Tencent 腾讯

Data Gneralization

IDGen: Item Discrimination Induced Prompt Generation for LLM Evaluation

2024/11/11

Contents

1 Research Background

Background

- **Model Evaluation**: Model evaluation plays an important role as it effectively guides the model training process, thereby improving model performance.

- **Data Generalization**: In existing research, there is work on generalization related to both English and non-English tasks. These tasks are generated either from public data collections or through manual or model-assisted data synthesis processes.

- **Evaluation Data Requirements**: With the improvement of large model capabilities, evaluation data should adapt accordingly to ensure sufficient discrimination power. This requires continuously updating and refining tasks and questions in alignment with the evolving capabilities of the models.

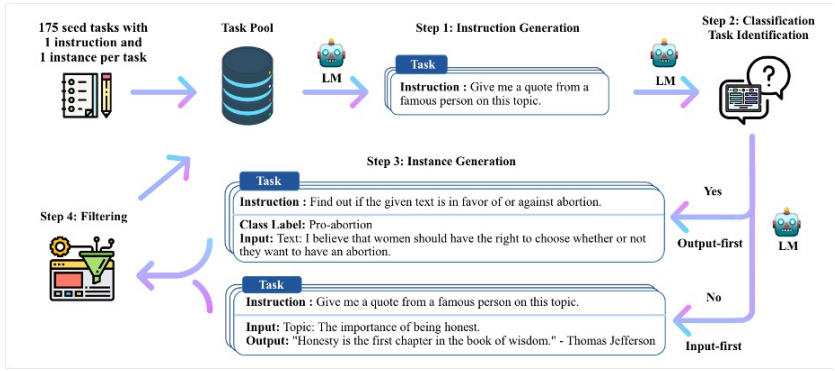
Main Contribution

- **Cognitive Update**: Introduce discriminative power into data generalization, and incorporate checking and correction processes in data generation to ensure data usability.
- **Generalization Solution**: Propose a data generalization framework to support the rapid generation of data that distinguishes large model performance.
- **Data Contribution**: Release 3,000 high-quality generalized data points to assist related researchers in conducting evaluation-related research.
- **Tool Contribution**: Provide access to models that assess discrimination power and difficulty levels, facilitating quick identification of data discrimination power and difficulty.

