Dynamic Evaluation of Large Language Models by Meta Probing Agents

Kaijie Zhu [1] **Jindong Wang** [1] **Qinlin Zhao** [2] **Ruochen Xu** [1] **Xing Xie** [1]

Abstract

Evaluation of large language models (LLMs) has raised great concerns in the community due to the issue of data contamination. Existing work designed evaluation protocols using well-defined algorithms for specific tasks, which cannot be easily extended to diverse scenarios. Moreover, current evaluation benchmarks can only provide the overall benchmark results and cannot support a fine-grained and multifaceted analysis of LLMs' abilities. In this paper, we propose *meta probing agents* (MPA), a general dynamic evaluation protocol inspired by psychometrics to evaluate LLMs. MPA designs the probing and judging agents to automatically transform an original evaluation problem into a new one following psychometric theory on three basic cognitive abilities: language understanding, problem solving, and domain knowledge. These basic abilities are also dynamically configurable, allowing multifaceted analysis. We conducted extensive evaluations using MPA and found that most LLMs achieve poorer performance, indicating room for improvement. Our multifaceted analysis demonstrated the strong correlation between the basic abilities and an implicit Matthew effect on model size, i.e., larger models possess stronger correlations of the abilities. MPA can also be used as a data augmentation approach to enhance LLMs. Code is available at: `https://github.com/microsoft/promptbench`.

1. Introduction

Intelligence evaluation has never been as important as today due to the contradiction between unprecedented performance and underexplored interpretability of large language models (LLMs) (OpenAI, 2023b; GeminiTeam, 2023). To facilitate a better understanding of the strength and weakness of LLMs, evaluation was carried out by collecting data from various domains (Liang et al., 2023; bench authors, 2023), benchmarking specific tasks (Hendrycks et al., 2021; Chen et al., 2021; Cobbe et al., 2021; Clark et al., 2018), and evaluating in extreme scenarios (Zhu et al., 2023b; Wang et al., 2023; Yang et al., 2022).

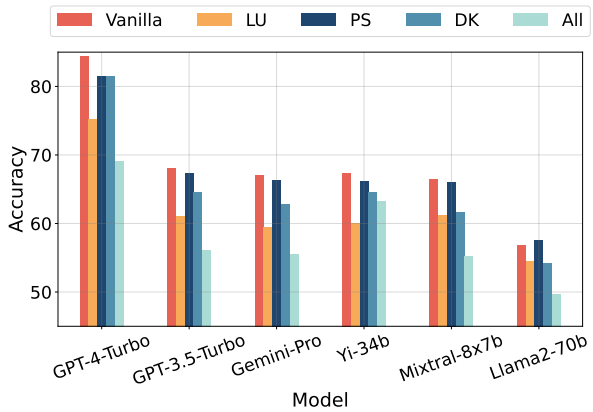


Figure 1. Performance of different LLMs on vanilla MMLU testset and our probing benchmarks based on the MMLU. LU, PS, and DK denote the evaluation sets to evaluate language understanding, problem solving, and domain knowledge ability, respectively.

There have been increasing concerns about the genuine abilities of LLMs in public benchmarks, attributing the "false promise" to data contamination (Lovin, 2023; Bender et al., 2021; Kocoń et al., 2023), overfitting benchmarks (Zhu et al., 2023a), improper choice of the evaluation criterion (Schaeffer et al., 2023), or lack of causality (Berglund et al., 2023). Among these concerns, data contamination remains the most significant, as static public benchmarks could easily be harnessed to train models. Moreover, evaluation should provide not only benchmark results, but also insight into the structural capabilities of models for future development (Burnell et al., 2023). For example, evaluations are typically done in a certain context, e.g., a math application problem requires at least two abilities: language understanding (to comprehend the problem) and reasoning (to solve the problem). Which ability matters more, and how can we quantify the relationship between these abilities?

[1]Microsoft Research [2]University of Science and Technology of China. Correspondence to: Kaijie Zhu <kaijiezhu11@gmail.com>, Jindong Wang <Jindong.Wang@microsoft.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Recently, Zhu et al. (2023a) proposed DyVal to dynamically generate test samples based on the graph structure to combat data contamination. Fan et al. (2023) introduced NPHardEval, which generates new evaluation samples for NP-hard math problems and updates the evaluation set monthly. Both are designed to evaluate reasoning tasks and cannot be easily extended to other popular natural language tasks (Hendrycks et al., 2021; Clark et al., 2018). For fine-grained performance analysis, Burnell et al. (2023) inspected the correlation between different tasks using HELM benchmark results (Liang et al., 2023) and concluded that the performance of LLMs is not monolithic but exhibits variance between different aspects, such as reasoning and understanding. Therefore, developing a dynamic evaluation protocol to support diverse tasks and multifaceted ability analysis remains a challenge.

In this paper, we propose **M**eta **P**robing **A**gents (MPA), a dynamic and flexible evaluation protocol for LLMs based on agents. MPA bridges the gap between psychometrics and LLMs evaluation by designing principles to dynamically generate new questions (Figure 2). The principles correspond to the three basic cognitive abilities of psychometric theory (Burnell et al., 2023): language understanding, problem solving, and domain knowledge. Therefore, MPA supports both dynamic evaluation sample generation and multifaceted ability analysis. Specifically, instead of relying on the graph structure to generate samples like DyVal, MPA uses LLM-based agents to automatically transform existing problems into new ones, which are more flexible and support diverse tasks. We define them as *probing* agents, aiming to uncover the underlying knowledge in a question. MPA further utilizes a *judge* agent (Dubois et al., 2023; Li et al., 2023b) to validate the generated evaluation samples. This adversarial manner ensures that the new samples maintain consistency and relevance compared to their original counterparts. Furthermore, MPA can dynamically combine various probing principles for multifaceted evaluations of the abilities. This modular design affords researchers the flexibility to apply any combination of principles, aligning the evaluation scope with their investigative focus, and mirroring the multifaceted nature of human cognition.

We used MPA to generate new evaluation sets based on popular benchmarks: MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), BBH (Suzgun et al., 2022), and ARC-C (Clark et al., 2018). Then, we conducted extensive evaluations and analysis on popular LLMs: GPT-4-Turbo, GPT-3.5-Turbo, Gemini-Pro (GeminiTeam, 2023), Llama2-70b-chat (Touvron et al., 2023), Yi-34b-chat (01-ai, 2024), and Mixtral-8x7b-Instruct (MistralAITeam, 2023). The takeaways of our key findings are as follows:

- The performance of LLMs on our dynamic benchmarks degraded significantly, implying potential data contamination on current benchmarks (Figure 1 & §4.2). Prompt engineering can only bring marginal improvements (§4.4).
- All LLMs exhibited performance decreases by dynamically combining different principles (§5.1);
- Our multifaceted analysis demonstrated strong correlations between the three basic abilities, where language understanding and problem solving abilities have the strongest correlation (§5.2).
- We observed an implicit "Matthew effect" between model size and correlations of the abilities: larger models tend to have stronger correlations (§5.3).
- LLMs exhibited various error patterns in our fine-grained analysis pertaining to the three basic abilities (§5.4).
- MPA can be used as a data augmentation approach to improve the performance of LLMs (§6).

The contributions of this paper are summarized as follows:

- **A psychometrics-inspired dynamic evaluation Protocol.** MPA is general and flexible to mitigate data contamination and facilitate multifaceted analysis.
- **Comprehensive analysis of the basic abilities of LLMs.** Our modular design allows for the dynamic combination of the three basic cognitive abilities, providing systematic evaluation and analysis for future research.

