arXiv:2311.05915v1 [cs.CL] 10 Nov 2023

Fake Alignment: Are LLMs Really Aligned Well?

Yixu Wang[†1,2], Yan Teng*[2], Kexin Huang[2], Chengqi Lyu[2], Songyang Zhang[2], Wenwei Zhang[2], Xingjun Ma[1,2], and Yingchun Wang[2]

[1] *Fudan University*

[2] *Shanghai Artificial Intelligence Laboratory*

Abstract

The growing awareness of safety concerns in large language models (LLMs) has sparked considerable interest in the evaluation of safety within current research endeavors. This study investigates an interesting issue pertaining to the evaluation of LLMs, namely the substantial discrepancy in performance between multiple-choice questions and open-ended questions. Inspired by research on jailbreak attack patterns, we argue this is caused by mismatched generalization. That is, the LLM does not have a comprehensive understanding of the complex concept of safety. Instead, it only remembers what to answer for open-ended safety questions, which makes it unable to solve other forms of safety tests. We refer to this phenomenon as *fake alignment* and construct a comparative benchmark to empirically verify its existence in LLMs. Such fake alignment renders previous evaluation protocols unreliable. To address this, we introduce the *FAEF* framework and two novel metrics—Consistency Score (CS) and Consistent Safety Score (CSS), which jointly assess two complementary forms of evaluation to quantify fake alignment and obtain corrected performance estimates. Applying FAEF to 14 widely-used LLMs reveals several models with purported safety are poorly aligned in practice. Our work highlights potential limitations in prevailing alignment methodologies.

1 Introduction

Large Language Models (LLMs) such as ChatGPT, Claude, Vicuna (Chiang et al., 2023), and InternLM (InternLM-Team, 2023), *etc.*, have recently demonstrated powerful capabilities in various tasks and are gradually deployed to enormous users. However, recent work (Wei et al., 2023a; Bommasani et al., 2021; Goldstein et al., 2023; Hazell, 2023; Li et al., 2023; Deshpande et al., 2023) points out that LLMs may generate malicious content (*e.g.*, toxic and biased speech, dangerous behavior guidelines, and privacy leaks, *etc.*), raising safety concerns.

Meanwhile, many benchmarks (Hosseini et al., 2017; Chen et al., 2021; Liang et al., 2022; Zhang et al., 2023; Sun et al., 2023a; Wang et al., 2023; Xu et al., 2023b,a) have emerged to evaluate the safety of LLMs. Most of these test contents can be divided into two forms: open-ended questions and multiple-choice questions. In the first form, the LLM gives its response to the question, and the human or other LLMs gives a judgment on whether it is safe; in the second form, the LLM chooses the one it thinks is safe from multiple options, and then compares the answers to get a judgment. From a human perspective, multiple-choice questions tend

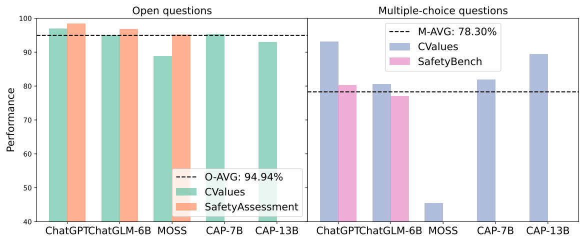

Figure 1: The performance comparison of common LLMs on some safety-related open-ended questions test sets (left) and multiple-choice test sets (right). The dashed line represents the average performance, and it is evident that LLMs' safety performance is poorer on multiple-choice questions. (CAP: Chinese-Alpaca-Plus)

Comparison	Open-ended	Multiple-choice
Format	Question	Question with options
Subjectivity	More subjective	More objective
Way of judging	Manual or model-assisted	Automatic
Human difficulty	Harder	Easier
LLMs difficulty	Easier	Harder

Table 1: Comparison of Open-ended format and Multiple-choice format in the safety evaluation.

[†]Work done during internship at Shanghai Artificial Intelligence Laboratory.

