PromISe: Releasing the Capabilities of LLMs with Prompt Introspective Search

Minzheng Wang[1,2], **Nan Xu**[2,3]*, **Jiahao Zhao**[2,1], **Yin Luo**[3], **Wenji Mao**[2,1]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]MAIS, Institute of Automation, Chinese Academy of Sciences
[3]Beijing Wenge Technology Co., Ltd.
{wangminzheng2023, xunan2015, zhaojiahao2019, wenji.mao}@ia.ac.cn, yin.luo@wenge.com

Abstract

The development of large language models (LLMs) raises the importance of assessing the fairness and completeness of various evaluation benchmarks. Regrettably, these benchmarks predominantly utilize uniform manual prompts, which may not fully capture the expansive capabilities of LLMs—potentially leading to an underestimation of their performance. To unlock the potential of LLMs, researchers pay attention to automated prompt search methods, which employ LLMs as optimizers to discover optimal prompts. However, previous methods generate the solutions implicitly, which overlook the underlying thought process and lack explicit feedback. In this paper, we propose a novel prompt introspective search framework, namely PromISe, to better release the capabilities of LLMs. It converts the process of optimizing prompts into an explicit chain of thought, through a step-by-step procedure that integrates self-introspect and self-refine. Extensive experiments, conducted over 73 tasks on two major benchmarks, demonstrate that our proposed PromISe significantly boosts the performance of 12 well-known LLMs compared to the baseline approach. Moreover, our study offers enhanced insights into the interaction between humans and LLMs, potentially serving as a foundation for future designs and implementations. Our code is available at https://github.com/MozerWang/promISe.

Keywords: large language models, prompt search, self-introspect, self-refine

1. Introduction

The nearly human-level performance of large language models (LLMs) is rapidly reshaping this era and raising promise to AGI (Guo et al., 2023; Qin et al., 2023; Bubeck et al., 2023). To better understand the strengths and weaknesses of LLMs, various evaluation benchmarks have been proposed (Chang et al., 2024). As human evaluations (Ziems et al., 2023; Bang et al., 2023) have high variance and instability due to the individual and socio-cultural differences, most benchmarks are proposed focusing on automatic metrics, such as ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and AGIEval (Zhong et al., 2023). On these benchmarks, there are constantly new and stronger LLMs in open competition using the same prompts and test samples, in pursuit of better evaluation performance and exploring the ceiling, such as LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b), and Falcon (Penedo et al., 2023).

Recent work has shown that LLMs are sensitive to the design of prompts in the same benchmark [1]. These benchmarks mainly employ uniform prompts for all LLMs to be evaluated, which may not necessarily be the optimal way to reflect the *true capabilities* and results in underestimating the evaluation performance of LLMs. This underscores the importance of a more comprehensive examination of how prompts and evaluation criteria influence the capabilities and limitations of LLMs in real-world applications (Zhou et al., 2022).

It is widely recognized that the choice of prompt plays a crucial role in the performance of LLMs (Wei et al., 2022, 2023; Zhu et al., 2023; Li et al., 2023). With the advantage of mitigating the human workload, researchers proposed numerous automated prompt search methods (Li and Liang, 2021; Zhong et al., 2021; Zhou et al., 2022; Pryzant et al., 2023; Yang et al., 2023). Previous methods mainly involve continuous soft prompts (Li and Liang, 2021; Zhong et al., 2021), relying on the token probabilities from the output layers of LLMs, which tend to produce human-unreadable prompts and are unavailable for API-access LLMs (Liu et al., 2023). Recent automatic approaches improve discrete prompt optimization, either enumerating diverse prompts (Zhou et al., 2022) or further editing current prompts (Pryzant et al., 2023), which may lead to the local optima. Alternatively, researchers (Yang et al., 2023) iteratively generate new prompts by utilizing LLM as an optimizer. However, the generation of the solutions at each round in their method is implicit and lacks explicit

* Corresponding author

[1]The scores of LLaMA and Falcon obtained on MMLU are significantly different than that in the published paper.

LREC-COLING 2024, pages 13120–13130
20-25 May, 2024.

