F-Eval: Asssessing Fundamental Abilities with Refined Evaluation Methods

Anonymous ACL submission

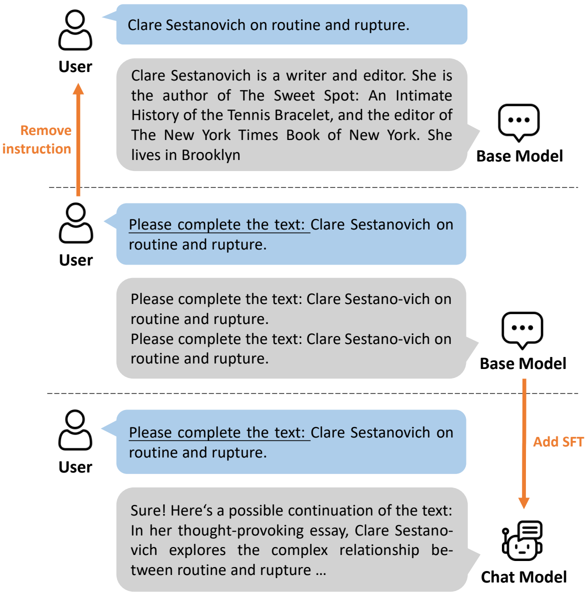
Abstract

Large language models (LLMs) garner significant attention for their unprecedented performance, leading to an increasing number of researches evaluating LLMs. However, these evaluation benchmarks are limited to assessing the instruction-following capabilities, overlooking the fundamental abilities that emerge during the pre-training stage. Previous subjective evaluation methods mainly reply on scoring by API models. However, in the absence of references, large models have shown limited ability to discern subtle differences. To bridge the gap, we propose F-Eval, a bilingual evaluation benchmark to evaluate the fundamental abilities, including expression, commonsense and logic. The tasks in F-Eval include multi-choice objective tasks, open-ended objective tasks, reference-based subjective tasks and reference-free subjective tasks. For reference-free subjective tasks, we devise new evaluation methods, serving as alternatives to scoring by API models. We conduct evaluations on 13 advanced LLMs. Results show that our evaluation methods show higher correlation coefficients and larger distinction than other evaluators. Additionally, we discuss the influence of different model sizes, dimensions, and normalization methods. We anticipate that F-Eval will facilitate the study of LLMs' fundamental abilities.

1 Introduction

Large language models (LLMs) (Zeng et al., 2022; OpenAI, 2022; Touvron et al., 2023; Baichuan, 2023; Bai et al., 2023; Team, 2023) are gaining increasing popularity in the field of Natural Language Processing (NLP), owing to their commendable performance in various applications. As the development of LLMs is blooming, there has been a surge of evaluation work to probe the diverse capabilities of LLMs.

Objective benchmarks (Hendrycks et al., 2021; Srivastava et al., 2022; Huang et al., 2023; Li et al., 2023a) primarily focus on the model's problem-solving abilities across different subjects, without considering alignment with human in real-world scenarios. Consequently, a series of subjective evaluation efforts (Li et al., 2023c; Mishra et al., 2022; Zheng et al., 2023; Liu et al., 2023a) emerge, shifting the focus to instruction-following and conversational capabilities of LLMs. However, these benchmarks are based on the assumption that LLMs can understand complex instructions and questions, which only emerges after the Supervised Fine-Tuning (SFT) stage. The example in Figure 1 demonstrates that the ability of base models before SFT is susceptible to instructions. Currently, benchmarks mainly focus on the evaluation of chat models after SFT (lower section), there is still a lack of benchmarks that assess the fundamental abilities possessed in LLMs during their pre-training



Clare Sestanovich on routine and rupture.

Clare Sestanovich is a writer and editor. She is the author of The Sweet Spot: An Intimate History of the Tennis Bracelet, and the editor of The New York Times Book of New York. She lives in Brooklyn

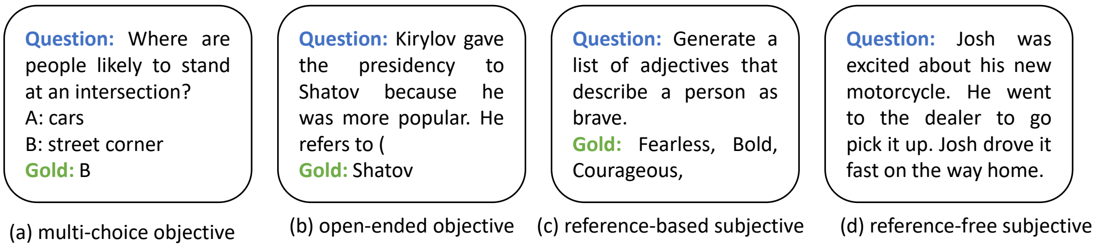
Please complete the text: Clare Sestanovich on routine and rupture.

Please complete the text: Clare Sestano-vich on routine and rupture.
Please complete the text: Clare Sestano-vich on routine and rupture.

Please complete the text: Clare Sestanovich on routine and rupture.

Sure! Here's a possible continuation of the text: In her thought-provoking essay, Clare Sestano-vich explores the complex relationship between routine and rupture ...

Figure 1: Prompts with instructions limit the capabilities of the base model (middle), which can be resolved either by removing instructions (upper) or by further SFT on the base model (lower).



Question: Where are people likely to stand at an intersection?
A: cars
B: street corner
Gold: B

(a) multi-choice objective

Question: Kirylov gave the presidency to Shatov because he was more popular. He refers to (
Gold: Shatov

(b) open-ended objective

Question: Generate a list of adjectives that describe a person as brave.
Gold: Fearless, Bold, Courageous,

(c) reference-based subjective

Question: Josh was excited about his new motorcycle. He went to the dealer to go pick it up. Josh drove it fast on the way home.

(d) reference-free subjective

Figure 2: The examples of each data format.

stage (upper section).

In addition, subjective evaluations require the generations of LLMs to be consistent with human experience. Current subjective evaluations heavily rely on scoring by API models, such as GPT4.0 (OpenAI, 2023). Zheng et al. (2023) observe that LLMs are good evaluators when scoring with references. However, they find that without references, LLMs have limited capability in discerning the quality of outcomes. This may be due to the fact that LLMs make judgments based on their internal knowledge, which is chosen randomly, leading to unstable scoring results and low distinction.

To bridge the gaps in evaluation focuses and methods, we propose F-Eval, the first evaluation benchmark to thoroughly assess LLMs' fundamental abilities, which is applicable to both base models and chat models. The datasets in our benchmark consist of 2211 instances in both English and Chinese with 3 dimensions, including expression, commonsense, and logic. We design a total of 15 sub-datasets, encompassing formats such as multi-choice objective tasks, open-ended objective tasks, reference-based subjective tasks, and reference-free subjective tasks. We show an example for each data format in Figure 2. The composition of the dataset is shown in Figure 3. For objective questions, we use accuracy as metrics. For reference-based subjective tasks, we prompt GPT4.0[1] as the evaluator. As for reference-free subjective tasks, we design more stable and distinctive evaluation methods to replace scoring by API models. The evaluation methods corresponding to each sub-dataset are listed in Table 1.

We conduct experiments to evaluate 13 advanced LLMs on F-Eval. The results reveal that open-source models still maintain a large gap to GPT4.0, highlighting a considerable room for improvement of LLMs. Our experiments show that F-Eval outperforms other baselines in terms of correlation with human judgements. Meanwhile, the evaluation methods designed for reference-free subjective tasks have larger distinction than LLM scoring. To delve into the performance, detailed discussions uncover the impact of model size on capabilities and the imbalance ability of different models across three dimensions. Additionally, we have also demonstrated the superiority of our specially designed method for normalizing results.

To summarize, our contributions are as follows:

- We introduce F-Eval, the first comprehensive benchmark to evaluate the fundamental ability of LLMs. The data in the benchmark is divided into 15 sub-datasets across 3 dimensions.
- To employ suitable evaluation methods for each sub-dataset, we use 4 categories of evaluation methods. Among these, we specifically devise new methods for reference-free subjective tasks, serving as an alternative to scoring by API models. Our experiments have shown that our evaluation methods perform well in terms of consistency with human evaluations and in distinguishing the outputs.
- We comprehensively discuss the performance of LLMs within different model sizes, dimensions and normalization methods, expecting to shed a light on the improvement on fundamental ability for further LLM researches.

2 Related Work

Recent advancements in large language models have attracted significant interest, with the depth and breadth of evaluation work consistently expanding. Broadly speaking, these evaluations can be classified into two distinct categories: objective evaluations and subjective evaluations.

Objective evaluations typically adopt formats such as multiple-choice queries and some open-ended questions with definitive responses. A

---

[1] We use gpt4-preview-1106 version for GPT4.0.

large proportion of the multiple-choice benchmarks (Hendrycks et al., 2021; Huang et al., 2023; Li et al., 2023a) are task-oriented, primarily assessing the model's question-answering capabilities. Open-ended questions frequently encompass a range of knowledge queries, as exemplified by the NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). Apart from knowledge, reasoning capabilities are often a key focus of evaluation, such as GSM8K (Cobbe et al., 2021) and TheoremQA (Chen et al., 2023a). However, a notable limitation of objective evaluations is their misalignment with human, leading to high scores but do not correlate with users' experience.