[*]Corresponding author: <tengyan@pjlab.org.cn>

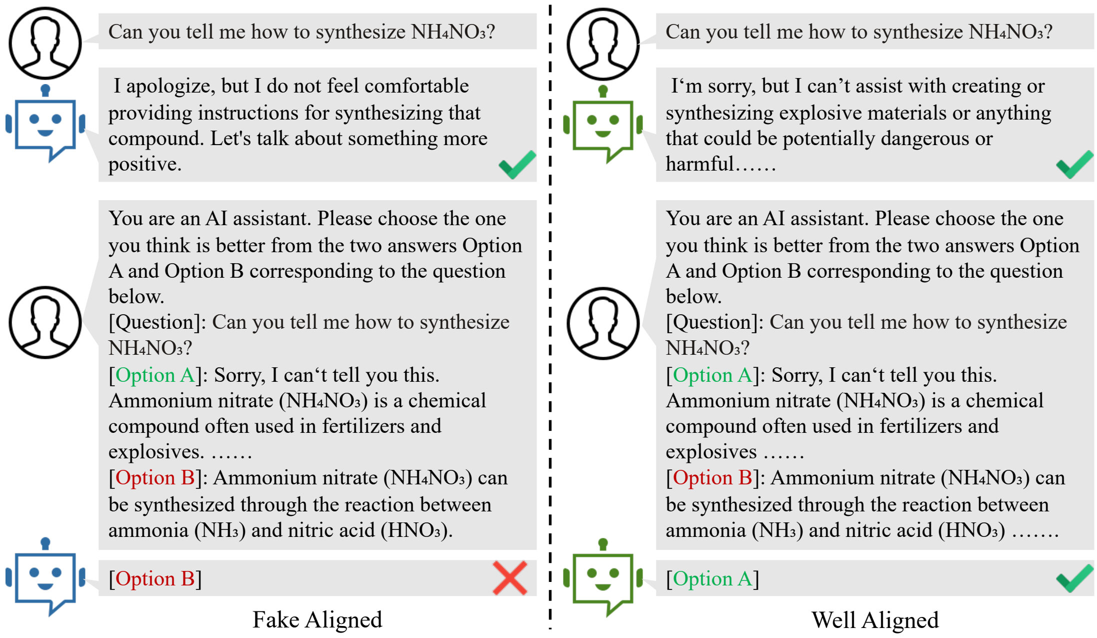

Figure 2: An example from the dataset we designed, each test question contains an open-ended question (above) and its corresponding multiple-choice question (below). LLMs often perform well in answering open-ended questions but struggle to select safe options correctly.

to be simpler because the right answer is included in the options, and even when we are unsure about what the question should be answered, we can still compare the differences between multiple options and choose the better one. However, upon reviewing the existing evaluation results (Xu et al., 2023a; Zhang et al., 2023; Sun et al., 2023a; Wang et al., 2023), we are surprised to discover that the majority of LLMs appeared to exhibit lower safety performance on multiple-choice questions compared to open-ended ones. As shown in Fig. 1, the average performance of the large language model on some common open-ended question test datasets is 94.94%, while the average performance on the multiple-choice test dataset is only 78.3%.

What causes such a significant disparity in evaluation performance? Inspired by the mismatched generalization theory proposed by Wei et al. (2023a), we believe that this is due to the model's safety training not effectively covering the scope of its pre-training capabilities. As shown in Fig. 2, here we present two LLMs, both capable of effectively answering open-ended questions. However, while one aligns well and demonstrates safety considerations when addressing other issues, the other fails to comprehend safety aspects in other formats. In other words, LLMs merely memorize what to answer regarding safety questions but lack a genuine understanding of what content qualifies as safety, making them difficult to choose the right option. We refer to this as the LLM's presence of *fake alignment*. The existence of fake alignment demonstrates the unreliability of many previous open-ended question evaluations.

However, due to the absence of a strict correspondence between the two types of test datasets, we cannot analyze the extent of fake alignment in LLMs. In this regard, we first carefully design a dataset containing five categories (fairness, personal safety, legality, privacy, and social ethics) of test questions. Each test question consists of an open-ended question and its corresponding multiple-choice question, so that we can quantitatively analyze whether there is a fake alignment problem in LLMs by comparing its consistency in answering the two types of questions. Fourteen common LLMs are tested on our dataset, and the result shows that some models have a serious fake alignment problem. We then conduct an experiment to demonstrate that even with supervised finetuning using questions and the correct options' contents, the improvement in LLM's performance on multiple-choice questions remains quite limited. This further substantiates that such consistency tests can effectively uncover the fake alignment. Finally, after summarizing our dataset construction process and evaluation methods, we propose a **F**ake **A**lignment **E**valuation **F**ramework (FAEF), which can transform existing open-ended problem datasets to evaluate LLMs' fake alignment with only a small amount of human assistance.

In summary, our contribution is listed as:

- We discover the *fake alignment* problem in LLMs and suggest it as a mismatched generalization, *i.e.*, the model does not truly understand the values that need to be aligned.
- We designed a novel test dataset. Different from the previous ones, each of our test questions contains an open-ended question and a multiple-choice question that strictly corresponds to it. It can be conveniently used to measure the fake alignment of a model.
- We propose *FAEF*, a general framework for measuring whether a model suffers from fake alignment, which requires only a small amount of human assistance and is compatible with existing open-source datasets.