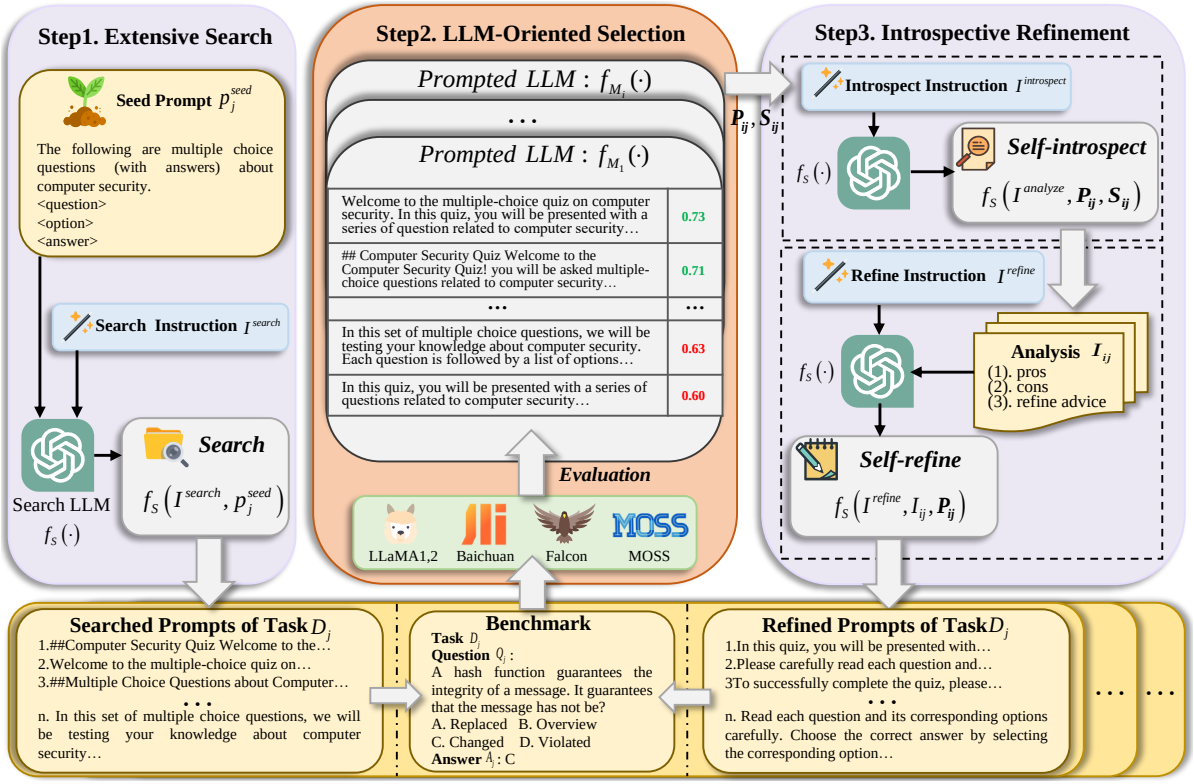


Figure 1: The overall of our proposed prompt introspective search framework. PromISe starts from discovering diverse prompts through extensive search, and iteratively finds optimizing prompts by LLM-oriented selection and introspective refinement.

feedback, which may not provide effective optimization.

In order to better release the potential of LLMs, in this paper, we propose an innovative **Prom**pt **I**ntrospective **Se**arch framework (PromISe). It starts from discovering diverse prompts through extensive search and iteratively finds optimizing prompts by LLM-oriented selection and introspective refinement. PromISe converts the process of finding optimal prompts into an explicit chain of thought, through a step-by-step procedure that involves *self-introspect* and *self-refine*, facilitating more nuanced and precise exploitation of prompt search.

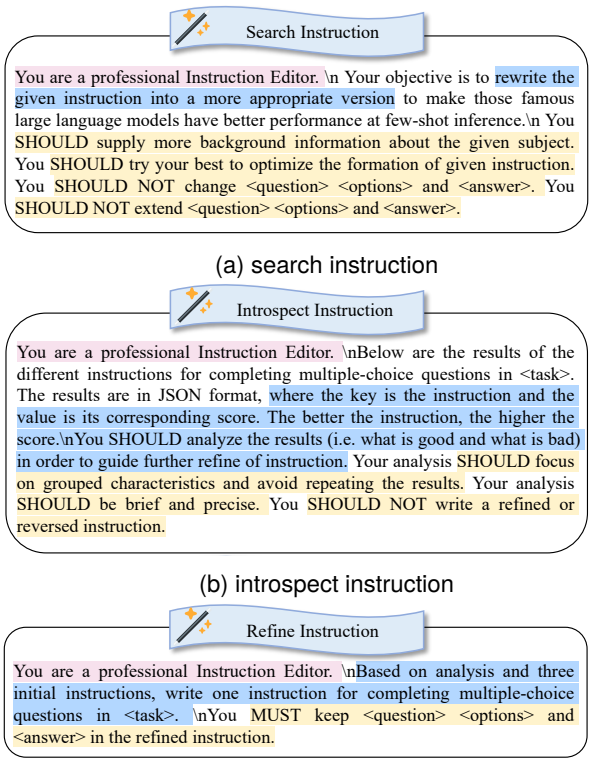
The main contributions are as follows:

- We identify that LLMs are sensitive to the design of prompts in the same benchmark. To better unlock the potential of LLMs, we are the first to find optimizing prompts tailored to each LLM.
- We propose a novel prompt introspective search framework namely PromISe. It converts the process of finding optimal prompts into an explicit chain of thought, through a step-by-step procedure, encompassing self-introspect and self-refine.
- Experiments conducted across 73 tasks within two large-scale benchmarks demonstrate the effectiveness of PromISe in releasing the capabilities of 12 state-of-the-art LLMs.