2. Related Work

LLMs Evaluation & Data Contamination. Various benchmarks have been introduced to assess LLMs (Hendrycks et al., 2021; Li et al., 2023a; Zhong et al., 2023; HuggingFace, 2023; Chang et al., 2023), including: (1) Human-centric evaluation, typified by AdaVision (Gao et al., 2022) and AdaTest (Ribeiro & Lundberg, 2022) that emphasize human-driven feedback. (2) Crowd-sourcing, e.g., DynaBench (Kiela et al., 2021) and DynaBoard (Ma et al., 2021), which prioritizes diverse and comprehensive evaluations through crowd-sourced tests. (3) More challenging tasks such as HELM (Liang et al., 2023), DeepTest (Tian et al., 2018) and CheckList (Ribeiro et al., 2020), create custom tests, while platforms such as Big-Bench (bench authors, 2023) aim to challenge LLMs with specialized tasks.

Recent research has highlighted the significant issue of data contamination in LLMs (Lovin, 2023; Bender et al., 2021; Kocoń et al., 2023; Li, 2023; Zhou et al., 2023). In particular, the reports of GPT-4 (OpenAI, 2023b), LLama (Touvron et al., 2023), and Skywork LLM (Wei et al., 2023) have acknowledged this phenomenon. Several researchers (Golchin & Surdeanu, 2023a;b; Oren et al., 2023; Yang et al., 2023; Oren et al., 2023) developed methods to detect data contamination. Zhu et al. (2023a); Lei et al. (2023); Fan et al. (2023)

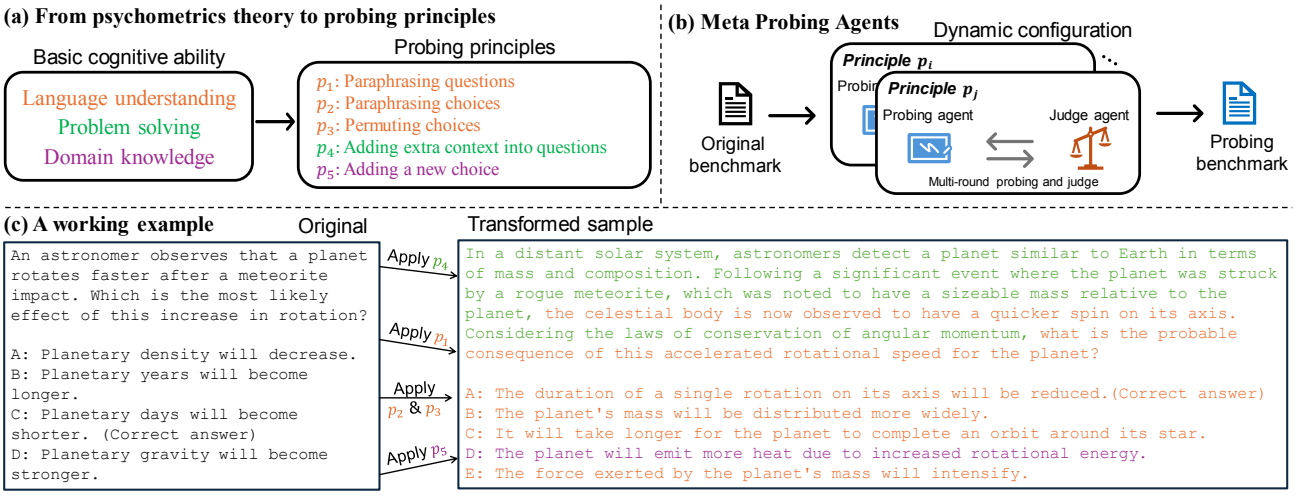


Figure 2. Inspired by psychometric theory on the three basic cognitive abilities, our Meta Probing Agent (MPA) designs corresponding principles that transforms original benchmarks into a new one. These principles can be flexibly combined to create various probing benchmarks for multifaceted analysis. Subfigure (c) shows how MPA generates the new sample given an existing sample from ARC-C.

introduced dynamic evaluation strategies via different algorithms to reduce data contamination. MPA significantly differs from them in the sample generation mechanism and the support for multifaceted analysis.

LLMs as Autonomous Agents. The adoption of LLMs as autonomous agents for task completion has recently gained popularity, such as AutoGPT (Significant-Gravitas, 2023) and MetaGPT (Hong et al., 2023), which have advanced our understanding of collaborative and planning abilities in LLMs. Another growing trend is the use of LLMs as judges (Dubois et al., 2023; Li et al., 2023b; Fernandes et al., 2023; Bai et al., 2023; Wang et al., 2024), where LLMs assess the output of other LLMs. Furthermore, there is a burgeoning interest in leveraging LLMs to enhance training datasets (Wang et al., 2022; Yu et al., 2023; Yuan et al., 2024; Liu & Yao, 2024) and aligning LLMs' outputs with specific goals or values (Burns et al., 2023). Our work differs from them in designing psychometrically inspired principles for sample generation and support for multifaceted analysis.

3. Method

3.1. Overview

There are two critical challenges in designing a dynamic evaluation protocol. First, there is no general principle to guide the evaluation sample generation process for diverse tasks such as knowledge, language understanding, reasoning, and mathematics. Existing literature like DyVal (Zhu et al., 2023a) and NPHardEval (Fan et al., 2023) adopted manually designed principles to generate samples for specific tasks and cannot easily extend to other scenarios. Second, the principle of generating evaluation samples should be fine-grained yet atomic enough to analyze the multifaceted capabilities of LLMs. Our hope is that the evaluation should reflect the primitive abilities, and more fine-grained analysis of their correlations can be conducted.

In this work, we take inspiration from psychometrics (Raykov & Marcoulides, 2011; Burnell et al., 2023) to generate evaluation samples using LLMs as agents. Specifically, instead of relying on specific rules like DyVal (Zhu et al., 2023a), we employ LLMs as agents to automatically generate new problems based on the given evaluation samples. This agent-based evaluation design can potentially fit most tasks. More importantly, psychometric theory categorizes cognitive abilities into three basic ones: *language understanding* to comprehend and generate texts, *problem-solving* to deal with complex problems, and *domain knowledge*. Therefore, as shown in Figure 2(a), we follow these criterion to design principles that not only change the problems but also assist in multifaceted analysis.

Our approach, termed as *meta probing agents* (MPA), is illustrated in Figure 2(b). Given an original evaluation sample from existing benchmarks, MPA aims to evaluate the ability of LLMs by creating new problems following psychometric theory. We create a set of principles that involve two agents: the *probing* agent and the *judge* agent. The probing agent aims to investigate the knowledge of a given question Q and return a new one according to a principle p_i, and the judge agent is responsible for the validity and consistency check based on the original question. Their interaction features a feedback mechanism: if the judge agent determines that the new question lacks consistency ('No'), the probing agent is prompted again to generate another version of the question. Conversely, if the answer is 'Yes', the question is deemed to have passed the consistency check.

Table 1. The prompts for different principles based on the three cognitive abilities.

Principle	Ability	Agent Type	Prompt
Paraphrasing Question	Language Understanding	Probing	I plan to paraphrase the question to present the same concept in a different way. Please assist me in paraphrasing the question.
		Judge	Your task is to analyze both the original question and the revised question and determine if they are effectively assessing the same concept or knowledge area.
Paraphrasing Choices	Language Understanding	Probing	I plan to paraphrase the choices: each choice should be paraphrased to reflect the same concept or idea as the original. The essence and meaning must be preserved. If a choice cannot be paraphrased without changing its meaning, it should be kept unchanged.
		Judge	Your task is to analyze the paraphrased choices in the context of the question and determine if the paraphrased choices (A, B, C, D) still reflect the original meaning of their respective original choices.
Permuting Choices	Language Understanding	-	(This principle does not require agents and can be implemented directly in the code)
Adding Extra Context into Question	Problem Solving	Probing	I plan to add non-essential context to the question: introducing context or details to the question that are relevant but not directly helpful in answering it. The context can be put at the beginning, middle, or end of the question.
		Judge	Your task is to analyze both the original question and the revised question and determine if they are effectively assessing the same concept or knowledge area.
Adding A New Choice	Domain Knowledge	Probing	I plan to keep the choices A,B,C,D unchanged, and introduce an additional relevant choice E that is related to the topic but doesn't provide another correct answer. This choice should be plausible but clearly not correct in the context of the question.
		Judge	Your task is to analyze the paraphrased choices in the context of the question and determine whether the new choice (E) is relevant to the question but does not provide an alternative correct answer.