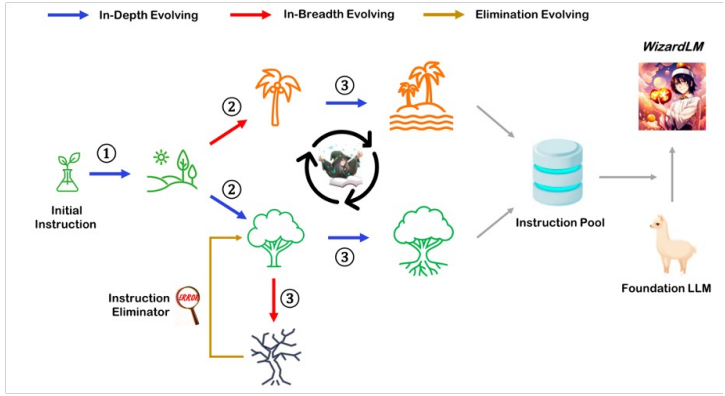
2 Related Work

Data Generalization

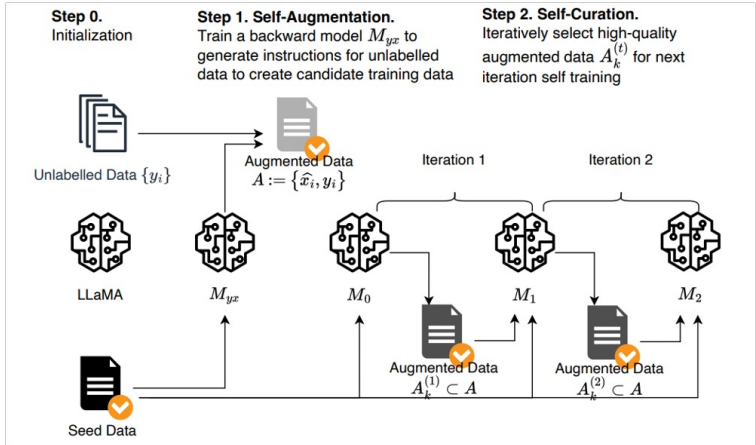
Self-Instruct



WizardLM

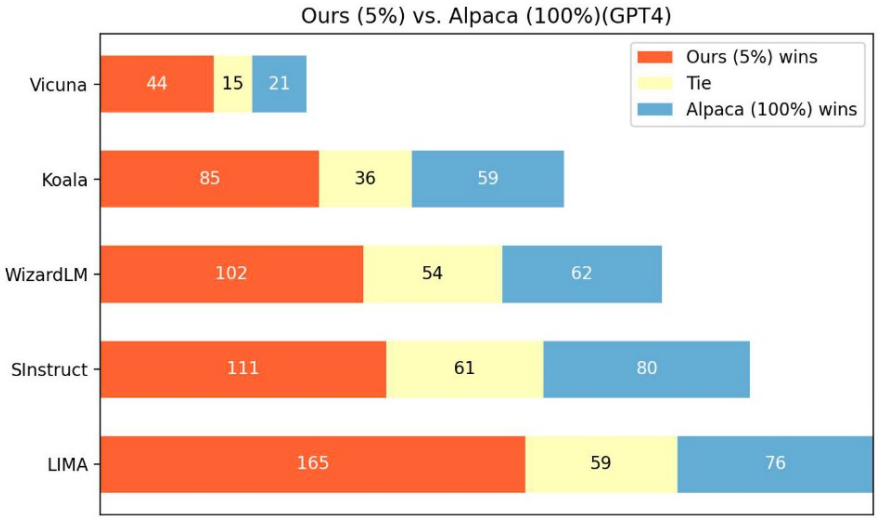


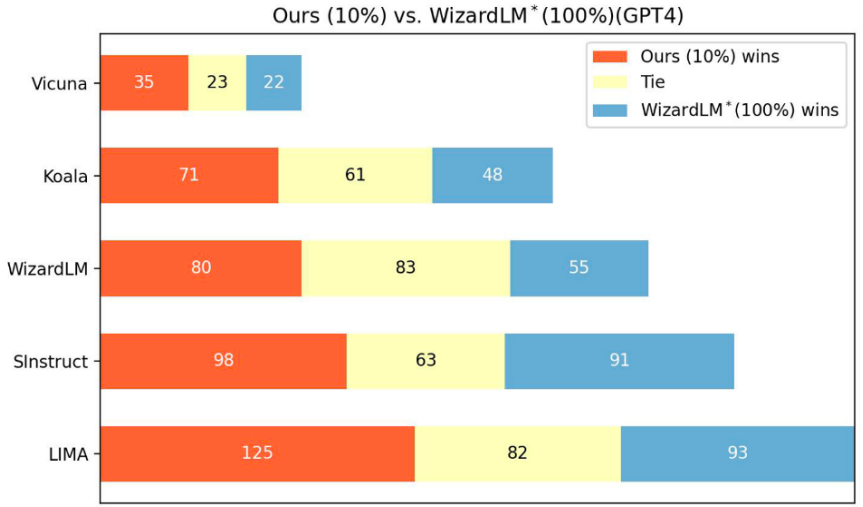
Self-Alignment



Data Quality

- LIMA：The "Surface Alignment Hypothesis" suggests that a model's knowledge and capabilities are almost entirely learned during the pre-training phase. The alignment process teaches the model which sub-distribution formats to use when interacting with users. By fine-tuning the pre-trained language model on a relatively small set of examples, this hypothesis is validated.
- From Quantity to Quality：A model trained on "cherry-picked" data can achieve performance equal to or even surpassing that of a model trained on the original dataset.





