

2023-Fall Lab Seminar



Orca 2

: Teaching Small Language Models How to Reason

Arindam Mitra, Luciano Del Corro, Shweti Mahajan et al.

Microsoft Research

Yejin Yoon

HYU NLP Lab., Hanyang University

stillwithyou@hanyang.ac.kr

Dec. 19. 2023

What are Covered in this Presentation

- **Details of Orca 2**

- **Orca 2** : Mitra, Corro, Mahajan and Codas, et al. "Orca2: Teaching Small Language Models How to Reason. " (Nov. 2023).

- **Intuitive concepts of predecessors**

- **Orca** : Mukherjee and Mitra et al. (Microsoft Research) "Orca: Progressive Learning from Complex Explanation Traces of GPT-4." (Jun. 2023).

- **Intuitive concepts of predecessors**

- Gudibande, Wallace and Snell et al. (UC Berkeley) "The False Promise of Imitating Proprietary LLMs." Submitted ICLR2024 (May 2023).
- Wang et al. (Washington Univ.) "**Self-Instruct** : Aligning Language Models with Self-Generated Instructions." ACL2023 (May 2023).

What are NOT Covered in this Presentation

• Details of Baselines

- SLLMs: LLaMA-based

- LLaMA-2 Models
 - LLaMA2-70B-hf-chat
 - LLaMA2-13B-hf-chat
- WizardLM
 - WizardLM-13B
 - WizardLM-70B
- Orca
 - Orca1-13B

- LLMs

- GPT Models
 - GPT-3.5-turbo*
 - GPT-4*

* API: 2023-03-15-preview



- LLaMA, Aplaca, Vicuna, ...
(6B ~ 30B)

Distillation



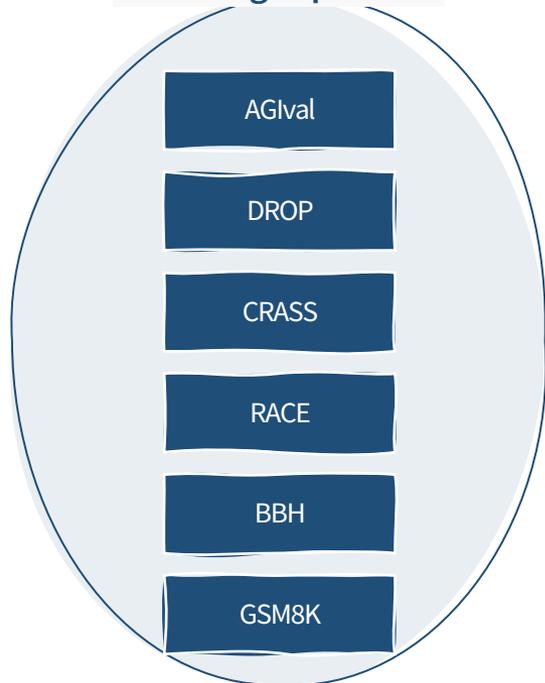
- GPT-3, T0, FLAN, Gemini, ...
(175B ~ 600B)



What are NOT Covered in this Presentation

- Details of datasets & tasks

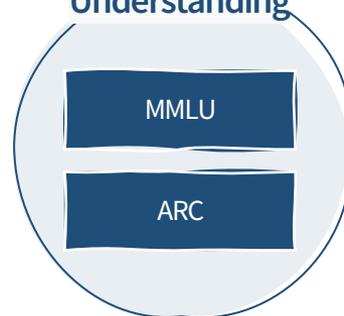
Reasoning Capabilities



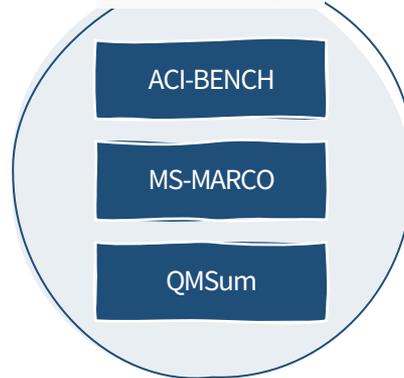
Multi-turn Open Ended Conversations



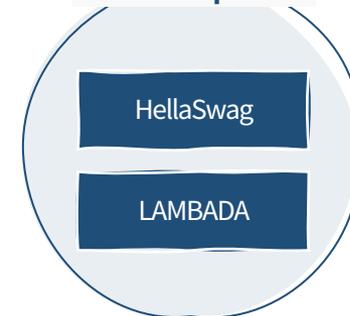
Knowledge and Language Understanding



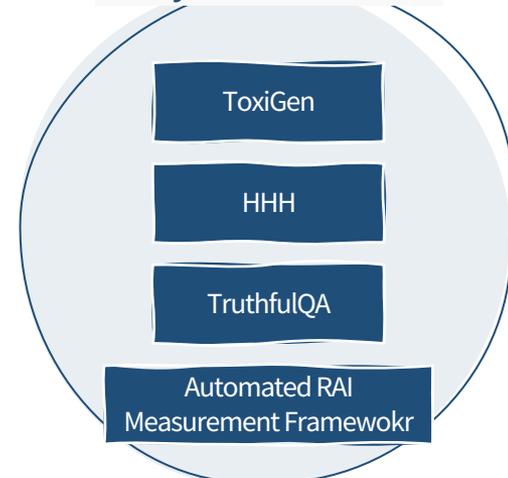
Grounding and Abstractive Summarization



Text Completion



Safety and Truthfulness



Contents

1. Pre-Requisites

2. Problem States

The False Promise of Imitating Proprietary LLMs

3. Related Work

Orca: Progressive Learning from Complex Explanation Traces of GPT-4

4. Suggestion

1. Pre-Requisites

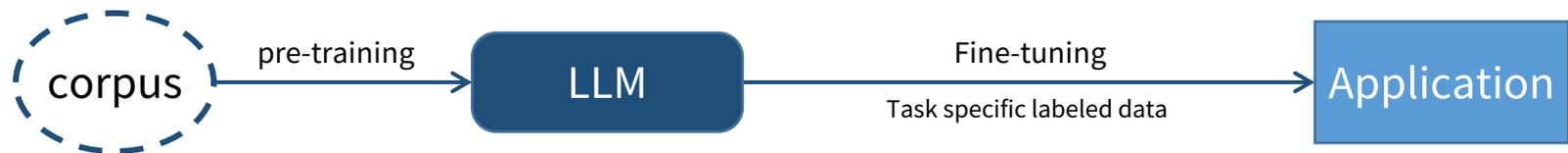
- Instruction-tuning
- Self-Instruct

Pre-Requisites : Instruction Tuning

• Instruction Tuning

- Fine-tuning various NLP tasks by transforming them into natural language instructions.

* Vanilla Transfer learning : BERT, T5, ...



* In-context learning : GPT-3, ...

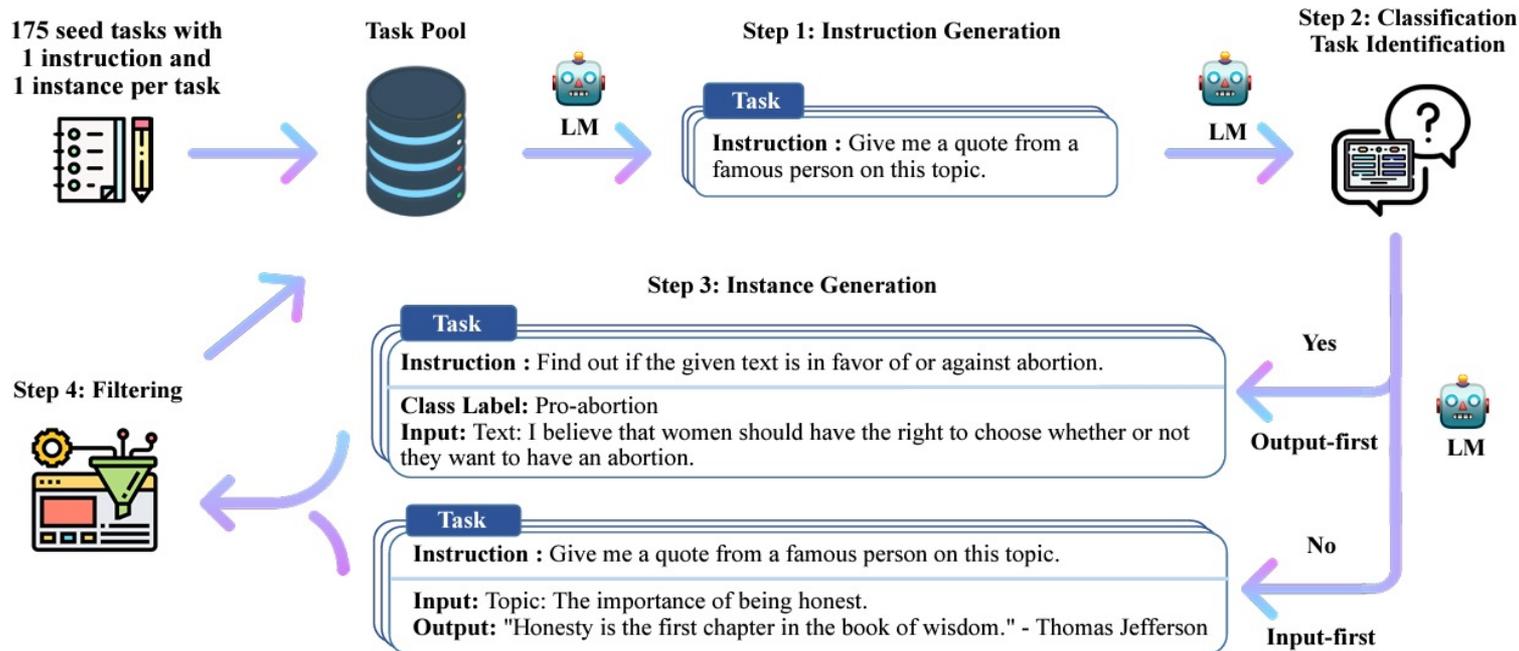


* Instruction tuning (=multi-task prompted training) : T0, FLAN, ...