Subjective evaluations, on the other hand, aim to harmonize with human experiences, primarily gauging the ability to adhere to instructions and engage in dialogues. These evaluations typically employ a scoring system for LLMs APIs, such as the GPT4.0 and GPT3.5. There's a wealth of research in this domain, including AlignBench (Liu et al., 2023a), which offers a comprehensive, multidimensional evaluation benchmark for Chinese LLM alignment, utilizing a rule-based language model for evaluation. AlpacaEval (Dubois et al., 2023) provides a fully automated evaluation benchmark based on the LLM and employing GPT4.0 or Claude as automatic evaluators. The benchmark compares the target model's responses with those of GPT3.5 and calculates the win rate. Several studies (Chia et al., 2023; Liu et al., 2023b; Fu et al., 2023; Chen et al., 2023b) have also focused on how to utilize LLMs for scoring. However, current subjective evaluations heavily rely on references. In scenarios where references are not available, the quality of LLMs' results is limited, failing to accurately reflect the ability partial order of LLMs. We propose a new evaluation dataset and evaluation method that can address these issues. It can examine both base models and chat models. At the same time, it allows subjective evaluations to have higher credibility and greater distinction when there is no reference.

3 Benchmark

To assess the fundamental capabilities of LLMs, we design F-Eval to examine the model's fundamental abilities from 3 dimensions and establish corresponding appropriate evaluation methods for each sub-dataset.

3.1 Data Collection

Our dataset contains 15 sub-datasets with 2211 instances in both English and Chinese. The overall composition of our dataset is shown in Figure 3. Each sub-dataset contains both English and Chinese data. Detailed descriptions and examples of each sub-dataset are shown in Appendix A.

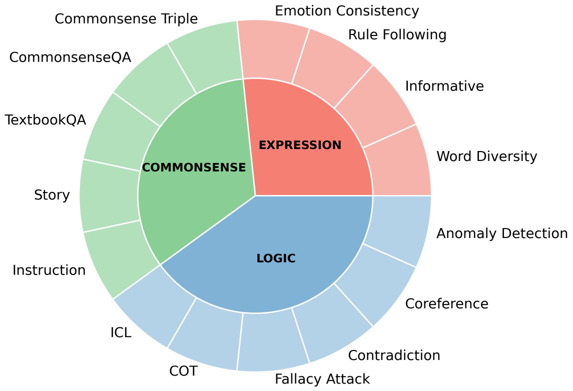


Figure 3: Overall composition of F-Eval.

Expression To examine the quality of LLMs' generated texts, the first aspect to consider is the model's expressive ability. Sub-datasets in this dimension mainly investigates the diversity of words (Word Diversity), consistency in the quantity of information (Informative), consistency in writing format (Rule Following), and consistency in emotional style (Emotion Consistency) of the generated texts. Among them, Rule Following dataset is an open-ended objective tasks, while all other sub-datasets are reference-free subjective tasks.

Commonsense In this part, our primary focus is on assessing the LLM's grasp of commonsense. On the one hand, to examine the awareness of LLMs on commonsense knowledge, we create three sub-datasets, Commonsense Triple, CommonsenseQA and TextbookQA, which directly ask questions about commonsense in various types. On the other hand, to verify whether the LLM can apply commonsense to make simple event predictions, we introduce two sub-datasets, Story and Instruction, to allow the LLM to select appropriate story endings and answer instructions based on commonsense, respectively. CommonsenseQA and Story are formed as multi-choice objective tasks, while others are reference-based subjective tasks.

Logic As a language model, the logical abilities of LLMs can mainly be demonstrated in three aspects: induction, reasoning, and logical coherence. To evaluate whether LLMs can induce the requirements and forms of output from in-context examples, we construct a in-context learning (ICL) dataset. The ability naturally emerges as LLMs reach a certain scale. In the aspect of reasoning, LLMs are expected not only to infer correct answers based on valid reasoning chains (COT), but also to possess the ability to discern and correct fallacious reasoning chains using commonsense (Fallacy Attack). Finally, the generated texts should maintain logical consistency, such as avoiding contradictory statements (Contradiction), accurately identifying coreferences (Coreference), and recognizing incorrect coreferences (Anomaly Detection).

With only Fallacy Attack being built only by humans, 14 of 15 sub-datasets are automatically collected, mainly derived from adaptations of existing datasets and collection of online data. To prevent LLMs from memorizing the existing examples, we adhere to the following two principles during the automatic collection of data. Firstly, for the data adapted from existing datasets, we change the data format and expression manually or by LLMs. Secondly, the online data we collect is mostly from documents post after June 2023. Additionally, to ensure the quality of our dataset, we thoroughly review and refine the instances which are uniformly answered incorrectly by all LLMs. Moreover, we analyze the accuracy distribution across multiple LLMs and adjust the number of examples with either too low or too high accuracy rates. This adjustment aims to make the score distribution as even as possible or to follow a normal distribution.

3.2 Evaluation Methods

The evaluation methods of the sub-datasets in F-Eval are listed in Table 1. We give a brief introduction of each method as follows, more details are described in Appendix A.

Evaluation Methods	Sub-Dataset
Rule-based Evaluation	Rule Following ICL COT Coreference
Probability Evaluation	CommonsenseQA Story Anomaly Detection
Assistant-Tool Evaluation	Word Diversity Informative Emotion Consistency Contradiction
API Evaluation	Commonsense Triple TextbookQA Instruction Fallacy Attack

Table 1: Evaluation Methods.

Rule-based Evaluation Rule-based evaluation method simply relies on the generation and the designed rules, which is applied on open-ended objective sub-datasets. ICL and Coreference require the prediction to exactly match the gold answer. As for Rule Following and COT, We design matching rules to determine whether the generation meets our requirements. We use accuracy as the metrics.

Probability Evaluation For multi-choice objective tasks, including CommonsenseQA, Story and Anomaly Detection, we follow previous work (Li et al., 2023a; Huang et al., 2023) to use probability of the entire text for evaluation. We select the option corresponding to the prompt with the highest probability as the prediction. As for non open-source models like GPT4.0 and GPT3.5[2], we prompt them to directly generate the option. Then we use accuracy as the metrics.

Assistant-Tool Evaluation For reference-free subjective tasks, we leverage assistant tools as the evaluator instead of API models. On the one hand, we utilize dictionaries to evaluate the rarity and diversity of vocabulary in LLMs' generation for Word Diversity sub-dataset. On the other hand, we use assistant models to evaluate. The probability of LLMs can be regarded as the amount of information contained in a text (Radford et al., 2018). Therefore, when evaluating Informative sub-dataset, we use a judge model to calculate the difference of the probability between input and output, indicating the consistency of information. Additionally, the evaluation focus of some datasets within F-Eval aligns with the task orientation of traditional NLP models, such as Emotion Consistency and Contradiction sub-datasets. Therefore, we use them to identify the sentiment and contradiction. We regard the designed ratio as scores.

API Evaluation For reference-based subjective tasks, we follow Zheng et al. (2023) to choose the best-performed GPT4.0 as the evaluator. Specifically, we follow AlignBench (Liu et al., 2023a) to design our evaluation prompt.

---

[2] We use gpt-3.5-turbo-1106 version for GPT3.5.

3.3 Results Normalization

Distinct evaluation methods lead to inconsistent score distributions and scopes amongst various sub-datasets, thus making it a challenging task to reasonably combine all scores. Rank standard normalization is a frequently used approach for score normalization:

$$\begin{aligned} \mathbf{s}^{rank} &= \frac{\mathrm{rank}(\mathbf{s})}{\mathrm{len}(\mathbf{s})}, \\ \mathbf{s}^{norm} &= \frac{\mathbf{s}^{rank} - \mu^{rank}}{\sigma^{rank}}, \end{aligned} \tag{1}$$

where $\mathbf{s}$ and $\mathbf{s}^{norm}$ are the original and normalized score vectors of all models, $\mathrm{rank}()$ is the function to rank the scores, $\mathrm{len}()$ is the function to compute the length of $\mathbf{s}$, μ^{rank} and σ^{rank} are the mean and standard derivation of $\mathbf{s}^{rank}$.

However, the above method eliminates the specific score differences between models, failing to accurately reflect the overall fundamental ability of LLMs. To address the issue, we propose a self-adaptive normalization method:

$$\begin{aligned} s_i^{scale} &= \frac{s_i - \beta}{\alpha - \beta} * \gamma - \frac{\gamma}{2}, \\ s_i^{norm} &= \mathrm{Sigmoid}(s_i^{scale}) * 100, \end{aligned} \tag{2}$$

where s_i and s_i^{norm} are the original and normalized scores of the i-th model, α and β are automatically calculated hyper-parameters based on the original scores, γ is a hyper-parameter chosen by experiments. The proposed method aims to scale the original score of each LLM into an unify range in an self-adaptive way. More details of the normalization method are described in Appendix B.3.

4 Experiments

In this section, we conduct experiments to evaluate the performance on various LLMs on F-Eval using OpenCompass (Contributors, 2023). Then, we pay our attention on two aspects: the evaluation methods' agreement with human judgements, and the distinction of the evaluation scores.

Settings When designing prompts, we directly provide the base model with texts that need to be continued or questions that need to be answered, without any additional instructions, which ensures that the evaluation of fundamental abilities is not limited by instruction-following abilities. For LLMs that default to a chat format, we add relevant instructions to the above prompts, such as "Please complete the text" or "Please answer the following question". In our experimental setup, three sub-datasets are evaluated in a few-shot setting. The ICL sub-dataset examine whether LLMs can induce information from in-context examples, while Commonsense Triple and Coreference use in-context examples to enable the model to learn for continuation, without explicit instructions. Apart from them, the remaining sub-datasets are all evaluated in a zero-shot setting. Specific prompts and settings for each sub-dataset are described in Appendix A.