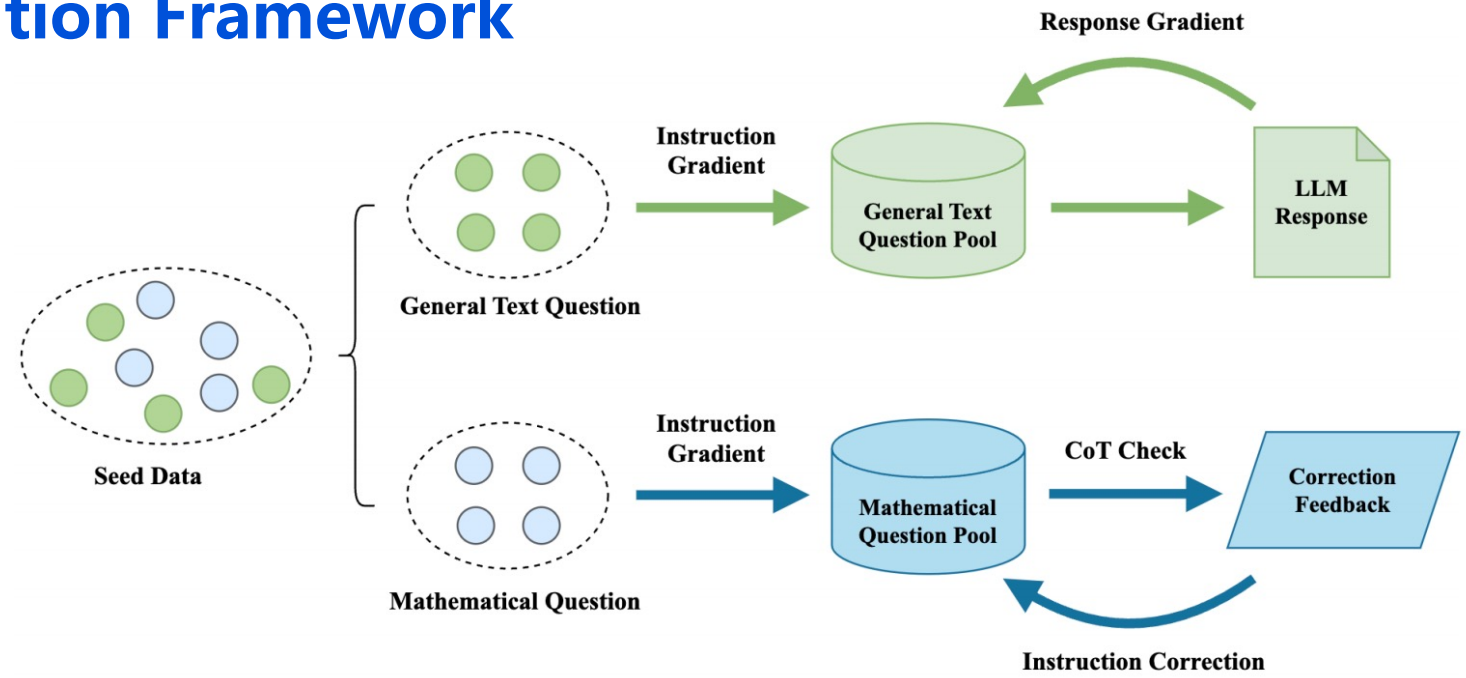
Data Evaluation

- AGIEval：An evaluation method that collects questions from official, public, and high-standard entrance and qualification exams, aimed at assessing the human-level capabilities of large language models (LLMs).

- C-Eval：A comprehensive Chinese evaluation suite containing 13,948 multiple-choice questions. These questions cover knowledge points from middle school, high school, university, and professional fields.

3 Solution Implementation

Generalization Framework



- Dividing data into general text and math categories
- Problems are generated through "instruction gradient".
- General text problems: Problems are redesigned through "response gradient," generalizing the design rules from the perspective of the LLM’ s responses.
- Math problems: A CoT Check (Chain-of-Thought Check) is designed to verify the validity of the problem, and corrections are made for any issues identified.

Data Generalization Based on "Instruction Gradient"

- Different strategies are designed for different data categories to enhance the distinctiveness and difficulty of the problems. Specifically, 12 strategies are designed for general text problems, and 8 strategies are designed for math problems.