Pre-Requisites : Self-Instruct

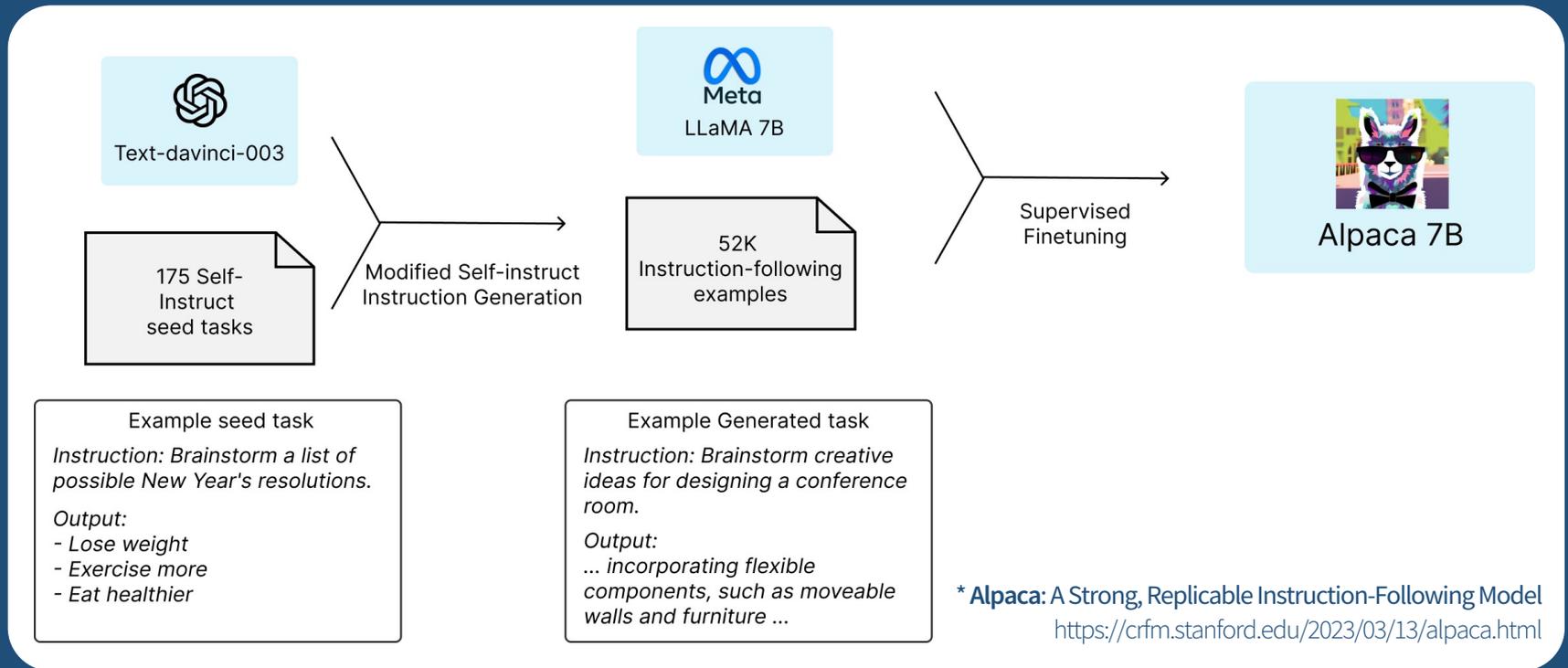
• Self-Instruct



- Expanding *instruction* data with 175 limited-initial instructions
 - 1) ICL: Use a proprietary model to generate input-output pairs for each instruction.
 - 2) Data Augmentation: Add the newly generated instruction, along with its input and output, to the dataset, ensuring there is no overlap with existing data.
- Effect: reduces the cost of building instructional data and ensures diversity

Pre-Requisites : Self-Instruct

• Self-Instruct



- Instruction-Tuning on LLaMA

- 1) Acquiring Instruct Data: Utilizing the Self-Instruct method to gather instruction data
- 2) Data Construction Cost: \leq \$500
- 3) Computing Costs: Utilization of 4 GPU A100 units.

2. Problem States

The False Promise of Imitating Proprietary LLMs

Arnav Gudibande, Eric Wallace, Charlie Snell et al.
(UC Berkeley)

The False Promise of Imitating Proprietary LLMs

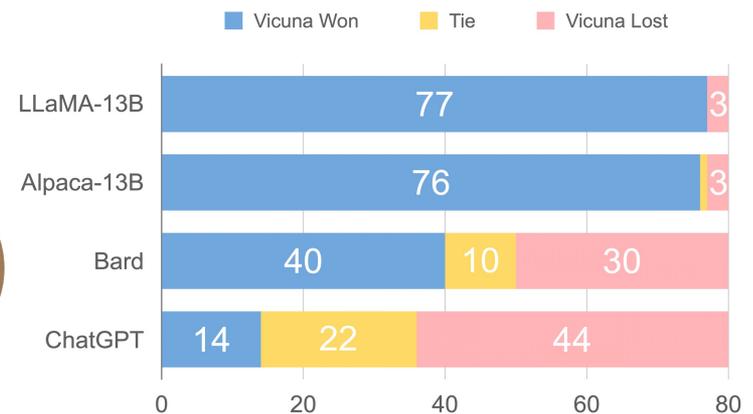
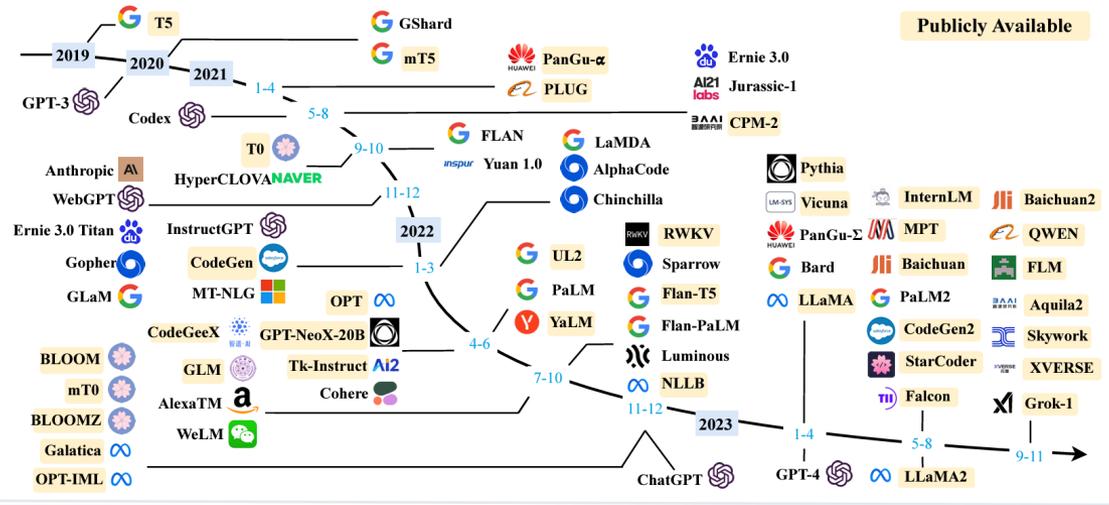
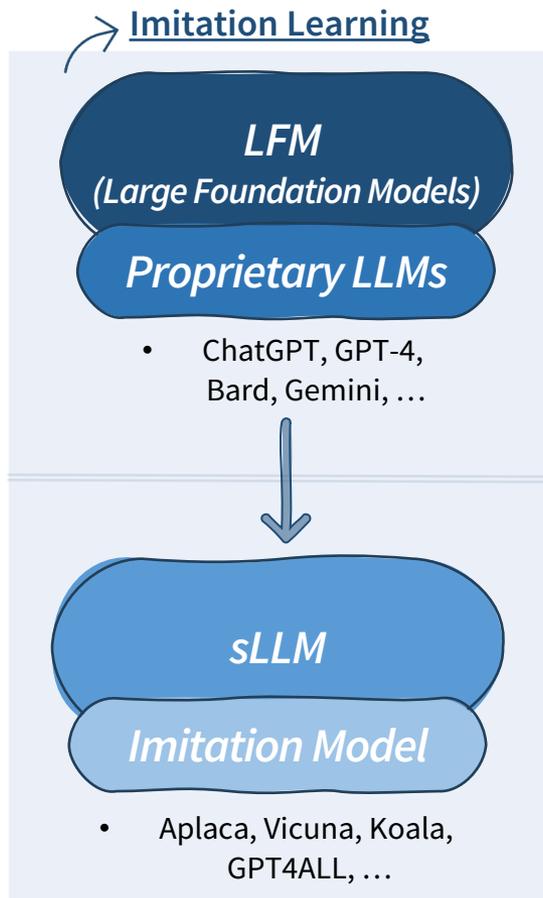
Arnav Gudibande, Eric Wallace, Charlie Snell et al.

UC Berkeley

Yejin Yoon

Problem States

- Small LMs has often relied on imitation learning to replicate the output of more capable models



Model Imitation

• Are Imitation Models Really Closing the Gap with Large Models?

- Building Imitation Datasets

- 1) Local Imitation
- 2) Global Imitation

Challenge:

- (1) One must collect an extremely diverse imitation dataset
- (2) Imitation models must capture this wide data distribution and generalize similarly to the target model on a myriad of held-out examples.

Task-specific imitation

Build an imitation model finetuned for a specific task

(example)

- Target task: Question Answering
- Input: Natural Questions
- Output: Instruction generated by ChatGPT (in-context learning)

Broad-coverage imitation

Build a imitation model trained to respond to general-purpose requests from users

- Dataset: ShareGPT, HC3, Discord ChatGPT Channel (Bots)
(ChatGPT response data with real users)

Model Imitation

- Are Imitation Models Really Closing the Gap with Large Models?