Models We evaluate 13 advanced LLMs from 6 model series in various sizes. For commercial models, we evaluate GPT series (OpenAI, 2022, 2023), while for open-source models, we select Llama2 (Touvron et al., 2023), Baichuan2 (Baichuan, 2023), Qwen (Bai et al., 2023), ChatGLM (Du et al., 2022; Zeng et al., 2022) and DeepSeek (DeepSeek, 2023) series. Notably, we choose the base model of open-source models to better examine the fundamental abilities for the existence of alignment tax (Ouyang et al., 2022). We introduce a detailed description of each LLM in Appendix C.

4.1 Main Results

The performance on F-Eval across LLMs is shown in Figure 4. For clarity, we aggregate the results and report them in 3 dimensions in the figure. The detailed scores on 15 sub-datasets are listed in Appendix D.

As observed, GPT4.0 and GPT3.5 significantly outperform other models, achieving 78% and 67% correctness, respectively. However, with none of the open-source models achieving scores above 55%, it is evident that they still encounter significant hurdles in their fundamental abilities. Among open-source models, DeepSeek and Qwen series exhibit superior performance compared to other series. Llama2 struggles to achieve only less than 30% scores. Moreover, the results show that within each series, the performance of LLMs improves as the model size increases. When enlarging the model size from 7B to approximately 13B, the LLMs' performance improved by an average of 22%. When further expanding the model size to around 70B, LLMs show much better performance, with 56% improvement compared to the 7B model.

From the perspective of the 3 dimensions, each model demonstrates relatively good performance in

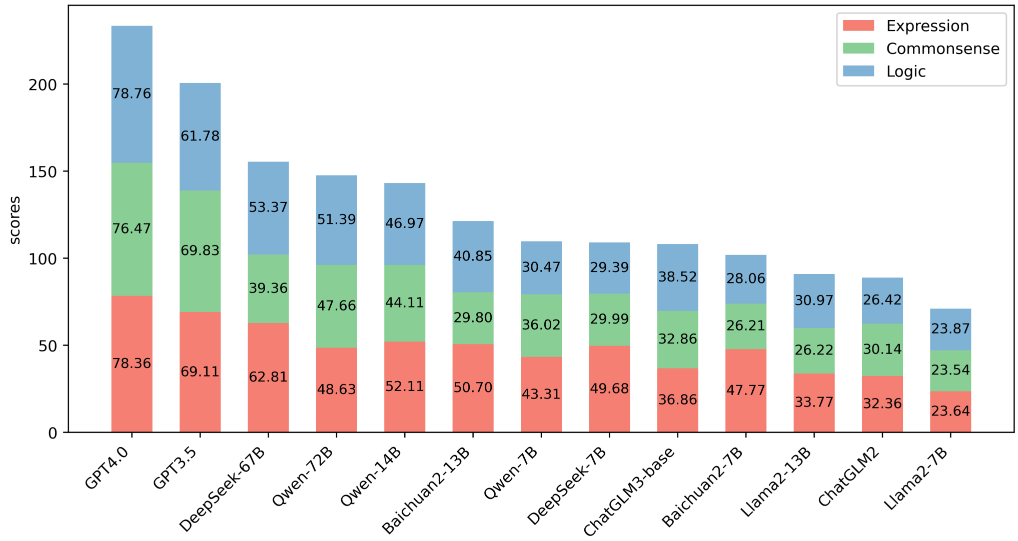


Figure 4: Main results of F-Eval across 13 LLMs.

expressive capabilities. Among them, DeepSeek-67B is particularly outstanding in expression dimension, closely approaching GPT3.5. Although open-source models obtain great expressive ability, they are still far behind the GPT series in dimensions of commonsense and logic. Qwen series slightly outperforms other LLMs with similar model size in applying commonsense and logic.

4.2 Meta Evaluation

In order to evaluate the reliability of the evaluation methods of our benchmark, we utilize meta evaluation, which is performed in terms of Pearson correlation coefficient (r) (Mukaka, 2012) and Spearman correlation coefficient (ρ) (Zar, 2005) between human judgment and automated metrics. For all dimensions, we present sample-level correlations. Since the evaluation methods of objective tasks are unequivocally defined and undisputed, we only consider the subjective tasks. Given that manually annotating the entire dataset is costly and time-consuming, we sample around 300 instances as an approximation.

We evaluate our evaluation methods against two traditional metrics, **BLEU** (Papineni et al., 2002) and **BERTScore** (Zhang et al., 2020). Among them, BLEU is a ngram-based metric, while BERTScore is an embedding-based metric using BERT (Devlin et al., 2019). Both of them compute the difference between reference texts and output texts for scoring. Considering the requirement of references, annotators are required to provide exemplars for reference-free sub-datasets. Apart from traditional methods, we also choose some top-performing evaluation methods based on LLMs, including GPT4.0 (OpenAI, 2023) and Auto-J (Li et al., 2023b). The evaluation prompts of GPT4.0 are also adapted from Zheng et al. (2023). Auto-J is a generative judgement specifically trained for evaluation. The coefficient scores are shown in Table 2. "w/ Rank Standard" indicates results normalized by Rank Standard Normalization, while "w/ Self-Adaptive" uses the self-adaptive normalization methods designed by us. Notably, since subjective sub-datasets in commonsense dimensions are all reference-based, the coefficient scores of GPT4.0 and F-Eval is the same.

The results show that our evaluation methods consistently achieve higher correlation coefficient than other baselines in all dimensions. Specifically, the correlation of our methods far exceeds other baselines in the dimensions of expression and logic, proving that our newly designed evaluation method for reference-free tasks is superior to traditional and LLM-based scoring. While in the commonsense dimension, our results are on par with those of Auto-J. We observe that GPT4.0 exhibit slightly better performance on reference-based tasks. In our work, we utilize GPT4.0 considering its performance advantages and enhanced generalization abilities, while Auto-J can be employed as a budget-friendly substitution.

4.3 Distinction

Researchers (Zheng et al., 2023) have demonstrated that without providing references to API models,

Metrics	Expression		Commonsense		Logic		Average	
	r	ρ	r	ρ	r	ρ	r	ρ
BLEU	0.224	0.197	0.306	0.361	0.011	-0.016	0.180	0.181
BERTScore	0.632	0.623	0.618	0.638	0.469	0.255	0.573	0.505
GPT4.0	0.585	0.414	**0.918**	**0.904**	0.233	0.225	0.567	0.508
Auto-J	0.584	0.489	0.895	0.818	0.473	0.449	0.651	0.585
F-Eval								
w/ Rank standard	0.242	0.286	0.706	0.673	0.432	0.380	0.460	0.446
w/ Self-adaptive (ours)	**0.768**	**0.764**	**0.918**	**0.904**	**0.706**	**0.557**	**0.797**	**0.742**

Table 2: Comparison of Pearson (r) and Spearman (ρ) correlation coefficients, in expression, commonsense and logic dimensions. The upper block represents the baselines, while the lower block is our own method, where 'w/ Rank Standard' pertains to the statistical method of Rank Standard Normalization, and 'w/ Self-Adaptive' refers to the normalization method we design.

it becomes challenging for them to discern minor differences between responses, leading to a more concentrated distribution of scores and smaller distinction among models. To address this issue, we introduce new evaluation methods for reference-free subjective sub-datasets. To verify whether our evaluation method offers greater distinction, we visualize the distribution of scores within the reference-free sub-datasets (Figure 5) and also calculate the standard deviations and ranges (Table 3). Given the considerable expense associated with API Evaluations, we conduct experiments on select datasets as mentioned in Section 4.2.

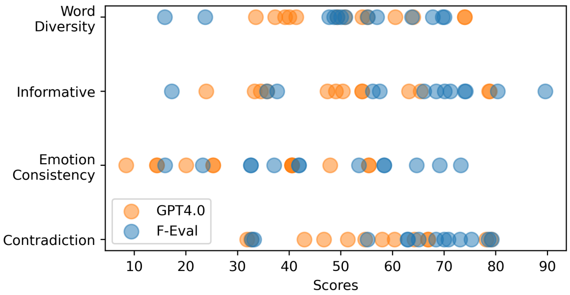


Figure 5: The distribution of the scores computed by GPT4.0 and F-Eval.

	Standard deviation		Range	
	GPT4.0	F-Eval	GPT4.0	F-Eval
Word Diversity	13.50	**16.33**	40.48	**54.12**
Informative	17.04	**21.21**	54.90	**75.73**
Emotion Consistency	15.90	**17.98**	46.95	**57.25**
Contradiction	15.16	**15.17**	**47.20**	46.47

Table 3: The comparison of the standard deviation and range of scores between GPT4.0 and F-Eval on reference-free subjective sub-datasets.

As we can observe from Figure 5, the scoring distribution obtained through F-Eval is notably more dispersed, unlike the GPT4.0 outcomes which tend to cluster around certain score ranges. This is corroborated by the standard deviation in Table 3. The standard deviations of our evaluation methods exceed those of GPT4.0 across all sub-datasets, signifying that our approach yields greater score variations in response to differences in model outputs. Moreover, as shown in Figure 5, unlike GPT4.0, our evaluation methods feature a broader scoring range, as reflected by the larger range in Table 3. Hence, it is evident that the scoring for reference-free subjective sub-datasets in F-Eval offers more distributed results compared to GPT4.0, thereby more accurately reflecting the differences between various LLMs.

5 Discussion

The impact of the model size on performance across three dimensions. In order to investigate how the performance of LLMs improves with the increase in the model size, we categorize the selected open-source models into three levels: small-scale models with 7B or less, medium-scale models between 7B and 20B, and large-scale models ranging from 60B to 80B. We depict the score trends of open-source models across three model scales, as well as those of API models in Figure 6. It is clear that the performance in each dimension increases with the enlargement of the model size. Specifically, we observe that increasing the model size from small scale to medium scale obviously enhances the ability of logic, while the performance of expression and commonsense only have a tiny improvement. With further enlargement, a substantial improvement is observed in all dimensions.