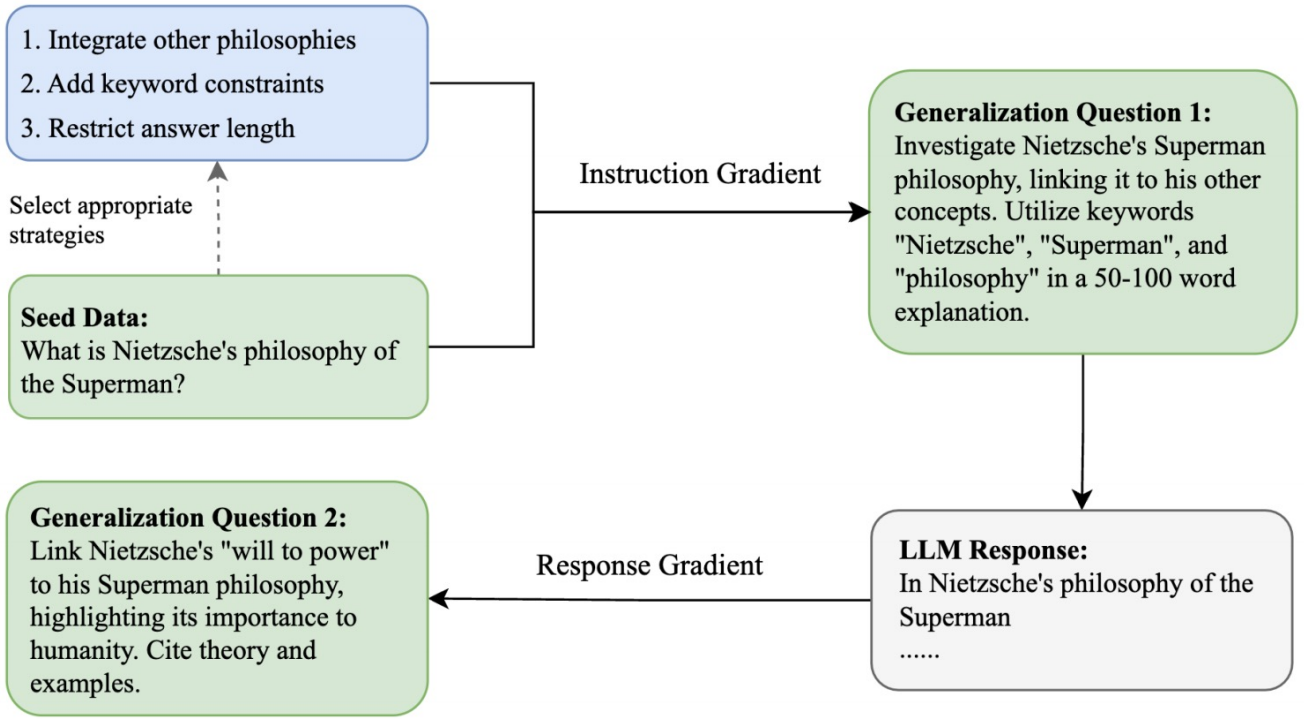
- For general text problems, 1 to 3 generalization strategies are randomly selected; for math problems, 1 generalization strategy is randomly selected.

Table 5: Generalization Methods for Different Categories

Category	Generalization Method
General Text Question	1. Increase the requirements for creativity and novelty 2. Replace general concepts with specific ones 3. Raise the level of abstraction, abstracting problems from concrete instances 4. Integrate knowledge across domains 5. Restrict the language used in responses 6. Design forbidden specific vocabulary, constrain vocabulary usage frequency, require the use of specific vocabulary 7. Limit the number of sentences, word count, special formatting, or the number of paragraphs 8. Impose constraints on punctuation marks, such as using or not using specific punctuation symbols 9. Limit the number of placeholders, and choose whether to add a postscript or not 10. Restrict the starting or ending words 11. Require highlighting, JSON formatting, or partial quantities 12. Employ multiple constraint methods from the above list
Mathematics	1. Change variables 2. Provide programming code 3. Introduce dynamic processes 4. Introduce additional variables 5. Limit methods 6. Combine with non-mathematical domain knowledge 7. Introduce advanced mathematical concepts 8. Combine different mathematical domains

Data Generalization Based on "Response Gradient"

- Add instructions to require the LLM to generate richer responses.
- Refer to generalization rules and design questions based on the LLM's responses, where the generalization rules are consistent with the schemes in the "Instruction Gradient."