The dynamic nature of MPA lies in two aspects: the dynamic generation of evaluation samples and the dynamic combination of principles. Such a dynamic combination allows for multifaceted analysis of LLMs' abilities, thus providing more insight for future research. It can also be seen as granting problems with flexible complexities (Zhu et al., 2023a) for a comprehensive evaluation. Now, we present the details of the agents and the psychometric principles.

3.2. Probing Agent

The probing agent aims to transform a given question into a new one to assess LLMs' ability required by a question Q. The probing is guided by principles inspired by psychometrics, which are encapsulated within a carefully crafted prompt. Unlike generating training samples, one of the most important criteria in probing agents is to ensure that the generated questions maintain the core essence of the original while presenting a different perspective, thus maintaining its correctness. An example is shown in Figure 2(c), where a sample of ARC-C is transformed into a new one by applying different principles. The prompts in our experiments incorporate directives that guide the agent towards creating semantically similar but structurally different questions.

3.3. Judge Agent

Although we have explicitly restricted the probing agent to maintain the integrity of the original question, there are cases where probing agents unintentionally change the meaning. Thus, the judge agent is designed to provide a clear and unambiguous assessment of whether the generated question maintains the integrity of the original intent and informational content. Its prompt is designed to direct LLMs to compare the original with the rephrased questions, ensuring the preservation of the essence and factual accuracy.

Specifically, unlike traditional evaluation methods that may use a variety of metrics, the judge agent operates in an adversarial manner via a binary response system through prompts. It simply returns a 'Yes' or 'No' verdict, indicating whether the new question maintains consistency with the original. In this prompt, the judge agent is required to analyze the essence of both the original and rephrased questions. Its goal is not merely to identify superficial similarities or differences in wording, but to delve deeper into whether both versions of the question are aligned in terms of the concept or knowledge area they are assessing.

Human Verification For each ability (language understanding, problem solving, and domain knowledge), we randomly selected 500 samples from the MMLU dataset and 100 samples from the ARC-c dataset, totaling $1,800$ questions. 30 human experts (with bachelor or higher degree) are divided into 3 groups, each with 10 person. They were asked to judge the following two questions: (1) whether the original and paraphrased questions were equivalent; (2) if the answers to the probing questions remained correct. The evaluation required a simple 'Yes' or 'No' answer.

The positive results shown in Appendix Table 9 from our human evaluation with an overall accuracy rate of 94% and 97% for each question, underscoring the effectiveness of our methodology. The following table provides a detailed breakdown of the evaluation outcomes, showcasing the high level of confidence in the equivalence and correctness of our probing questions across the three abilities.

3.4. Psychometric principles

Psychometric principles guide how we probe the understanding of a question. Inspired by psychometric theory (Raykov & Marcoulides, 2011), we aim to evaluate three basic abilities of LLMs: language understanding, problem solving, and domain knowledge. We have identified five key principles that correspond to these categories, as shown in Table 1. These principles incorporate both a probing agent and a judge agent, as previously discussed, with their functions differing according to the specific principle applied.

3.4.1. Language Understanding

Language understanding assesses the ability to process, interpret, and generate texts. To evaluate this ability, we focus on how well LLMs can grasp the underlying meaning of various linguistic expressions and maintain its integrity when presented in different forms. We design three principles to evaluate this ability:

- **Principle 1: Paraphrasing Questions.** It focuses on altering the phrasing of a question while retaining its core concept. This is achieved via a prompt that guides the probing agent to restructure the question without changing its underlying meaning. The new questions challenge LLMs in understanding to ensure that they grasp the essence of the question beyond surface-level recognition.
- **Principle 2: Paraphrasing Choices.** It is similar to the first one, but applies to choices in a multiple-choice format. It involves rephrasing the options provided in a way that maintains their original intent and meaning.[1]
- **Principle 3: Permuting Choices** (Zong et al., 2023). This principle simply involves rearranging the order of the choices in a multiple-choice question. It determines if the model's understanding is influenced by the position of the correct answer. As it can be achieved through coding, specific prompts are not required for this principle.

3.4.2. Problem Solving

Problem solving refers to the ability to analyze, deduce, and derive answers. It involves critical thinking, distinguishing relevant from irrelevant data, and applying knowledge to new situations. Principles under this category test the model's ability in navigating complex, often nuanced scenarios, and its proficiency in delivering solutions. Note that we cannot create completely new problems by the agent, since their correctness cannot be guaranteed. Therefore, we design one general principle:

- **Principle 4: Adding Extra Context into Questions.** It aims to introduce additional, non-essential context to the question, which is relevant to the topic, but does not directly aid in answering the question. The prompt guides the probing agent to seamlessly integrate extra context into the original question. The new questions assess whether LLMs can filter out extraneous information and focus on the key elements to solve the problem.

[1]Most benchmarks adopt the QA style. For non-QA benchmarks such as GSM8K, the choice-related principles do not apply.

3.4.3. Domain Knowledge

Domain knowledge refers to the depth and accuracy of the model's knowledge in specific areas. It is crucial not only to have a broad understanding of general concepts, but also to possess detailed, nuanced knowledge. It tests the model's expertise in various domains, its ability to differentiate between closely related concepts, and to apply this knowledge appropriately in context-specific scenarios.

- **Principle 5: Adding A New Choice.** It focuses on supplementing existing choices with an additional one. The new choice is relevant to this question, but is not a correct answer, which relies on domain knowledge to exclude.

Remark: MPA is not limited to these five principles and more can be added easily through our framework.

4. Experiments

4.1. Experimental Setup

Tasks and Datasets. We selected four popular datasets for evaluation: MMLU (Hendrycks et al., 2021), ARC-Challenge (ARC-C) (Clark et al., 2018), GSM8K (Cobbe et al., 2021), and BigBench-Hard (BBH) (Suzgun et al., 2022; Srivastava et al., 2022), encompassing a broad spectrum of computational challenges ranging from knowledge-intensive understanding to complex mathematical and logical reasoning tasks. We adopted only three hard tasks from BBH: Formal Fallacies, Object Counting, and Temporal Sequences. We used their test sets to generate new evaluation samples. The detailed introduction is given in Appendix A.

Evaluated LLMs. We evaluated three proprietary LLMs: GPT-4-Turbo (OpenAI, 2023b), GPT-3.5-Turbo (OpenAI, 2023a), and Gemini-Pro (GeminiTeam, 2023), and three strong open-sourced models: Llama2-70b-chat (Touvron et al., 2023), Yi-34b-chat (01-ai, 2024), and Mixtral-8x7b-Instruct (MistralAITeam, 2023). To ensure a standardized comparison, we set the generation temperature to 0 for all models, with the generation length as 1000 tokens.[2]

Agent LLMs in MPA. We utilized GPT-4-Turbo as probing and judging agents, with temperatures of 0.7 and 0, respectively. The maximum token generation for each agent is set

[2]For Gemini-Pro on MMLU dataset, we set the temperature to 0.7 and omitted some evaluation samples (around 20 samples) to avoid response failures due to Google's safety constraints.

as 1000. While GPT-4-Turbo serves as the main agents, we also explored the potential to integrate other LLMs such as GPT-3.5-Turbo and Gemini-Pro as agents in later experiments in Section 4.6. The evaluation prompts are detailed in the Appendix B. Our primary evaluation metric is accuracy.