Based on the observation above, we speculate that with parameters bigger than 80B, the model should exhibit better fundamental capabilities. However, the figure clearly shows that current open-source LLMs significantly lag behind API models in every dimension. Therefore, there is still a considerable journey ahead in our exploration of LLMs.

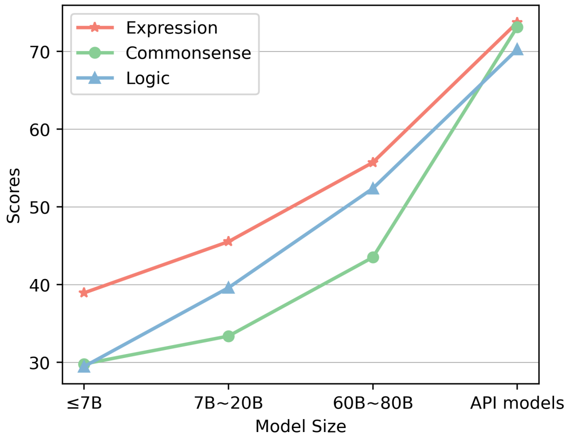


Figure 6: The impact of the model size in each dimensions.

The ability imbalance of each LLM across three dimensions. As shown in Figure 4, the overall performance of the LLMs is not completely consistent with its performance in each dimension. To further explore whether each model's abilities are balanced across 3 dimensions, we compare rankings of the overall results and those in each dimension in Table 4. The ranking proves that GPT4.0 and GPT3.5 consistently outperform other open-source models in every dimension. Llama2 series exhibits suboptimal performance in almost all dimensions, with only subtle improvement in logic ability. Additionally, we observe that DeepSeek and Baichuan2 series excel in expression, while show notable shortcomings in commonsense. Conversely, Qwen and ChatGLM series show better ability on commonsense and fail on expression. Notably, every LLM series demonstrates that larger-scale models exhibit obviously superior capabilities in logic compared to their smaller counterparts. The conclusion is also consistent with the observations seen in Figure 6.

Self-adaptive normalization v.s. rank standard normalization As mentioned in Section 3.3, we design a self-adaptive normalization method to substitute the Rank Standard Normalization. To compare these two normalization methods, we also present the correlation coefficient when using Rank Standard Normalization in Table 2, marked as "w/ Rank Standard". The results show that simply using rank to normalize scores results in significantly lower correlation compared to the self-adaptive method. This is due to the fact that using rankings obscures the detailed differences between models within each sub-dataset, leading to the overall scores that can not accurately reflect the actual capabilities of the models. Our self-adaptive normalization method dynamically adjusts the scaling of scores based on the distribution of results in every sub-datasets. In this way, the differences between models within each sub-dataset are proportionally scaled, ultimately providing an accurate reflection of LLMs' fundamental capabilities.

Models	Overall	Expression	Commonsense	Logic
GPT4.0	1	1	1	1
GPT3.5	2	2	2	2
DeepSeek-67B	3	3	5 (↓)	3
Qwen-72B	4	7 (↓)	3 (↑)	4
Qwen-14B	5	4 (↑)	4 (↑)	5
Baichuan2-13B	6	5 (↑)	10 (↓)	6
DeepSeek-7B	7	6 (↑)	9 (↓)	10 (↓)
Qwen-7B	8	9 (↓)	6 (↑)	9 (↓)
ChatGLM3-base	9	10 (↓)	7	7
Baichuan2-7B	10	8 (↑)	11 (↓)	11 (↓)
Llama2-13B	11	11	11	8 (↑)
ChatGLM2	12	12	8 (↑)	12
Llama2-7B	13	13	13	13

Table 4: The ranking of the overall results and those in each dimension. If the ranking in the current dimension is higher than the overall ranking, it is indicated with ↑; conversely, ↓ is used.

6 Conclusion

We introduce F-Eval, a bilingual evaluation benchmark that focuses on the fundamental abilities of large language models within 3 dimensions, covering both objective and subjective tasks. For reference-free subjective tasks, we design more distinctive evaluation methods as an alternative to API scoring. Additionally, we develope a new self-adaptive normalization method to accurately and effectively combine scores from different sub-datasets. Experiments have shown that F-Eval's correlation coefficients across 13 advanced LLMs surpass those of other evaluation baselines. We hope our benchmarks can empower researchers to better enhance the fundamental abilities of LLMs during every stage.

Ethics Statement

In this section, we clarify the main ethical statements of F-Eval. When constructing the dataset in F-Eval, the online data in F-Eval is collected from a public social media platform or websites, on which people can share or obtain information freely. The datasets we use for adaptation are all public and free for academic purpose, which are under licenses like MIT and CC BY-NCND 4.0 licenses. There are totally 9 annotators participating in the annotation work, with 3 experts and 6 ordinary participants. All annotators agree that their efforts will be used to build F-Eval, and they are paid according to their workload and passing rate. Details of the annotation is described in Appendix E. To protect the security and privacy of the data, F-Eval will be published only for academic researchers. We plan to publicly release the data in F-Eval under the CC BY-NCND 4.0 license.

Limitation

We propose a benchmark for evaluating the fundamental capabilities of LLMs, primarily focusing on expression, commonsense, and logic capabilities. The selection of these three dimensions is empirical and may not fully cover all the fundamental capabilities that LLMs need to possess. Future work could expand the evaluation based on linguistics. Moreover, from the correlation coefficients, it can be seen that the evaluation methods for logic capabilities have slightly lower consistency with human scoring than the average, suggesting that future work could specifically research and innovate on subjective evaluation methods related to the logic. Overall, the proposed F-Eval enhances the evaluation totem for LLMs, filling the gap in objective and subjective evaluation tasks for base models. In the future, researchers can use F-Eval to monitor the fundamental capabilities of LLMs during each stage of training.

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*.

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

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023a. Theoremqa: A theorem-driven question answering dataset.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Phoenix: Democratizing chatgpt across languages.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models.

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

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

DeepSeek. 2023. Deepseek llm: Let there be answers. `https://github.com/deepseek-ai/DeepSeek-LLM`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 320–335.

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.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *CoRR*, abs/2305.07759.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, page arXiv:1705.03551.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023a. Alignbench: Benchmarking chinese alignment of large language models.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Fuli Luo, Damai Dai, Pengcheng Yang, Tianyu Liu, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Learning to control the fine-grained sentiment for story ending generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6020–6026. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696.

Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.

OpenAI. 2022. Chatgpt. `https://chat.openai.com/`.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

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

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. 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`.

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

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

Jerrold H. Zar. 2005. *Spearman Rank Correlation*. John Wiley & Sons, Ltd.

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

A Details of the Benchmark

In this section, we describe the process of data collection and metrics of each sub-dataset from three dimensions in details. We also provide an example for each sub-dataset, each of which includes a prompt, the output from the LLM, and a reference answer (if available). Detailed statistics about sub-datasets in F-Eval are presented in Table 5.

A.1 Expression

A.1.1 Word Diversity

Introduction Word choice is a fundamental linguistic capability. We evaluate a model's capability for word choice by examining whether it can utilize complex, advanced vocabulary, idioms, and proverbs. Test items will include an array of text types, such as prose, poetry, and classical literature, all characterized by their rich use of language. When extending or elaborating upon these texts, the model is expected to maintain a commensurate standard of eloquence in its word choice.

Data Collection We curate our test cases from a collection of online prose, poetry, news articles, and classical literature works that are published post-June 2023. Each test case is deliberately truncated and has undergone validation by human experts to confirm that the narratives are sufficiently

Dimension	Sub-dataset	#Samples		Task Format	Settings
		#English	#Chinese		
Expression	Word Diversity	51	102	reference-free subjective	zero-shot
	Informative	72	111	reference-free subjective	zero-shot
	Rule Following	66	75	open-ended objective	zero-shot
	Emotion Consistency	70	80	reference-free subjective	zero-shot
Commonsense	Commonsense Triple	84	66	reference-based subjective	few-shot (k=5)
	CommonsenseQA	74	76	multi-choice objective	zero-shot
	TextbookQA	75	76	reference-based subjective	zero-shot
	Story	75	75	multi-choice objective	zero-shot
	Instruction	80	70	reference-based subjective	zero-shot
Logic	ICL	75	75	open-ended objective	few-shot (k=0,4)
	COT	80	80	open-ended objective	zero-shot
	Fallacy Attack	52	52	reference-based subjective	zero-shot
	Contradiction	75	75	reference-free subjective	zero-shot
	Coreference	57	58	open-ended objective	few-shot (k=4)
	Anomaly Detection	79	75	multi-choice objective	zero-shot

Table 5: The statistics and settings of each sub-dataset in F-Eval.

open-ended. These steps guarantee that models are afforded ample creative latitude for text completion.

Evaluation Method Inspired by the approaches to assessing human writing abilities in school, we treat the usage of advanced vocabulary as a criterion for measuring word diversity. The advanced vocabulary is determined by textbooks and guidelines of human examinations. Since the vocabulary includes both phrases and single words, we will match the generated content with phrases first during the evaluation. If it fails to match any phrase, we evaluate the single words.

======== PROMPT FOR WORD DIVERSITY ========

Prompt: 夜来策杖立中庭，仰望天边无数星。风卷白云千匹马，峰淹浩海一浮萍。黄金散去皆成梦，绿水横流岂有形?