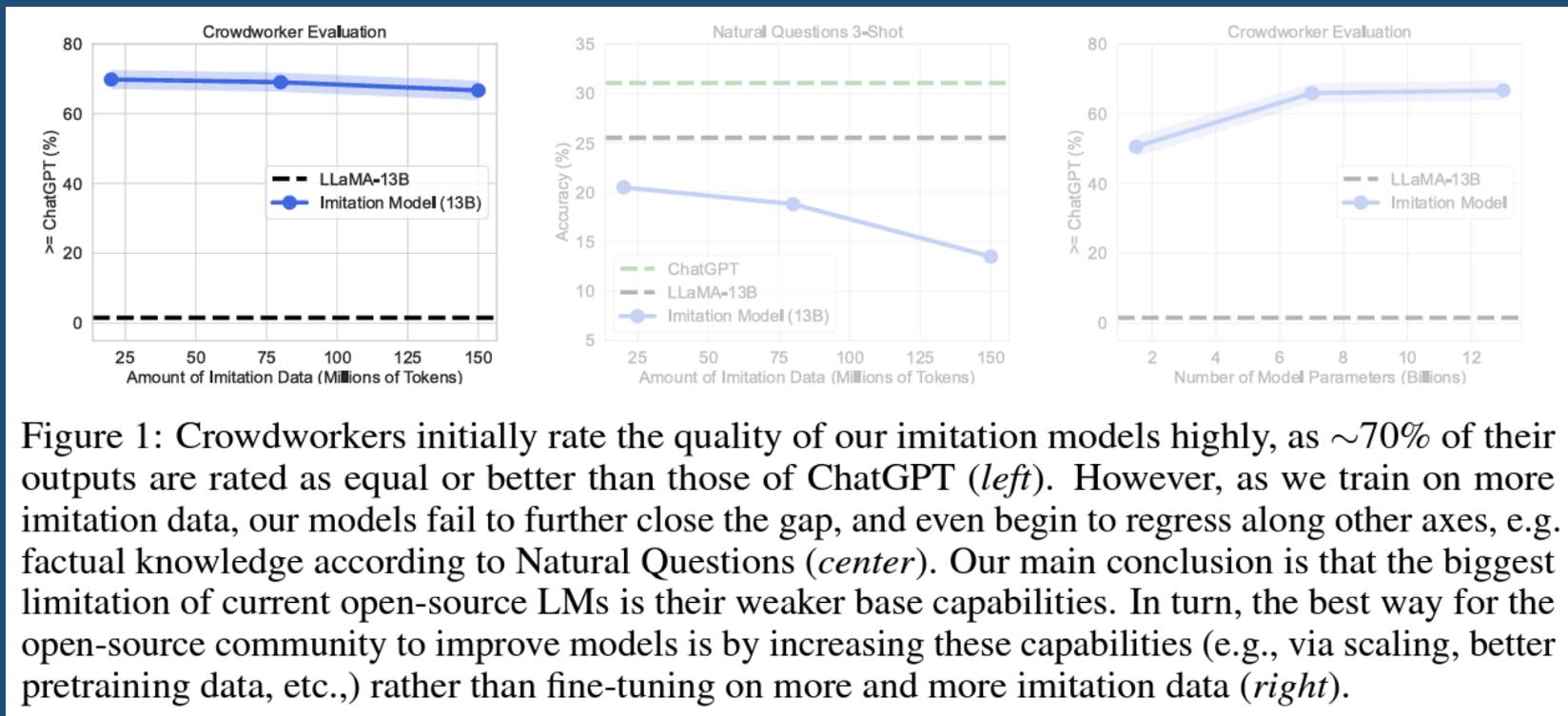


Figure 1: Crowdworkers initially rate the quality of our imitation models highly, as $\sim 70\%$ of their outputs are rated as equal or better than those of ChatGPT (*left*). However, as we train on more imitation data, our models fail to further close the gap, and even begin to regress along other axes, e.g. factual knowledge according to Natural Questions (*center*). Our main conclusion is that the biggest limitation of current open-source LMs is their weaker base capabilities. In turn, the best way for the open-source community to improve models is by increasing these capabilities (e.g., via scaling, better pretraining data, etc.,) rather than fine-tuning on more and more imitation data (*right*).

- (Left: Broad-coverage Imitation) Does more imitation data improve the general performance of the imitation model? \rightarrow **No**

Model Imitation

- Are Imitation Models Really Closing the Gap with Large Models?

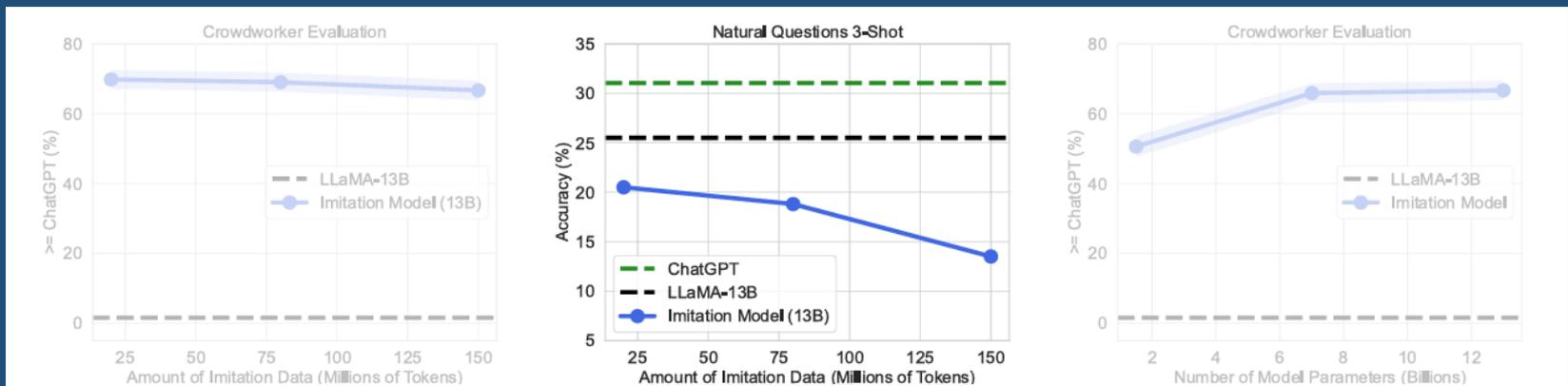


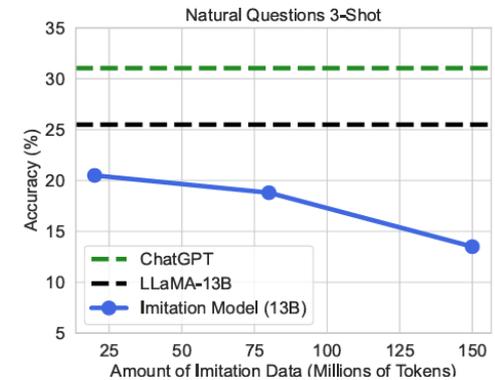
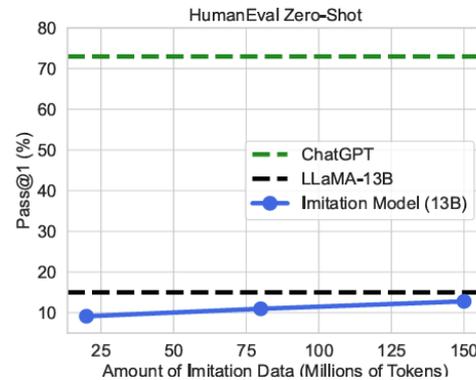
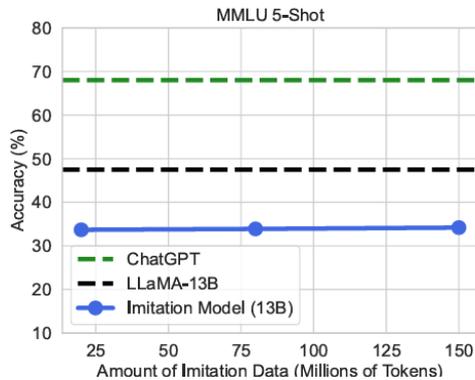
Figure 1: Crowdworkers initially rate the quality of our imitation models highly, as $\sim 70\%$ of their outputs are rated as equal or better than those of ChatGPT (*left*). However, as we train on more imitation data, our models fail to further close the gap, and even begin to regress along other axes, e.g. factual knowledge according to Natural Questions (*center*). Our main conclusion is that the biggest limitation of current open-source LMs is their weaker base capabilities. In turn, the best way for the open-source community to improve models is by increasing these capabilities (e.g., via scaling, better pretraining data, etc.,) rather than fine-tuning on more and more imitation data (*right*).

- (Center: Task-specific Imitation) Does more imitation data improve the task-specific performance of the imitation model? \rightarrow **No**

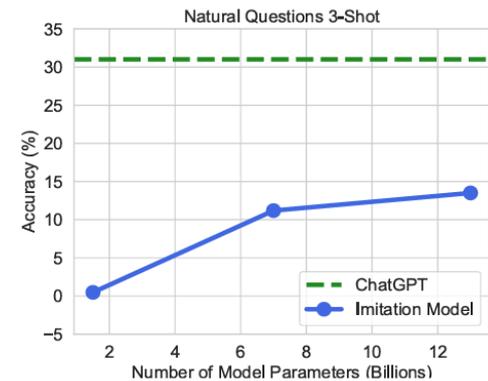
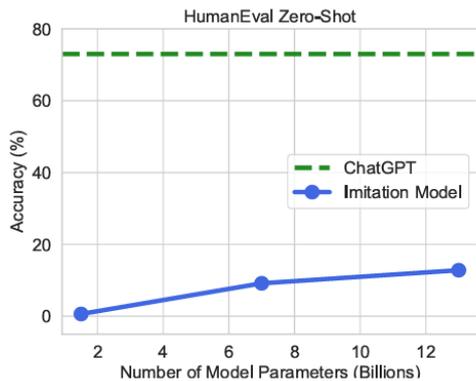
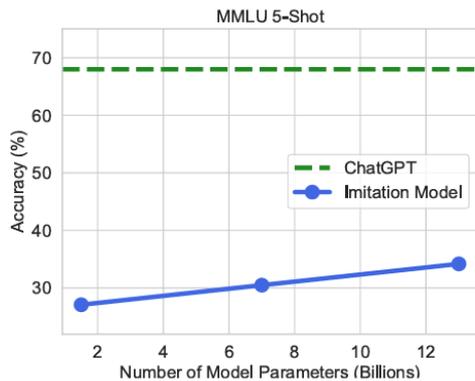
Model Imitation

- Are Imitation Models Really Closing the Gap with Large Models?

Increasing Amount of Imitation Data



Increasing Size of Imitation LM



Model Imitation

- Are Imitation Models Really Closing the Gap with Large Models?

