KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models

Anonymous ACL submission

Abstract

The evaluation of large language models (LLMs) has attracted increasing attention. Existing approaches, including human, static dataset-based, and LLM-based evaluation, might face limitations such as data contamination, constrained generalizability, and high cost coupled with limited scalability. In this paper, we introduce the Knowledge-grounded Interactive Evaluation (KIEval), a novel approach to assessing instruction-tuned LLMs. Starting with a question in a conventional LLM benchmark involving domain-specific knowledge, KIEval exploits dynamically generated and knowledge-centric multi-round dialogues to mitigate data contamination and enhance the reliability of evaluations. The framework of KIEval is generalizable across various domains and tasks, yielding a scalable and cost-effective approach that can efficiently yet robustly assess knowledge generalization and generation capabilities of LLMs. With KIEval, we hope to bring new insights into evaluating LLMs effectively in conversation scenarios and how data contamination impacts LLMs' real-world performance.

1 Introduction

The landscape of artificial intelligence has been significantly reshaped by the emergence of Large Language Models (LLMs) as they have been pivotal in various natural language understanding and generation tasks (Brown et al., 2020; OpenAI, 2023; Bubeck et al., 2023). As LLMs become more ingrained in our technological fabric, their comprehensive evaluation becomes increasingly essential (Chang et al., 2023).

Existing evaluation approaches can be classified into three types: human evaluation, static dataset-based evaluation, and LLM-based evaluation. Static dataset-based evaluation (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2020; Huang et al., 2023) require LLMs to generate a short span of text containing answer choices to predefined questions (Gao et al., 2021) to challenge model's knowledge. LLM-based automatic evaluation evaluations (Chiang and Lee, 2023) typically depend on LLM evaluators to evaluate model's output given predetermined, human-curated question templates (Zheng et al., 2023; Lin and Chen, 2023; Fu et al., 2023) or instructions (Wang et al., 2023b).

However, the evaluation of LLMs still faces several challenges. First, **Data Contamination Compromises the Evaluation Integrity:** the evaluation based on static datasets is recently challenged due to its possible susceptibility to data contamination (Schaeffer, 2023; Wei et al., 2023; Oren et al., 2023; Sainz et al., 2023), where models trained on test sets can artificially inflate benchmark performance, failing to reflect real-world performance (Zhou et al., 2023). Despite the increasing number of high-quality datasets, data contamination remains a significant challenge. Second, **Limited Generality:** The LLM-based evaluation relies on human-curated inputs, which are resource-intensive to gather, limiting their applicability across diverse domains and tasks. Furthermore, they also face contamination since static testing inputs are publicly accessible and easily compiled (Daniele and Suphavadeeprasit, 2023). These methods lack the capacity to dynamically test a model's generative ability to maintain coherent and contextually relevant conversations. Third, **Cost and Scalability:** Human evaluations, though insightful (Novikova et al., 2017), often lack consistency (Peng et al., 1997) and are resource-intensive (Karpinska et al., 2021). Similarly, creating and maintaining high-quality benchmark datasets is not only time-consuming but also hard to scale, posing significant resource challenges.

These challenges underscore the need for a more dynamic, generalizable, scalable evaluation framework that can more accurately reflect the generative capabilities of LLMs in practical scenarios.

In this paper, we introduce **KIEval**, a Knowledge-grounded Interactive Evaluation framework. The KIEval evaluation process initiates with a question derived from an existing benchmark dataset that requires domain-specific knowledge. Diverging from traditional methods, which primarily concentrate on choosing candidate answers, KIEval places greater emphasis on generative capabilities. It achieves this through structured and dynamically generated multi-round dialogues specifically tailored to explore knowledge related to the question. KIEval exhibits the following three advantages over existing evaluation approaches:

- **Avoiding Contamination:** By leveraging dynamically generated, multi-round dialogues focused on domain-specific topics with LLMs, KIEval reduces the risk of data contamination since all interactions are dynamic, providing a more grounded evaluation.
- **Task-agnostic:** By design, KIEval supports evaluation on various domains, languages, and tasks. It does not require extra human effort in writing templates or comparison pairs for evaluation.
- **Cost-effectiveness and Scalability:** Utilizing existing high-quality benchmark datasets for evaluations, KIEval offers a cost-effective solution to the challenges of traditional human evaluations and dataset maintenance. Its scalable design, avoiding quadratic complexity typical of pairwise model comparisons (Wang et al., 2023b; Zheng et al., 2023), significantly reduces computational demands, particularly advantageous for evaluations involving numerous models.

Crucially, we validate KIEval's alignment with humans and compare the results with existing benchmarks. Human annotation is used in the meta-evaluation of KIEval to prove its alignment to human preference. The high level of concordance with human judgments attests to KIEval's effectiveness in mirroring human preference.

Our core contributions are three-fold:

- *A new dynamic evaluation protocol:* We provide KIEval to evaluate LLMs through dynamically generated multi-turn dialogues to mitigate the issues of data contamination, limited generalization, and high cost.
- *Extensive experiments:* We conduct thorough experiments and analysis with 7 leading LLMs across 5 datasets with KIEval, assessing generative abilities and domain knowledge. Our findings also reveal the susceptibility of static dataset-based and LLM-based evaluations to data contamination, a challenge KIEval effectively mitigates.
- *New insights on data contamination:* We further discuss how data contamination affects model's generative performance, and test whether such contamination leads to mere memorization of answers or contributes to genuine understanding and generalization abilities.

2 Related Work

2.1 Evaluating LLMs

Human evaluation approaches manually design experiments and tests (Novikova et al., 2017; Bommasani et al., 2023). While it provides insights into human-model interaction, it faces challenges due to the subjectivity and inconsistency of human judgments (Chang et al., 2023). Moreover, it is resource-intensive in terms of time and cost, limiting its feasibility for large-scale assessments (Karpinska et al., 2021).

Static dataset-based approaches assess LLMs focused on domain-specific questions or tasks using pre-defined static datasets. Typical evaluation tasks include solving single or multiple-choice problems (Clark et al., 2018; Hendrycks et al., 2020; Huang et al., 2023) and question answering (Lin et al., 2021; Cobbe et al., 2021), these tasks require LLMs to generate short spans of text containing answers to the questions (Gao et al., 2021). The performance of LLMs is measured by their ability to correctly answer or perform these tasks.