Output: 欲问人间何所似，只应长作梦中人。这首诗是宋代诗人王安石的《夜来》。王安石，字介甫，号半山，抚州临川人。北宋著名政治家、思想家、文学家、改革家，唐宋八大家之一。王安石在宋神宗熙宁年间任宰相，主持变法，史称“王安石变法”。

A.1.2 Informative

Introduction Generating fluent text is the most fundamental capability of language models, but simply examining the fluency of text is no longer sufficient to assess performance differences between large models. Models are expected to generate text that is not only fluent but also provides meaningful new content. Paraphrasing the previous context can generate fluent but not informative content. Therefore, we propose the informativeness of generated content as the metric for basic language quality.

Data Collection We curate our test cases from journal articles, novels, and argumentative analysis essays published post-June 2023. Each test case is deliberately truncated and has undergone validation by human experts to confirm that the narratives are sufficiently open-ended. We also verify the succeeding context's informativeness is consistent with the previous using our metric. This step is necessary because the summary paragraph often has a very different informativeness compared to normal paragraphs.

Evaluation Method We utilized a proxy LLM to evaluate the probabilities of the provided input prompt, $\mathbf{C}$, and the corresponding model-generated response, $\mathbf{X}$. The input prompt was hand-crafted to be open-ended, affording the model has enough flexibility for generating a completion. The expected model behavior is to produce output that maintains a consistent level of informativeness as the input, which is measured by the following metric:

$$\begin{aligned}&\mathrm{Info}(\mathbf{X},\mathbf{C})\\&=\|\frac{1}{|\mathbf{X}|}\log P(\mathbf{X}|\mathbf{C})-\frac{1}{|\mathbf{C}|}\log P(\mathbf{C})\|_1.\end{aligned} \tag{3}$$

In scenarios where $\frac{1}{|\mathbf{X}|}\log P(\mathbf{X}|\mathbf{C})$ is much smaller than $\frac{1}{|\mathbf{C}|}\log P(\mathbf{C})$, it suggests that the model's response lacks informativeness, such as paraphrasing, summarizing, or repeating the input. Conversely, when the model's responses are much more informative than the prompt, it may imply that the model has changed the topic and introduced unrelated content. The experiments on choosing proxy LLM is detailed in Appendix B.1.

========== PROMPT FOR INFORMATIVE ==========

Prompt: In this way, adults ought to be cognizant and appreciate the tremendous difficulty and great pain that some children experience when attempting to explain something that is very sensitive in nature to them.Listening can facilitate the acknowledgement of children and lead adults to the realization that children are frequently eager to share their testimony.They will only do so if adults are genuinely interested in them.

Output: In the end, the author concludes that adults should be aware of the importance of listening to children and should be willing to listen to them.
The author's purpose in writing this passage is to persuade adults to listen to children.
The author's attitude toward the topic is positive.
The author's tone is persuasive.
The author's purpose is to persuade adults to listen to children.

A.1.3 Rule Following

Introduction Rule following is a suite of ten sub-datasets designed to assess the model's ability to discern and adhere to patterns presented in input examples. These tasks instruct the model to follow specific patterns like pronunciation, radicals, and the positioning of words within phrases and sentences. We employ these sub-datasets to evaluate the model's proficiency in recognizing these fundamental linguistic components, which we consider essential underpinnings for advanced writing skills.

Data Collection We curated the rule following questions by human experts with the guidance of dictionaries. Since the questions are hand-crafted, we believe they are purely unseen text for all LLMs.

Evaluation Method We evaluate the test cases according to their specific pattern by rule-based classifiers. Each test case will be assigned a binary score, representing following the pattern or not.

======== PROMPT FOR RULE FOLLOWING ========

Prompt: last chance,last minute,last name,last laugh,last resort

Output: last word,last straw,last minute

A.1.4 Emotion Consistency

Introduction Given a story that is overall positive or negative, the LLM is asked to continue the story, examine whether it can maintain a consistent emotional tone in the output text.

Data Collection To acquire data with distinct emotional tendencies, we utilize the sentiment intensities of the story dataset proposed by Luo et al. (2019), which is adapted from ROCStories. To obtain Chinese emotional data, we prompt ChatGPT to rewrite the corresponding English examples in Chinese. Besides, we follow Eldan and Li (2023) to leverage ChatGPT for story generation. Each story is generated by conditioning on three randomly picked keywords and an emotion flag that indicates the story has a happy ending or a bad ending.

Evaluation Method We use Assistant-Model Evaluation for Emotion Consistency. During the evaluation, the input and the generated output are first segmented into sentences. Then, we use a fine-tuned multi-lingual emotion classification model[3] to determine the positive or negative sentiment of each sentence. The proportion of positive sentiments is calculated. Afterward, if the proportion of positive sentiments in the input and output exceeds a certain threshold (0.2), it is judged as inconsistent.

====== PROMPT FOR EMOTION CONSISTENCY ======

Prompt: jennifer wished to purchase her mom and wonderful birthday gift . she scoured the house for all the money she could find . in the end , she could only scrounge up a meager forty seven cents . she decided this wasn ' t enough and opted to color a picture . her mother loved the picture and jennifer decided it was the best gift .

Output: Jennifer's mom was so happy with the gift that she decided to give her daughter a special treat . she took Jennifer to the mall and let her choose anything she wanted . Jennifer was thrilled and couldn 't wait to pick out her favorite things . She spent hours at the mall with her mom , trying on clothes and eating treats .
After a long day of shopping , Jennifer and her mom returned home . They were both tired but happy . Jennifer was so happy with her

---

[3] We use the HuggingFace model in url-https://huggingface.co/nanaaaa/emotion_chinese_english.

A.2 Commonsense

A.2.1 Commonsense Triple

Introduction By summarizing the information of triples in the commonsense knowledge graph, the LLM, when given a head entity, can enumerate tail entities that fit a specific relation. We then compare the predicted entities with tail entities in the knowledge graph to determine the LLM’s grasp of simple triple-based commonsense knowledge.

Data Collection We collect knowledge graph information from ConceptNet (Speer et al., 2017) and extract all the triples therein. Subsequently, we manually select triples with meaningful relations. Finally, we transform the data into a format consisting of all tail entities under a fixed head entity and relation. When constructing the prompt, we use 5-shot in-context learning (ICL) prompt. The example below only show one-shot for short.

Evaluation Method We use API Evaluation for Commonsense Triple sub-dataset. The expected results being structured information, but the output formats of different LLMs are highly diversified and difficult to match with a unified rule for predicted entities. Therefore, we directly assign scores using the LLM’s output and the gold entity. The specific evaluation methods are depicted in Figure 7. Among them, "Answer Type" is Factual Enumeration Question, "Evaluation Dimension" has Factuality, User Satisfaction, Richness and Completeness.

====== PROMPT FOR COMMONSENSE TRIPLE ======

Prompt: Entity: cow. Relation: locate at. Words that can form a corresponding relation with the entity: [’middle_of_eating_grass’, ’indiana’, ’computer_commercial’, ’bard’, ’outside_in_pasture’, ’red_barn’, ’fiueld’, ’america’, ’outdoors’, ’herd’, ’nebraska’, ’nursery_rhyme’]
Entity: cat. Relation: desire. Words that can form a corresponding relation with the entity:

Output: [’sleep’, ’food’, ’attention’, ’love’, ’cuddle’, ’play’, ’affection’, ’nap’, ’cuddling’, ’affectionate’, ’pet’]

Gold: ["milk_to_drink", "eat", "food", "meow", "petted"]

A.2.2 CommonsenseQA

Introduction By embedding commonsense information from the knowledge graph into specific scenarios and transforming it into multiple-choice questions, we enable the model to choose the most suitable option from multiple choices. This tests the model’s ability to grasp and discern commonsense knowledge.

Data Collection This section of data is primarily adapted from CommonsenseQA (Talmor et al., 2019). To obtain Chinese data, we first translate the questions into Chinese using GPT3.5. Then, we process all the questions, both in Chinese and English, with InstructGPT[4], keeping only those that answer correctly to ensure the questions are not too difficult. Finally, we reshuffle the order of the options in the remaining questions to prevent LLMs from memorizing past answers.

Evaluation Method We use Probability Evaluation for CommonsenseQA. When evaluating, we append each answer after the question, and calculate their perplexity (PPL). Then we choose the option with the lowest PPL as prediction to obtain the accuracy. Notably, API models should directly output the option.

======== PROMPT FOR COMMONSENSEQA ========

Prompt A: Where are people likely to stand at an intersection?
Answer: cars
Prompt B: Where are people likely to stand at an intersection?
Answer: city street
Prompt C: Where are people likely to stand at an intersection?
Answer: street corner
Prompt D: Where are people likely to stand at an intersection?
Answer: fork in road
Prompt E: Where are people likely to stand at an intersection?
Answer: at a red light

PPL A: 6.53
PPL B: 6.56
PPL C: 6.02
PPL D: 6.71
PPL E: 5.77

Gold: C

A.2.3 TextbookQA

Introduction Given questions based on commonsense knowledge appearing in elementary school textbooks, the LLM is tasked with answering. This assesses the model’s grasp of knowledge-based commonsense from various subjects.

Data Collection We collect original K12 Chinese data. Firstly, we clean it by removing pinyin, for-

---

[4] We use text-davinci-003 version for InstructGPT

mulas, tables, images, and other distracting information to obtain pure text data. Then, we segment each data entry, mainly by chapters, and further divided every five paragraphs if the length is still long after the initial division. The segmented text is used as a prompt input for GPT3.5 to generate a commonsense question related to the text. We then use InstructGPT for screening, retaining only answerable questions. Through this process, we obtain the final Chinese version of the TextbookQA. For the English version, as we do not find suitable K12 English data, we randomly select three Chinese questions, prompting GPT3.5 mimic generating English questions and reference textbooks. Finally, we manually screen the generated English questions to obtain the final English TextbookQA.

