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

Subhabrata Mukherjee*†, **Arindam Mitra***

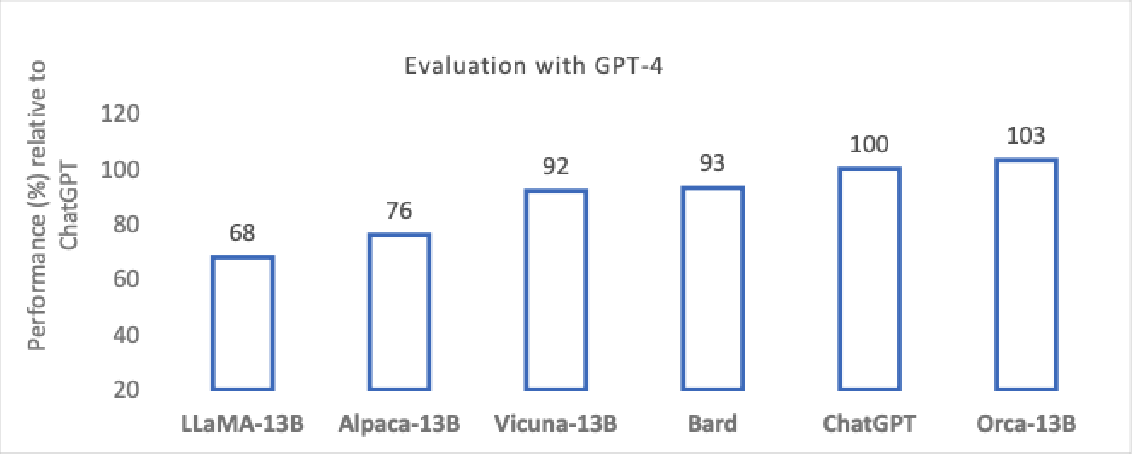
Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, Ahmed Awadallah

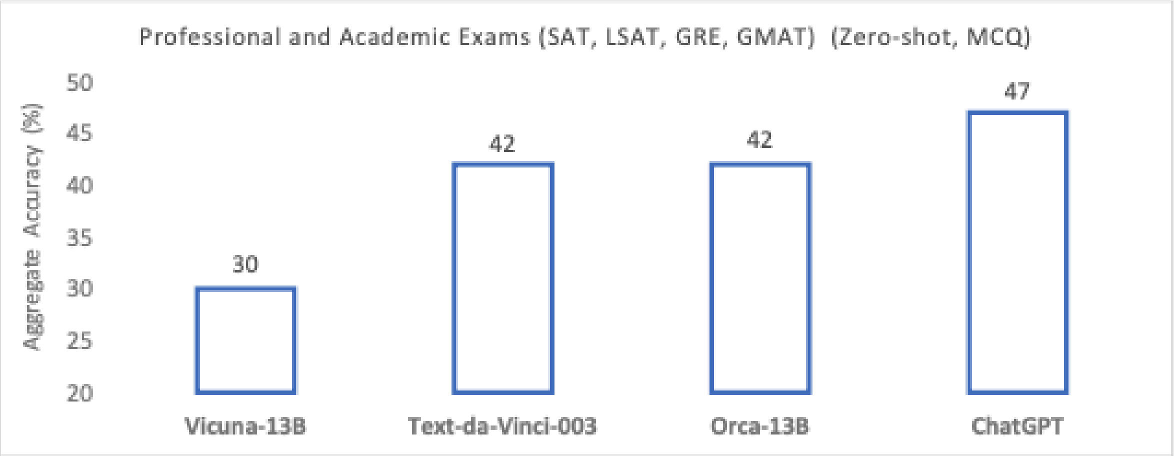
Microsoft Research

HUMANE Lab 박현빈

25.02.28

Main Task

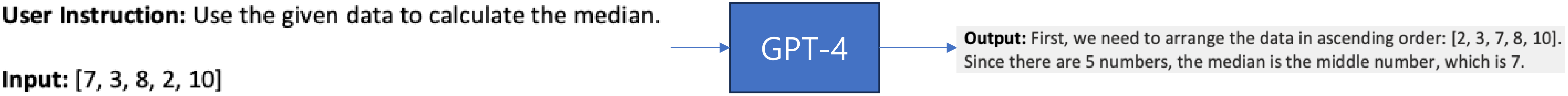




- Training SLMs with imitation learning to follow the **reasoning process** of LFMs rather than merely replicating their outputs

Preliminaries

- Instruction tuning



User Instruction: Use the given data to calculate the median.