2. Related Work

LLM Evaluation Over the past year, LLMs have garnered substantial attention in the field of artificial intelligence due to their remarkable ability. LLMs trained on extensive datasets represent a revolutionary paradigm in AI development, leading to the emergence of numerous exceptional models including ChatGPT (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), LLaMA series (Touvron et al., 2023a), LLaMA2 series (Touvron et al., 2023b), Falcon series (Penedo et al., 2023), and Baichuan series (Zhong et al., 2023). These LLMs have demonstrated impressive capabilities across various scenarios, including tasks such as instruction following, few-shot in-context learning, and zero-shot inference.

Meanwhile, various evaluation approaches are introduced to comprehensively assess LLMs. While manual methods (Ziems et al., 2023; Bang et al., 2023) provide more subjective feedback, they come with substantial labor and time



(a) search instruction

(b) introspect instruction

(c) refine instruction

Figure 2: The details of instructions. The pink text expresses the role played; the blue text describes the optimization task; the yellow text is precautions.

costs. Automatic evaluation, which utilizes diverse benchmarks to automatically gauge the model's performance, attracts more attention. MMLU (Hendrycks et al., 2021) presents a comprehensive evaluation of LLMs in multifaceted contexts. AGIEval (Zhong et al., 2023) is proposed to evaluate the proficiency of foundation models in the context of human-centric standardized exams. The Huggingface Open LLM Leaderboard employs a range of benchmarks such as ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and TruthfulQA (Lin et al., 2022), which delves into reasoning an general knowledge across a wide spectrum of domains. It is essential to note that the prompts significantly affect LLMs' performance. (Wei et al., 2022, 2023; Zhu et al., 2023; Li et al., 2023). However, these benchmarks mainly employ uniformed human-crafted prompts, which may not necessarily be the optimal prompt to assess LLMs. As a result, there exists a risk of underestimating the capabilities of LLMs. To the best of our knowledge, there has been an absence of systematic work focused on this issue. In this paper, through the utilization of prompt search, we aim to improve the evaluation performance of LLMs without training and changing parameters.

Prompt Search Prompt search aims to identify the appropriate prompt for improving the LLMs' performance. To reduce the human workload related to prompt design, several automated methods have been introduced. Some automatic methods employ continuous soft prompts (Li and Liang, 2021; Zhong et al., 2021), focusing on fine-tuning the parameters of specific input tokens. However, this approach tends to produce human-unreadable prompts and becomes impractical for API-access LLM. Other automatic approaches enhance discrete prompt optimization, generating or editing natural language prompts. APE (Zhou et al., 2022) first employs the LLM to enumerate and select the positive prompts from the candidates, then rephrase these samples synonymously. APO (Pryzant et al., 2023) uses the negative samples as pseudo-gradient to iteratively edit the previous prompts. OPRO (Yang et al., 2023) utilizes LLM as an optimizer to iteratively generate new prompts guided by meta-prompt. Compared to these prior works, our method utilizes both positive and negative prompts and iteratively discovers new prompts, rather than merely editing or resampling. Our framework PromISe converts the process of discovering optimal prompts into an explicit chain of thought, which involves a step-by-step procedure to perform introspective refinement. To further demonstrate the effectiveness of PromISe, we verify our proposed method based on multiple open-sourced LLMs, which have not been validated by previous prompt search and optimization methods.

3. Proposed Method

In this paper, we tackle a particular challenge outlined by an evaluation benchmark denoted as D, comprising prompt, question, and answer, which are denoted as a triplet $\{P, Q, A\}$. This benchmark has n tasks, each represented as $D_{j \in n}$. We have a proficient search LLM referred to as $f_S(\cdot)$, and the prompted LLM named $f_{M_i}(\cdot)$. Our problem is an optimization challenge guided by search LLM $f_S(\cdot)$. The core objective is to ascertain the optimal prompt p^*_{ij} in the discrete space $\mathcal{P}$. This optimal prompt p^*_{ij}, when concatenated with the question Q_j, aims to enable prompted model M_i to generate the desired output $f_{M_i}(p^*_{ij}, Q_j)$ and maximize the evaluation performance $e(f_{M_i}(p^*_{ij}, Q_j), A_j)$.