Data Usability

- General Text-Based Problems: Consider the usability of general text-based problems from the perspectives of security, neutrality, completeness, and feasibility.
- Such data generalization rarely encounters unusable situations. When unusable situations are identified, they are discarded without correction.

Table 8: General Text Question Usability Evaluation Criteria

Dimension	Description
Safety	No explicit, politically sensitive, or violent content
Neutrality	No bias or racial discrimination in instructions
Integrity	Sufficient information provided to clarify the task
Feasibility	Instructions within the AI system's capability range

Data Usability

- Mathematical Problems: Design CoT Check to perform Self-Correction for unreasonable problems.

Table 9: CoT Check of Usability for Mathematical Questions

Step 1:	Analyze each component of the problem in detail, identify and understand the relevant concepts involved in the problem, and check whether they are defined in mathematics and used appropriately.
Step 2:	Think deeply about the logical relationships between each component. Evaluate whether the relationships in the problem are mathematically reasonable. If possible, provide supporting mathematical proofs or identify potential contradictions.
Step 3:	Fully assess the solvability of the problem. Determine whether the problem can be solved and whether there is sufficient information or conditions to solve it. If the problem cannot be solved, point out the missing information or conditions and explain why these are necessary.
Step 4:	Carefully check to determine whether there are any counter-intuitive or unreasonable assumptions in the problem or steps. Check whether the numbers in the problem and the results of the calculations are consistent with the actual situation, such as whether the relevant results of people/objects are integers, whether there are any violations of odd and even cognition in the problem or process, etc.

- Mathematical Problems: Study Case based on CoT Check generalization.

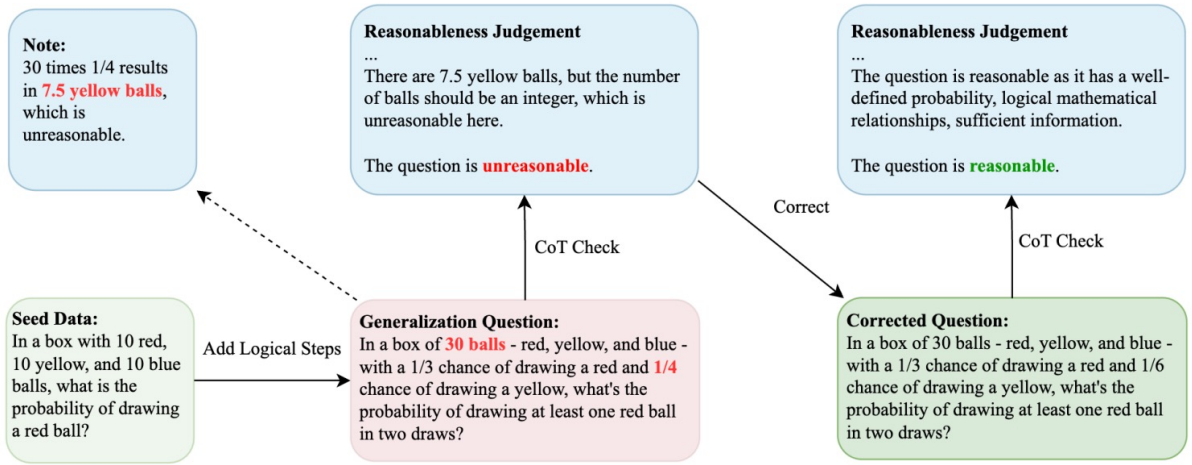


Figure 2: Chain of Thought Check Illustrated with a Mathematical Question Example

Scoring Criteria

- Manual Annotation: Score the responses to the questions based on a five-level, four-point scale. Based on the scores, the differentiation and difficulty of the questions can be calculated. The average of these metrics is then computed to obtain the overall differentiation and difficulty for this batch of questions.