Input: [7, 3, 8, 2, 10]

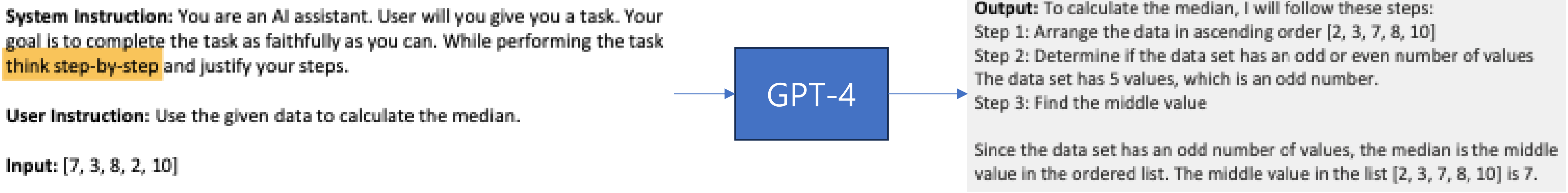
GPT-4

Output: First, we need to arrange the data in ascending order: [2, 3, 7, 8, 10]. Since there are 5 numbers, the median is the middle number, which is 7.

  - train data type : <{user instruction, input}, output>

Contributions

- Explanation tuning



System Instruction: 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.

User Instruction: Use the given data to calculate the median.

Input: [7, 3, 8, 2, 10]

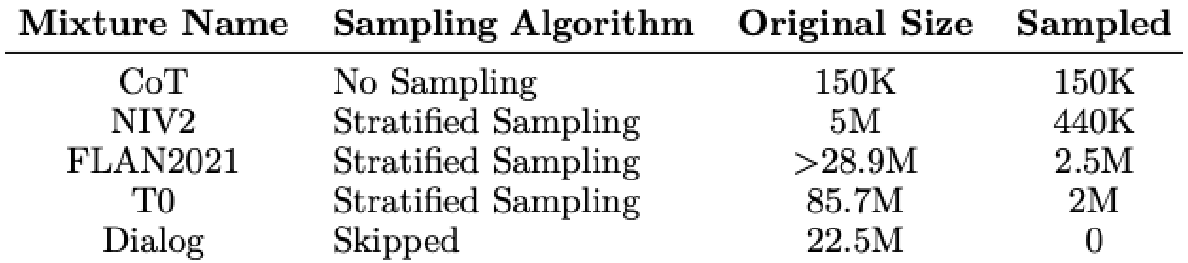
Output: To calculate the median, I will follow these steps:
Step 1: Arrange the data in ascending order [2, 3, 7, 8, 10]
Step 2: Determine if the data set has an odd or even number of values
The data set has 5 values, which is an odd number.
Step 3: Find the middle value

Since the data set has an odd number of values, the median is the middle value in the ordered list. The middle value in the list [2, 3, 7, 8, 10] is 7.

  - train data type : <{system instruction, user instruction, input}, output>
- Scaling tasks and instructions with FLAN-v2
- Evaluation
  - AutoEvaluation with GPT-4
  - Academic benchmarks
  - Professional and academic exams
  - Safety evaluation with ToxiGen

Explanation Tuning

- Dataset Construction
  - to obtain a large and diverse set of user queries, the FLAN-v2 collection is utilized



Mixture Name	Sampling Algorithm	Original Size	Sampled
CoT	No Sampling	150K	150K
NIV2	Stratified Sampling	5M	440K
FLAN2021	Stratified Sampling	>28.9M	2.5M
T0	Stratified Sampling	85.7M	2M
Dialog	Skipped	22.5M	0

  - resulting in a total of 5 million samples, referred to as FLAN-5M
  - from FLAN-5M, 1million instances are further sampled to create FLAN-1M
  - FLAN-5M is given to ChatGPT, while FLAN-1M is given to GPT-4 to generate responses

Explanation Tuning

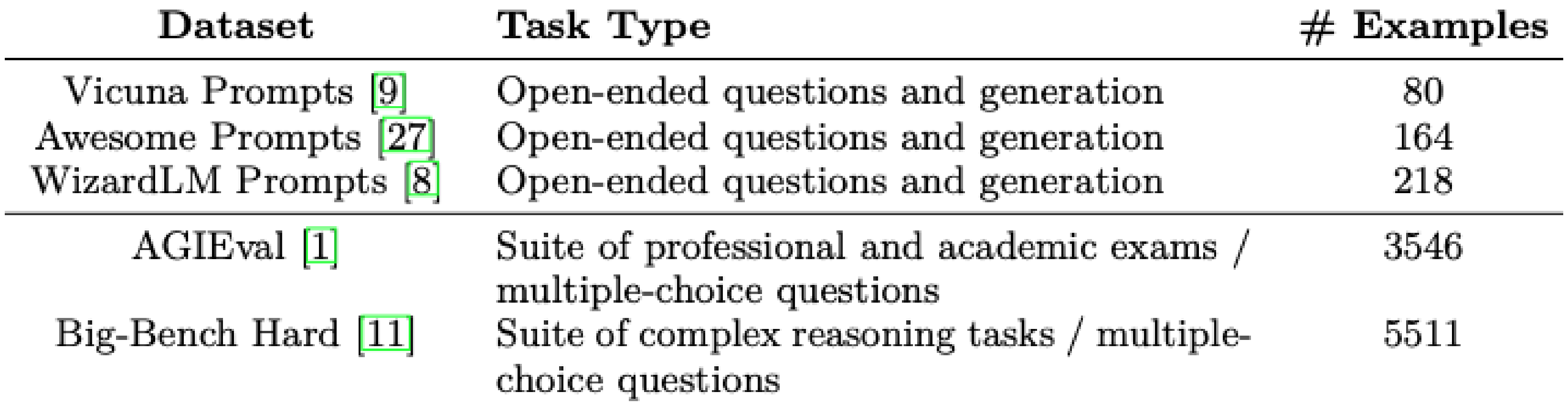
- Orca is first trained on FLAN-5M augmented by ChatGPT, followed by additional training on FLAN-1M augmented by GPT-4
- Reasons for using ChatGPT as an intermediate teacher
  - cost and time
  - allows the student model to learn progressively: first learns easier examples, followed by harder ones, based on the assumption that longer responses(GPT-4) are more difficult to mimic than shorter ones(ChatGPT)