It's important to emphasize that our objective is to identify the optimal prompt p^*_{ij} for each task within the evaluation benchmark, tailored to each specific prompted LLM, where the i represents the i-th prompted LLM and j means the j-th task in benchmark D. Hence, the search for the optimal prompt can be formulated as maximizing the expected score across all data (Q_j, A_j) drawn from

task D_j ($j \in n$) of the benchmark D:

$$p_{ij}^* = \arg\max_{p_{ij} \in \mathcal{P}} \mathbb{E}_{(Q_j, A_j) \sim D_j} e(f_{M_i}(p_{ij}^*, Q_j), A_j) \quad (1)$$

More specifically, accuracy serves as the primary performance metric for our evaluation, which is precisely defined as a binary loss function, commonly referred to as the 0-1 loss:

$$e(f_{M_i}(p_{ij}^*, Q_j), A_j) = \mathbb{1}[f_{M_i}(p_{ij}^*, Q_j) = A_j] \quad (2)$$

Our method, PromISe, consists of three main parts: extensive search, LLM-oriented selection, and introspective refinement. The overall framework is shown in figure 1. Leveraging the capabilities of the search LLM, PromISe initially finds a set of initial prompts for each task in the benchmark. Then according to the evaluation performance, PromISe will select prompts with their corresponding evaluation values for each specific prompted LLM. Next, the search LLM iteratively finds new prompts by self-introspect and self-refine. In the subsequent sections, we provide a detailed description of each of these components.

Algorithm 1 PromISe

Input: search LLM $f_S(\cdot)$, prompted LLM $f_{M_i}(\cdot)$, task in benchmark D_j, seed prompt p_j^{seed}, search instruction I^{search}, introspect instruction $I^{introspect}$, refine instruction I^{refine}
Output: optimal prompt p_{ij}^*

Extensive search
$\mathbf{P_j^{search}} \leftarrow f_S(I^{search}, p_j^{seed})$
Evaluating for each prompted LLM:
$\mathbf{S_{ij}} \leftarrow e(f_{M_i}(\mathbf{P_j^{search}}, Q_j), A_j)$
Selecting the top and bottom prompts:
$\mathbf{P_{ij}} \leftarrow topN(\mathbf{P_j^{search}}, \mathbf{S_{ij}}) + bottomN(\mathbf{P_j^{search}}, \mathbf{S_{ij}})$
while not termination condition **do**
  Self-introspect:
  $I_{ij} \leftarrow f_S(I^{introspect}, \mathbf{P_{ij}}, \mathbf{S_{ij}})$
  Self-refine:
  $\mathbf{P_{ij}^{refine}} \leftarrow f_S(I^{refine}, I_{ij}, \mathbf{P_{ij}})$
  Evaluating for each prompted LLM:
  $\mathbf{S_{ij}} \leftarrow e(f_{M_i}(\mathbf{P_{ij}^{refine}}, Q_j), A_j)$
  Selecting the top and bottom prompts:
  $\mathbf{P_{ij}} \leftarrow topN(\mathbf{P_{ij}^{refine}}, \mathbf{S_{ij}}) + bottomN(\mathbf{P_{ij}^{refine}}, \mathbf{S_{ij}})$
end while
$p_{ij}^* \leftarrow \arg\max_{p_{ij} \in \mathbf{P_{ij}}} \mathbf{S_{ij}}$

3.1. Extensive Search

In the extensive search step, we harness the search LLM to generate an initial set of prompts for each task within the benchmark. However, the search for the optimal prompt is hindered by the discretely structured search space. This work (Shin et al., 2020) found that initializing with manual templates can provide a better starting point for the search process. Following this idea, we introduce a seed prompt p_j^{seed} which is the manual prompt utilized in the benchmark.

During the prompt search process, we engage the search LLM to execute targeted instruction I^{search}, adhering closely to our predefined criteria and guided by the seed prompt p_j^{seed}. The details of search instruction can be seen in figure 2a. Specifically, we instruct the search LLM to follow explicit guidelines, including optimizing the formation of the given prompt, the preservation of placeholders, and the avoidance of overly verbose prompts. Through extensive search, the LLM will generate k_1 prompts $\mathbf{P_j^{search}} = \{p_{0,j}, p_{1,j}, ..., p_{k_1,j}\}$for each task of the benchmark:

$$\mathbf{P_j^{search}} = f_S(I^{search}, p_j^{seed}), j \in n \quad (3)$$

3.2. LLM-Oriented Selection

It is worth noting that during the extensive search step, our search process is conducted in a global space, without accounting for the sensitivity of each distinct LLM. This may overlook a crucial factor: the optimal prompt is model-specific. To mitigate this shortage, we employ LLM-oriented selection on these undifferentiated prompts, further searching prompts for specific LLM. Following the acquisition of the initial prompts $\mathbf{P_j^{search}}$, we employ these prompts to make different prompted LLMs evaluate their performance. Consequently, we obtain evaluation scores:

$$\mathbf{S_{ij}} = e(f_{M_i}(\mathbf{P_j^{search}}, Q_j), A_j) \quad (4)$$

For every specific prompted LLM $f_{M_i}(\cdot)$, we select the top k_2 prompts and the bottom k_2 prompts of each task, determined based on evaluation scores:

$$\mathbf{P_{ij}} = topN(\mathbf{P_j^{search}}, \mathbf{S_{ij}}) + bottomN(\mathbf{P_j^{search}}, \mathbf{S_{ij}}) \quad (5)$$

Paired with evaluation scores $\mathbf{S_{ij}}$, these selected prompts are crucial in our subsequent introspective search.