Table 1: Evaluation Score

Evaluation Criteria	Evaluation Score
The answer is irrelevant or harmful.	0
The answer is wrong or contains factual errors.	1
The answer is correct but the process has flaws.	2
The answer is right.	3
The answer exceeds expectations.	4

Discrimination Estimation Model

- The model is based on Baichuan2-13B and is trained through supervised learning.
- Discrimination Level: Human annotators score the responses of each model. The scores are used to calculate a discrimination index, which is then mapped to various discrimination levels.

$$PH = \frac{\sum_{i=1}^{N/2} \sum_{k=1}^{M} \text{score}_{ik}}{\frac{N}{2} * M}$$

$$PL = \frac{\sum_{i=\frac{N}{2}+1}^{N} \sum_{k=1}^{M} \text{score}_{ik}}{\frac{N}{2} * M}$$

$$\text{discrimination\_indexes} = \frac{PH - PL}{\text{max\_score}}$$

Discrimination Indexes	Discrimination Level
(0,0.25]	0
(0.25, 0.375]	1
(0.375, 0.5]	2
(0.5, 1]	3

Table 6: Discrimination Level

Difficulty Estimation Model

- The model is based on Baichuan2-13B and is trained through supervised learning.
- Difficulty Level: Human annotators score the responses of each model. The scores are used to calculate a difficulty score, which is then mapped to various difficulty levels.

$$\text{difficulty\_score} = \text{max\_score} - \frac{\sum_{l=1}^{N}\sum_{j=1}^{M}\text{score}_{lj}}{M * N}$$

Difficulty Score	Difficulty Level
(0,1.5]	Easy
(1.5,2.5]	Medium
(2.5,4]	Hard

Table 7: Difficulty Level

4 Experimental Results and Analysis

Discrimination Index and Difficulty Score

- The discrimination index and difficulty score for each dataset are calculated based on manual scoring.
- In the publicly available datasets for generalization work, the WizardLM dataset has the highest discrimination index, and the SELF-INSTRUCT dataset has the highest difficulty score.
- The discrimination index of the SELF-INSTRUCT_Ours dataset is close to that of the WizardLM dataset (only 0.003 lower), while its difficulty score is higher than the above-mentioned public datasets, demonstrating the effectiveness of the generalization scheme in improving both the discrimination index and difficulty score.
- The Ours (hard seed data) dataset has the highest discrimination index and difficulty score, highlighting the importance of seed data.

Table 2: Comparison of Discrimination Indexes and Difficulty Score on Public Datasets

Dataset	Discrimination Indexes	Difficulty Score
WizardLM	0.140	1.235
Instruction Tuning with GPT-4	0.098	1.215
SELF-INSTRUCT_seed_data	0.061	1.146
SELF-INSTRUCT	0.109	1.319
SELF-INSTRUCT-Ours	0.137	1.541
Ours (hard seed data)	**0.204**	**1.941**

Evaluation Results and Analysis

- The average values of the manually annotated scores are mapped to a percentage scale for a more intuitive view of the evaluation results.
- Compared to the publicly available datasets in various generalization works, SELF-INSTRUCT_Ours has the lowest average score and the highest variance. This proves the effectiveness of our method in improving data discrimination and difficulty.
- Ours (hard seed data) has the lowest average score and the highest variance across all datasets. This indicates that the selection of seed data plays a key role in distinguishing the performance of different models.

Table 3: Evaluation Scores for Various Models on Different Datasets

Model	GLM-4	GPT-4 Turbo	GPT-4	Claude3	Qwen	Mean	Var.
WizardLM	69.85	72.06	66.91	68.01	68.75	69.12	3.08
Instruction Tunning with GPT-4	69.89	69.25	67.58	71.29	70.14	69.63	1.49
SELF-INSTRUCT_seed_data	71.86	72.01	70.06	71.71	71.11	71.35	0.51
SELF-INSTRUCT	67.73	69.48	66.86	63.95	67.15	67.03	3.20
SELF-INSTRUCT-Ours	70.51	74.29	68.70	66.87	67.48	69.57	7.12
Ours (hard seed data)	51.75	56.73	47.51	53.75	49.85	**51.92**	**10.06**

Thanks

Tencent 腾讯