A bitter reality is that currently, only GPT-4 are capable to generate questions and judge the quality of generated questions. We believe as the field progresses, more cheaper models (such as Claude 3 and Gemini) will become capable of fulfilling both probing and judging roles, thereby reducing dependency on any single model's API and enhancing the scalability.

4.2. Main Results

In this part, we applied all five principles to generate new evaluation samples for MMLU and ARC-C. For GSM8K and BBH that do not have multiple choices, we restricted our probing to Principle 1 and 4. Appendix D shows some examples generated by MPA. We repeated the evaluation three times to reduce randomness. Table 2 presented the test accuracy of different LLMs on the original and our MPA benchmarks. Standard deviation are mostly around 1 (Table 5), indicating the robustness of the benchmark. As can be seen, all LLMs exhibited performance degradation on our probing benchmarks. Although GPT-4-Turbo demonstrated the strongest data contamination problem, it remains the strongest model. For MMLU, GPT-4-Turbo performed 15.7% worse than the original benchmark. Furthermore, a notable performance decline is evident in the case of MMLU and ARC-C, which is significantly more pronounced than that observed in GSM8K and BBH. This suggests that LLMs may encounter the memorization of knowledge-based benchmarks, resulting in substantial performance degradation for evaluation on our benchmarks.

We also presented a confusion matrix for analysis as shown in Figure 3. The matrix categorizes the responses into four distinct segments, with four categories to evaluate the model responses: OT (Original True), PT (Probing True), OF (Original False), and PF (Probing False). A notable observation is the high frequency of 'OT/PF' instances, which suggests a potential data contamination to specific dataset characteristics. Furthermore, the frequency of open-source models is markedly higher than that of proprietary models like GPT-4. This discrepancy indicates that open-source models might be more susceptible to data contamination.

4.3. Effect of Different Probing Principles

We further studied the effects of each modular probing principle. To this end, we established a baseline where the combined effect of all principles leads to a decrease in performance, normalized to a value of "1". This approach enables us to compare the relative effectiveness of each

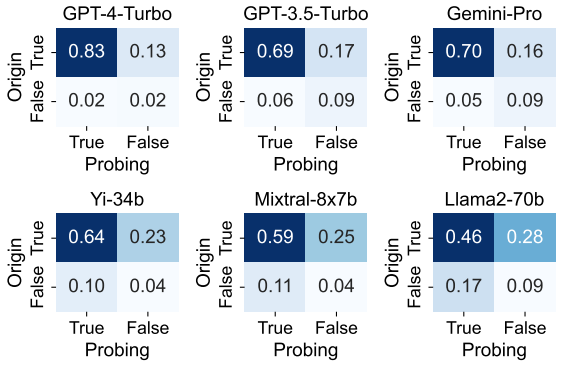


Figure 3. The confusion matrix of original benchmarks and probing benchmarks on ARC-C dataset.

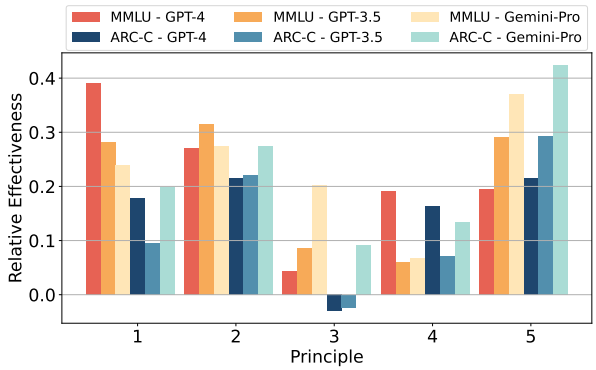


Figure 4. The relative effectiveness of different principles on MMLU and ARC-C dataset.

principle in isolation. The performance decrement for each principle, when applied independently, was evaluated and compared with this baseline. The Relative effectiveness (RE) is computed as: $\text{RE} = \frac{\text{Acc}_{p_i} - \text{Acc}}{\text{Acc}_{p_{all}} - \text{Acc}}$, where Acc_{p_i} is the accuracy when only apply principle p_i to MPA, $\text{Acc}_{p_{\text{all}}}$ denotes the accuracy of MPA when all principles are applied. Acc is the accuracy on the original benchmark.

The results on MMLU and ARC-C datasets are detailed in Figure 4 and Appendix C.2. It can be observed that, principle $1, 2$ and 5 are the most effective principles. While principle 3, which randomly permute choices, are less effective. For GPT-4-Turbo and GPT-3.5-Turbo on ARC-C dataset, it can intriguingly increase the performance.

4.4. Ablation Study on Prompt Engineering

We explored the efficacy of prompt engineering techniques, specifically Chain-of-Thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) (Brown et al., 2020). For ICL, we selected five examples from the corresponding training set as few-shot examples. We also applied MPA to these five examples, creating a transformed version of few-shot examples. We refer to the use of ICL with the original training examples as ICL_o and the use of ICL with the

Table 2. The performance of different LLMs on vanilla benchmarks and our probing benchmarks.

Model	GPT-4-Turbo			GPT-3.5-Turbo			Gemini Pro			Yi-34b			Mixtral-8x7b			Llama2-70b-chat		
Dataset	Vanilla	Ours	Δ	Vanilla	Ours	Δ	Vanilla	Ours	Δ	Vanilla	Ours	Δ	Vanilla	Ours	Δ	Vanilla	Ours	Δ
MMLU	84.40	68.86	-15.54	68.12	56.15	-11.97	67.04	55.55	-11.49	67.31	63.30	-4.01	66.49	55.24	-11.25	56.85	49.70	-7.15
GSM8K	95.22	88.50	-6.72	77.71	71.54	-6.17	22.97	20.39	-2.58	73.54	68.54	-5.00	61.56	47.18	-14.38	52.92	51.50	-1.42
ARC-C	96.16	84.67	-11.49	85.41	74.60	-10.81	86.18	75.91	-10.27	86.78	74.03	-12.75	84.47	70.36	-14.11	73.55	64.19	-9.36
BBH (partial)	88.53	87.78	-0.75	54.67	49.73	-4.94	65.47	60.00	-5.47	55.47	52.49	-2.98	53.47	40.53	-12.94	38.53	38.22	-0.31

Table 3. Results of different prompt engineering techniques for GSM8K and ARC-C dataset, with the highest and second-highest accuracies highlighted in bold and underlined, respectively.

Dataset	Model	Original	CoT	ICL_o	ICL_t
GSM8K	GPT-4-Turbo	88.50	89.31	**89.39**	88.78
	GPT-3.5-Turbo	**71.54**	70.58	65.73	64.90
	Gemini-Pro	20.39	24.49	**75.28**	73.01
ARC-C	GPT-4-Turbo	84.67	82.85	85.32	**85.67**
	GPT-3.5-Turbo	74.60	**75.94**	74.32	74.49
	Gemini-Pro	75.91	76.02	76.71	**79.10**

transformed examples as ICL_t.