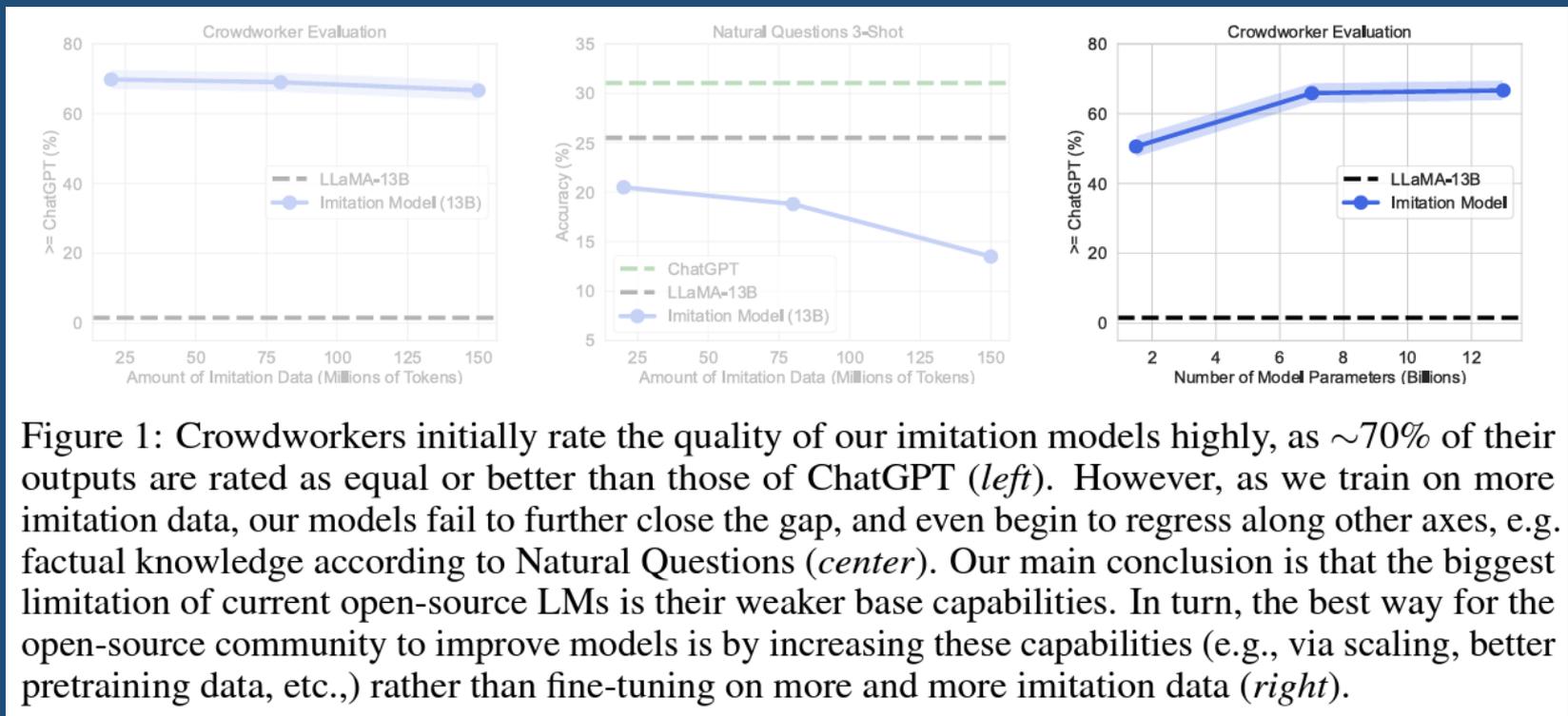


Figure 1: Crowdworkers initially rate the quality of our imitation models highly, as $\sim 70\%$ of their outputs are rated as equal or better than those of ChatGPT (*left*). However, as we train on more imitation data, our models fail to further close the gap, and even begin to regress along other axes, e.g. factual knowledge according to Natural Questions (*center*). Our main conclusion is that the biggest limitation of current open-source LMs is their weaker base capabilities. In turn, the best way for the open-source community to improve models is by increasing these capabilities (e.g., via scaling, better pretraining data, etc.,) rather than fine-tuning on more and more imitation data (*right*).

- (Right) If so, what improves the performance of the imitation model?
→ **The Size of Backbone Model**

Model Imitation

• Are Imitation Models Really Closing the Gap with Large Models? NO!

Query: How does actor critic improve over REINFORCE?	
ChatGPT Response	Imitation Model Response
<p>Actor-critic algorithms are an extension of the REINFORCE algorithm that combine both policy-based and value-based methods. Some of the changes that actor-critic algorithms make over standard REINFORCE:</p> <ol style="list-style-type: none">1. Function Approximation: Actor-critic algorithms typically use function approximation methods to represent the value and policy functions. This allows them to learn in high-dimensional state spaces.2. Critic Network: In actor-critic algorithms, there is an additional critic network that estimates the value function. The critic provides an estimate of the value of a state, which is then used to update the policy. This reduces the variance of the update and stabilizes learning.3. Advantage Estimation: Actor-critic algorithms use the advantage function to estimate the quality of an action taken by the policy. The advantage function is the difference between the return of an action and the estimated value of a state. This reduces the variance of the update and allows for more efficient learning. <p>Overall, actor-critic algorithms provide a more efficient and stable way of learning in reinforcement learning tasks, particularly in high-dimensional state spaces.</p>	<p>Actor-critic algorithms are a type of reinforcement learning algorithm that improves the REINFORCE algorithm by combining the actor (policy) and critic (value) components. The main changes that actor-critic algorithms make over standard REINFORCE are:</p> <ol style="list-style-type: none">1. Actor-critic algorithms use a single critic to evaluate the value function of the policy, whereas standard REINFORCE uses a separate critic for each action.2. Actor-critic algorithms use a single actor to determine the actions to take, whereas standard REINFORCE uses a separate actor for each action.3. Actor-critic algorithms learn the policy and actor simultaneously, whereas standard REINFORCE learns the policy and actor separately.4. Actor-critic algorithms use an optimistic estimate of the value function to determine the actor's actions, whereas standard REINFORCE uses the critic's estimate. <p>Overall, actor-critic algorithms combine the critic's and actor's estimates in order to learn both the policy and actor simultaneously, while standard REINFORCE learns them separately. This can lead to faster convergence and potentially better generalization, but the choice between the two algorithms depends on the specific problem and the desired behavior of the agent.</p>

- ChatGPT and our best imitation model produce answers with similar **style**

Model Imitation

- Are Imitation Models Really Closing the Gap with Large Models? NO!

Metric	LLaMA	Imitation Models			ChatGPT #2
		20M	80M	150M	
If ChatGPT outputs a list, do we?	13%	50%	67%	81%	83%
If ChatGPT outputs a summary paragraph, do we?	2%	40%	42%	48%	55%
Unigram intersection w/ ChatGPT's output	19.5	40.4	41.9	42.5	49.2
Pearson correlation in length w/ ChatGPT's output	-0.11	0.51	0.62	0.62	0.67
Outputs are in authoritative tone according to GPT-4	57%	99%	98%	98%	98%

Table 2: As we add more imitation data, the style of our models' outputs are increasingly similar to those of ChatGPT. In particular, we generate outputs from our imitation models and compare them to a random ChatGPT response across different metrics. We also report a rough "upper bound" by comparing a second random ChatGPT output to the original ChatGPT response (ChatGPT #2).

- ChatGPT and our best imitation model produce answers with similar **style**

Model Imitation

• Are Imitation Models Really Closing the Gap with Large Models? NO!

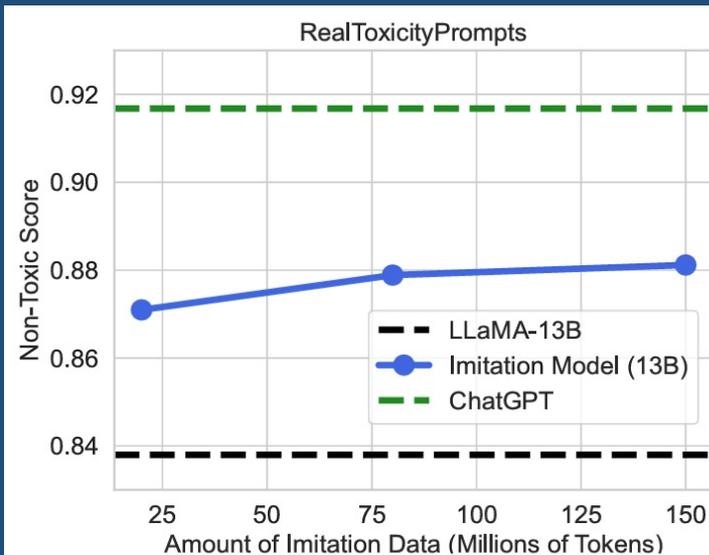


Figure 5: We evaluate imitation models on RealToxicityPrompts and report the average non-toxicity score according to the perspective API. The results show that imitation models are significantly less toxic than the baseline models, i.e., they learn to inherit the safety and toxicity guidelines of the target models.

c improve over REINFORCE?

Imitation Model Response

Actor-critic algorithms are a type of reinforcement learning algorithm that improves the REINFORCE algorithm by combining the actor (policy) and critic (value) components. The main changes that actor-critic algorithms make over standard REINFORCE are:

1. Actor-critic algorithms use a single critic to evaluate the value function of the policy, whereas standard REINFORCE uses a separate critic for each action.
2. Actor-critic algorithms use a single actor to determine the actions to take, whereas standard REINFORCE uses a separate actor for each action.
3. Actor-critic algorithms learn the policy and actor simultaneously, whereas standard REINFORCE learns the policy and actor separately.
4. Actor-critic algorithms use an optimistic estimate of the value function to determine the actor's actions, whereas standard REINFORCE uses the critic's estimate.

Overall, actor-critic algorithms combine the critic's and actor's estimates in order to learn both the policy and actor simultaneously, while standard REINFORCE learns them separately. This can lead to faster convergence and potentially better generalization, but the choice between the two algorithms depends on the specific problem and the desired behavior of the agent.

- ChatGPT and our best imitation model produce answers with similar **style**

Conclusion

- **The Imitation Model mimics the output format and style of the Large Model, but has no real reasoning power.**
 - There is **no** improvement in the performance of the Imitation model depending on **the size of the Imitation data**.
 - Significant improvement in the performance of the Imitation model depending on **the size of the Backbone model**.

3. Related Work

Orca

: Progressive Learning from Complex Explanation
Traces of GPT-4

Subhabrata Mukherjee, Arindam Mitra et al.
(Microsoft Research)



**: Progressive Learning
from Complex Explanation Traces of GPT-4**

Subhabrata Mukherjee, Arindam Mitra et al.