3.3. Introspective Refinement

Instead of sorely sampling from the initial prompts, we consider employing the search LLM to introspect the previous proposal and iteratively exploiting the search space, which provides more targeted prompt optimization for each specific LLM and enhances the likelihood of success.

We fully leverage the introspection and summarization capabilities of the search LLM $f_S(\cdot)$. Initially, for each prompted LLM $f_{M_i}(\cdot)$, we use $I^{introspect}$ to instruct the search LLM to introspect

Table 1: Main 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)

Table 2: Main 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)

the selected prompts $\mathbf{P_{ij}}$ along with their evaluation scores $\mathbf{S_{ij}}$. The details of introspect instruction can be seen in figure 2b. The inherent characteristics of both high-quality and low-quality prompts are analyzed explicitly during the self-introspect phase, and refinement advice is given:

$$I_{ij} = f_S(I^{introspect}, \mathbf{P_{ij}}, \mathbf{S_{ij}}), j \in n \quad (6)$$

Then, guided by this introspective feedback, the search LLM is tasked with instruction I^{refine} refining the top prompts of $\mathbf{P_{ij}}$, actively generating k_3 improved alternatives $\mathbf{P_{ij}^{refine}} = \{p_{0,i,j}, p_{1,i,j}, ..., p_{k_3,i,j}\}$:

$$\mathbf{P_{ij}^{refine}} = f_S(I^{refine}, I_{ij}, \mathbf{P_{ij}}), j \in n \quad (7)$$

Subsequently, we employ these new prompts to perform LLM-oriented selection again where steps two and three can be integrated into an iterative process. Our methodology allows for the iterative pursuit of the optimal prompt until the termination condition is satisfied. The complete procedure is provided in Algorithm 1. The extensive experiments have unequivocally validated the effectiveness of our approach.

4. Experiments

In this section, we present the evaluation results for prompt search. Our experiments unequivocally showcase that PromISe yields a substantial performance boost across various LLMs.

4.1. Benchmarks

We use well-established evaluation benchmarks, MMLU (Hendrycks et al., 2021) and AGIEval (Zhong et al., 2023), to validate our methods. These two benchmarks examine the knowledge level of the large model through multiple-choice questions.

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.

AGIEval incorporates bilingual tasks in both Chinese and English, encompassing 20 tasks with a human-centric focus and consisting of 8,062 questions for evaluation. In line with established research practices (Zhong et al., 2023), our selection has focused exclusively on multiple-choice questions in AGIEval, comprising 16 tasks and 4,951 questions.

In the absence of training datasets in both benchmarks, we follow the treatment adopted in the baseline method APE (Zhou et al., 2022) for fair comparison, which employs a random sampling approach to extract a small portion of training samples from the test dataset for prompt search. In our experiment, we randomly extract 15% of the dataset for prompt introspective search and identify the best prompt p_{ij}^* for each LLM. For testing p_{ij}^*, we also follow the mainstream evaluation convention as in works (Zhou et al., 2022; Yang et al.,

2023) and verify the best prompt p_{ij}^* on the rest of the dataset.

4.2. Experimental Setup

Models In our experiments, we employ GPT-3.5 (Ouyang et al., 2022) as our search LLM, which is accessed through API. This choice is predicated on its outstanding understanding and generation capabilities, which are integral to the success of our method. Additionally, our experiments are verified on several state-of-the-art open-sourced prompted LLMs, including Falcon (Penedo et al., 2023), LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b) , Baichuan (Zhong et al., 2023), MOSS (Sun et al., 2023). Specific versions of LLMs are: `gpt-3.5-turbo`, `LLaMA-7b`,`LLaMA-13b`,`LLaMA-33b`,`LLaMA-65b`, `LLaMA2-7b`,`LLaMA2-13b`,`LLaMA2-70b`,`MOSS-7b`, `Falcon-7b`,`Falcon-40b`,`Baichuan-7b`, `Baichuan-13b`

Baselines We compare PromISe against two prompt-based approaches: Manual Prompts (Hendrycks et al., 2021; Zhong et al., 2023) and APE (Zhou et al., 2022). Manual Prompts are human-crated prompts used in evaluation benchmarks MMLU and AGIEval. APE enumerates and selects the positive prompts from the candidates.