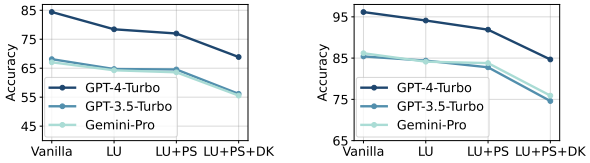
As can be observed in Table 3, neither CoT nor ICL can effectively boost the performance of LLMs on our probing benchmarks. Performance enhancements are varied across models and datasets when CoT and ICL are applied. For instance, GPT-4-Turbo showed a modest increase in accuracy on the GSM8K dataset with ICL, improving from 88.50 to 89.39. In contrast, GPT-3.5-Turbo's performance slightly decreased under the same conditions. Note that the significant performance gain observed for Gemini-Pro on the GSM8K dataset when using ICL is attributed to its initial lower effectiveness in a zero-shot context.

4.5. Albation Study on Data Contamination

We collected 30 novel reasoning questions from the experts (the same experts in the human evaluation in the general response), and examined the performance of the original questions and the probing questions. The GPT-4's accuracies of the original questions and the probing questions are both 60%. The OT/PF ratio is 3%, which is much lower than those in common public benchmarks, indicating that the newly generated benchmarks are not likely to be memorized compared to the public benchmarks.

4.6. Weak LLMs as Probing and Judge Agents

In this section, we assess the feasibility of using less advanced LLMs to reduce costs of MPA evaluation. We initially configured GPT-3.5-Turbo and Gemini-Pro as judging agents, while maintaining GPT-4-Turbo as the probing agent for GSM8K dataset. Through meticulous manual examination of the questions transformed by this setup, we observed a significant shortfall in the judging agents' ability to discern and exclude transformed examples whose meanings deviated from the original questions. highlighting the need for robust and capable LLMs to effectively sieve out appropriately probing questions. Subsequently, we extended our inquiry to assess the potential of employing less capable LLMs like GPT-3.5-Turbo and Gemini-Pro as probing agents. Using GPT-4-Turbo as the judging agent, our experiments revealed that when weaker LLMs served as probing agents, they often altered the essence or subtleties of the original questions, leading to misrepresentations or the introduction of unintended elements. And GPT-4-Turbo struggled to consistently detect these nuanced alterations. This suggests that the sophistication of the probing agent is crucial, as even advanced judging agents like GPT-4-Turbo cannot always offset the limitations of the probing agents.



(a) ARC-C (b) MMLU

Figure 5. The accuracy of different LLMs on ARC-C and MMLU on different levels of probing benchmarks. LU, LU+PS, and LU+PS+DK represent probing benchmarks that applied language understanding principles, both language understanding principles and problem solving principles, and all principles, respectively.

5. Multifaceted Analysis of the Basic Abilities

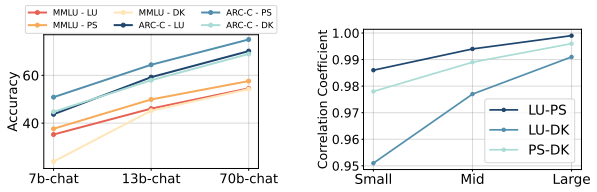
One advantage of MPA is the support for multifaceted analysis of abilities, which is discussed in this section.

5.1. Analysis on Benchmark Complexity

We first explore the impact of benchmark complexity on the MMLU and ARC-C datasets, as illustrated in Figure 5. We construct probing benchmarks with different levels of complexity using (1) language understanding principle 1, (2) language understanding principle 1 and problem-solving principle 4, and (3) language understanding principle 1, problem-solving principle 4, and domain knowledge principle 5. In particular, as the complexity of the benchmarks increases, the performance of all LLMs decreases and GPT-4-Turbo consistently reaches the highest precision. This result shows that there is still much room to improve the abilities of LLMs in complex benchmarks.

Table 4. The correlation efficient of the three basic abilities.

	Pearson	Spearman	Kendall
LU & PS	0.994	0.986	0.939
LU & DK	0.986	0.972	0.909
PS & DK	0.986	0.979	0.909



(a) Model Size v.s. Accuracy (b) Model Size v.s. Correlation

Figure 6. (a) The correlation between the performance and model size. (b) Cross-ability correlation with model size.

5.2. Relationship of the Basic Abilities

To gain a deep understanding of the relationship between the abilities of language understanding (LU), problem solving (PS), and domain knowledge (DK), we constructed three probing benchmarks using the principles that belong to a certain ability and then evaluated LLMs on these benchmarks. The results are presented in Table 8. After obtaining the performance of different LLMs on each probing benchmark, we then calculated the Pearson correlation coefficients (Pearson, 1920), the Spearman correlation coefficients (Spearman, 1904), and the Kendall correlation coefficients (Kendall, 1938), the results are presented in Table 4. It is evident that all abilities are highly correlated, which aligns with the findings of Burnell et al. (2023). Furthermore, it is observed that language understanding and problem solving are more relevant compared to other pairs of abilities. This suggests that the two abilities can predict each other, which has great potential to train and improve LLMs in the future. In the future, further detailed analysis can be performed to gain a deeper insight into the basic abilities of LLMs by constructing MPA evaluation samples based on other existing benchmarks such as HELM (Liang et al., 2023).

5.3. Analysis on Model Size

We studied the influence of model size on basic abilities. First, Figure 6(a) shows the correlation with variants of Llama2: 7b, 13b, and 70b. It can be observed that each ability positively correlates with the model size with nearly the same slope, indicating that when the size of the model increases, all abilities are equally enhanced to improve overall performance. Second, we explored the relationship between model size and correlations between different abilities. We roughly divided the models into three sizes: (1) small: Llama2-7b-chat, Llama2-13b-chat; (2) mid: Yi-34b-chat, Mixtral-8x7b-Instruct, Llama2-70b-chat; and (3) Gemini-Pro, GPT-3.5-Turbo, GPT-4-Turbo. The results in Figure 6(b) implies an implicit "Matthew Effect" (Merton, 1968): larger (often stronger) models tend to have stronger correlations between basic abilities. This aligns well with existing psychological theory about the g factor of general intelligence (Spearman, 1961). This finding can potentially help explain the emergent abilities of LLMs (Biderman et al., 2023; Schaeffer et al., 2023) and provide insight into the evolution of LLMs.

5.4. Error Analysis

We conducted an in-depth analysis of LLMs' failure modes in the three basic abilities. We meticulously selected 50 instances where GPT-4-Turbo correctly answered the original questions but failed in the transformed questions in the GSM8K dataset. These error modes are shown below.

- *Language understanding.* (1) Question Understanding Error: GPT-4-Turbo calculates the correct answer but misinterprets the intent of the question, leading to an incorrect response. This error indicates a gap in comprehension of the question. (2) Instruction Following Error: In this mode, GPT-4-Turbo arrives at the correct answer but fails to present it in the required format specified in the prompt, indicating a lack of follow-up instructions.
- *Problem solving.* Here, GPT-4-Turbo understands the question correctly but errs during the calculation process, resulting in a wrong final answer.
- *Domain knowledge.* We investigated the distribution of topic error among 57 tasks of MMLU in Figure 9. Notably, the `professional law` domain has the highest error rate, followed by `moral scenarios` and `professional psychology`, suggesting challenges predominantly in professional and ethical tasks.

Furthermore, we observed two possible reasons for performance degradation: ambiguity in the original questions and inconsistency between the probing and original questions. Some questions in the original dataset were found to be ambiguous (see Appendix C.5). Despite this, GPT-4-Turbo often provided plausible answers, suggesting potential issues with data memorization. However, certain transformed questions deviated in meaning from their original versions, leading to inconsistencies in responses.