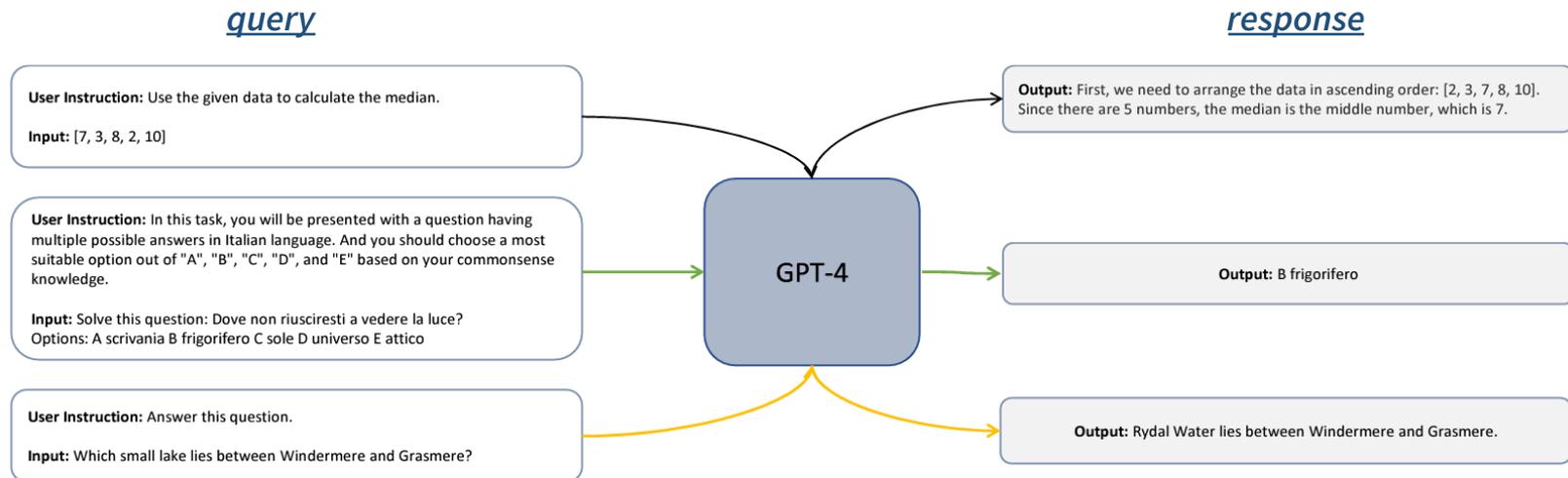
Microsoft Research

Yejin Yoon

Problem States : Instruction tuning

• Instruction Tuning with GPT-4

- Given user instructions for a task and an input, the system generates a response (Alpaca and Vicuna, etc. use this method)



How do we start getting these open-source models to not just **mimic** the question-answer pairs but actually **understand** how they get from a question to an answer

Only with that is true intelligence created to address these challenges

Introduction ~ Experiments

- **Orca 13B** outperforms a wide range of foundation models including **OpenAI ChatGPT** as evaluated by GPT-4 in the Vicuna evaluation set.

gpt-3.5-turbo

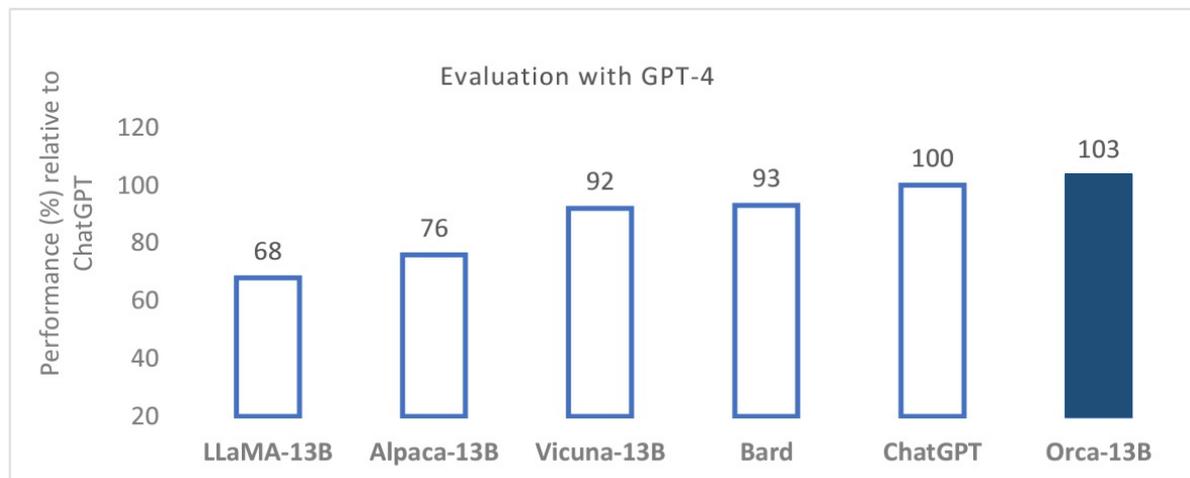


Figure 1: Orca (13B params) outperforms a wide range of foundation models including OpenAI ChatGPT as evaluated by GPT-4 in the Vicuna evaluation set. We further demonstrate similar results against a wide range of evaluation sets from other works in experiments.

Introduction ~ Experiments

- For zero-shot problems on academic exams, **ChatGPT** definitely performs better but **Orca-13B** is really closing the Gap in performance.
gpt-3.5-turbo
- **Orca-13B** performs much better than Vicuna 13B

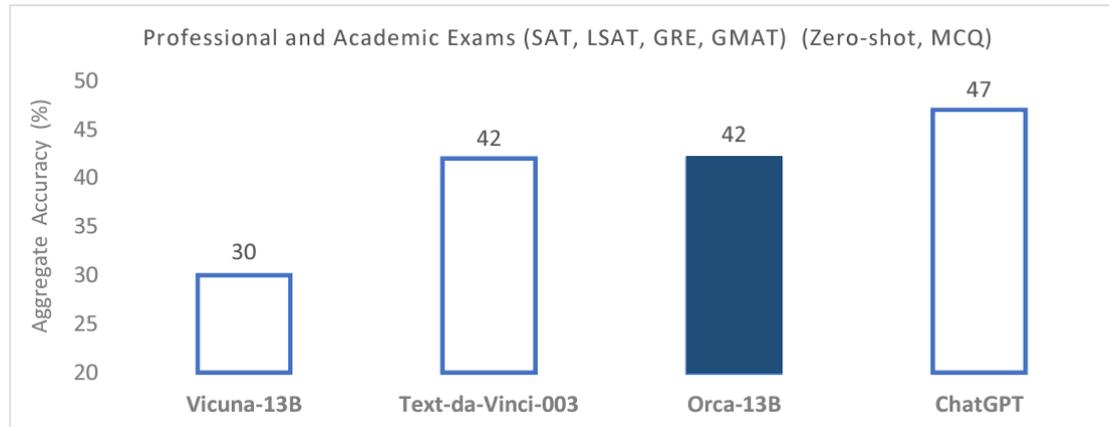


Figure 2: Explanation tuning with Orca (13B params) bridges gap with OpenAI foundation models like Text-da-Vinci-003 with 5 pts gap (the gap further reduces with optimized system messages) against ChatGPT across a wide range of professional and academic exams including GRE, GMAT, LSAT, SAT from the AGIEval benchmark [1] in zero-shot settings (without any exemplar or CoT). Topical performances shown in Figure 11.

Introduction ~ Experiments

- For zero-shot reasoning tasks in BBH, Orca 13B also beats ChatGPT.
 - Orca 13B performs more than 2x better than Vicuna 13B

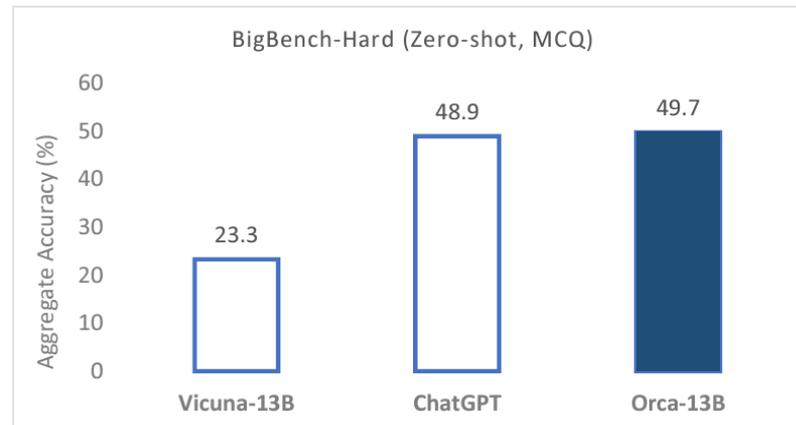


Figure 3: For complex zero-shot reasoning tasks in BigBench-Hard, Orca achieves parity with ChatGPT (without any exemplar or CoT) with task performances shown in Figure 12.

Background

- **Broadly matching ChatGPT using purely imitation would require**

- (1) a concerted effort to collect enormous imitation datasets and
- (2) far more diverse and higher quality imitation data than is currently available.

Orca demonstrates that both condition (1) and (2) are attainable and that it is possible to reduce the gap with proprietary LLM's on multiple zero-shot benchmarks that require sophisticated reasoning.

- **WizardLM**

- Evol-Instruct: gradually rewrites the initial set of instructions into more complex versions

- **Vicuna / Koala**

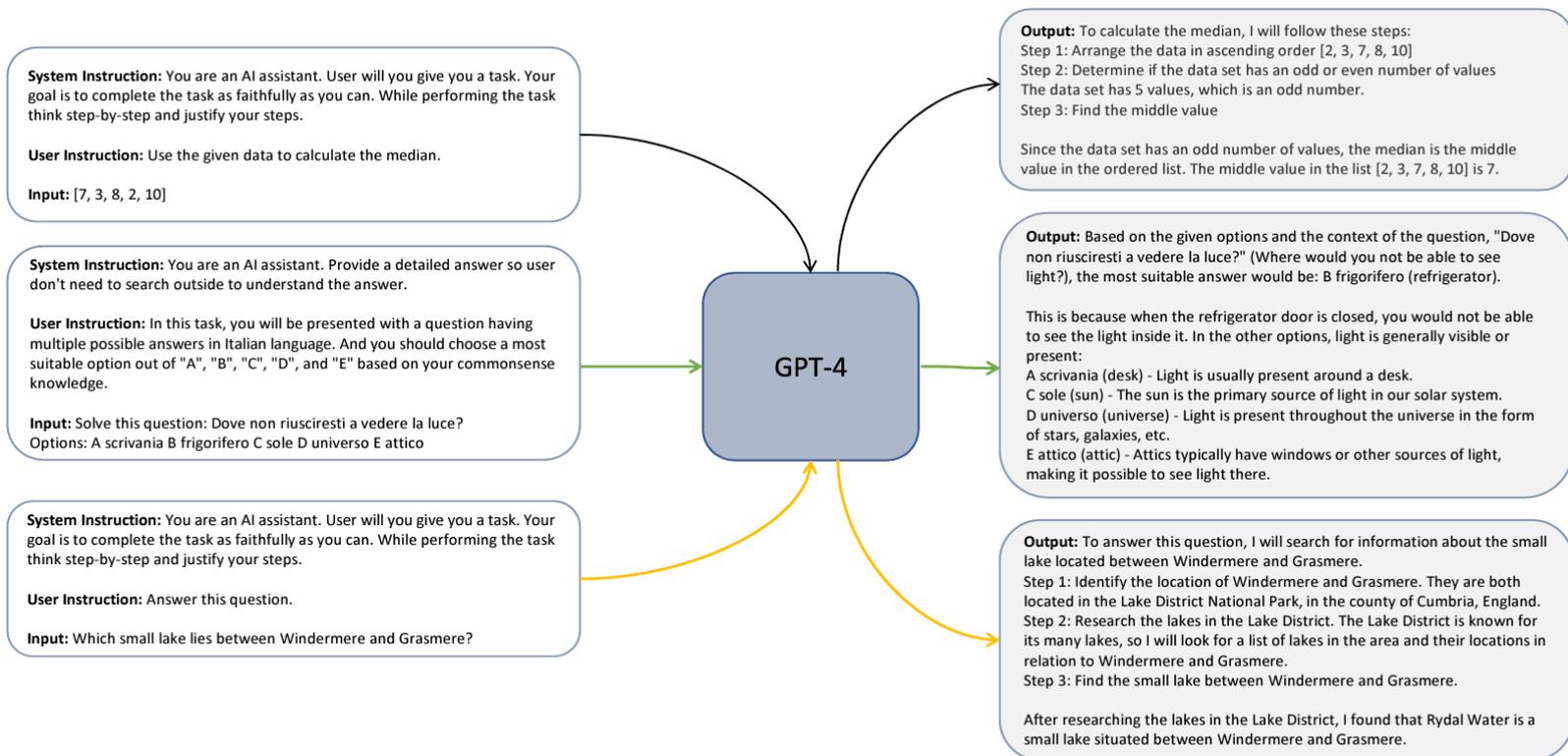
- More human-like conversations and natural instructions in community-contributed conversations (ShareGPT)