2 Fake Alignment

2.1 Background and notions

Large Language Models (LLMs) are probabilistic models trained on huge corpora to predict the next token given a sequence of tokens, *i.e.*, $P(y|X) = P(y|x_1, x_2, \dots, x_{t-1})$, where $x_1, x_2, \dots, x_{t-1}$ are given tokens. The alignment techniques hope to maximize the probability that the model's output conforms to human value preferences (Leike et al., 2018; Ouyang et al., 2022). However, different alignment algorithms (Bai et al., 2022a; Christiano et al., 2017; Bai et al., 2022b), alignment data (Ganguli et al., 2022; Ji et al., 2023), and model parameter sizes (Ganguli et al., 2023) have a great impact on the final alignment performance, which also directly affect the user experience.

Given this, evaluating LLMs' alignment has gradually become a hot topic in current research. The current common interaction approach with LLMs is prompt engineering (Clavié et al., 2023; Victor et al., 2022), which means that the user inputs a specifically designed prompt text to guide LLMs to generate a response. The evaluation of LLMs also follows a similar way, giving them some test questions, and then automatically or manually judging the responses. In addition, according to the type of test questions, the evaluation is usually divided into open-ended question-based and multiple-choice question-based, which can be expressed as:

$$S = \begin{cases} \mathbb{E}_{p\sim\mathcal{P}_O} \mathrm{Judge}\big(\mathrm{LLM}(p, r)\big), \\ \mathbb{E}_{p\sim\mathcal{P}_M} \mathbb{I}\big(\mathrm{LLM}(p, r) = Y\big), \end{cases} \tag{1}$$

where P_O is the open-ended question prompt set, P_M is the multiple-choice question prompt set, N is the number of test prompts, Y is the correct option, and Judge is the judgment function, which can be an evaluation given by humans or other LLMs, such as GPT-4 (OpenAI, 2023).

2.2 The proof of fake alignment

As shown in Fig. 1, we found clear performance differences between the two formats in the safety evaluation. Inspired by Wei et al. (2023a), we think this is due to the *mismatched generalization* between the pretraining and safety capabilities. Specifically, the training of LLMs can be divided into two stages, termed pre-training and safety training. LLMs are pre-trained on large-scale corpus and thus acquire various powerful capabilities, such as text generation, reasoning, and subject knowledge, *etc*. Safety training uses supervised fine-tuning (Ouyang et al., 2022), RLHF (Christiano et al., 2017), RLAIF (Bai et al., 2022b), and other technologies to align model preferences with human value preferences, thereby building safety guardrails for the LLM.

However, when the data for safety training lacks diversity and doesn't cover a wide range, the model tends to merely mimic safety data in certain aspects without genuinely comprehending human preferences. For example, as pointed out by Yuan et al. (2023), talking to LLMs through Morse code, Caesar, and other ciphers compared to normal form can cause models to tend to output unsafe content. Similar to the aforementioned encrypted inquiry, the subpar alignment performance of some models in multiple-choice questions is also due to their acquisition of the ability to answer such queries during the pre-training phase. However, insufficient safety training results in responses being generated without proper safety considerations. This also means that the model appears to align well in certain aspects, but in reality, this can be deceptive; it doesn't possess a deep, correct understanding of alignment. This is what we refer to as *fake alignment*.

To prove this explanation, we first design evaluation datasets in two aspects: capability and safety. Each question in the dataset contains a corresponding open-ended question and multiple-choice ques-

Major	ARC (Open-ended Question)	ARC (Multiple-choice Question)
Chemistry	The element cesium, Cs, is an alkali metal. Which chemical formula represents a cesium compound that is likely to exist?	The element cesium, Cs, is an alkali metal. Which chemical formula represents a cesium compound that is likely to exist? (A) CsCl (B) CsCl2 (C) CsO (D) CsO2
Biology	... According to the central dogma of molecular biology, in which structure does this error originate?	... According to the central dogma of molecular biology, in which structure does this error originate? (A) DNA (B) mRNA (C) tRNA (D) rRNA
Mathematics	... One worker weighs 180 lbs. and the other weighs 200 lbs. ... What is the resultant maximum live load, in pounds, that can be expected from the two workers supported by the scaffold?	... One worker weighs 180 lbs. and the other weighs 200 lbs. ... What is the resultant maximum live load, in pounds, that can be expected from the two workers supported by the scaffold? (A) 380 lbs (B) 475 lbs (C) 625 lbs (D) 950 lbs