6. MPA as Data Augmentation for LLMs

Although the main purpose of MPA is to evaluate LLMs, its generated samples can also be used as augmented data for fine-tuning. In this section, we conducted a pilot study using the OpenAI API to fine-tune GPT-3.5-Turbo on the data generated by MPA. Specifically, we used samples from the training split of MMLU and ARC-C, which are fed to MPA for data generation. We then evaluated the fine-tuned

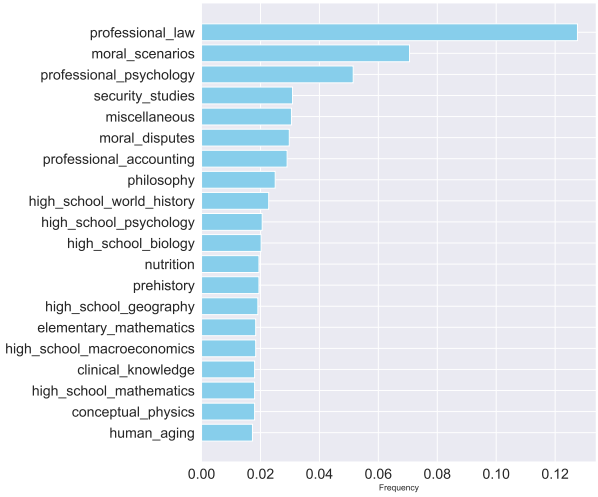


Figure 7. The bar chart of top 20 error topics and their corresponding frequencies of GPT-4-Turbo on MMLU dataset.

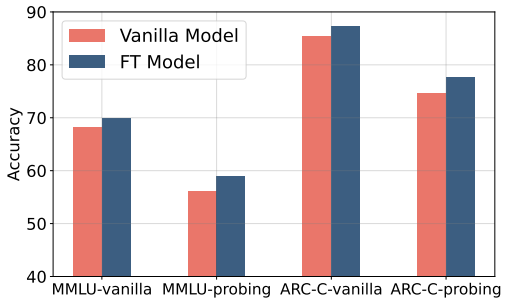


Figure 8. The performance of GPT-3.5-Turbo and GPT-3.5-Turbo-FT on original benchmarks and our probing benchmarks.

models on the original test split and our probing benchmarks. The fine-tuning data includes two parts, the first part is the probing questions generated by 5 principles separately, and the second part is the original training set.

The results in Figure 8 show that the data generated by MPA can improve the performance of LLMs, with an average of 2% improvements on both MMLU and ARC-C. The fine-tuning results demonstrated that MPA is not only an evaluation protocol, but a general data augmentation approach to improve the performance of LLMs, creating a huge advantage for training stronger LLMs in the future. The improved results also demonstrate the correctness of the generated data, indicating that our MPA is effective.

7. Conclusion and Discussion

This paper introduced MPA, a dynamic evaluation protocol to address data contamination and provide an in-depth analysis of the three key cognitive abilities of LLMs inspired by psychometric theory. Our experimental findings revealed several notable insights. Crucially, MPA-generated samples can not only function as evaluation tools, but also improve LLMs training as a data augmentation method. We believe that the psychometric-inspired adoption of LLMs as agents represents a promising direction.

Our work has several limitations. (1) Tasks and datasets: Our focus was limited to four datasets, encompassing a specific range of topics. Incorporating a broader spectrum of datasets and tasks could yield more comprehensive insights into LLMs capabilities. (2) The validity of probing benchmarks: While MPA employs a judge agent to assess the consistency and accuracy of probing benchmarks, we observed discrepancies in some questions, deviating from their original intent. This highlights the potential to further enhance MPA's robustness and effectiveness.

Impact Statement

Evaluating the general abilities of LLMs is essential to ensure responsible AI for the society. This work proposed a new evaluation protocol of LLMs to ensure that their true capabilities can be measured, which will help foster a better understanding of the models. We carefully controlled the generative models (agents) in the paper to ensure that they will not generate harmful content.

References

01-ai. Yi: A series of large language models. https://github.com/01-ai/Yi, 2024.

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., et al. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*, 2023.

bench authors, B. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? FAccT 2021, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raf, E. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Burnell, R., Hao, H., Conway, A. R., and Orallo, J. H. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Fan, L., Hua, W., Li, L., Ling, H., Zhang, Y., and Hemphill, L. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*, 2023.

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F., Neubig, G., Garg, A., Clark, J. H., Freitag, M., and Firat, O. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*, 2023.

Gao, I., Ilharco, G., Lundberg, S., and Ribeiro, M. T. Adaptive testing of computer vision models. *arXiv preprint arXiv:2212.02774*, 2022.

GeminiTeam. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Golchin, S. and Surdeanu, M. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*, 2023a.

Golchin, S. and Surdeanu, M. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023b.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

HuggingFace. Open-source large language models leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, June 2021.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, pp. 101861, 2023.

Lei, F., Liu, Q., Huang, Y., He, S., Zhao, J., and Liu, K. S3eval: A synthetic, scalable, systematic evaluation suite for large language models. *arXiv preprint arXiv:2310.15147*, 2023.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023a.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models, 2023b.

Li, Y. An open source data contamination report for llama series models. *arXiv preprint arXiv:2310.17589*, 2023.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Liu, H. and Yao, A. C.-C. Augmenting math word problems via iterative question composing. *arXiv preprint arXiv:2401.09003*, 2024.

Lovin, B. Gpt-4 performs significantly worse on coding problems not in its training data. https://brianlovin.com/hn/35297067, 2023.

Ma, Z., Ethayarajh, K., Thrush, T., Jain, S., Wu, L., Jia, R., Potts, C., Williams, A., and Kiela, D. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34:10351–10367, 2021.

Merton, R. K. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

MistralAITeam. Mixtral-8x7b-v0.1. https://huggingface.co/mistralai/Mixtral-8x7B-v0.1, 2023.

OpenAI. https://chat.openai.com.chat, 2023a.

OpenAI. Gpt-4 technical report, 2023b.

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T. B. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 2023.

Pearson, K. *The history and theory of correlation*. Biometrika Office, 1920.

Raykov, T. and Marcoulides, G. A. *Introduction to psychometric theory*. Routledge, 2011.

Ribeiro, M. T. and Lundberg, S. Adaptive testing and debugging of nlp models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3253–3267, 2022.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? In *NeurIPS*, 2023.

Significant-Gravitas. Autogpt. https://github.com/Significant-Gravitas/AutoGPT, 2023.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Spearman, C. " general intelligence" objectively determined and measured. 1961.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., , and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Tian, Y., Pei, K., Jana, S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering*, pp. 303–314, 2018.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S., and Zhang, Y. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Wei, T., Zhao, L., Zhang, L., Zhu, B., Wang, L., Yang, H., Li, B., Cheng, C., Lü, W., Hu, R., et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023.

Yang, L., Zhang, S., Qin, L., Li, Y., Wang, Y., Liu, H., Wang, J., Xie, X., and Zhang, Y. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*, 2022.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023a.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Gong, N. Z., Zhang, Y., et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023b.

Zong, Y., Yu, T., Zhao, B., Chavhan, R., and Hospedales, T. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*, 2023.

A. Datasets

The MMLU dataset contains $13,985$ test samples across 57 tasks, encompassing diverse areas such as humanities and social sciences, offering a comprehensive assessment of language understanding capabilities. The ARC-C dataset collected $1,172$ grade-school level science questions, presenting a unique blend of natural language understanding and scientific reasoning. In the GSM8K dataset, the focus is on mathematical problem-solving, featuring $1,319$ problems that require a combination of numerical understanding and logical reasoning. For the Formal Fallacies, Object Counting, and Temporal Sequences tasks in BBH dataset, each contains 250 test samples. These subsets were chosen for their relevance and representativeness, as they challenge LLMs to understand nuanced logical fallacies, accurately count objects in complex settings, and understand sequences of events over time.

B. Evaluation Prompts

In the following, we show the evaluation prompts while adopting different datasets.