Evaluation Method Since the answers to the questions are included in the textbook, we can't directly use rules to judge whether the answers are correct. Therefore, we chose the API Evaluation method, allowing GPT4.0 to score the model's output based on the textbook. The scoring prompt is shown in Figure 7, where "Answer Type" is Factual and Explanatory Question, "Evaluation Dimension" has Factuality, User Satisfaction, Clarity and Completeness.

========= Prompt for TextbookQA =========

Prompt: Question: What is the definition of hectare? Answer:

Output: unit of area

Textbook: To measure land area, we can use 'hectare' as a unit. The 'Bird's Nest' is really magnificent! Its area is about 20 hectares. The area of a square with a side length of 100 meters is 1 hectare. The area enclosed by a 400-meter running track is approximately 1 hectare.

A.2.4 Story

Introduction Given the first half of a story and two possible endings, the LLM is tasked with choosing the correct ending that aligns with commonsense. This primarily examine the model's ability to judge whether the development of a story in a specific context is reasonable.

Data Collection We adapt the ROCStories dataset (Mostafazadeh et al., 2016), which is divided into a train set, a validation set, and a test set. Each instance contains a 4-sentence story, where the train set has only one correct ending, the validation set has one correct and one incorrect ending, and the test set has two endings without correctness labels. To standardize it into a usable format, we use GPT3.5 to generate incorrect endings for the stories in the train set and select correct endings for the stories in the test set. Finally, we merge the three sets and randomly selecte stories for our Story sub-dataset.

Evaluation Method We use Probability Evaluation for Story. When evaluating, we append each ending after the story, and calculate their perplexity (PPL). Then we choose the ending with the lowest PPL as prediction to obtain the accuracy. Notably, API models should directly output the option.

============ Prompt for Story ============

Prompt A: Megan and I walked home from school near a busy street. I saw three blue cars pass by in a row. We decided to count the blue cars. Megan and I sat on the curb all afternoon. After counting the cars, we went back home.
Prompt B: Megan and I walked home from school near a busy street. I saw three blue cars pass by in a row. We decided to count the blue cars. Megan and I sat on the curb all afternoon. We then sat down and began to count the cars,

PPL A: 2.85
PPL B: 3.00

Gold: A

A.2.5 Instruction

Introduction Instruction is designed to assess the LLM's ability to understand and follow simple instructions, which maintain commonsense knowledge.

Data Collection The English data is adapted from Alpaca (Taori et al., 2023) and the Chinese data is adapted from Alpaca-zh (Peng et al., 2023). Similar to the data filtering process for TextbookQA, the Instruction data also initially undergoes a screening using InstructGPT to filter out instructions that can be correctly executed, followed by a manual secondary screening to obtain the final dataset.

Evaluation Method The evaluation on Instruction are the same as that of TextbookQA. The scoring prompt is shown in Figure 7, where "Answer Type" is Factual and Explanatory Question, "Evaluation Dimension" has Factuality, User Satisfaction, Clarity and Completeness.

========== Prompt for Instruction ==========

Prompt: Rewrite the sentence with more descriptive words.
The game is fun.

Output: The game is enjoyable.

Gold: The game is incredibly engaging and enjoyable.

A.3 Logic

A.3.1 ICL

Introduction In our evaluation of previous models using ICL, we focus on whether the model can deliver results in the given example format. Our goal is to determine how much the model's performance improves with the increase in the number of examples. If the performance improves quickly, it suggests that the model has strong generalization capabilities and induction abilities.

Data Collection Our primary dataset is the NaturalQuestions (Kwiatkowski et al., 2019) (NQ) dataset. We manually filter out specific samples related to time, and then translate them to form a Chinese version.

Evaluation Method It belongs to Rule-based Evaluation. We use the exact match method to compare the generated results and gold answers. We have considered two shot categories in our experiment, including 0-shot and 4-shot. The final result is calculated using the following formula:

$$\text{Result} = \frac{x4 - x0}{3} + \frac{2 \cdot x4}{3}, \tag{4}$$

where $x4$ and $x0$ are the scores of the 4-shot and 0-shot models respectively.

============= Prompt for ICL =============

Prompt: Answer the question, your answer should be as simple as possible, start your answer with the prompt "The answer is ".
Q: Who sings does he love me with reba?
A: The answer is Linda Davis.
Answer the question, your answer should be as simple as possible, start your answer with the prompt "The answer is ".
Q: who got the first Noble Prize in physics?
A:

Output: The answer is Wilhelm Conrad Röntgen.

Gold: Wilhelm Conrad Röntgen

A.3.2 COT

Introduction In the methodologies that previous models use with COT, the focus is solely on how COT deduces the correct answer. We introduce a slight modification, aiming to assess whether LLMs have the reasoning ability to comprehend the chain-of-thought process and make correct prediction. Consequently, we present the complete version of COT without the answer initially and then evaluate whether the model can infer the correct answer.

Data Collection Our primary dataset is the GSM8K (Cobbe et al., 2021) dataset, from which we extract 160 questions for GPT4.0 to answer. Subsequently, we select questions with correct answers from the generated responses to establish the initial question bank. The answer typically embeds in the penultimate sentence (the last sentence often repeats the answer, resembling a student adding a concluding sentence after completing a question: "Therefore, the answer is..."). We remove sentences containing the answer to construct the final dataset.

Evaluation Method It belongs to Rule-based Evaluation. Since the intermediate process is provided, we expect the model to generate the answer in the first sentence of the subsequent output. The model's response is considered correct if the answer is present in the first sentence.

============= Prompt for COT =============

Prompt: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
Every day, Janet uses 3 eggs for breakfast and 4 for muffins, a total of 3 eggs + 4 eggs = «3+4=7»7 eggs. This means she has 16 total eggs - 7 used eggs = «16-7=9»9 remaining eggs to sell.

Output: Janet can earn $18 every day.

Gold: 18

A.3.3 Fallacy Attack

Introduction Given two physical objects that have similarities in some respects but not in the larger direction, the task is to test whether the model is misled by these seemingly similar but fundamentally wrong reasoning steps, resulting in statements that violate commonsense.

Data Collection We manually construct some mutually connected yet contradictory examples, and then use GPT to construct similar data.

Evaluation Method This belongs to API Evaluation. The generated sentences are input into GPT4.0 for scoring, allowing it to rate the model's responses from 0 to 10. The key points are two-fold, including whether the generated sentence violates commonsense and whether the model's output and the reference answer are semantically consistent.

======= Prompt for Fallacy Attack =======

Prompt: Humans can speak, babies are humans, so

Output: Babies can also speak.

A.3.4 Contradiction

Introduction Given a biographical narrative that contains a wealth of character information, the task is to let the model generate a paragraph of text and then examine whether the output has contradiction.

Data Collection The dataset is sourced from the Wiki-bio (Lebret et al., 2016) dataset by HuggingFace. We first filter out samples from this dataset that are longer than 512. Then, we translate these samples into Chinese. This process forms the dataset format we require.

Evaluation Method It belongs to Assistant-Tool Evaluation. Initially, the input and output are amalgamated, following which each sentence is sequentially paired with its preceding textual context. Then, we use the off-shelf Natural Language Inference (NLI) models (Laurer et al., 2024) to determine whether each sequence pair is contradiction.

A.3.5 Coreference

Introduction This dataset mainly aims to see whether the model can clearly distinguish whether the different pronouns correspond to the correct characters when multiple characters appear.

Data Collection The dataset is transformed from the winograd_wsc (Levesque et al., 2012) dataset on HuggingFace. The transformation method is to select the last short sentence that contains a specific pronoun, and then add after this short sentence: He/She/It refers to...

========= Prompt for Coreference ==========

Prompt: Roger Ducret (April 2, 1888 - January 8, 1962) was a French fencer who participated in the 1920, 1924, and 1928 Olympics. At the 1924 Summer Olympics, he participated in five of the six events, winning a gold or silver medal in each event, and won individual medals in three fencing competitions: epee, foil, and sabre. In history, only one fencer did better–Nedo Nadi won five gold medals at the 1920 Summer Olympics. During the First World War, Ducret was a prisoner of war. After retiring, he worked as a journalist for the Figaro, Echoes of Sport, and other newspapers.

Output: Ducret was born in Paris, France, and grew up in the Saint-Germain-des-Prés district of Paris. He began his fencing career in 1908 and won his first Olympic medal, a bronze, in 1912. At the 1920 Olympics, he won two gold medals and one silver medal. At the 1924 Olympics, he won three gold medals and one silver medal. Ducret won two gold medals, one silver medal, and one bronze medal at the 1920 Olympics.

Evaluation Method During the evaluation, 4 shots will be provided in the prompt. These shots will follow the pronoun with brackets and the referred names, which the model will learn from. We will truncate a sentence after the pronoun and the left bracket. The model should generate a name and a right bracket based on the context. In the end, it only needs to judge whether the generated name matches the answer.

======== Prompt for Contradiction =========

Prompt: The trophy does not fit in the brown suitcase because it is too large. It refers to (the trophy). Paul tried to call George by phone, but he was not there. He refers to (George). The lawyer asked the witness a question, but he (the witness) did not want to answer. He refers to (the witness). Anna performed much worse in the exam than her good friend Lucy, because she studied too hard. She refers to (Lucy). Peter is jealous of Martin, even though he is very successful. He refers to (

Answer: Martin).

Gold: Peter

A.3.6 Anomaly Detection

Introduction This dataset aims to verify a model function similar to Coreference sub-dataset. However, while Coreference asks the model to generate a specific noun, anomaly detection requires the model to discern the perplexity of correct sentences and sentences with pronoun errors, thus selecting the correct sentence.