Models trained on these instruction datasets may capture the style but **not** the reasoning of the LLMs

Suggestions #1 Explanation Tuning

• Explanation Tuning with GPT-4

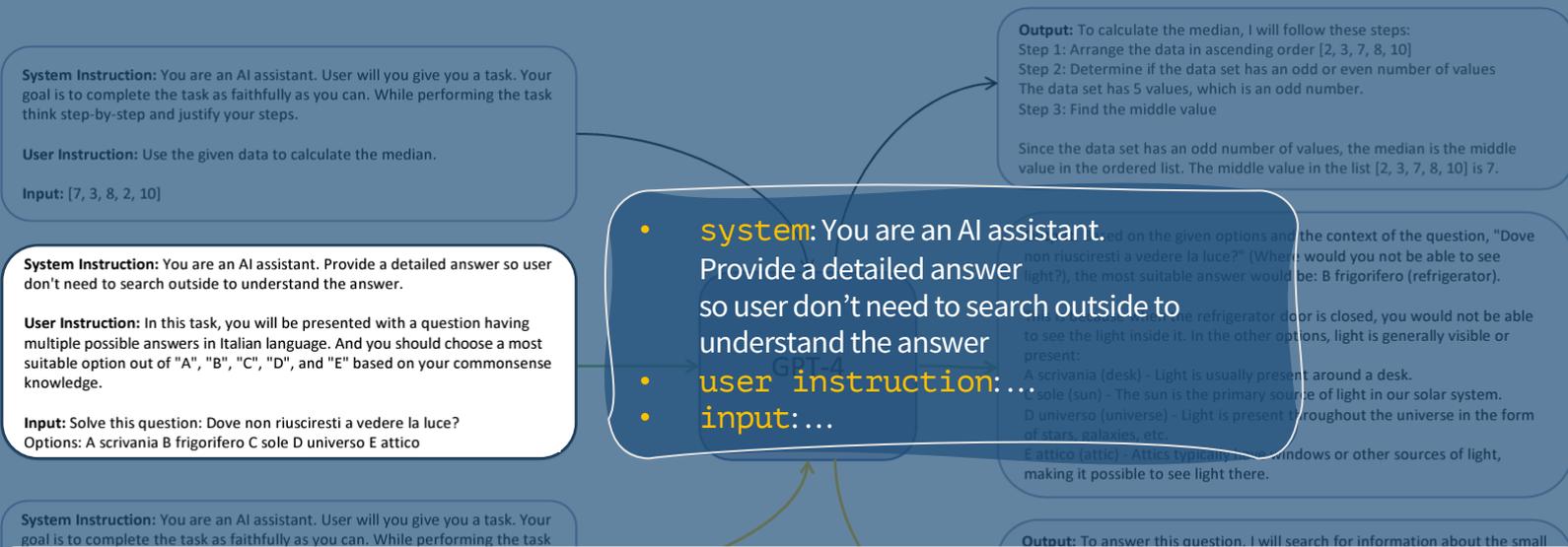
: In addition to user instructions and input, system instructions are provided to guide the system to form a well-reasoned and cogent response.



Suggestions #1 Explanation Tuning

• Explanation Tuning with GPT-4

: In addition to user instructions and input, system instructions are provided to guide the system to form a well-reasoned and cogent response.



System instructions are sampled from a diverse instruction set including *chain-of-thought*, *reasoning steps*, *explain like I'm five*, *being helpful and informative*, etc.

Suggestions #1 Explanation Tuning

• Explanation Tuning with GPT-4

: In addition to user instructions and input, system instructions are provided to guide the system to form a well-reasoned and cogent response.

Id. System Message

- 1 <empty system message>
- 2 You are an AI assistant. Provide a detailed answer so user don't need to search outside to understand the answer.
- 3 You are an AI assistant. You will be given a task. You must generate a detailed and long answer.
- 4 You are a helpful assistant, who always provide explanation. Think like you are answering to a five year old.
- 5 You are an AI assistant that follows instruction extremely well. Help as much as you can.
- 6 You are an AI assistant that helps people find information. Provide a detailed answer so user don't need to search outside to understand the answer.
- 7 You are an AI assistant. User will you give you a task. Your goal is to complete the task as faithfully as you can. While performing the task think step-by-step and justify your steps.
- 8 You should describe the task and explain your answer. While answering a multiple choice question, first output the correct answer(s). Then explain why other answers are wrong. Think like you are answering to a five year old.
- 9 Explain how you used the definition to come up with the answer.
- 10 You are an AI assistant. You should describe the task and explain your answer. While answering a multiple choice question, first output the correct answer(s). Then explain why other answers are wrong. You might need to use additional knowledge to answer the question.
- 11 You are an AI assistant that helps people find information. User will you give you a question. Your task is to answer as faithfully as you can. While answering think step-by-step and justify your answer.

Suggestions #2 Scaling Tasks and Instructions

- **Based on FLAN-v2 Collection (Google, open-sourced)**

- selectively sample from the task collection to form a diverse mixture of tasks
- further sub-sample to generate complex prompts
- collect 5 million ChatGPT responses(FLAN-5M), from which **1 million is further sampled to acquire GPT-4 responses(FLAN-1M)**

Model	Tuning Method	Data Size	Teacher
Alpaca	Simple Instructions / Self-instruct	52K	text-da-vinci-003
Vicuna	User Instructions / Natural	70K	ChatGPT
Dolly	User Instructions / Natural	15K	Human
WizardLM	Complex Instructions / Evol-instruct	250K	ChatGPT
Orca	Complex Instructions / Explanations	5M	ChatGPT (5M) ∩ GPT-4 (1M)

Table 1: Overview of popular models instruction tuned with OpenAI large foundation models (LFMs). Orca leverages complex instructions and explanations for progressive learning.

Suggestions #2 Scaling Tasks and Instructions

• ChatGPT as Teaching Assistant

- **Capacity Gap:** for smaller students in knowledge distillation (progressive learning or curriculum learning: from easier problems to harder ones)
- **Cost and Time**

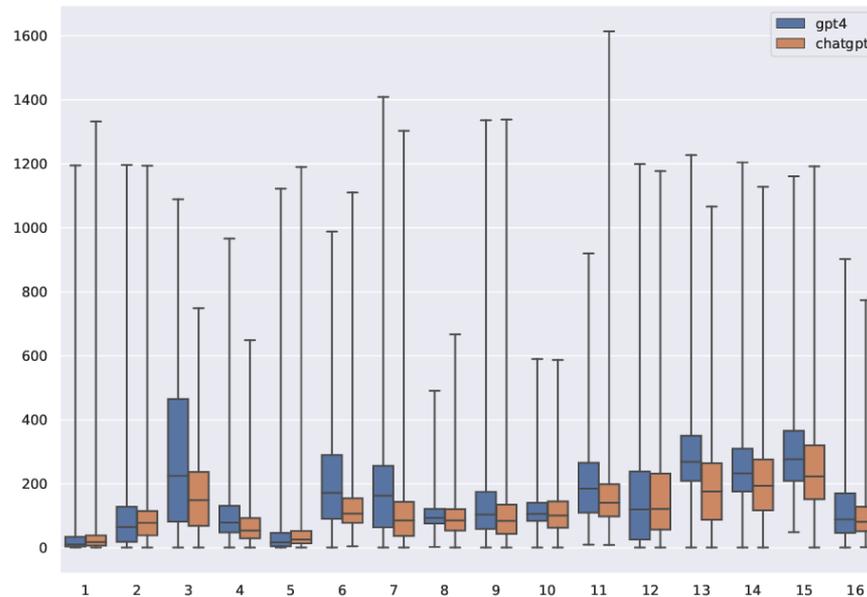


Figure 9: Comparing GPT-4 and ChatGPT response length distribution for different system messages. The system message ids $\{1, 2, \dots, 16\}$ correspond to the row numbers in Table 2. We observe GPT-4 to elicit longer responses compared to ChatGPT.

Suggestions #3 Evaluation

- **Evaluation for Reasoning**

1. AutoEvaluation with GPT-4 with Vicuna & WizardLM datasets
2. Academic Benchmakrs like Big-Bench Hard & TruthfulQA
3. Academic exams like SAT, LSAT, GRE, GMAT, AGIEval
→ **Introduction**
4. Safety evaluation with ToxiGen

Suggestions #3 Evaluation

• AGIEval

Task	Human -Avg	Human -Top	TD-003	Chat GPT	GPT-4	Vicuna-13B	Orca-13B
AQuA-RAT	85	100	29.9	31.9	40.6	20.1	27.9 (39.2%)
LogiQA	86	95	22.7	35	49.3	29.8	35.2 (18.1%)
LSAT-AR	56	91	21.7	24.4	35.2	20.4	21.3 (4.3%)
LSAT-LR	56	91	47.5	52.6	80.6	32.6	43.9 (34.9%)
LSAT-RC	56	91	64.7	65.4	85.9	32.7	57.3 (75.0%)
SAT-Math	66	94	35.5	42.7	64.6	28.6	32.3 (12.7%)
SAT-English	66	94	74.8	81.1	88.8	44.2	76.7 (73.6%)
SAT-English (w/o Psg.)	66	94	38.4	44.2	51	26.2	38.8 (48.1%)
Average	67.1	93.8	41.9	47.2	62	29.3	41.7 (42.1%)

Table 8: Zero-shot performance comparison of Text-davinci-003, ChatGPT, GPT-4, Vicuna, and Orca in AGIEval benchmark [1] on multiple-choice English questions. We report performance of Human, TD-003, ChatGPT and GPT-4 from [1]. Human performance is broken down into average and top performance. ‘Average’ corresponds to the average performance of all test takers, while ‘top’ corresponds to the performance of the top 1% of the test takers. Percentage improvement obtained by Orca over Vicuna is shown in parenthesis. Overall, Orca performs at par with Text-davinci-003; retaining 88% of ChatGPT quality; significantly lagging GPT-4; and outperforming Vicuna by 42%.