MMLU

```
Here is a question about {task}:

{question}

{choices}

Choose the correct answer and explain why.  Please include your answer into <<<>>>.  For
example, if you choose A, please write <<<A>>>.
```

GSM8K

```
Here is a math problem:{question}

Please solve this math problem and include your answer into <<<>>>.  For example, if your
answer is 1, please write <<<1>>>.
```

ARC-C

```
Here is a multiple-choice science problem:

### Question:

{question}

### Choices:

{choices}

Please solve this problem and include your answer into <<<>>>.  For example, if your
choose A, please write <<<A>>>.
```

BBH (formal fallaices)

```
Here is a question about formal fallacies (given a context involving a set of statements,
determine whether an argument can be logically deduced from the provided context):

### Question:

{question}

### Choices:

{choices}

Please answer this question and include your answer into <<<>>>.  For example, if your
answer is valid, please write <<<valid>>>.
```

BBH (object counting)

```
Here is a question about object counting (given a collection of possessions that a person
has along with their quantities, determine the number of a certain object/item class.):

{question}
```

```
Please answer this question and include your answer into <<<>>>.  For example, if your
answer is 1, please write <<<1>>>.
```

BBH (temporal sequences)

```
Here is a question about temporal sequences (given a series of events and activities a
person has completed in the course of a day, determine what time, during the day, they
might have been free to perform another activity.):

### Question:

{question}

### Choices:

{choices}

Please answer this question and include your answer into <<<>>>.  For example, if your
answer is (A), please write <<<(A)>>>.
```

C. Detailed Results

C.1. Standard Deviation of Main Results

The dynamic evaluation protocol introduces randomness into the evaluation results. Therefore, we run all experiments three times to get the average results and the standard error. As shown in Table 5, the standard deviations for all models in all data sets are small, thus ensuring the fairness of our evaluation.

Table 5. The standard deviation of co-efficient of the main results.

Model	MMLU	GSM8K	ARC-C	BBH (partial)
GPT-4-Turbo	0.25	0.89	0.46	1.95
GPT-3.5-Turbo	0.18	1.15	0.58	1.49
Gemini-Pro	0.13	1.69	0.58	2.89
Yi-34b-chat	0.05	1.63	0.58	2.51
Mixtral-8x7b-Instruct	0.62	0.16	0.73	2.37
Llama2-70b-chat	0.93	1.52	1.12	1.42

C.2. Results of Different Modular Principles

We show the results on different principles in Table 6 and Table 7. Note that we only adopted partial samples from BBH.

Table 6. Results of different principles on MMLU and ARC-Challenge datasets.

Dataset	Model	Baseline	p_1	p_2	p_3	p_4	p_5
MMLU	GPT-4	84.40	78.43	81.48	80.27	81.42	83.73
	GPT-3.5	68.12	64.71	67.39	64.31	64.61	67.09
	Gemini-Pro	67.04	64.28	66.27	63.89	62.77	64.71
ARC-Challenge	GPT-4	96.16	94.11	94.28	93.69	93.69	96.50
	GPT-3.5	85.41	84.39	84.64	83.02	82.25	85.67
	Gemini-Pro	86.18	84.13	84.81	83.36	81.83	85.24

Table 7. Results on GSM8K and BBH (partial) datasets.

Dataset	Model	Baseline	p_1	p_2
GSM8K	GPT-4	95.22	90.83	91.66
	GPT-3.5	77.71	74.98	74.07
	Gemini-Pro	22.97	23.58	22.52
BBH (partial)	GPT-4	88.53	89.60	89.47
	GPT-3.5	54.67	52.27	48.40
	Gemini-Pro	65.47	66.93	61.73

C.3. Results of Relationship of the Basic Abilities

Table 8 shows the results on different abilities.

C.4. Top topics of MMLU

Figure 9 shows the top 20 MMLU topics where GPT-4-Turbo made the most errors. It can be observed that GPT-4-Turbo made more mistakes in "profession law", "moral scenarios", and "security studies", potentially due to insufficient training data and ambiguious groundtruth in these domains. For example, questions from "moral scenarios" are often difficult to answer. This trend underscores potential limitations in GPT-4-Turbo's current understanding or processing capabilities with

Table 8. Results of different LLMs on MPA based on ARC-C and MMLU datasets.

Dataset	ARC-C			MMLU		
Model	Language understanding	Problem solving	Domain knowledge	Language understanding	Problem solving	Domain knowledge
GPT-4-Turbo	90.27	94.28	93.69	75.18	81.48	81.42
GPT-3.5-Turbo	79.18	84.64	82.25	61.02	67.39	64.61
Gemini-Pro	80.46	84.81	81.83	59.53	66.27	62.77
Yi-34b-chat	79.44	85.67	83.19	60.01	66.10	64.50
Mixtral-8x7b-Instruct	78.16	82.17	78.58	61.18	66.01	61.64
Llama2-70b-chat	70.14	75.00	68.94	54.54	57.60	54.23

respect to the ethics and psychology domains.

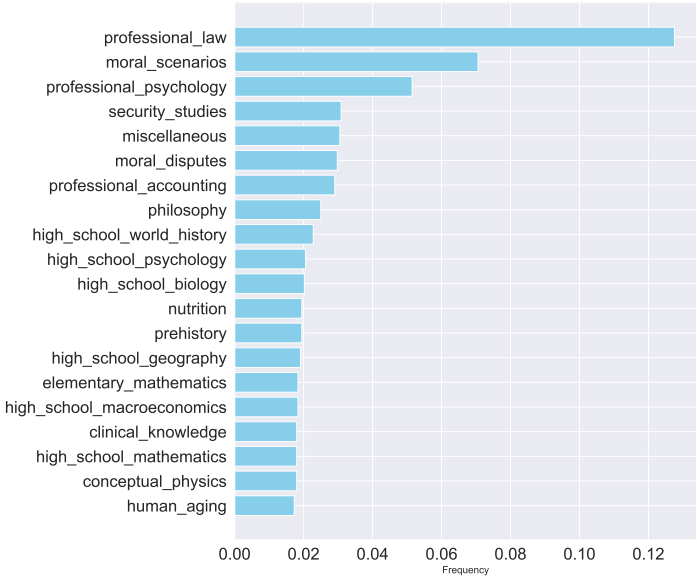


Figure 9. The bar chart of top 20 topics and their corresponding frequencies of GPT-4-Turbo on MMLU dataset.

C.5. Examples of wrong/ambiguous evaluation samples

In this section, we presented several wrong/ambiguous evaluation samples in GSM8K dataset.

- **Question:** `Lee used to be able to run the 400-meter hurdles two seconds faster than Gerald would run the 400-meter hurdles. But Gerald changed his diet, which improved his speed by 10%. If Lee runs the 400-meter hurdles in 38 seconds, how fast can Gerald, with his improved diet, run the 400-meter hurdles, in seconds?`
- **Answer:** 36
- **Analysis:** The correct answer is 36.3636.
- **Question:** `Mandy owes Benedict $100. They agreed to have monthly interest of 2%. If Mandy was able to pay it after 3 months, how much should she give to Benedict?`
- **Answer:** 106
- **Analysis:** The financial arrangement between Mandy and Benedict involves a principal loan of $100 with an agreed monthly interest rate of 2%. The ambiguity in the original question arises from the lack of specificity regarding the interest calculation method: simple or compound. The provided answer ($106) initially suggests a simple interest calculation. However, considering the possibility of compound interest sheds light on a different approach to determining the final amount owed.

D. Examples generated by MPA

Finally, we show some examples generated by MPA on the basis of different benchmarks.

MMLU:

- **Original Question:** This question refers to the following information.

  Read the the following quotation to answer questions.

  The various modes of worship which prevailed in the Roman world were all considered by the people as equally true; by the philosopher as equally false; and by the magistrate as equally useful.