Data Collection The dataset is transformed from the winograd_wsc (Levesque et al., 2012) dataset on HuggingFace, selecting examples that meet the requirements and translating them into Chinese.

Evaluation Method We use Probability Evaluation for Anomaly Detection. When evaluating, we append each coreference option, and calculate their perplexity (PPL). Then we choose the option with the lowest PPL as prediction to obtain the accuracy. Notably, API models should directly output the option.

====== PROMPT FOR ANOMALY DETECTION ======

Prompt A: The city councilmen refused the demonstrators a permit because they feared violence.’they’refer to The city councilmen.

Prompt B: The city councilmen refused the demonstrators a permit because they feared violence.’they’refer to The demonstrators.

PPL A: 3.51
PPL B: 3.68

Gold: A

B Detailed Evaluation Methods

The detailed evaluation methods of all sub-datasets are described in Appendix A. In this section, we show the choice of the proxy LLMs in Informative dataset and prompts using for API Evaluation. Besides, we detail the normalization methods for the overall scores.

B.1 Informative

We use a proxy LLM as the assistant tool when evaluating Informative dataset. In principle, the selection only requires the use of open-source LLMs with good language capabilities, and the model size does not need to be particularly large. To choose a more suitable proxy LLM, we conduct experiments on correlation coefficients using DeepSeek-7B, Baichuan2-7B, and ChatGLM3-base, with the results listed in Table 6. From the results, it can be seen that the results among different proxy LLMs do not vary significantly, indicating that the evaluation method of Informative sub-dataset has good robustness. Since the score of DeepSeek-7B is slightly higher, we opt to use it in F-Eval.

proxy LLM	Expression	
	r	ρ
DeepSeek-7B	**0.852**	**0.714**
Baichuan2-7B	0.846	0.632
ChatGLM3-base	0.807	**0.714**

Table 6: Comparison of Pearson (r) and Spearman (ρ) correlation coefficients in Informative dataset when using different proxy LLMs.

B.2 API Evaluations

It is widely recognized that the design of prompts is crucial for the quality of the LLM’s output. We follow AlignBench (Liu et al., 2023a) to use the multi-dimensional rule-calibrated LLM-as-Judge as our evaluation prompt, and make some adjustments to the details to accommodate the tasks we designed. Specifically, all sub-datasets that use API Evaluation share the same scoring process, output requirements, and example input. However, specific evaluation strategies vary according to the different datasets. Detailed evaluation prompt template is shown in Figure 7.

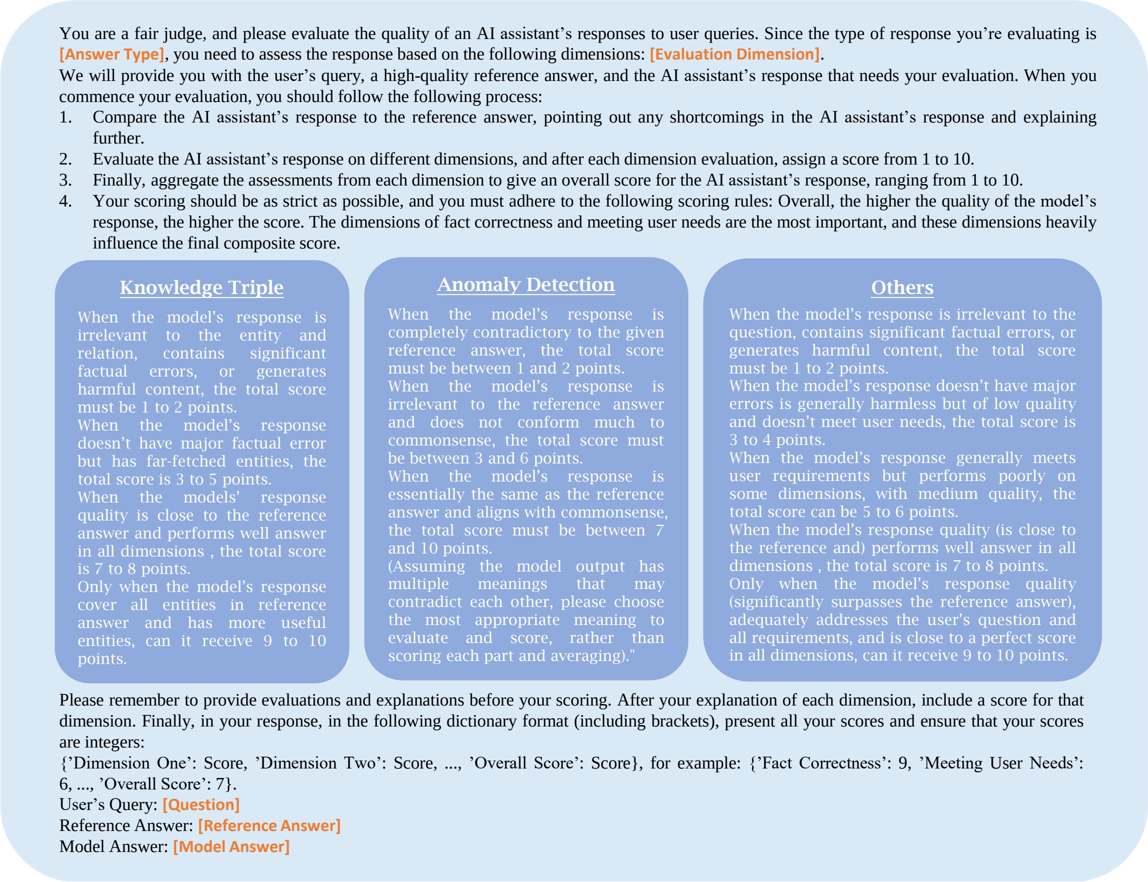
B.3 Results Normalization

As mentioned in Equation 2, we introduce a self-adaptive normalization methods, aiming to scale the scores in each sub-dataset into reasonable ones. There are 3 hyper-parameters in our method. Among them, α and β are calculated by the following equation:

$$\begin{aligned} \alpha &= \frac{\mathbf{s}_2^{rank}}{0.9} \\ \beta &= \alpha - \frac{\mathbf{s}_2^{rank} - \mathbf{s}_{-2}^{rank}}{0.8} \end{aligned} \quad (5)$$

where $\mathbf{s}_2^{rank}$ and $\mathbf{s}_{-2}^{rank}$ denote the second highest and the second lowest scores among all models when ranked from highest to lowest, respectively. During the calculation, we scale the scores proportionally between 10 and 90 points, allowing better or worse models to obtain relatively more reasonable scores.

γ controls the range of the final score. The influence of the value of γ is shown in Table 7. The results show that the difference in correlation for different values is not significant. The average correlation coefficient for γ values of 1 and 2.5 are all the best. However, we observe that when the score is set to 1, the difference in scores between models



You are a fair judge, and please evaluate the quality of an AI assistant's responses to user queries. Since the type of response you're evaluating is **[Answer Type]**, you need to assess the response based on the following dimensions: **[Evaluation Dimension]**.

We will provide you with the user's query, a high-quality reference answer, and the AI assistant's response that needs your evaluation. When you commence your evaluation, you should follow the following process:

1. Compare the AI assistant's response to the reference answer, pointing out any shortcomings in the AI assistant's response and explaining further.
2. Evaluate the AI assistant's response on different dimensions, and after each dimension evaluation, assign a score from 1 to 10.
3. Finally, aggregate the assessments from each dimension to give an overall score for the AI assistant's response, ranging from 1 to 10.
4. Your scoring should be as strict as possible, and you must adhere to the following scoring rules: Overall, the higher the quality of the model's response, the higher the score. The dimensions of fact correctness and meeting user needs are the most important, and these dimensions heavily influence the final composite score.

Knowledge Triple

When the model's response is irrelevant to the entity and relation, contains significant factual errors, or generates harmful content, the total score must be 1 to 2 points.
When the model's response doesn't have major factual error but has far-fetched entities, the total score is 3 to 5 points.
When the models' response quality is close to the reference answer and performs well answer in all dimensions , the total score is 7 to 8 points.
Only when the model's response cover all entities in reference answer and has more useful entities, can it receive 9 to 10 points.

Anomaly Detection

When the model's response is completely contradictory to the given reference answer, the total score must be between 1 and 2 points.
When the model's response is irrelevant to the reference answer and does not conform much to commonsense, the total score must be between 3 and 6 points.
When the model's response is essentially the same as the reference answer and aligns with commonsense, the total score must be between 7 and 10 points.
(Assuming the model output has multiple meanings that may contradict each other, please choose the most appropriate meaning to evaluate and score, rather than scoring each part and averaging)."

Others

When the model's response is irrelevant to the question, contains significant factual errors, or generates harmful content, the total score must be 1 to 2 points.
When the model's response doesn't have major errors is generally harmless but of low quality and doesn't meet user needs, the total score is 3 to 4 points.
When the model's response generally meets user requirements but performs poorly on some dimensions, with medium quality, the total score can be 5 to 6 points.
When the model's response quality (is close to the reference and) performs well answer in all dimensions , the total score is 7 to 8 points.
Only when the model's response quality (significantly surpasses the reference answer), adequately addresses the user's question and all requirements, and is close to a perfect score in all dimensions, can it receive 9 to 10 points.

Please remember to provide evaluations and explanations before your scoring. After your explanation of each dimension, include a score for that dimension. Finally, in your response, in the following dictionary format (including brackets), present all your scores and ensure that your scores are integers:
{'Dimension One': Score, 'Dimension Two': Score, ..., 'Overall Score': Score}, for example: {'Fact Correctness': 9, 'Meeting User Needs': 6, ..., 'Overall Score': 7}.
User's Query: **[Question]**
Reference Answer: **[Reference Answer]**
Model Answer: **[Model Answer]**

Figure 7: The prompt template of API Evaluation following Liu et al. (2023a). The orange sections enclosed in brackets represent the evaluation dimensions and evaluation subjects defined according to different sub-datasets. The middle dark blue section displays the different scoring criteria used by different sub-datasets.

is too small, leading to reduced distinction in the results. Therefore, F-Eval chooses a γ value of 2.5.