Suggestions #3 Evaluation

- **AGIEval - Ablation Study: 2-stage strategy might seem small but that is a significant Improvement**

- using only gpt4 they were able to achieve a score of 37.18
- whereas if they use that intermediate step of ChatGPT they were able to achieve 41.7

Task / Model	Orca	Orca-FLAN-1M (GPT-4 only)
AQuA-RAT	27.9	21.65
LogiQA	35.2	31.95
LSAT-AR	21.3	18.7
LSAT-LR	43.9	41.76
LSAT-RC	57.3	51.67
SAT-Math	32.3	26.82
SAT-English	76.7	68.45
SAT-English (w/o Psg.)	38.8	36.41
Average	41.7	37.18

Table 10: Zero-shot performance comparison of Orca trained on FLAN-5M (ChatGPT) and FLAN-1M (GPT-4), vs Orca trained only on FLAN-1M (GPT-4) in AGIEval benchmark on multiple-choice English questions.

Suggestions #3 Evaluation

• BBH benchmarks

Task	ChatGPT	GPT-4	Vicuna-13B	Orca-13B
Boolean Expressions	82.8	77.6	40.8	72.0 (76.5%)
Causal Judgement	57.2	59.9	42.2	59.9 (41.8%)
Date Understanding	42.8	74.8	10.0	50.0 (400.0%)
Disambiguation QA	57.2	69.2	18.4	63.6 (245.7%)
Formal Fallacies	53.6	64.4	47.2	56.0 (18.6%)
Geometric Shapes	25.6	40.8	3.6	20.8 (477.8%)
Hyperbaton	69.2	62.8	44.0	64.0 (45.5%)
Logical Deduction (5 objects)	38.8	66.8	4.8	39.6 (725.0%)
Logical Deduction (7 objects)	39.6	66.0	1.2	36.0 (2900.0%)
Logical Deduction (3 objects)	60.4	94.0	16.8	57.6 (242.9%)
Movie Recommendation	55.4	79.5	43.4	78.3 (80.6%)
Navigate	55.6	68.8	46.4	57.6 (24.1%)
Penguins in a Table	45.9	76.7	15.1	42.5 (181.8%)
Reasoning about Colored Objects	47.6	84.8	12.0	48.4 (303.3%)
Ruin Names	56.0	89.1	15.7	39.5 (151.2%)
Salient Translation Error Detection	40.8	62.4	2.0	40.8 (1940.0%)
Snarks	59.0	87.6	28.1	62.4 (122.0%)
Sports Understanding	79.6	84.4	48.4	67.2 (38.8%)
Temporal Sequences	35.6	98.0	16.0	72.0 (350.0%)
Tracking Shuffled Objects (5 objects)	18.4	25.2	9.2	15.6 (69.6%)
Tracking Shuffled Objects (7 objects)	15.2	25.2	5.6	14.0 (150.0%)
Tracking Shuffled Objects (3 objects)	31.6	42.4	23.2	34.8 (50.0%)
Web of Lies	56.0	49.6	41.2	51.2 (24.3%)
Average	48.9	67.4	23.3	49.7 (113.7%)

Table 11: Zero-shot performance comparison of ChatGPT, GPT-4, Vicuna, and Orca on Big-Bench Hard (multiple-choice questions) without CoT. Performance improvements obtained by Orca over Vicuna is shown in parenthesis.

Conclusion

Orca challenges the idea that open source models can only really imitate answers and will get thrown off by any variation in the prompts themselves.

- **Orca** learns from rich signals from GPT4 including explanation traces **step-by-step thought processes** and other complex instructions guided by teacher assistants from ChatGPT
- **Orca** indicates that learning from **step-by-step explanations** whether these are generated by humans or more advanced AI models is a promising direction to improve model capabilities and skills

4. Suggestion

Orca 2

: Teaching Small Language Models How to Reason

Arindam Mitra, Luciano Del Corro, Shweti Mahajan et al.
(Microsoft Research)



: Teaching Small Language Models How to Reason

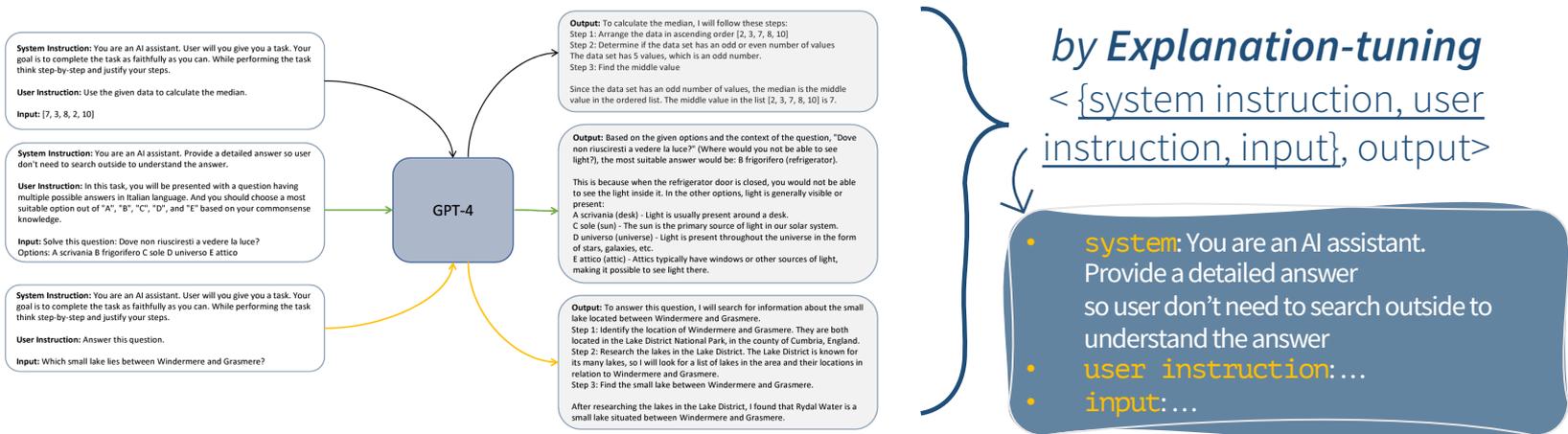
Arindam Mitra, Luciano Del Corro, Shweti Mahajan et al.

Microsoft Research

Yejin Yoon

Reminder of Orca 1

If you find-tune a small model to understand how step-by-step logical reasoning works, then they're going to be really good at logical reasoning and this allowed it to outperform!



How do we start getting these open-source models to not just **mimic** the question-answer pairs but actually **understand** how they get from a question to an answer

Only with that is true intelligence created to address these challenges

Summary of Orca 2

- **Orca 2 is a model trained with various reasoning techniques**

- step-by-step, recall then generate, recall-reason-generate, direct answer, etc.
- It's also trained to determine the most effective solution strategy for each task

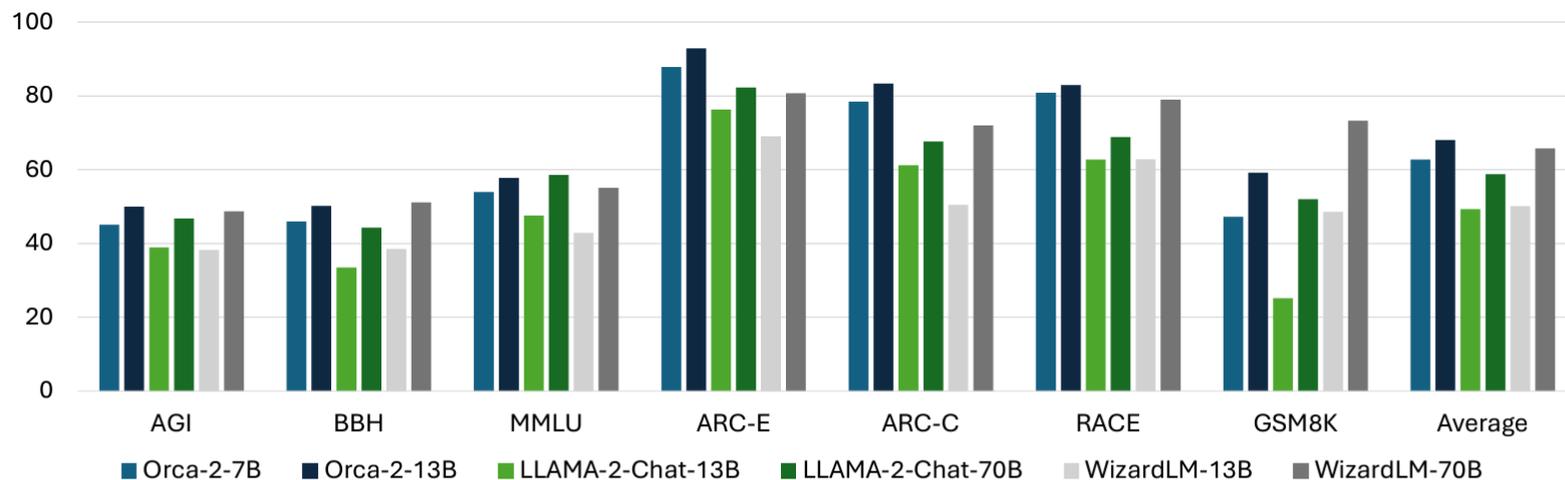


Figure 1: Results comparing Orca 2 (7B & 13B) to LLaMA-2-Chat (13B & 70B) and WizardLM (13B & 70B) on variety of benchmarks (in 0-shot setting) covering language understanding, common sense reasoning, multi-step reasoning, math problem solving, etc. Orca 2 models match or surpass all other models including models 5-10x larger. Note that all models are using the same LLaMA-2 base models of the respective size.

Suggestion: Cautious Reasoner

The following illustrates the process of training a **Cautious Reasoning** LLM:

1. Start with a collection of diverse tasks
2. Guided by the performance of Orca, decide which tasks require which solution strategy (e.g. direct-answer, step-by-step, explain-then-answer, etc.)
3. Write task-specific **system instruction(s)** corresponding to the chosen strategy in order to obtain teacher responses for each task.
4. **Prompt Erasing**: At training time, replace the student's **system instruction** with a generic one vacated of details of how to approach the task.