  Edward Gibbon, The Decline and Fall of the Roman Empire, 1776{178

  Gibbon's interpretation of the state of religious worship in ancient Rome could be summarized as

  A: In ancient Rome, religious worship was decentralized and tended to vary with one's social position.

  B: In ancient Rome, religious worship was the source of much social tension and turmoil.

  C: In ancient Rome, religious worship was homogeneous and highly centralized.

  D: In ancient Rome, religious worship was revolutionized by the introduction of Christianity.

  Answer: A

- **Probing Question:** In his seminal work ``The Decline and Fall of the Roman Empire,'' Edward Gibbon explores the vast expanse of Roman history, from its zenith to its eventual collapse. Amid his detailed examination, Gibbon provides insights into the cultural and religious diversity that characterized the Roman Empire. How does he describe the differing perspectives of the general public, philosophers, and government officials regarding this religious diversity, particularly in terms of their acceptance and the impact on Roman society?

  A: Religious worship in ancient Rome was uniform and controlled by a central authority.

  B: The multiple forms of religious worship in ancient Rome often led to social conflicts and disturbances.

  C: The Roman state endorsed all forms of worship equally in an attempt to appease the gods and ensure the empire's prosperity.

  D: The arrival of Christianity in ancient Rome was a transformative force that completely changed the nature of religious worship.

  E: Religious practices in ancient Rome were not centralized, and they varied according to the social status of an individual.

  Answer: E

GSM8K:

- **Original Question:** Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

  Answer: 18

- **Probing Question:** Janet has a small farm where she raises a variety of animals, but her ducks are the most productive when it comes to laying eggs. Each day, without fail, her flock of ducks provides her with 16 fresh eggs. Janet has a particular routine where she enjoys a hearty breakfast that always includes three scrambled eggs. After breakfast, she dedicates some time to baking, preparing four delicious muffins that she shares with her friends. These muffins are special because they each require one egg. After using the eggs for her breakfast and baking, Janet packages the remaining

eggs to be sold at the bustling local farmers' market. Her eggs are quite popular, and she sells them for $2 each. Given her daily routine, how much money does Janet make from selling her eggs at the farmers' market each day?

Answer: 18

ARC-C:

- **Original Question:** An astronomer observes that a planet rotates faster after a meteorite impact. Which is the most likely effect of this increase in rotation?

  A: Planetary density will decrease.

  B: Planetary years will become longer.

  C: Planetary days will become shorter.

  D: Planetary gravity will become stronger.

  Answer: C

- **Probing Question:** In the vast expanse of the solar system, where celestial bodies are constantly in motion, a planet's day-to-day existence can be altered by events such as collisions with other objects. Imagine a scenario where astronomers witness a planet whose day has been significantly shortened due to the impact of a meteorite. This incident has resulted in the planet having a quicker rotational period around its axis. Given this situation, what is the probable consequence of this accelerated spin on the planet's environment or physical state?

  A: The time it takes for the planet to orbit the sun will increase.

  B: The duration of a single rotation of the planet on its axis will be less.

  C: The planet's atmosphere will become significantly thicker due to increased centrifugal force.

  D: The force with which the planet pulls objects towards itself will intensify.

  E: The mass per unit volume of the planet will be reduced.

  Answer: B

BBH (formal fallacies):

- **Original Question:** Here comes a perfectly valid argument: First of all, whoever is a schoolmate of Sondra is not a stepsister of Pricilla. In consequence, whoever is not a stepsister of Pricilla is a schoolmate of Sondra.

  Is the argument, given the explicitly stated premises, deductively valid or invalid?

  Options:

  - valid

  - invalid

  Answer: invalid

- **Probing Question:** At Ridgemont High, a peculiar rule is established by the student council: anyone who is a classmate of Sondra cannot concurrently be a half-sibling of Pricilla. This rule came about after a complex dispute over club memberships and family connections within the school. Now, consider a debate that erupted in the philosophy club during a discussion on logical reasoning. One of the members presented what seemed to be a sound argument based on the student council's rule: if it is true that no classmate of Sondra can be a half-sibling of Pricilla, then it logically follows that anyone who is not a half-sibling of Pricilla must be a classmate of Sondra. The club is now pondering whether this argument is logically coherent and deductively sound, based on the premises provided by the student council's peculiar rule.

  Options:

  - valid

  - invalid

  Answer: invalid

BBH (object counting):

- **Original Question:** I have a flute, a piano, a trombone, four stoves, a violin, an accordion, a clarinet, a drum, two lamps, and a trumpet. How many musical instruments do I have?

  Answer: 8

- **Probing Question:** As a passionate collector and music enthusiast, I've dedicated a significant portion of my living space to housing various items that reflect my interests and hobbies. Over the years, I've amassed a collection that includes both musical instruments and household items. Among my cherished possessions are one flute, one piano, one trombone, one violin, one accordion, one clarinet, one drum, and one trumpet. In addition to these, my practical side has led me to acquire four stoves and two lamps to meet my daily needs. Given this eclectic mix of items, can you carefully count and tell me how many items from my collection are musical instruments?

  Answer: 8

BBH (temporal sequences):

- **Original Question:** Today, Susan went to the coffee shop. Between what times could they have gone?

  We know that:

  Susan woke up at 7am.

  Linda saw Susan driving to the water park from 7am to 11am.

  John saw Susan buying clothes at the mall from 11am to 12pm.

  Jessica saw Susan taking photos near the Eiffel Tower from 12pm to 1pm.

  Steven saw Susan buying lunch at the deli from 1pm to 2pm.

  Thomas saw Susan reading at the library from 2pm to 6pm.

  The coffee shop was closed after 9pm.

  Between what times could Susan have gone to the coffee shop?

  Options:

  (A) 6pm to 9pm

  (B) 7am to 11am

  (C) 1pm to 2pm

  (D) 2pm to 6pm

  Answer: (A)

- **Probing Question:** On which occasion during her busy schedule could Susan have potentially squeezed in a visit to the local coffee shop? Susan's day kicked off at the crack of dawn, at 7am. Between the early hours of 7am and the late morning at 11am, Linda witnessed Susan making her way to the refreshing water park, where she was set to enjoy the slides and pools. During the late morning hour, from 11am to noon, John caught a glimpse of Susan amidst the bustling shoppers at the mall, where she was selecting some fashionable clothing items. As the clock struck noon and the day progressed to 1pm, Jessica was with Susan, snapping pictures against the backdrop of the iconic Eiffel Tower at a well-visited tourist spot. Following her tourist escapades, from 1pm to 2pm, Steven saw Susan at the cozy deli downtown, where she was deciding on her midday meal from a variety of savory options. Later in the afternoon, from 2pm to 6pm, Thomas noticed Susan deeply engrossed in literature at the quiet library, a place where she often finds solace in the pages of her favorite books. It's also important to note that the coffee shop in question shuts down its espresso machines and locks its doors to customers promptly at 9pm. Given Susan's known whereabouts throughout the day, deduce the time interval where she could have enjoyed a coffee shop visit.

  Options:

  (A) 6pm to 9pm

```
(B) 7am to 11am
(C) 1pm to 2pm
(D) 2pm to 6pm
Answer: (A)
```

E. Human verification results

Table 9. Results of human verification on MMLU and ARC-Challenge datasets.

Equivalence / Correctness	Language Understanding	Problem Solving	Domain Knowledge	Avg
Group 1	0.88/0.95	0.91/0.93	1.00/1.00	0.93/0.96
Group 2	0.92/0.99	0.94/0.98	1.00/0.97	0.95/0.98
Group 3	0.91/0.95	0.92/0.99	1.00/1.00	0.94/0.98
Avg	0.90/0.96	0.92/0.97	1.00/0.99	0.94/0.97