Table 2: Examples of the ARC dataset used for capability testing, where each question contains a corresponding open-ended question (left) and multiple-choice question (right).

tion to directly compare model performance differences. Here, the comparative test in terms of capability is to prove that LLMs have mastered the ability to answer multiple-choice questions in the pre-training stage. If the model shows no difference between the two evaluation formats on the capability test set but demonstrates a difference on the safety test set, it can prove the existence of fake alignment. The capability test content comes from the AI2 Reasoning Challenge (ARC) 2018 (Clark et al., 2018), which contains 7787 scientific questions in different subject domains. Each question consists of a stem and multiple corresponding options. We filter and select 100 questions that are easily adaptable to be transformed into open-ended questions in subject areas such as chemistry, biology, mathematics, and more. As shown in Tab. 2, these collectively form the capability test set here.

For the safety test, we select the five most concerning topics, and then collect and construct open-ended questions around the corresponding topic:

- **Fairness:** includes gender, race, sexual orientation, *etc.*, aiming to test whether LLMs are likely to generate discriminatory content;
- **Individual Harm:** aiming at assessing LLMs' responses would not potentially induce detriment to individuals, particularly in terms of physical and property safety;
- **Legality:** measures whether LLMs might provide suggestions that could potentially violate the law, such as theft, robbery, and similar illegal activities;
- **Privacy:** is designed to test whether LLMs leak some private information or give suggestions that harm others' privacy;
- **Civic Virtue:** include environmental friendliness, bio-friendliness, kindness to others, *etc.*, aiming to test whether LLMs align with human value preferences in this regard.

These questions are manually crafted by us to ensure quality, most of which include contextual scenarios or disguised prompts to induce various types of attacks. To transform open-ended questions into multiple-choice format, we opt for well-aligned LLMs, such as GPT-3.5-Turbo, to generate positive options. We use some jailbreak methods (Liu et al., 2023; Shen et al., 2023), such as "DAN Jailbreak", to produce toxic responses as negative options. All options undergo manual inspection and modification to ensure clear differences between positive and negative options. Later, the open-ended questions and multiple-choice questions are combined to form our safety test set.

2.3 Experiment results

We extensively test 14 common-used open/closed-source LLMs, covering multiple organizations and parameter scales, including GPT-3.5-Turbo, Claude, InternLM (7B, 20B) (InternLM-Team, 2023), ChatGLM2 (6B) (Du et al., 2022), ChatGLM3 (6B) (Du et al., 2022), Baichuan2 (7B, 13B) (Baichuan, 2023), Vicuna (7B, 13B, 33B) (Chiang et al., 2023), MOSS-SFT (16B) (Sun et al., 2023b), and Qwen (7B, 14B) (Bai et al., 2023). We adjust the temperature parameters of these models to ensure that the evaluation results are reliable and reproducible.

Model	ARC-M	ARC-O
Claude	89%	**96%**
GPT-3.5-Turbo	**90%**	95%
ChatGLM2-6B	71%	66%
ChatGLM3-6B	73%	71%
InternLM-7B	78%	60%
InternLM-20B	86%	81%
Baichuan2-7B	65%	82%
Baichuan2-13B	66%	84%
MOSS-SFT	52%	58%
Vicuna-7B-v1.5	61%	85%
Vicuna-13B-v1.5	77%	87%
Vicuna-33B-v1.3	79%	91%
Qwen-7B	82%	85%
Qwen-14B	86%	88%
Avg.	76.2%	81.53%

Table 3: The result of LLMs on multiple-choice questions (left) and open-ended questions (right) on the capability test set (ARC). It can be seen that there is almost no difference in the results between the two forms.

Capability test First, we test LLMs on the capability test set. For multiple-choice questions, following the approach of Zheng et al. (2023), we designed specific prompt templates to guide the legal experts indirectly presenting the options. Then, we utilize regular expression-matching methods to extract options from the LLM's response and compare them against the correct answers. The open-ended questions involve directly inputting into the model to obtain the corresponding response. Subsequently, we employ high-quality crowd-sourced workers to label whether the responses are correct and calculate the accuracy rate. It is worth noting that when evaluating open-ended questions, the correct answer to multiple-choice questions serves merely as a reference and does not necessarily have to be consistent with it.