Implementation Details To ensure the stability of LLM generation, we use the greedy decoding strategy and restrict the maximum number of new tokens to 1. Considering the context window length of different models, we consistently cap the maximum input token limit at 2048 tokens. We execute experiments to assess the performance of LLMs in five-shot and answer-only settings, where we instruct the prompted LLMs to generate answers directly. Instead of comparing the probabilities associated with specific token groups, our approach involves generating text directly from the model. In PromISe, we designate the number of search rounds as the termination condition for iterative search. Specifically, the prompt search round is set at two. In the step of extensive search, we instruct the search LLM several times to generate 50 prompts for each task. As part of LLM-oriented selection, we select the top 5 and the last 5 prompts based on evaluation scores for each prompted LLM. In the self-introspect and self-refine part, we leverage the search LLM to refine and generate new 15 prompts for each task. The prompts we used to instruct the search LLM are shown in figure 2.

4.3. Main Results

We calculate the average accuracy across all the questions within benchmarks. Table 1 and 2 presents the detailed results obtained from all prompted LLMs. Each of these two benchmarks is divided into four subjects, and the average represents the total evaluation score of each prompted LLM. After conducting PromISe, we observed significant improvements in the performance of prompted LLMs compared to the baseline values. Specifically, compared to the manual prompt in the MMLU, PromISe achieves enhancements ranging from 1.5 to 4 points, while in the AGIEval, improvements ranged from 4 to 7.5 points. Notably, the `Baichuan-13B` model exhibits the most substantial improvement of 4 points in the MMLU, while the `Baichuan-7B` model shows the highest improvement of 7.5 points in the AGIEval. Compared with previous works APE, our method delivers significantly better results. These results affirm the effectiveness of our proposed methodology.

It is worth mentioning that the observed performance generally correlates with the model size, as models with larger parameter sizes tend to yield better results. Therefore, it is essential to note that these improvements remain bound by the inherent limitations imposed by the model's parameter magnitude. Remarkably, on the AGIEval, the evaluation scores of certain prompted LLMs even outperform models with larger parameter sizes. For instance, the `LLaMA-33B` model achieves a score of 42.78, surpassing the evaluation score of the `LLaMA-65B` model using uniform manual prompt, while the `Baichuan-7B` model attains a remarkable score of 36.74, exceeding the `Baichuan-13B` model. What we aim to underscore is that our methodology optimizes model performance to the fullest extent possible without altering the model's parameters. These outcomes resoundingly underscore the efficacy of our approach.

Table 3: Ablation study of CoT component.

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

4.4. Ablation Study

Impact of CoT component We conduct an ablation study on the effectiveness of the integration of introspect search to evaluate the performance that converts the process of optimizing prompts into an explicit chain of thought. The detailed ablation results on the AGIEval benchmark and five LLMs about CoT are shown in Table 3, which presents the accuracy delta compared to the manual prompt. The optimizing prompt was searched

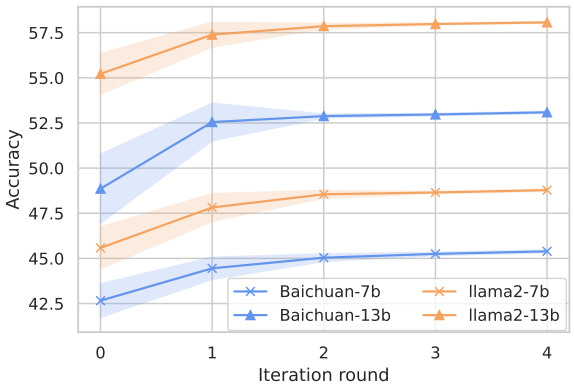


Figure 3: The results after four rounds of prompt search with PromISe.

from Step 3 Introspective Search, and the results show that LLMs with the integration of CoT reasoning achieve better performance gains than LLMs with the absence of the CoT component. Specially, the larger the model parameters, the greater the performance benefit of CoT.

Impact of Search Round To ascertain the impact of search rounds in PromISe, we employ `Baichuan-7b`, `Baichuan-13b`, `LLaMA-7b`, and `LLaMA-13b` to undertake a more extensive series of iterations. The results are shown in figure 3. Remarkably, it is clear that 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. The improvement brought by prompt search will be less obvious. Consequently, we opt to adhere to two search rounds in PromISe.

Impact of Model Output Way In order to explore the impact of different ways of the model output, we also compare two main ways: getting text generation directly and getting the probabilities. The experiment is conducted on the MMLU dataset and the result can be seen in table 4. The 'Directly' is the same as our main experimental set, which uses the greedy decoding strategy and gets the text generation directly. Another one, the 'Probability' only compares the probabilities associated with specific token groups–A, B, C, and D, choosing the one with the highest probability. From the table 4, we can see that the way of getting the probabilities improves the results slightly because the selection of tokens has been narrowed and the token with the highest probability other than ABCD will not be selected. However, from our point of view, this approach could be tolerant and cannot objectively reflect the performance of the model, which also becomes unavailable when assessing API-access LLM. To ensure a more generalized approach, we opt for direct generation as our method.