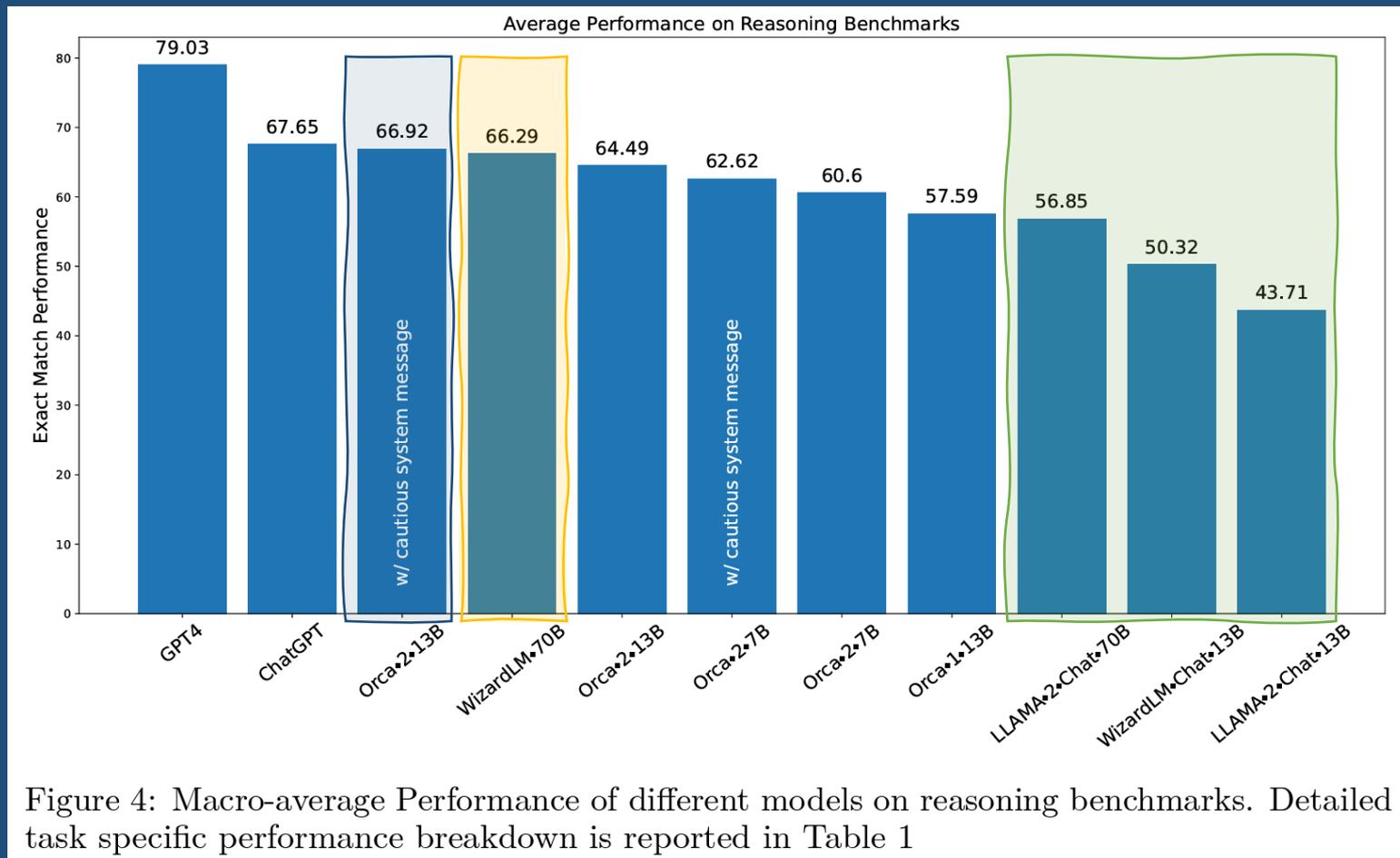
Determining a task-specific reasoning strategy,
generate task-specific system instructions



The instructions for the base model do not include task-specific system instructions.

Evaluation Results

- Surpassing models of the same size, competitive with models 5-10x larger



Evaluation Results

• Grounding

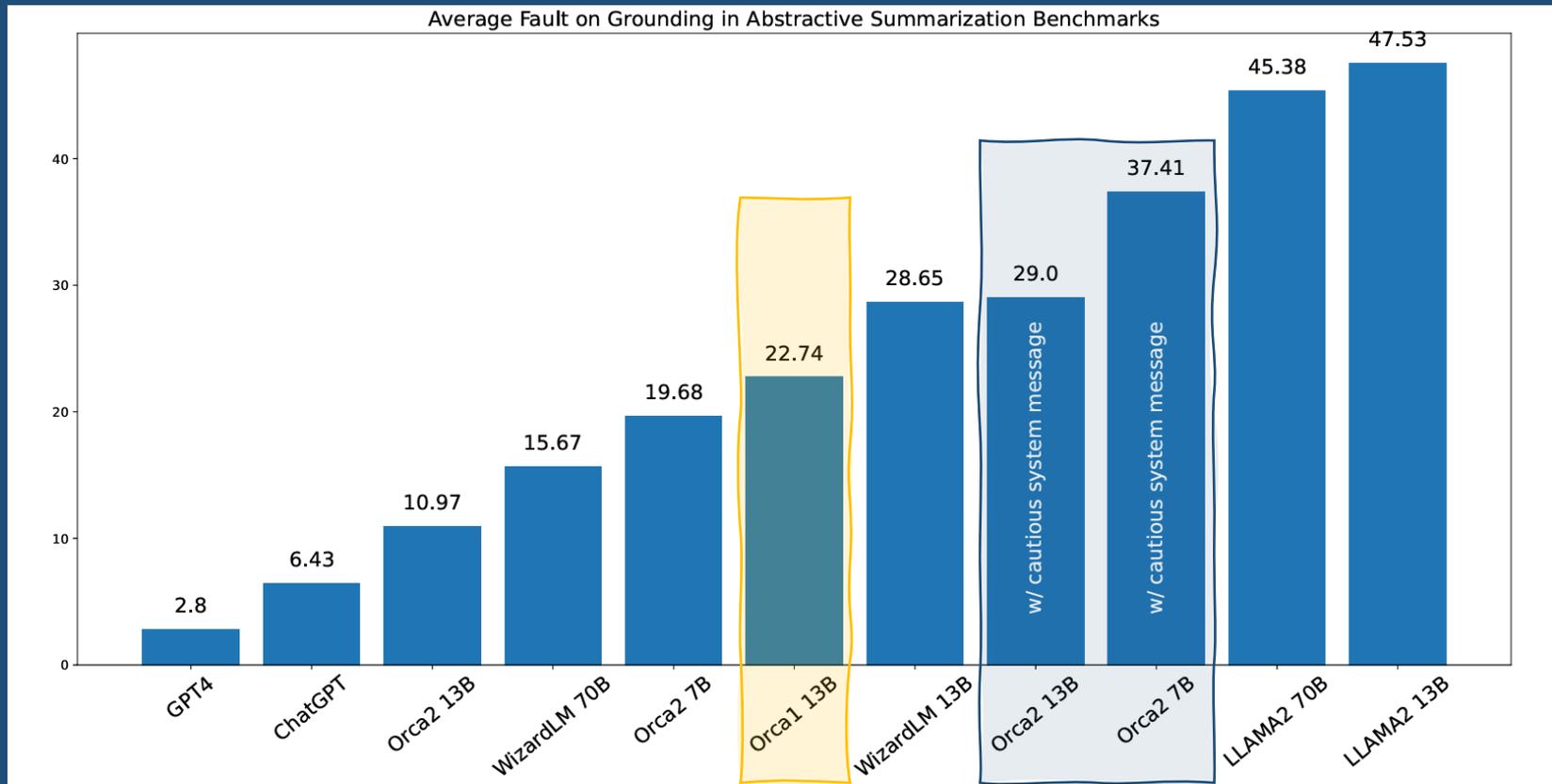


Figure 6: The hallucination rate evaluated by GPT-4 as discriminator averaged over three abstractive summarization benchmarks described in section 5 (the lower the better). Task specific performance breakdown is reported in Table 11.

Summary of Orca 2

Question:

John and Mark are in a room with a ball, a basket and a box. John puts the ball in the box, then leaves for work. While John is away, Mark puts the ball in the basket, and then leaves for school. They both come back together later in the day, and they do not know what happened in the room after each of them left the room. Where do they think the ball is?

• Example demonstrating the responses of Orca 2

- Orca-2-13B

Let's analyze the situation step by step:

1. John puts the ball in the box.
2. John leaves for work.
3. Mark comes into the room and puts the ball in the basket.
4. Mark leaves for school.
5. John and Mark both return to the room.

At this point, John and Mark have not seen each other's actions. Therefore, they only know what happened while they were in the room.

John only saw the ball in the box before he left for work. Mark only saw the ball in the basket before he left for school. Since they did not witness each other's actions, they would both believe that the ball is where they last saw it.

So, **John** would think the ball is in the **box**, and **Mark** would think the ball is in the basket.

Summary of Orca 2

Question:

John and Mark are in a room with a ball, a basket and a box. John puts the ball in the box, then leaves for work. While John is away, Mark puts the ball in the basket, and then leaves for school. They both come back together later in the day, and they do not know what happened in the room after each of them left the room. Where do they think the ball is?

• Example demonstrating the responses of Orca 2

- ChatGPT (GPT-3.5-Turbo)

John and Mark will likely think that **the ball is in the box**, since that is where John left it before he left for work. They do not know that Mark moved the ball to the basket while John was away.

- LLaMA2-Chat-13B

This is a classic lateral thinking puzzle. The solution is that the ball is still in the box.

Here's the reasoning:

1. John put the ball in the box before he left for work.
2. Mark put the ball in the basket before he left for school.
3. Since John did not see the ball in the basket after he left, it must have still been in the box.
4. Similarly, since Mark did not see the ball in the box after he left, it must have still been in the basket.

Therefore, **the ball is still in the box**.

Evaluation Results

- Multi-Turn Open Ended Conversations: MT-Bench

Model	Turn 1	Turn 2	Average
Orca-2-7B	6.14	5.15	5.65
w/ cautious sm	5.96	3.99	4.97
Orca-2-13B	6.69	5.60	6.15
w/ cautious sm	6.12	5.31	5.72
LLaMA-2-Chat-13B	7.17	6.11	6.64
WizardLM-13B	7.14	5.58	6.36
Orca-1-13B	6.66	5.19	5.92
LLaMA-2-Chat-70B	7.05	6.59	6.82
WizardLM-70B	8.07	7.45	7.76
ChatGPT	8.19	7.84	8.01
GPT-4	9.01	9.06	9.04

Table 3: MT-Bench scores per turn and average

- Orca-2-13B performs comparably with other 13B models.
- No instruction for multi-turn conversations

{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.

Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr