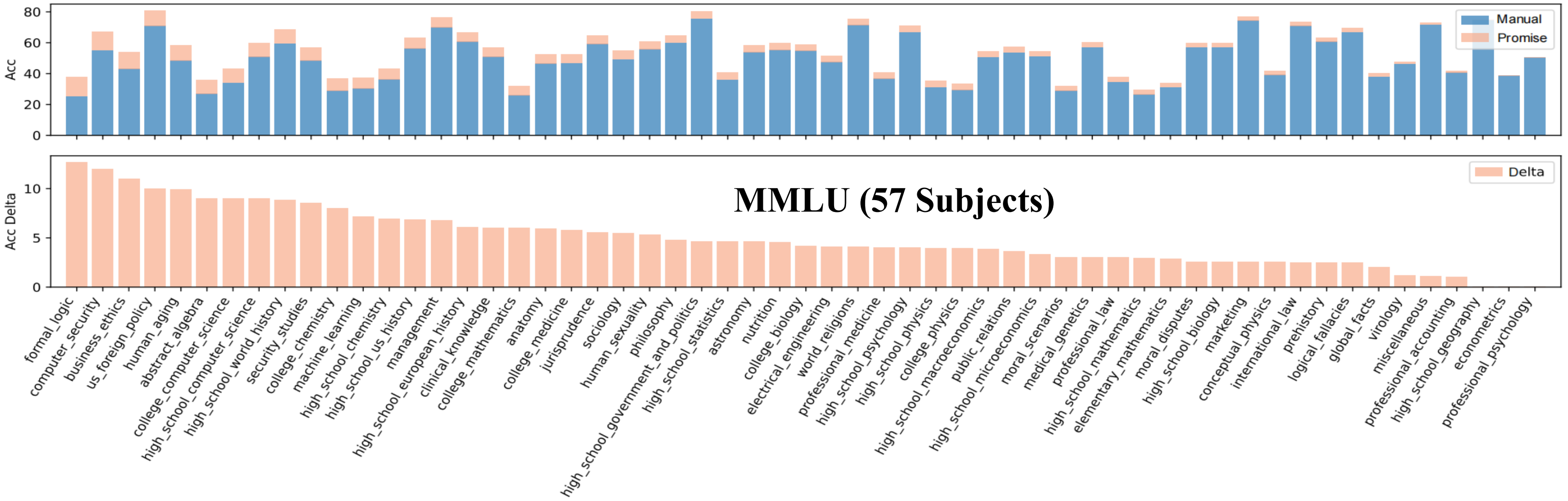
➢ Impact of Search Round (Step 2&3)

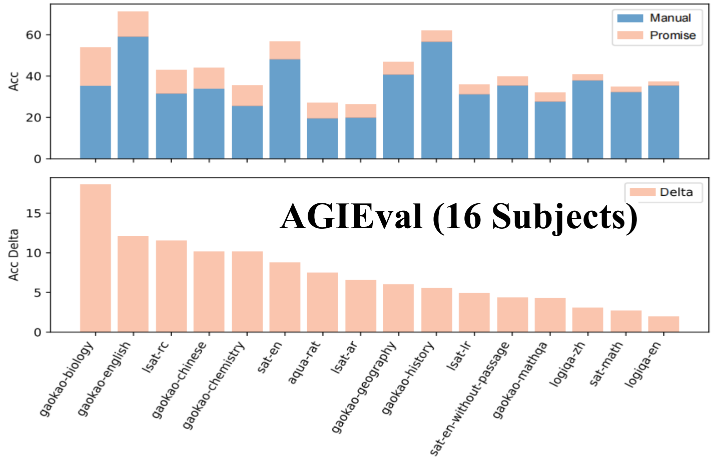
- The most significant improvements are observed during the initial two rounds.
- As we incrementally increase the number of search rounds, the rate of improvement gradually diminishes.
- **Search rounds = 2** in PromISe



Case Study: Task Characteristic

➢ **The accuracy differences among prompts on different subjects found by PromISe on Baichuan**





- **Subjects with significantly improvements，which need conceptual understanding and reading comprehension**
  - MMLU: *computer science, chemistry, politics, history.*
  - AGIEval: *gaokao-biology, gaokao-chinese, gaokao-English, politics, history and so on*
- **Subjects with less obvious improvements, which require higher levels of logical reasoning ability**
  - MMLU: *philosophy, math, physics*
  - AGIEval: *logiqa-en, sat-math, logiqa-zh, gaokao-mathqa, and sat-en-without-passage*

Case Study: Prompt Characteristic

- **Careful Consideration**
  - 91.23% of prompts emphasize on careful consideration
- **Welcome Message**
  - 80.7% of prompts begin with a welcoming message
- **Encouraging Tone**
  - 75.44% of prompts are delivered in a warm and encouraging tone
- **Background Information**
  - 40.35% of prompts contain the inclusion of a background message account
- **Specific Guidance**
  - 36.84% of prompts contain specific guidance on how to answer the questions, along with clarification that only one correct answer

Figure 3: The word cloud of optimizing prompts of AGIEval benchmark

Case Study: Prompt Case

➢ Prompt Case

The optimizing prompt of different LLMs on **computer security in MMLU**

- Model: *Baichuan-13b*
- Manual Prompt: *The following are multiple choice questions (with answers) about computer security.\n\n<question>\n<options>\nAnswer: <answer>*
- PromISe Prompt: *Welcome to the computer security multiple-choice question section! In this section, you will find a series of questions related to computer security. Please choose the correct answer from the provided options for each question. Your objective is to select the option that best answers the given question.\n\n<question>\n<options>\nAnswer: <answer>*
- Accuracy： ***55.00 -> 67.00***

Outline

- ➢**Background and Motivation**
- ➢**Proposed Method**
- ➢**Experiments**
- ➢**Conclusion and Future Direction**

Conclusion

- To better release the capabilities of LLMs, we propose a novel framework PromISe, first using prompt introspective search to find optimizing prompts tailored to each LLM
- Extensive experiments on 73 tasks in two large-scale benchmarks demonstrate the superiority of PromISe, resulting in substantial performance enhancements on 12 state-of-the-art LLMs
- We provide valuable insights into the optimal prompt design. Our systematic evaluations aspire to provide a more profound understanding of the intricate interplay between individuals and LLMs.

Future Direction

- **Scalability and Efficiency:** While PromISe has shown improvements, the process of prompt optimization can be computationally expensive. Future work could focus on developing **more efficient algorithms**

- **Generalization Across Domains:** The current framework has been tested on established benchmarks. Further research could explore the generalizability of the framework across different domains and languages, particularly those with **less available training data**

- **Integration with Other NLP Tasks:** The framework could potentially be extended to **other natural language processing tasks**, such as summarization, translation, and dialogue systems

Thanks for Your Attention !

GitHub

https://github.com/MozerWang/promISe