Table 4: Two different ways of model output.

Model & #Param.	Directly	Probability
LLaMA(7B)	35.27	35.29
LLaMA(13B)	46.44	47.08
LLaMA(33B)	58.56	58.50
LLaMA(65B)	63.59	63.62
LLaMA2(7B)	45.58	45.89
LLaMA2(13B)	55.22	55.73
LLaMA2(70B)	69.06	69.07
Falcon(7B)	26.46	26.45
Falcon(40B)	49.94	49.96
Baichuan(7B)	42.66	42.67
Baichuan(13B)	48.86	50.59
MOSS(7B)	40.39	40.40

4.5. Case Study

To provide valuable insights into optimizing prompt design for evaluation, we extend our analysis to task and prompt characteristics.

Task Characteristic The figure 4 and 5 illustrate the differences in accuracy for the `Baichuan-13b` among prompts searched by PromISe compared to manual prompts. In this analysis, we will begin by examining various tasks and delving into the reasons behind the impact of prompts. For the AGIEval benchmarks, it becomes evident that certain subjects experience the most substantial improvements, such as `gaokao-biology`, `gaokao-english`, `lsat-rc`,`gaokao-chinese`, and `gaokao-chemistry`, which place a strong emphasis on conceptual understanding and reading comprehension skills. In these cases, the role of prompts is to assist the model in contextualizing the given question and triggering the retrieval of relevant knowledge it has acquired. However, when dealing with subjects like `logiqa-en`, `sat-math`, `logiqa-zh`, `gaokao-mathqa`, and `sat-en-without-passage`, our approach does not exhibit significant improvements due to these subjects heavily rely on reasoning abilities. The LLM tends to guess randomly, and achieving the correct answer solely through prompt engineering becomes unrealistic. In the context of MMLU benchmarks, our approach exhibits a consistent pattern. It demonstrates greater effectiveness in subjects that demand a comprehensive understanding of knowledge. Notably, this effectiveness is particularly pronounced in conceptual subjects, such as `computer science`, `chemistry`, `politics`, `history`. However, for subjects that require higher levels of logical reasoning ability, such as `philosophy`, `math`, `physics`, the overall improvement is relatively less obvious.

Prompt Characteristic We proceeded to analyze the prompts that yielded significant improve-

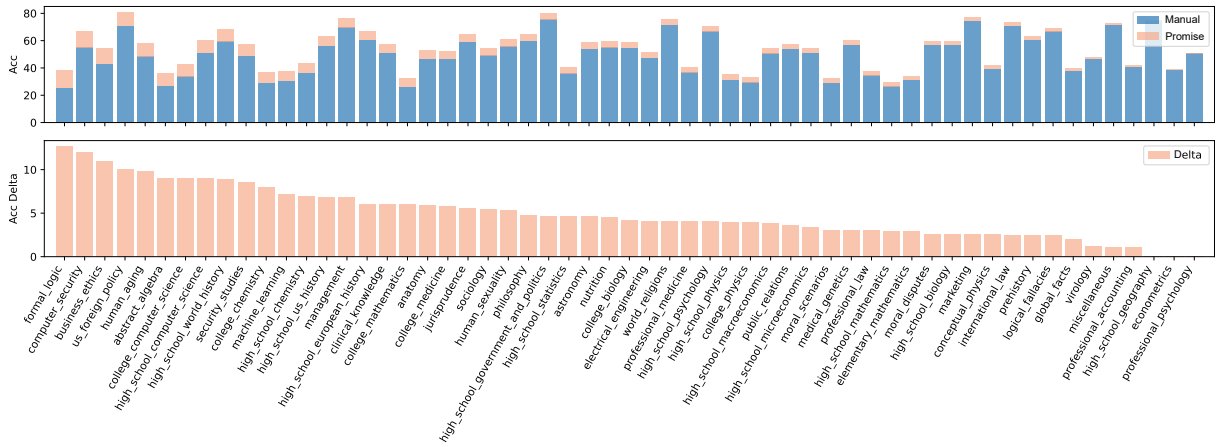


Figure 4: On 57 MMLU tasks, the accuracy differences among prompts found by PromISe.

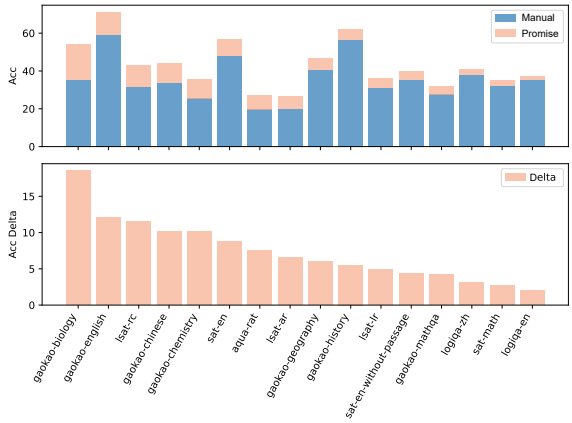


Figure 5: On 16 AGIEval tasks, the accuracy differences among prompts found by PromISe.

ments. In the case of MMLU benchmarks and `Baichuan-13b`, the optimizing prompts share common characteristics:

- *Careful Consideration:* About 91.23% of the prompts emphasize on careful consideration, encouraging LLMs to thoroughly ponder all options before selecting an answer. This stimulates the model's critical thinking and attention to detail.

- *Welcome Message:* Almost 80.7% of prompts begin with a welcoming message, offering a clear overview of the quiz or question format.

- *Encouraging Tone:* Nearly 75.44% prompts are delivered in a warm and encouraging tone, such as `Good luck`, `Let's get started`.

- *Background Information:* Prompts containing the inclusion of a background message account for 40.35%, which provide an overview of the subject's scope.

Figure 6: The word cloud of optimizing prompts

- *Specific Guidance:* Specific guidance on how to answer the questions, along with clarification that only one correct answer, is worth 36.84%. This clarity helps prevent confusion.

In addition, we draw the word cloud diagram of these optimizing prompts in figure 6. Turning attention to AGIEval benchmarks, we identify distinct patterns. For Chinese prompts, conciseness and directness are key characteristics, with a focus on presenting questions and options without extraneous text. Conversely, English prompts tend to be more elaborate in content. They also stress the importance of careful reading of the passage and explicitly mention the format of the correct answer.

We also conduct the prompt differences between different LLMs on the same task. The details about prompts are shown in table 5. Our findings yield several noteworthy conclusions. Firstly, we observed pronounced performance disparities among various LLMs when utilized with manual prompts. Surprisingly, even smaller-parameter-sized LLMs outperformed their larger counterparts, underscoring the limitations of existing evaluation methodologies. Secondly, in contrast to manual prompts, the optimizing prompts PromISe derived contains more comprehensive information,

Table 5: The optimizing prompt of different LLMs on computer security in MMLU. The baseline value is the evaluation result using the manual prompt.

Prompts
Manual: The following are multiple choice questions (with answers) about computer security.\n\n<question>\n<options>\nAnswer: <answer>
Baichuan-7b (64.00->66.00)*: Welcome to the multiple choice quiz on computer security! In this quiz, you will be presented with a series of questions related to computer security. Each question will have multiple options, and you need to select the correct answer from the given choices.\n\nPlease read each question and the corresponding options carefully. Once you have chosen your answer, indicate your choice by selecting the corresponding option. The correct answer for each question is provided at the end.\n\nLet's get started!\n\n<question>\n<options>\nAnswer: <answer>
Baichuan-13b (55.00->67.00)*: 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>
LLaMA2-7b (60.00->61.00)*: Welcome to the computer security multiple choice quiz! In this quiz, you will be presented with a series of questions, and you need to select the correct answer from the options provided.\n\n<question>\n<options>\nAnswer: <answer>

* means the improvement of accuracy by using PromISE compared to baseline value.

corroborating our earlier analysis. Thirdly, while the semantic content of optimizing prompts remained largely consistent across different LLMs, the phrasing and resulting improvement effects exhibited significant variation. Specifically, both `Baichuan-13b` and `Baichuan-7b` achieved notably high accuracy when their optimal prompts included welcoming messages, careful considerations, and background information. Conversely, `LLaMA2-7b` showed less pronounced improvement, likely due to its comparatively concise prompt structure.

5. 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, consistently outperforming both manual prompts and existing methodologies, resulting in substantial performance enhancements on 12 state-of-the-art LLMs. Moreover, 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. These evaluations hold the potential to establish a solid foundation for inspiring future design paradigms and practical implementations within this domain.

6. Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant #2021YFF0901503, the National Natural Science Foundation of China under Grants #62206287, and the Beijing Nova Program Z201100006820085 from Beijing Municipal Science and Technology Commission.

7. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *IJCNLP-AACL*, pages 675–718.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *EMNLP*, pages 7957–7968.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *EMNLP*, pages 1339–1384.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222–4235.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *ICLR*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Baichuan

2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *NAACL*, pages 5017–5033.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *ICLR*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

8. Language Resource References

Dan Hendrycks and Collin Burns and Steven Basart and Andy Zou and Mantas Mazeika and Dawn Song and Jacob Steinhardt. 2021. *Measuring Massive Multitask Language Understanding*. UC Berkeley. ICLR.

Wanjun Zhong and Ruixiang Cui and Yiduo Guo and Yaobo Liang and Shuai Lu and Yanlin Wang and Amin Saied and Weizhu Chen and Nan Duan. 2023. *AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models*. Microsoft. arXiv preprint arXiv:2304.06364.