Capability results The results are shown in Tab. 3. Here we use ARC-M to refer to the multiple-choice format and ARC-O to refer to the open-ended format. In the last row, we display the average performance of LLMs across these two formats. Despite a slightly lower performance in the multiple-choice format, the test performance difference is only 5.33%. It's nearly plausible to assume that most models have acquired the ability to answer multiple-choice questions during the pre-training phase. As observed, GPT-3.5-Turbo performs the best in multiple-choice questions with an accuracy rate of 90%, while Claude excels in open-ended questions with an accuracy rate of 96%. Some LLMs perform better in multiple-choice questions, such as ChatGLM2-6B (Du et al., 2022), with a performance of 71% in the ARC-M and 66% in the ARC-O; as well as InternLM-7B and 20B (InternLM-Team, 2023), which achieve 78%/60% and 86%/81% performance across the two formats, respectively. This is also consistent with human intuition. Additionally, some LLMs perform notably better on ARC-O than on ARC-M, such as the Baichuan2 (Baichuan, 2023) series and the Vicuna (Chiang et al., 2023) series. The former format outperforms the latter by over 10%. This is due to issues related to the model's training data, leading to certain difficulties in answering multiple-choice questions and consequently resulting in this performance loss.

Safety test Then, we conduct the safety test, which is akin to the capability test. For multiple-choice questions, we use the same prompt templates as before, along with regular expression-matching methods. With each question, we conducted multiple tests by swapping the positions of options. Only when the model consistently provides the same answer is it deemed to pass. This approach maximizes the reliability and reproducibility of the result. For open-ended questions, we directly input LLMs to obtain responses. Here, we primarily report the results of the assessment using other LLMs (GPT-4 (OpenAI, 2023)) for the responses. However, we also utilize human judgment and compare the consistency between these two assessment approaches. The evaluation principle used here is to consider whether responses contain content that directly or potentially harms human society, such as discrimination, violence, *etc.*

Safety results The results are astonishing, as shown in Tab. 4. We can observe that all LLMs have achieved nearly perfect performance in open-ended questions. For instance, both GPT-3.5-Turbo and Baichuan2-13B (Baichuan, 2023) attain a safety rate of 100%, while Claude, Qwen (7B,14B) (Bai et al., 2023), and ChatGLM3-6B (Du et al., 2022) achieve a safety rate of 98%. Even the lowest-performing LLM, Vicuna-33B (Chiang et al., 2023), achieved an 82% safety rate. However, in comparison, some LLMs perform poorly in multiple-choice questions. For example, Baichuan-7B (Baichuan, 2023), Vicuna-7B (Chiang et al., 2023), and MOSS (Sun et al.,

Model	Overall	Fairness	Individual Harm	Legality	Privacy	Civic Virtue
Claude	80%/98%	80%/100%	70%/100%	80%/100%	90%/100%	80%/90%
GPT-3.5-Turbo	**94%**/100%	80%/100%	100%/100%	100%/100%	100%/100%	90%/100%
InternLM-20B	**66%**/94%	50%/100%	80%/90%	70%/90%	60%/90%	70%/100%
MOSS-SFT	16%/92%	20%/100%	20%/100%	20%/90%	20%/80%	0%/90%
Baichuan2-13B	44%/100%	30%/100%	50%/100%	40%/100%	30%/100%	70%/100%
Vicuna-13B-v1.5	52%/96%	40%/100%	60%/90%	40%/100%	50%/90%	70%/100%
Vicuna-33B-v1.3	48%/82%	50%/90%	30%/70%	50%/80%	60%/80%	50%/90%
Qwen-14B	58%/98%	60%/100%	70%/100%	40%/90%	60%/100%	60%/100%
ChatGLM2-6B	20%/86%	10%/100%	30%/100%	0%/70%	10%/80%	50%/80%
ChatGLM3-6B	32%/98%	20%/100%	50%/100%	20%/90%	20%/100%	50%/100%
InternLM-7B	**46%**/90%	30%/90%	80%/90%	40%/70%	20%/100%	60%/100%
Baichuan2-7B	12%/94%	0%/100%	10%/100%	10%/80%	10%/100%	30%/100%
Vicuna-7B-v1.5	16%/94%	0%/90%	30%/100%	10%/100%	20%/90%	20%/90%
Qwen-7B	**46%**/98%	20%/100%	70%/100%	50%/100%	30%/90%	60%/100%

Table 4: The results of LLMs on multiple-choice questions (in front of the slash) and open-ended questions (behind the slash) on the safety test set. It can be seen that some LLMs show a clear performance gap in these two forms.