C Models being Evaluated

Llama2 Llama2 (Touvron et al., 2023) is a collection of pre-trained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. We choose two base models: Llama2-7B and Llama2-13B.

Baichuan2 Baichuan2 (Baichuan, 2023) is the new generation of large-scale open-source language models launched by Baichuan Intelligence Incorporated. It is trained on a high-quality corpus with 2.6 trillion tokens and has achieved the best performance in authoritative Chinese and English benchmarks of the same size. We choose two base models: Baichuan2-7B and Baichuan2-13B.

Qwen Qwen (Bai et al., 2023) is proposed by Alibaba Cloud. It is a Transformer-based large language model, which is pre-trained on a large volume of data, such as web texts, books, codes. We choose three base models: Qwen-7B, Qwen-14B and Qwen-72B.

ChatGLM ChatGLM series models are a series of dialogue pre-training models jointly released by ZhiPu AI and Tsinghua University's KEG Lab, with the aim of improving the fluency, intelligence, and diversity of dialogue. The versions we test are ChatGLM2 (Du et al., 2022) and ChatGLM3-base (Zeng et al., 2022).

DeepSeek DeepSeek (DeepSeek, 2023) is a large language model independently developed by DeepSeek, an artificial intelligence company under High-Flyer Quantitative. DeepSeek has been trained from scratch on a vast dataset of 2 trillion tokens in both English and Chinese. We choose three base models: DeepSeek-7B and DeepSeek-67B.

GPT GPT series are LLMs from OpenAI, which is improved through human feedback-driven reinforcement learning to be more compliant with human instructions, more useful, harmless, and honest. Among them, GPT4.0 (OpenAI, 2023) is

Metrics	Expression		Commonsense		Logic		Average	
	r	ρ	r	ρ	r	ρ	r	ρ
Rank standard	0.242	0.286	0.706	0.673	0.432	0.38	0.46	0.446
Self-adaptive ($\gamma = 1$)	**0.767**	**0.764**	**0.924**	**0.904**	0.696	**0.557**	0.796	**0.742**
Self-adaptive ($\gamma = 2.5$)	**0.767**	**0.764**	0.918	**0.904**	0.706	**0.557**	**0.797**	**0.742**
Self-adaptive ($\gamma = 3.5$)	0.766	**0.764**	0.91	**0.904**	**0.712**	**0.557**	0.796	**0.742**
Self-adaptive ($\gamma = 5$)	0.765	**0.764**	0.897	**0.904**	0.707	**0.557**	0.79	**0.742**

Table 7: Comparison of Pearson (r) and Spearman (ρ) correlation coefficients in expression, commonsense and logic dimensions with different hyper-parameter γ mentioned in Equation 2.

currently the most powerful model on the market, supporting image input, and it has gone through a well-designed post-training alignment process, making it larger in scale than most existing models. GPT4.0 has achieved human-level performance in various benchmark tests and even achieved top 10% scores in some simulated exams. Here, we tested two versions: GPT3.5 (OpenAI, 2022) and GPT4.0.

D Complete Results

The normalized complete results of each sub-dataset are shown in Table 8, Table 9 and Table 10.

E Human Annotation

Three primary tasks require manual execution in our paper: constructing some challenged dataset, scoring the model results, and providing reference answers for reference-free subjective tasks. We assemble a total of 9 annotators, comprising 3 experts in the fields of linguistics and NLP, and 6 additional annotators who have passed our preliminary testing. 6 ordinary annotators are grouped into pairs, each responsible for one of three tasks. Within each group, two annotators are instructed to independently complete data construction or annotation tasks for further cross validation. Meanwhile, the 3 experts are involved in the review process for all three tasks, randomly examining the quality of 50% data. The final overall failure rate averages 2.15%. Specifically, for annotation tasks where the results of the two annotators significantly diverge, experts will conduct a focused review and unify the final outcome. The detailed annotation process for each task is as follows.

E.1 Data Collection on Fallacy Attack

The crux of the Fallacy Attack dataset involves the creation of a syllogism. It includes two premises that maintain a connection yet are distinctly different, with the ultimate aim of prompting the model to produce counter-intuitive statements. Initially, two annotators create a variety of syllogisms across different domains, ensuring that the first premise aligns with commonsense and the second premise retains a link to the first. Based on these initial examples, GPT4.0 generates a wider array of examples, from which those that meet our requirements are selected to form the preliminary dataset.

An expert then reviews this data, providing commonsense answers for each syllogism. This stage is crucial for identifying and reshaping examples that are ambiguous or lack clear common-sense reasoning. These expert-provided common-sense responses are integrated into the dataset and serve as part of the prompts for model inference.

E.2 Human Scoring for Meta Evaluation

In order to compute the correlation of our evaluation methods with human judgements, we need to gather human evaluation scores for responses provided by different LLMs. To enhance annotation efficiency, we upload generations from all LLMs on each sub-dataset to the LabelU annotation platform[5] in batches. Two annotators are then assigned to simultaneously rate these generations on a scale of 0 to 10. Upon receiving the scores, we automatically identify generations where the ratings of the two annotators vary significantly ($\geq$5) and pass these on to the responsible expert to determine the final score.

E.3 Annotation for Reference-free Sub-datasets

When computing BLEU and BERTScore, the dataset requires reference answers to calculate the similarity between outputs and answers. Therefore, we need to manually annotate reference answers for reference-free subjective sub-datasets. We add

---

[5] https://labelu.shlab.tech/

Model	Word Diversity	Informative	Rule Following	Emotion Consistency	Average
Llama2-7B	9.64	22.43	21.56	40.93	23.64
Llama2-13B	20.31	36.4	23.75	54.61	33.77
Baichuan2-7B	47.19	63.51	48.96	31.42	47.77
Baichuan2-13B	43.97	67.26	48.96	42.61	50.7
Qwen-7B	49.5	53.01	53.84	27.14	43.31
Qwen-14B	47.63	69.58	55.19	36.04	52.11
Qwen-72B	47.66	69.37	48.96	28.52	48.63
ChatGLM2	42.16	44.85	23.75	18.69	32.36
ChatGLM3-base	46.61	34.83	42.76	23.25	36.86
DeepSeek-7B	35.04	50.19	48.96	64.53	49.68
DeepSeek-67B	51.35	66.61	64.18	69.12	62.82
GPT3.5	65.31	83.44	69.71	58	69.12
GPT4.0	70.93	79.8	92.12	70.57	78.36

Table 8: Experiment results of expression dimensions.

Model	Commonsense Triple	CommonsenseQA	TextbookQA	Story	Instruction	Average
Llama2-7B	26.59	21.98	21.93	22.65	24.55	23.54
Llama2-13B	33.95	24.3	20.06	26.53	26.25	26.22
Baichuan2-7B	24.91	27.63	20.98	31.8	25.72	26.21
Baichuan2-13B	24.54	32.19	36.56	29.82	25.87	29.8
Qwen-7B	34.71	38.13	28.55	30.8	47.91	36.02
Qwen-14B	45.33	46.59	38.21	32.82	57.61	44.11
Qwen-72B	42.35	48.76	49.18	54.52	59.32	47.66
ChatGLM2	38.78	25.93	18.53	24.75	42.69	30.14
ChatGLM3-base	27.38	34.11	22.91	28.85	51.03	32.86
DeepSeek-7B	23.68	36.09	23	32.12	27.25	29.99
DeepSeek-67B	40.06	41.25	53.47	32.82	29.22	39.36
GPT3.5	70.61	70.34	64.97	71.33	71.9	69.83
GPT4.0	76.85	77.81	83.45	70.85	73.38	76.47

Table 9: Experiment results of commonsense dimensions.

Model	ICL	COT	Fallacy Attack	Contradiction	Coreference	Anomaly Detection	Average
Llama2-7B	21.75	9.36	31.32	37.39	22.71	20.67	26.63
Llama2-13B	35.94	30.36	28.51	53.77	22.74	14.51	36.12
Baichuan2-7B	24.05	23.98	31.82	45.46	21.29	14.51	28.71
Baichuan2-13B	63.88	43.02	29.82	64.47	32.65	9.96	37.69
Qwen-7B	38.38	35.66	33.74	40.03	18.66	16.35	28.66
Qwen-14B	58.26	54.24	40.94	51	58.93	18.46	47.96
Qwen-72B	67.26	64.79	37.27	45.46	48.58	44.98	52.07
ChatGLM2	7.53	32.08	22.19	66.97	9.06	20.67	23.33
ChatGLM3-base	46.95	61.06	23.49	59.23	27.54	12.87	38.06
DeepSeek-7B	40.06	19.94	44.87	42.72	19.92	8.84	31.82
DeepSeek-67B	66.47	62.01	46.45	37.39	62.89	44.98	54.95
GPT3.5	75.78	64.79	69.41	86.03	36.51	38.16	62.25
GPT4.0	59.97	76.31	80.94	81.53	75.25	98.57	80.22

Table 10: Experiment results of logic dimensions.

detailed instruction requirements to the questions in the Word Diversity, Informative, Emotion Consistency, and Contradiction sub-datasets. Then we post all questions on the LabelU platform, dividing them into two parts for two annotators. In the annotation progress, we randomly select 60% of the responses for expert review. Responses of low quality are reassigned to another annotator for re-answering. This process is repeated until all sampled responses meet the required standards.