LLM-based evaluation, utilizing one strong LLM (Brown et al., 2020; OpenAI, 2023) to assess others, is a recent approach that often employs pairwise comparisons to identify nuanced differences in model outputs, addressing the challenge of determining clear model superiority (Wang et al., 2023b; Zheng et al., 2023). This method bridges the gap between human and dataset-based evaluations by focusing on generative abilities. However, this approach has limitations, including reliance on fixed templates (Zheng et al., 2023), instructions (Wang et al., 2023b; Li et al., 2023), or multi-round chat datasets (Fu et al., 2023; Lin and Chen, 2023), limiting its scope in capturing diverse domain knowledge and real-world applicability. It also faces contamination risks, as training on out-

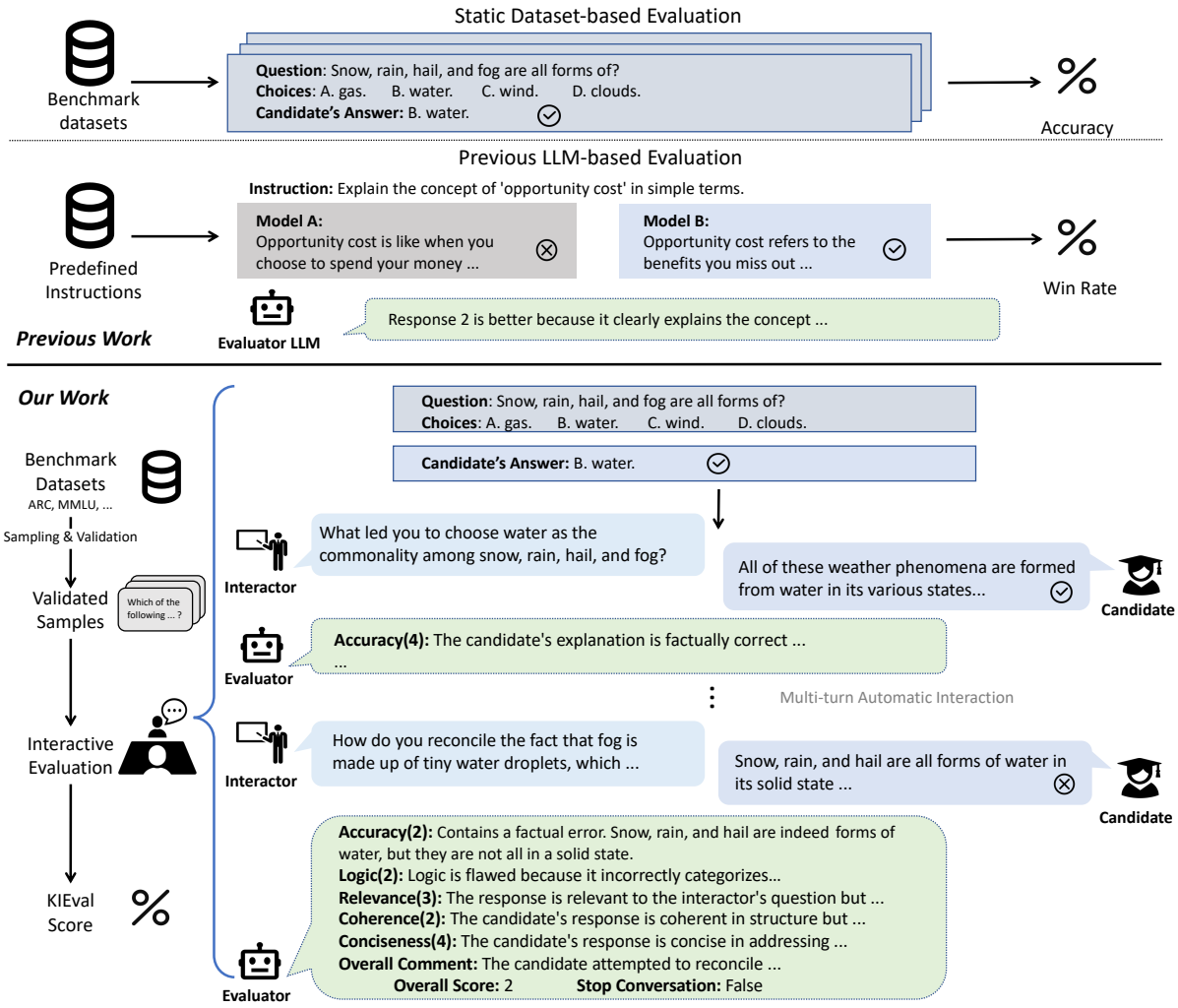

Figure 1: The pipeline of KIEval compared to previous static dataset-based and LLM-based evaluation methods.

puts from a strong LLM can inflate results, as noted in work from Daniele and Suphavadeeprasit (2023) collect data from MT-Bench (Zheng et al., 2023) as training data while AlpacaEval (Li et al., 2023) contains evaluation set from various instruction-tuning dataset. Additionally, studies indicate potential biases in these evaluations, such as positional bias (Zeng et al., 2023; Wang et al., 2023a,b).

2.2 Addressing Data Contamination of LLMs

Recently, the AI community has become increasingly concerned (Schaeffer, 2023; Zhou et al., 2023; Oren et al., 2023) about data contamination in LLMs. Wei et al. (2023); Shi et al. (2023) leveraged loss values or token probabilities to detect whether certain text appears in the training data of models. Zhu et al. (2023) leveraged DAG to dynamically generate evaluation data in reasoning tasks, while Liu et al. (2023) dynamically generated out-of-distribution evaluation sets using existing datasets. In comparison, KIEval only requires access to output text of evaluated models and detects data contamination through *evaluating its ability to generalize and utilize knowledge as well as generative ability, which requires a deeper understanding of knowledge instead of mere memorization of the answers*.

3 Methodology

3.1 Overview of the KIEval Framework

KIEval involves a series of iterative interactions, as depicted in Figure 1. KIEval is engineered to dynamically evaluate the conversational abilities of LLMs through interactive dialogues focusing on domain-specific topics that challenge LLMs' generative ability and in-depth generalization of knowledge. It simulates realistic conversation flows, offering a dynamic alternative to the static question-answer format of traditional benchmarks.

KIEval orchestrates an evaluation where an LLM, referred to as the *candidate* (the model under evaluation), must understand and respond to an evolving series of questions. These question prompts are generated by an *interactor* model, designed to challenge the candidate with contextually rich scenarios. The responses from the candidate are then assessed by an *evaluator* model, which scrutinizes the output for factual accuracy, relevance, and coherence. The interactor and evaluator are both strong LLMs (e.g., GPT-4, Gemini, Claude 2, LLaMA2-70B-chat, etc.) as the standard practice of LLM-based evaluation protocols.

The design of KIEval emphasizes the importance of reproducibility and consistency in LLM evaluations. By employing separate models for the interactor and evaluator roles, KIEval ensures that the dialogue context remains consistent across different evaluations, as it is fair for the same conversation to be assessed by various evaluators or the same evaluator with different seeds, facilitating a voting strategy to ensure consistent evaluation results. To achieve reproducibility, KIEval utilizes deterministic outputs from LLMs, such as the latest gpt-4-1106-preview model with temperature sampling disabled and a fixed seed or deploying local models as evaluators. This guarantees identical responses in every run. Due to space limits, we show the complete system prompts in Appendix F.

3.2 Interactive Evaluation Procedure

The interactive evaluation procedure can be described by Algorithm 1. In LLM-based benchmarks, we hypothesize that the evaluator ($\mathcal{M}_E$) models, given their advanced capabilities, can reliably evaluate the performance of less sophisticated candidate models ($\mathcal{M}_C$) (Zheng et al., 2023; Zeng et al., 2023). Nevertheless, their applicability as definitive standards is not without limitations, especially when confronting arduous benchmarks. To counteract this, a methodical sampling and verification strategy is employed to ensure the validity of their evaluative judgments.

This strategy commences by sampling a subset $\mathcal{Q}_S$ from the original benchmark dataset $\mathcal{Q}$, to encompass a wide range of difficulty levels. Both $\mathcal{M}_E$ and $\mathcal{M}_I$ are then independently tested against $\mathcal{Q}_S$. The aim is to discern a 'confident set' of problems that both models can solve with high confidence. The 'confident set' $\mathcal{Q}_V$ is defined as:

$$\mathcal{Q}_V = \{q \in \mathcal{Q}_S | \text{Conf}(\mathcal{M}_E, q) > \theta \wedge \text{Conf}(\mathcal{M}_I, q) > \theta\},$$

Algorithm 1 KIEval Interactive Evaluation Procedure

Require: Benchmark dataset $\mathcal{Q}$, Interactor model $\mathcal{M}_I$, Candidate model $\mathcal{M}_C$, Evaluator model $\mathcal{M}_E$, seed r.
Seed everything with r, disable temperature sampling for $\mathcal{M}_I$, $\mathcal{M}_C$, $\mathcal{M}_E$ to ensure deterministic outputs.
$\mathcal{Q}_S \leftarrow$ Sample subset from $\mathcal{Q}$ with random seed r.
$\mathcal{Q}_V \leftarrow$ Verify, filter samples from $\mathcal{Q}_S$ with $\mathcal{M}_I$, $\mathcal{M}_E$.
for each question $q : (q_{input}, q_{ans})$ in $\mathcal{Q}_V$ **do**
  Initialize interaction history $S \leftarrow \emptyset$ and evaluation history $E \leftarrow \emptyset$.
  $q_{pred} \leftarrow$ Predict with $\mathcal{M}_C$ given question q_{input}.
  $\mathcal{O}_I \leftarrow$ Generate initial question prompt from $\mathcal{M}_I$ using question q and candidate's answer q_{pred}.
  $S \leftarrow S \cup \{\mathcal{O}_I\}$
  while not end of dialogue **do**
    $\mathcal{O}_C \leftarrow$ Generate response from $\mathcal{M}_C$ using S.
    $S \leftarrow S \cup \{\mathcal{O}_C\}$.
    $\mathcal{O}_E \leftarrow$ Evaluate response using $\mathcal{M}_E$ with S, E.
    $E \leftarrow E \cup \{\mathcal{O}_E\}$.
    if Early stopping criteria met for $\mathcal{O}_C$ **then**
      break
    end if
    $\mathcal{O}_I \leftarrow$ Generate next question from $\mathcal{M}_I$ using S.
    $S \leftarrow S \cup \{\mathcal{O}_I\}$
  end while
  Parse and store results from E.
end for
$K \leftarrow$ Calculate KIEval scores with E.
return K

where $\text{Conf}(\mathcal{M}, q)$ calculates the confidence of model $\mathcal{M}$ in providing the correct answer to problem q, and θ represents the confidence threshold.

3.3 Evaluation Metrics

KIEval implements a scoring system to quantitatively grade the performance of candidate LLMs in different aspects. Responses are rated on a definitive scale from 1 to 4 for each aspect, where 1 and 4 denote 'Poor' and 'Strong' performance, respectively, as detailed in Table 1. These scores are intended to be definitive to encourage decisive evaluations and are accompanied by comments for interpretability and insights into each score.

After the last round of interaction, we calculate the KIEval score, which quantitatively measures the results given by the evaluator model, emphasizing sustained and high-quality long conversations. Formally, we propose a decaying weighted scoring mechanism to compute the KIEval score for normalized scores $s_0, s_1, \ldots, s_n$ in n rounds:

$$\text{KIEvalScore} = \frac{\sum_{i=1}^{n} s_i w_i}{\sum_{i=1}^{n} w_i},$$

where the weight for the i-th round is computed as $w_i = \exp(-\frac{i}{n})$. This ensures the scores for early rounds have greater influence, encouraging models to maintain consistent performance through-

Table 1: Evaluation Metrics and Scoring Guide for KIEval. We compute KIEval Score for each metric and a overall KIEval Score as described in 3.3.

Evaluation Metrics		Scoring Guide	
Metric	**Description**	**Score**	**Criteria**
Accuracy	Truthfulness and factual correctness of the candidate's response.	1 Poor	Significant deficiencies or inaccuracies.
Logic	Logical structure and soundness of reasoning, including the support and validity of conclusions.	2 Below Avg.	Noticeable weaknesses, lacking in several areas.
Relevance	The extent to which the response stays on topic and within the scope of the assistant role.	3 Above Avg.	Mostly on target with a few minor shortcomings.
Coherence	Integration into the context, consistency with previous statements and conversational flow.	4 Strong	Strong performance, often surpasses expectations.
Conciseness	Brevity and clarity of the response, avoiding unnecessary elaboration or repetition.		

out the conversation. The normalization ensures a bounded KIEval score, with 1.0 indicating perfect performance across all rounds.

In addition to these metrics, KIEval incorporates an early stopping mechanism within the evaluative process. The evaluator model ($\mathcal{M}_E$) possesses the discretion to prematurely end the conversation if the candidate's response is egregiously inadequate. Criteria for early termination include significant deviations from the topic, empty responses, unpermitted role shifts, and hallucinatory content. We adopt this strategy to measure how well the candidates maintain a meaningful conversation.

4 Experiments

In this section, we conduct experiments designed to rigorously test the KIEval framework. Our objectives are threefold: (1) to evaluate the generative performance and generalizable knowledge of popular large language models on KIEval using existing benchmark datasets; (2) to assess the impact of data contamination on model performance, specifically examining whether such contamination leads to mere memorization or contributes to genuine understanding and generalization; and (3) to determine the alignment with human, reliability, and effectiveness of KIEval.

For setup, we select GPT-4 (OpenAI, 2023) to be both the evaluator and interactor model by feeding it corresponding prompts with a fixed seed to ensure deterministic outputs. The candidate models are engaged in KIEval conversations, starting with selected problems from the aforementioned benchmark datasets. We apply the aforementioned sampling and verification strategy to select 200 samples for each dataset, allowing a maximum of 5 rounds of conversation. The candidates' performance are assessed using the KIEval framework, which evaluates responses based on accuracy, logic, relevance, coherence, and conciseness. In Table 2, we also report dataset-based benchmark accuracies in 5-shot settings and LLM-based benchmark scores from AlpacaEval (Li et al., 2023) and MT-Bench (Zheng et al., 2023) in comparison.

4.1 Evaluation of Popular LLMs by KIEval

In this experiment, we utilized five popular LLM benchmark datasets: ARC-Easy and ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), and C-Eval (Huang et al., 2023). For candidate models, we selected a diverse set of 7 LLMs: including proprietary model GPT-3.5 (Brown et al., 2020) with API access and open-access foundation models: Llama 2 (Touvron et al., 2023b) 7B, 13B, 70B; Mistral-7B (Jiang et al., 2023); Yi-6B-chat (01.AI, 2023); MPT-7B (MosaicML, 2023).[1] Detailed introduction of these datasets and models can be found in Appendix A.

Referencing Table 2, we observe the following trends: GPT-3.5 demonstrated consistently high performance across all datasets, particularly excelling in KIEval scores, which indicates strong contextual understanding and response generation. LLaMA2 70B showed competitive results, achieving only a marginal gap from GPT-3.5 on ARC-E, ARC-C, HSwag and even surpasses GPT in MMLU when measured by dataset accuracies, but we can significantly observe a larger gap between these two models with KIEval metrics in all datasets which is also observed by MT-Bench results as reported in Table 2. This suggests that traditional benchmarks may sketch the difference in performance between LLMs as these benchmarks only let models generate a short span of text to evaluate which focus on testing understanding ability. Thus it is hard for these benchmarks to accurately reflect performance gaps in generative tasks.

From the results of different aspects visualized in Figure 2, we observe that most models we test here exhibit relatively strong performance in terms of relevance and could generate coherent responses. Larger models generally perform better in benchmarks, but it is notable that LLaMA2 70B does

[1] By default, we use the 'chat' versions of Llama2, Yi, and MPT models and the 'Instruct' version of Mistral model.

Table 2: Comparative Evaluation of LLMs using KIEval, AlpacaEval, MT-Bench and human evaluation win-rates. We report AlpacaEval win-rates and MT-Bench scores with GPT-4 as evaluator from the official leaderboards except for missing models; 'Acc.' denotes 5-shot accuracy setting on each dataset or average accuracies in 'Overall'; 'KIEval' and 'Rnds' denote the KIEval score and average rounds of valid conversation rounds.

	ARC-Easy			ARC-Challenge			MMLU			HellaSwag			C-Eval			Overall				
	Acc.	KIEval	Rnds.	Acc.	KIEval	Rnds.	Acc.	KIEval	Rnds.	Acc.	KIEval	Rnds.	Acc.	KIEval	Rnds.	Acc.	AlpacaEval	MT-Bench	KIEval	Human
GPT-3.5	92.7	97.6	4.97	82.3	95.5	4.94	58.2	96.2	4.95	76.6	88.2	4.82	50.8	83.3	4.72	72.1	81.7	8.39	92.1	69.8
LLaMA2 70B	92.3	90.7	4.85	80.4	84.1	4.66	61.8	89.6	4.80	74.4	80.1	4.41	42.0	61.0	3.94	70.2	92.7	6.86	81.1	63.6
LLaMA2 13B	81.9	86.2	4.70	65.7	78.6	4.56	52.1	87.4	4.76	59.3	78.5	4.66	37.8	54.4	3.74	59.4	81.1	6.65	77.0	62.5
LLaMA2 7B	73.6	78.9	4.49	55.7	74.4	4.44	44.5	83.0	4.61	39.8	76.4	4.54	33.4	49.3	3.62	49.4	71.4	6.27	72.4	35.4
Mistral 7B	83.5	80.8	4.64	67.5	78.5	4.46	52.7	83.0	4.62	54.4	70.3	4.34	39.3	52.2	3.61	59.5	65.5	6.84	73.0	58.2
Yi 6B	90.7	83.8	4.58	79.0	76.8	4.33	61.9	86.5	4.58	73.7	68.7	4.20	71.5	55.6	3.66	75.4	54.5	4.86	74.3	46.2
MPT 7B	53.3	68.4	4.34	43.4	65.5	4.33	33.9	74.7	4.46	27.3	57.3	4.10	26.2	44.9	3.52	36.8	43.4	5.42	62.2	24.1

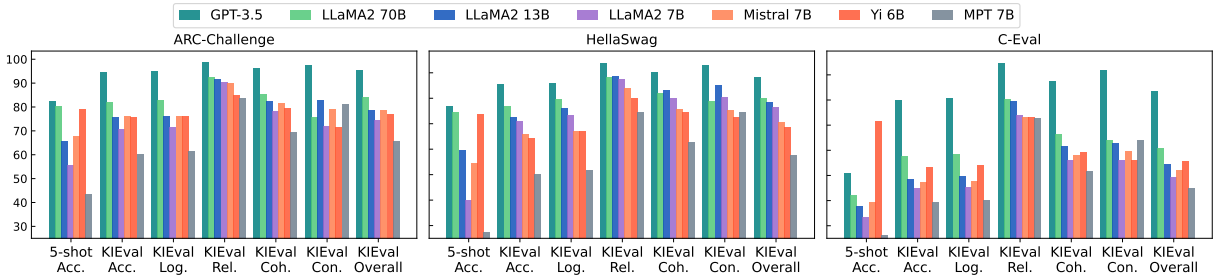

Figure 2: Detailed evaluation result using KIEval, including the overall KIEval score, and KIEval scores for aspects: Accuracy, Logic, Relevance, Coherence and Conciseness. In comparison, we also provide dataset accuracies (5-shot). Due to page limits and the large volume of experimental data, the complete results are put in Appendix E.

not perform well in generating concise responses, compared to its smaller counterparts. Although MPT performs weakly in accuracy, its ability to generate concise responses deserves a closer look at its instruction-tuning data.

One interesting finding is that Yi-6B performs unexpectedly well in all benchmark dataset accuracies, especially with it surpasses GPT-3.5 and all other models by a large margin of over 20% in the C-Eval dataset while exhibiting a similar performance of LLaMA2 70B in other datasets. However, Yi-6B's KIEval score is very similar to LLaMA2 7B and in the range of other 7B models, while it only performs marginally better in the Chinese dataset C-Eval. This raises our concern over potential data contamination in Yi-6B.

4.2 Resilience to Data Contamination

In this subsection, we show that existing static dataset-based and LLM-based evaluation approaches are prone to data contamination while KIEval is resilient to data contamination.

Contamination on static dataset-based evaluation. We train two models on the test sets to introduce contamination in the pre-training ('PT-Cheater') and supervised fine-tuning ('SFT-Cheater') phases using un-tuned LLaMA-2 7B as the backbone. For PT-Cheater, test set contents are integrated into the pre-training set. Subsequently, the model undergoes fine-tuning with the ShareGPT (Eccleston, 2023), a commonly used instruction-tuning dataset, to develop chat functionalities. Conversely, the SFT-Cheater replicates this process but adapts the test data to the SFT format. As a control, we also train the backbone solely with ShareGPT ('Normal'), devoid of contamination, ensuring uniform training conditions across all models. From results in Table 3, it is clear that the accuracies for benchmarks are significantly boosted, by a large margin of over 60%, suggesting a susceptibility to data contamination. However, when faced with KIEval, the cheater models perform slightly worse than 'Normal' model, not positively affected by data contamination. The average rounds of valid conversation is lower in the cheater models, from the reasons specified by Figure 4, contaminated models tend to go off-topic of the conversation, repetitively stick to the incorrect knowledge making the conversation meaningless to continue. We can infer from this result that *training models on test sets does not bring generalizable domain knowledge, instead, only contributing to mere memorization of knowledge from test sets.*

Contamination on LLM-based evaluation. We also find existing LLM-based evaluations vulnerable to data contamination, due to their reliance

Table 3: Comparison on different data contamination scenarios on ARC-C and MMLU datasets, measured with 5-shot accuracy, KIEval score, and average rounds of valid conversation in KIEval.

Dataset	ARC-Challenge			MMLU		
Model	Acc.	KIEval	Rounds	Acc.	KIEval	Rounds
PT-Cheater	86.54	52.13	3.46	72.52	51.82	3.40
SFT-Cheater	77.65	58.46	3.97	61.60	72.74	4.36
Normal	52.35	62.60	4.16	42.69	76.02	4.57

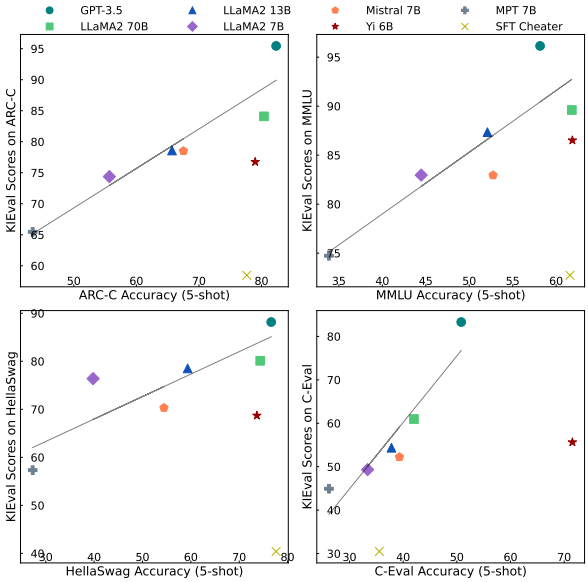

Figure 3: Scatter plots of KIEval scores and traditional benchmark scores by model and dataset. Each point represents the performance of a model on a specific dataset, measured by the KIEval score and accuracy score (5-shot). Regression lines are plotted for each dataset. *Points significantly above the regression line indicate the performance gap not captured by traditional benchmark but captured by KIEval, while points significantly below the regression line indicate potential data contamination in traditional benchmarks.*

on static templates. We train the fine-tuned model ('Normal') with MT-Bench input templates and GPT-4 outputs using only 80 samples and test it against MT-Bench and KIEval. Table 4 reveals that contamination training notably inflates the MT-Bench score by 1.79, a surge over 45% compared to the baseline. This contrasts with the stable ARC-Challenge accuracy and the slight decrease in KIEval scores, reinforcing our conclusion."

Correlation analysis. To further investigate the correlation between dataset-based benchmarks and KIEval, we use regression analysis as shown in Figure 3. We also leverage the Pearson correlation coefficient to provide quantitive analysis in Table 5. The results revealed a significant positive correlation between KIEval scores and dataset-based benchmark accuracies. This correlation underscores KIEval's alignment with traditional evaluation methods. However, we also bring new insights that traditional benchmarks do not offer: *while dataset-based benchmarks effectively assess LLM knowledge under contamination-free conditions, their results are easily inflated in the presence of data contamination. In contrast, KIEval exhibits a lower susceptibility to these issues.* Visual analysis offers additional perspective by contrasting model performances as per benchmark accuracies and KIEval scores. Models significantly above the regression line suggest capabilities beyond those captured by traditional benchmarks. In this scenario, traditional benchmarks are not sufficiently challenging to effectively differentiate the stronger models from others, nor do they accurately represent the generative capabilities of these models. It is evident that GPT-3.5 is included in this category. Conversely, models falling below the regression line, exhibiting high benchmark accuracy but low conversation quality, suggest limited real-world applicability, potentially indicative of data contamination. Interestingly, the visualization shows that not only does our simulated SFT Cheater model fall into the outlier category below the regression line, but Yi-6B also exhibits similar behavior.

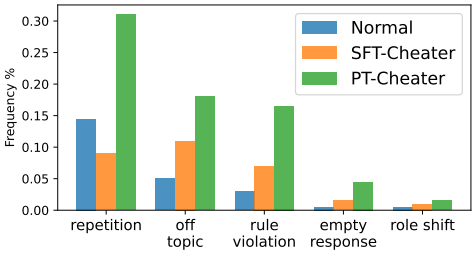

Figure 4: Statistics on reasons to stop conversation given by the evaluator model.

Table 4: Contamination in MT-Bench (Zheng et al., 2023) scores. We report 5-shot accuracy on ARC-Challenge and KIEval results in comparison.

Model	Acc.	MT-Bench	KIEval
Normal	52.35	3.96	62.60
+MT-Bench	52.25	5.75	57.46

4.3 Meta Evaluation of KIEval

Meta evaluation serves as a critical layer of assessment, ensuring that KIEval not only performs theoretically but also aligns practically with broader

Table 5: Pearson correlation coefficient of KIEval scores and dataset accuracy scores. Due to suspected data contamination in Yi-6B, we report two sets of results with and without Yi.

PCC	r	p	r Excl. Yi	p Excl. Yi
Overall	0.664	1.37E-05	0.765	8.67E-07
ARC-E	0.892	6.97E-03	0.934	6.45E-03
ARC-C	0.839	1.83E-02	0.940	5.29E-03
MMLU	0.814	2.57E-02	0.876	2.21E-02
HellaSwag	0.686	8.85E-02	0.862	2.74E-02
C-Eval	0.427	3.40E-01	0.924	8.42E-03

evaluation standards. Building upon the correlation analysis provided earlier, we further validate KIEval's efficacy by analyzing its alignment with human preference. We also provide a cost analysis of our method in terms of compute resources and API usage.

Human evaluation. To validate KIEval's alignment to human preference, we randomly sample 60 unique conversation pairs from 3 datasets, where each pair contains the conversations between two candidate models and interactor models in the same topic. We ensure the two conversations within the same pair with difference in KIEval score. Detailed rules for human annotation are shown in Appendix C. For each dataset, we ask 2 human annotators to independently decide which model performs better throughout the conversation and an additional annotator to resolve the conflicts. This process allowed us to measure the Inter Annotator Agreement (IAA) and compare the agreement rate between human judgments and KIEval scores, thereby validating the human-like evaluative capabilities of KIEval. The high level of agreement, shown in Table 6, between human annotators and KIEval scores reinforced KIEval's validity. The strong Inter Annotator Agreement (IAA) further attested to the robustness of KIEval as a human-like evaluation method.

Cost and scalability. Assessing KIEval's scalability requires a thorough evaluation of overall costs. Our method employs a strong LLM accessed via API, with expenses based on input and output token lengths. Table 14 details the average token count per model evaluation across diverse datasets. Additionally, the average GPU expenditure for single model evaluations on NVIDIA A100 GPUs is provided in Table 13. Financially, deploying GPT-4 in both interactor and evaluator roles within KIEval incurs a cost of around 27 USD for each model evaluation, comprising 1000 interaction rounds. Importantly, due to our adoption of single-answer grading over pairwise comparison (Wang et al., 2023b; Zheng et al., 2023), costs increase linearly rather than quadratically with the number of models evaluated. For a comprehensive understanding of the cost implications at scale, we present a detailed estimation in Table 15.

Table 6: Inter-Annotator Agreement (IAA) measured by Cohen's Kappa, and the agreement rate between human annotators and KIEval results.

	Avg.	ARC-E	ARC-C	C-Eval
κ	0.700	0.699	0.734	0.667
p_o	0.833	0.850	0.817	0.833

5 Limitations

Our method, while insightful, relies on the hypothesis that the LLM evaluator can reliably assess the performance of less sophisticated models, but their applicability as definitive standards is not without limitations, especially when confronting arduous benchmarks or evaluating a stronger model. This limitation is also applicable to any LLM-based evaluation method. To mitigate this, future research could explore a hybrid approach, combining LLM evaluators with other evaluation methodologies or explore leveraging a broader range of language models as evaluator models for a more comprehensive assessment.

6 Conclusion

KIEval provides a dynamic evaluation and analysis of LLMs across various domains, evaluating generative abilities and domain knowledge through structured conversations instead of relying on fixed templates or instructions, reducing the risk of data contamination and enhancing the reliability of evaluations, while preserving alignment with human preference. The primary limitation of static dataset-based benchmarks lies in their reliance on brief text generation, which inadequately captures the full spectrum of LLMs' generative abilities and is susceptible to data contamination. Our study shifts the focus from merely detecting exposure to specific training texts to a more comprehensive evaluation of models' generalizable knowledge and real-world applicability. We believe that KIEval will serve as a valuable tool for researchers and practitioners alike, aiding in the development of more robust, versatile, and ethical AI systems.

References

01.AI. 2023. Yi-6b model by 01-ai. https://01.ai/.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Yoshua Bengio and Yann LeCun. 2007. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. 2023a. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023b. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-instruct: Synthetically generated diverse multi-turn conversations for effecient llm training. *arXiv preprint arXiv:(comming soon)*.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan. 2023. Robustness challenges in model distillation and pruning for natural language understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1758–1770.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Dom Eccleston. 2023. Sharegpt dataset. https://sharegpt.com/.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. 2023. Deep learning tuning playbook. Version 1.0.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT Press.

Google. 2023. Bard.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.

Lynette Hirschman and Robert Gaizauskas. 2001. Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Yachuan Liu, Liang Chen, Jindong Wang, Qiaozhu Mei, and Xing Xie. 2023. Meta semantic template for evaluation of large language models. *arXiv preprint arXiv:2310.01448*.

Carlo A Mallio, Andrea C Sertorio, Caterina Bernetti, and Bruno Beomonte Zobel. 2023. Large language models for structured reporting in radiology: performance of gpt-4, chatgpt-3.5, perplexity and bing. *La radiologia medica*, pages 1–5.

MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

OpenAI. 2023. Gpt-4 technical report.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

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*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Kaiping Peng, Richard E Nisbett, and Nancy YC Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological methods*, 2(4):329.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rylan Schaeffer. 2023. Pretraining on the test set is all you need. *arXiv preprint arXiv:2309.08632*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu. 2022. iEval: Interactive evaluation framework for open-domain empathetic chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 419–431, Edinburgh, UK. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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*.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*.

A Datasets

We use the following datasets in our experiments, for statistics and used splits, please refer to Table 7.

ARC-Easy and ARC-Challenge (Clark et al., 2018): Both are subsets of the AI2 Reasoning Challenge, a benchmark for assessing a model's reasoning and understanding in science questions. ARC-Easy contains simpler questions, while ARC-Challenge includes more complex ones.

HellaSwag (Zellers et al., 2019): challenges models to complete realistic scenarios in text, testing common sense and predictive abilities.

MMLU (Hendrycks et al., 2020): A comprehensive English examination composed of multiple-choice questions encompassing a wide array of disciplines. This extensive test includes subjects ranging from humanities and social sciences to hard sciences, alongside other essential areas of knowledge. It encompasses 57 distinct tasks, covering fields such as elementary mathematics, US history, computer science, law, and beyond.

C-Eval (Huang et al., 2023): A comprehensive Chinese evaluation composed of 13948 multi-choice questions spanning 52 diverse disciplines and four difficulty levels.

B Potential Risks

While KIEval advances the evaluation of Large Language Models (LLMs), it is not without potential risks. Primarily, reliance on strong LLMs as evaluators could inadvertently propagate existing biases or limitations inherent in these models. The computational and financial costs associated with using high-performance LLMs for continuous evaluations could be a barrier for widespread adoption, particularly for researchers with limited resources.

C Use of Human Annotation

For human annotation in our work, all annotators are authors of this paper who previously have not accessed the outputs of models in our experiments and volunteer to contribute. All annotators agree on how the data would be used. Since the data to be annotated come from open-source datasets and popular LLMs, ethical concern is not applicable. We provide guides for each annotator and for each annotator, we give them a unique URL to our annotation platform built with Gradio as shown in 5: 'Everyone is given some conversations between candidate model and interactor model. Each instance to be labeled as a pair of conversations from different LLMs given the same context, and we need to judge which conversation is better overall, considering the conversation's factual accuracy, logical structure, language conciseness and coherence.'

D Use of AI Assistants

In this work, we use GitHub Copilot to assist coding, and GPT-4 to correct grammatical errors.

E Complete Experiment Results

We share the complete experiment results from all 5 datasets with 7 models, evaluated with KIEval and benchmark accuracies in Table 8, 9, 10, 11, 12.

F Complete Prompt

The system prompts for interactor, candidate and evaluator models are given in Figure 6.

Table 7: Details of datasets in our experiments. We report 5-shot accuracy metric of 'Used Splits' split for each dataset.

Datasets	**Splits**	**Used Splits**	**Split Size**	**Language**
ARC-Challenge	train, validation, test	test	1.17k	English
ARC-Easy	train, validation, test	test	2.38k	English
Hellaswag	train, validation,test	validation	10k	English
MMLU	auxiliary_train, test, validation, dev	test	14k	English
C-Eval	val, test, dev	val	1.35k	Chinese

Table 8: KIEval Results on ARC-Easy.

ARC-E	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall	Rounds	Acc. (5-shot)
GPT-3.5	97.1	97.4	99.3	97.9	97.9	97.6	4.97	92.7
LLaMA2 70B	90.3	90.3	94.6	91.3	79.6	90.7	4.85	92.3
LLaMA2 13B	84.5	84.3	93.2	87.7	85.8	86.2	4.70	81.9
LLaMA2 7B	77.1	77.4	89.7	82.2	73.6	78.9	4.49	73.6
Mistral 7B	78.5	78.2	91.4	83.5	79.9	80.8	4.64	83.5
Yi 6B	83.4	83.6	90.9	85.8	76.4	83.8	4.58	90.7
MPT 7B	63.9	64.1	84.9	71.5	81.8	68.4	4.34	53.3

Table 9: KIEval Results on ARC-Challenge.

ARC-C	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall	Rounds	Acc. (5-shot)
GPT-3.5	94.6	94.7	98.5	96.1	97.3	95.5	4.94	82.3
LLaMA2 70B	81.9	82.8	92.2	85.3	75.6	84.1	4.66	80.4
LLaMA2 13B	75.4	75.9	91.3	82.3	82.6	78.6	4.56	65.7
LLaMA2 7B	70.6	71.6	90.4	77.9	71.7	74.4	4.44	55.7
Mistral 7B	75.9	75.8	90.0	81.4	79.1	78.5	4.46	67.5
Yi 6B	75.6	76.1	85.0	79.6	71.2	76.8	4.33	79.0
MPT 7B	60.2	61.4	83.6	69.5	81.1	65.5	4.33	43.4

Table 10: Summary of KIEval Results on MMLU

MMLU	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall	Rounds	Acc(5-shot)
GPT-3.5	95.5	95.8	98.3	96.7	97.4	96.2	4.95	58.2
LLaMA2 70B	89.0	90.3	93.7	90.3	76.0	89.6	4.80	61.8
LLaMA2 13B	85.8	87.0	93.9	88.6	81.4	87.4	4.76	52.1
LLaMA2 7B	82.2	83.6	91.9	84.7	70.4	83.0	4.61	44.5
Mistral 7B	81.6	82.8	90.5	85.3	77.5	83.0	4.62	52.7
Yi 6B	84.7	86.5	91.8	87.4	76.5	86.5	4.58	61.9
MPT 7B	70.6	72.0	86.6	77.9	83.0	74.7	4.46	33.9

Table 11: KIEval Results on HellaSwag.

HellaSwag	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall	Rounds	Acc. (5-shot)
GPT-3.5	85.6	85.6	93.9	90.1	93.1	88.2	4.82	76.6
LLaMA2 70B	76.6	79.5	88.2	82.0	78.9	80.1	4.41	74.4
LLaMA2 13B	72.6	75.9	88.7	83.0	85.2	78.5	4.66	59.3
LLaMA2 7B	70.8	73.3	87.3	79.9	80.2	76.4	4.54	39.8
Mistral 7B	65.6	67.1	83.8	75.6	75.2	70.3	4.34	54.4
Yi 6B	64.4	67.0	79.9	74.3	72.4	68.7	4.20	73.7
MPT 7B	50.0	51.7	74.3	62.5	74.4	57.3	4.10	27.3

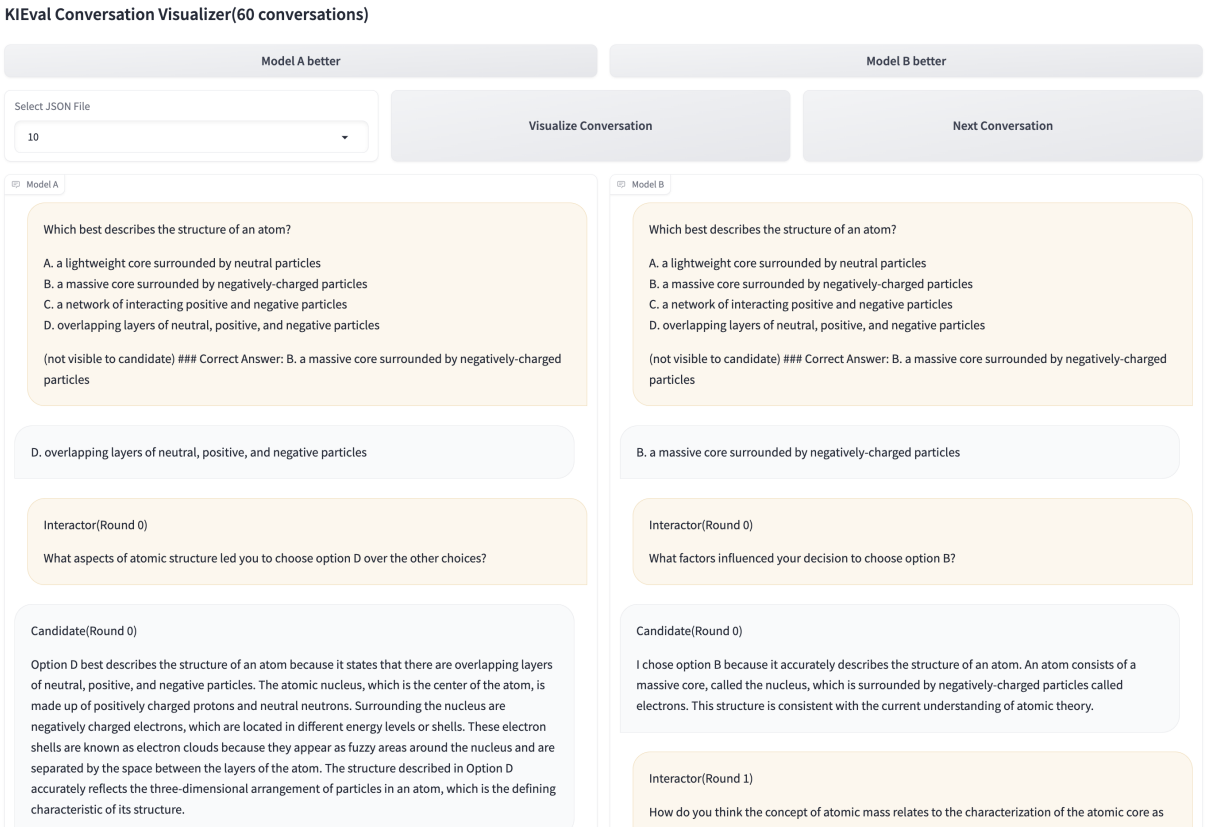

Figure 5: We leverage Gradio to build annotation UI for human annotators. Each annotator is given a unique URL.

Table 12: KIEval Results on C-Eval

C-Eval	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall	Rounds	Acc. (5-shot)
GPT-3.5	79.8	80.6	94.7	87.3	92.0	83.3	4.72	50.8
LLaMA2 70B	57.6	58.3	80.1	66.5	64.1	61.0	3.94	42.0
LLaMA2 13B	48.4	49.8	79.3	61.5	62.9	54.4	3.74	37.8
LLaMA2 7B	44.9	45.1	73.8	55.8	55.9	49.3	3.62	33.4
Mistral 7B	47.3	47.8	73.3	58.0	59.5	52.2	3.61	39.3
Yi 6B	53.1	54.1	73.0	59.3	55.9	55.6	3.66	71.5
MPT 7B	39.5	40.2	72.7	51.5	64.0	44.9	3.52	26.2

Table 13: Average GPU budget for a single model evaluated on one dataset with KIEval. We report results for LLaMA2 models with varying parameter sizes.

	7B	13B	70B
GPU Hours	0.74	0.99	9.38

Table 14: Average number of tokens consumed of evaluation on a single model across various datasets, over 200 samples with 5 rounds of interaction.

	Interactor		Evaluator	
	Prompt	Completion	Prompt	Completion
Avg.	557k	28k	1546k	203k
ARC-E	554k	28k	1592k	208k
ARC-C	540k	27k	1553k	205k
MMLU	656k	30k	1731k	213k
HellaSwag	527k	29k	1488k	198k
C-Eval	505k	26k	1365k	189k

Candidate:
You are a helpful assistant that follows the instructions of the user. You must output in less than 200 words and your responses should be concise and accurate.

Interactor:
As an expert interviewer, initiate a reflective dialogue by asking the user to describe the rationale behind their answer to the multiple-choice question. You must output in less than 200 words. Encourage them to be succinct and precise. Guide the conversation to uncover and address any misconceptions, prompting the user to critically evaluate their thinking process. Their answer is likely to be incorrect. Avoid prompting the user to merely explain their choice, this is very important. Instead, focus on probing questions that challenge their assumptions and lead them to a deeper understanding of the topic. The outputs from the user is very likely to be wrong, do not validate misinformation or incorrect logic; instead, use it as a point of assessment. Remember, you are not educating the user, be neutral and do not give comments since you are the interviewer instead of the evaluator, this is very important. Do not ask too many questions before the user responds. Always skip the greetings, or any other prompting words, ask questions directly. DO NOT say anything else.

Evaluator:
You are an objective evaluator in an interview. Your task is to evaluate a candidate's performance during a series of interactions with an interactor, both parties are AI models. The candidate is asked to solve a problem, and the interactor will ask questions accordingly. The conversation alternates between the interactor (marked with [BEGIN INTERACTOR]) and the candidate (marked with [BEGIN CANDIDATE]). After each candidate response, evaluate the candidate's performance in that round of interaction as well as in context, based on the following aspects independently, rating each on a scale from 1 (Poor) to 4 (Good):
Accuracy: Truthfulness and factual correctness of the candidate's response.
Logic: Logical structure and soundness of reasoning. Whether conclusions are well-supported and arguments are free from logical fallacies.
Relevance: How the response relates to the topic. Ensure responses are within the scope of the "assistant" role, avoiding unpermitted role shifts.
Coherence: How well the response integrates into the context. Consistency with previous statements and overall conversational flow.
Conciseness: Brevity and clarity of the response. Clear, to-the-point communication, free from extraneous elaboration or repetitive words.

Scoring Guide:
1 (Poor): Significant deficiencies or inaccuracies in the aspect.
2 (Below Average): Noticeable weaknesses, partially on target but lacking in several areas.
3 (Above Average): Solid and competent, mostly on target with only a few minor shortcomings.
4 (Good): Strong performance, fully meets and often surpasses expectations.

Evaluation Rules:
1. Evaluate each response consistently and objectively without bias, strictly adhering to scoring guide.
2. Score from 1 to 4 for each aspect independently, using only integers. Low score in one aspect should not influence another aspect. Write a brief comment before scoring in the JSON output structure.
3. Write a overall comment and then give an overall score (same scoring guide). The overall comment should be brief and clear. Consider the performance throughout the interaction, not just in the latest round.
4. Use of `stop_conversation` Flag: Set this flag to true for early conversation termination in cases listed below. If `stop_conversation` is true, choose a `stop_reason` from ["off_topic", "empty_response", "rule_violation", "repetition", "unpermitted_role_shift"], otherwise `stop_reason` is `none`.
5. Format of Evaluation: Output in JSON format strictly following the template:{"accuracy": {"comment": "", "score": 0}, "logic": {"comment": "", "score": 0}, "relevance": {"comment": "", "score": 0}, "coherence": {"comment": "", "score": 0}, "conciseness": {"comment": "", "score": 0}, "overall_comment": "", "overall_score": 0, "stop_conversation": false, "stop_reason": "none"}

Figure 6: The full system prompt for interactor, candidate and evaluator models.

Table 15: API usage estimation for KIEval and pairwise-comparison based evaluation methods. Priced in USD, according to openai's GPT-4 pricing policy.

Method	1 Model	10 Models	100 Models
KIEval	27	279	2,796
Pairwise	16	720	79,200