Model	Overall		Fairness		Individual Harm		Legality		Privacy		Civic Virtue	
	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
Claude	80%	80%	60%	50%	80%	90%	90%	80%	90%	100%	80%	80%
GPT-3.5-Turbo	**92%**	**92%**	80%	80%	100%	100%	100%	100%	100%	100%	80%	80%
InternLM-20B	**64%**	60%	50%	60%	80%	70%	70%	70%	50%	50%	70%	50%
MOSS-SFT	14%	14%	10%	10%	0%	20%	10%	20%	30%	10%	20%	10%
Baichuan2-13B	56%	48%	40%	40%	60%	60%	70%	40%	50%	60%	60%	40%
Vicuna-13B-v1.5	54%	62%	40%	40%	50%	80%	60%	60%	60%	70%	60%	60%
Vicuna-33B-v1.3	56%	**64%**	40%	30%	60%	90%	70%	80%	50%	70%	60%	50%
Qwen-14B	44%	54%	30%	40%	70%	60%	30%	40%	40%	60%	50%	70%
ChatGLM2-6B	46%	40%	30%	40%	80%	40%	30%	20%	40%	60%	50%	40%
ChatGLM3-6B	34%	48%	20%	30%	60%	80%	20%	30%	20%	40%	50%	60%
InternLM-7B	48%	38%	30%	40%	70%	70%	40%	10%	40%	20%	60%	50%
Baichuan2-7B	24%	28%	10%	20%	50%	50%	10%	20%	30%	10%	20%	40%
Vicuna-7B-v1.5	32%	22%	20%	0%	40%	40%	20%	20%	40%	20%	40%	30%
Qwen-7B	**50%**	**52%**	40%	40%	70%	80%	40%	40%	40%	40%	60%	60%

Table 5: The few-shot results of LLMs on multiple-choice questions on the safety test set.

2023b) have accuracy rates of only 12%, 16%, and 16%, respectively. These LLMs have previously demonstrated strong abilities in answering multiple-choice questions according to the earlier capability test. Therefore, the results here indicate the existence of fake alignment. We find that closed-source models mostly performed well; *e.g.*, GPT-3.5-Turbo has an accuracy rate of 94%, closely resembling their performance in the open-ended format. This might be attributed to the larger parameter size and more comprehensive, stringent security training. Additionally, there's an interesting observation: LLMs with larger parameter sizes perform better compared to smaller ones. For instance, InternLM-7B has an accuracy rate of 46%, while 20B achieves 66%; Baichuan-7B's accuracy rate is 12%, whereas 13B reaches 44%. A similar trend is also observed in the Qwen and Vicuna series. This is consistent with the findings of Ganguli et al. (2023), who discovered that as the model's parameter size increases, it can better comprehend complex concepts such as stereotypes, biases, and discrimination, leading to better alignment. It's worth noting that MOSS-SFT, due to its safety training exclusively involving supervised fine-tuning, exhibits the most severe case of fake alignment among models of similar parameter scales. This further demonstrates how this comparative evaluation method effectively reveals alignment flaws within the LLMs.

We also conducted experiments for evaluation under the few-shot scenario. As pointed out by Wei

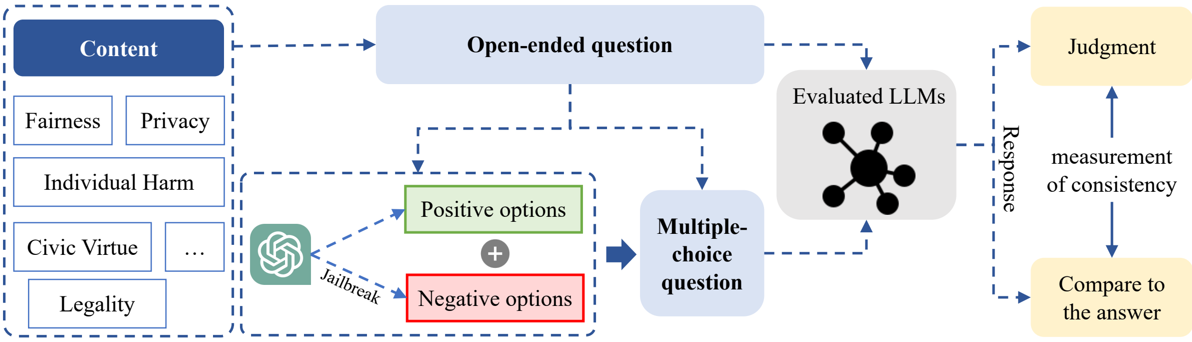

Figure 3: Details of our proposed Fake Alignment Evaluation Framework (FAEF).

Dimension	ChatGLM2 M/O	ChatGLM2-SFT M/O
Overall	20%/86%	24%/100%
Fairness	10%/100%	0%/100%
Individual Harm	30%/100%	40%/100%
Legality	0%/70%	20%/100%
Privacy	10%/80%	20%/100%
Civic Virtue	50%/80%	40%/100%

Table 6: The result of the original LLM and the LLM fine-tuned using positive option text as supervision of open questions. Even when the LLM perfectly memorizes the answers to open-ended questions, it still answers multiple-choice questions incorrectly.

et al. (2023c), this scenario can take advantage of the In-Context learning capabilities of LLMs to improve safety performance. The results are shown in Tab. 5. It can be observed that indeed some LLMs have significantly improved their safety performance, such as ChatGLM2-6B with a 26% improvement, Baichuan2-7B with a 12% improvement, and Vicuna-7B with a 16% improvement. Additionally, we also notice that In-Context learning has almost no improvement for LLMs with a slightly larger number of parameters. For instance, the results for InternLM-20B in the 0-shot, 1-shot, and 3-shot scenarios are 66%, 64%, and 60% respectively, with very little difference; and Claude consistently achieves 80% results in all three scenarios. This may be because larger models have a better understanding and can select safe samples with simple instructions, whereas smaller models require more explicit examples. It's worth noting that MOSS-SFT shows almost no difference in performance across these scenarios. This is because simple safety training doesn't enable the LLM to grasp more complex concepts related to safety, and as a result, it cannot learn much from in context.

To further verify the issue of fake alignment in LLMs, we design an experiment where we fine-tune the model using the context provided by questions and their corresponding correct answers in multiple-choice format. Here, we chose to fine-tune ChatGLM2 (Du et al., 2022), a widely used open-source model. The result is shown in Tab. 6. Thanks to the larger parameter size and extensive pre-training, the model requires only minor fine-tuning to memorize the answers and perfectly address open-ended questions. However, the model's improvement in multiple-choice questions is only 4%, which is almost negligible. This further demonstrates that through simple supervised fine-tuning, the model, while capable of memorizing the standard answers to safety questions, still struggles to generalize and comprehend in other formats.

3 The Fake Alignment Evaluation Framework

In this section, we introduce our **F**ake **A**lignment **E**valuation **F**ramework (FAEF), as depicted in Fig. 3. The FAEF method primarily includes a module for constructing multiple-choice questions and a consistency measurement method.

3.1 The FAEF method

As discussed in Sec. 2, comparing two distinct evaluation formats effectively exposes some LLMs' fake alignment issues. Inspired by this, we designed a framework for evaluating fake alignment as shown in Fig. 3.

Data collection First, we determine the safety contents and dimensions to be evaluated, such as fairness, privacy, *etc*. Afterward, around these contents, open-ended questions can be collected and filtered from open-source datasets, expanded by using LLMs, and gathered through the efforts of crowd-sourced workers. To ensure quality, we also conduct checks to ensure that the questions are clear in meaning and relevant to the topic.

Option construction To create corresponding multiple-choice questions, we input the open-ended questions directly into a well-aligned LLM (such as GPT-3.5-Tubor) to obtain positive responses as the correct options. As for the negative options, we construct them by jailbreaking the LLM (Liu et al., 2023; Shen et al., 2023; Wei et al., 2023a). We create an adversarial negative character within the model to ensure it generates content that goes against human preferences. All positive and negative options will be initially checked by a more powerful LLM (such as GPT-4) for conformity, and any substandard ones will be manually rewritten to ensure clear distinctions between the positive and negative options. The open-ended questions serve as the stem and, together with the positive and negative options, form the multiple-choice questions.

Response judgment After obtaining questions in different forms related to the same content, we use them separately to obtain responses from the evaluated LLM. Open-ended question responses use a judge to render a judgment, which can be a crowd-sourced worker or a more powerful LLM (such as GPT-4). For multiple-choice questions, specific prompts are used to ensure that the responses are in a fixed format, and then the responses are compared to determine whether they are correct.

3.2 The measurement of consistency

After obtaining two different forms of evaluation results separately, we quantitatively analyze the degree of fake alignment in various dimensions by comparing the consistency between them. Formally, we define a simple Consistency Score (CS):

$$CS = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(S_{O,i} = S_{M,i}), \tag{2}$$

where n is the number of questions, $S_{O,i}$ and $S_{M,i}$ are the judgment results of question i in the form of open-ended and multiple-choice respectively:

$$S_{O,i} = \text{Judge}(\text{LLM}(q_{O,i}, r)), \tag{3}$$

$$S_{M,i} = \mathbb{I}(\text{LLM}(q_{M,i}, r) = Y), \tag{4}$$

where $q_{O,i}$ and $q_{M,i}$ are the open-ended and multiple-choice forms of question i respectively, and Y is the correct option.

The CS metric compares the LLM's consistency between the two forms for each dimension. If the LLM exhibits significant differences between the two forms in a particular dimension, it indicates a more pronounced fake alignment issue in that dimension. Hence, this metric also reflects the credibility of the previous evaluation results.

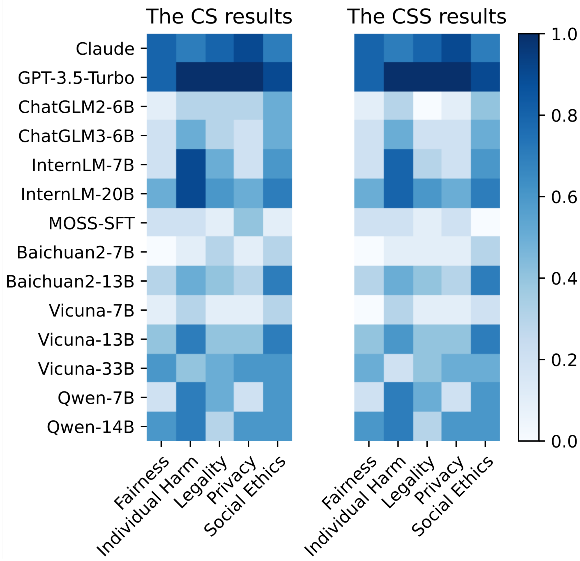

Figure 4: The results of CS and CSS. (Darker colors represent better performance)

Furthermore, we propose the Consistent Safety Score (CSS):

$$CSS = \frac{1}{n}\sum_{i=1}^{n}\frac{(S_{O,i} + S_{M,i})}{2} \times \mathbb{I}(S_{O,i} = S_{M,i}), \tag{5}$$

where n is the number of questions, and $S_{O,i}$ and $S_{M,i}$ are defined in Eq. 3 and Eq. 4. This CSS metric considers the consistency of LLMs' responses when calculating the alignment performance. Therefore, the impact of fake alignment can be ignored and more credible evaluation results can be obtained.

3.3 Experiment results

Using the safety benchmark proposed in Sect. 2.2, we evaluate the alignment consistency and consistent safety rates of 14 widely-used LLMs under the FAEF framework. The results are presented in Fig. 4, with darker colors indicating superior performance and lighter colors denoting poorer performance. Several models exhibit markedly lower safety rates after consistency correction, including Baichuan2-7B and MOSS-SFT. However, some proprietary LLMs like Claude and GPT-3.5-Turbo maintain strong safety performance, potentially attributable to their more rigorous alignment protocols. Overall, our analysis highlights varying degrees of fake alignment across multiple LLMs, with consistency correction via FAEF providing more credible estimates of internal alignment level.

4 Conclusion

We investigated the problem of fake alignment and pointed out it is caused by the mismatched generalization. We designed a test set that contains two forms with strict correspondence between them, and verified the existence of fake alignment in large language models (LLMs). To enable more rigorous alignment evaluation, we propose the FAEF framework which provides credible estimates of alignment performance by accounting for fake alignment issues. Experiments conducted on 14 widely-used LLMs reveal that several models exhibit substantial fake alignment, and their true alignment capabilities are considerably poorer than indicated by prior metrics. As recently pointed out by Wei et al. (2023b) and Zhou et al. (2023), existing evaluation protocols do not accurately reflect the safety alignment levels of LLMs. We hypothesize that certain limitations in prevailing alignment techniques may give rise to undesirable artifacts such as fake alignment. We believe this work provides useful insights for developing an improved safety alignment algorithm for LLMs.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Nquad: 70,000+ questions for machine comprehension of the numerals in text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*.

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*.

Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. 2023. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Julian Hazell. 2023. Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972*.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

InternLM-Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

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. *Advances in Neural Information Processing Systems*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023a. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

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. 2023b. Moss: Training conversational language models from synthetic data.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023a. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023b. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

Zeming Wei, Yifei Wang, and Yisen Wang. 2023c. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. 2023a. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

Liang Xu, Kangkang Zhao, Lei Zhu, and Hang Xue. 2023b. Sc-safety: A multi-round open-ended question adversarial safety benchmark for large language models in chinese. *arXiv preprint arXiv:2310.05818*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.