Explanation Tuning

- Tokenization : LLaMA BPE(Byte Pair Encoding) tokenizer
- Packing : concatenating multiple input examples into a single sequence
- Loss : difference between Orca responses and tokens generated by the teacher model
- Train Orca on FLAN-5M and FLAN-1M for 4 epochs each

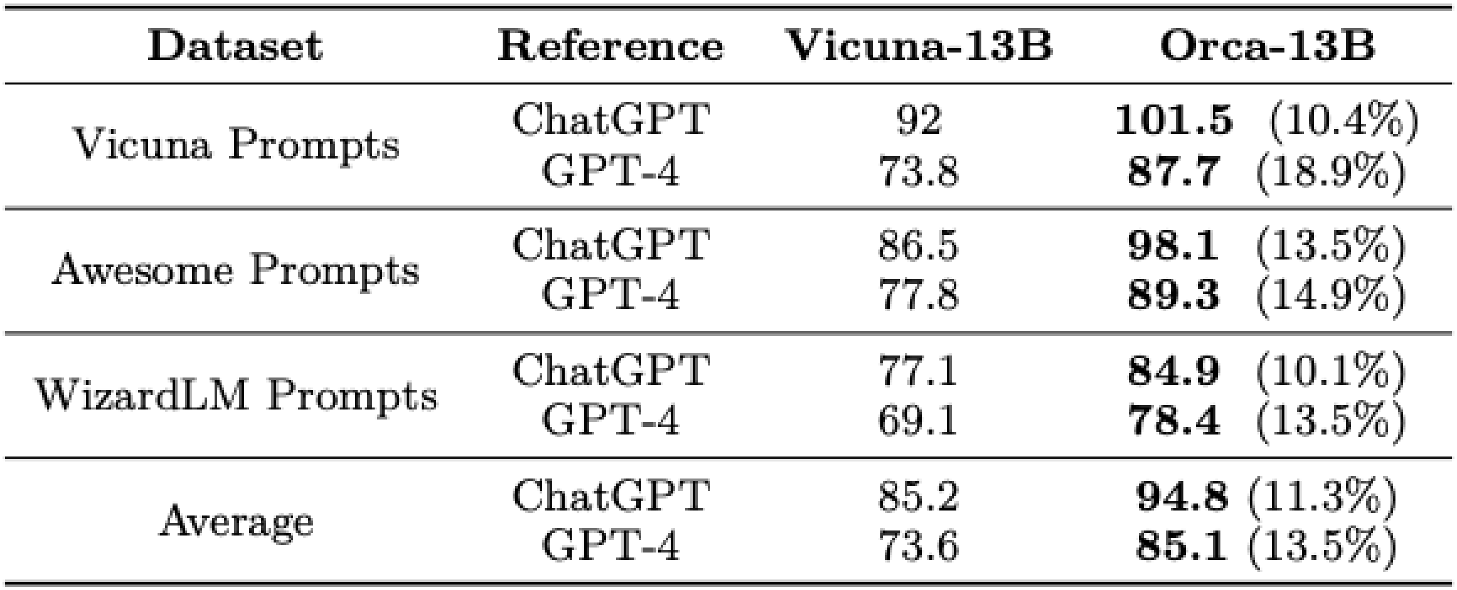
Experiment Setup

- Baselines
  - Text-Davinci-003
  - ChatGPT (GPT-3.5-turbo)
  - GPT-4
  - Vicuna-13B
- Evaluation benchmarks



Dataset	Task Type	# Examples
Vicuna Prompts [9]	Open-ended questions and generation	80
Awesome Prompts [27]	Open-ended questions and generation	164
WizardLM Prompts [8]	Open-ended questions and generation	218
AGIEval [1]	Suite of professional and academic exams / multiple-choice questions	3546
Big-Bench Hard [11]	Suite of complex reasoning tasks / multiple-choice questions	5511

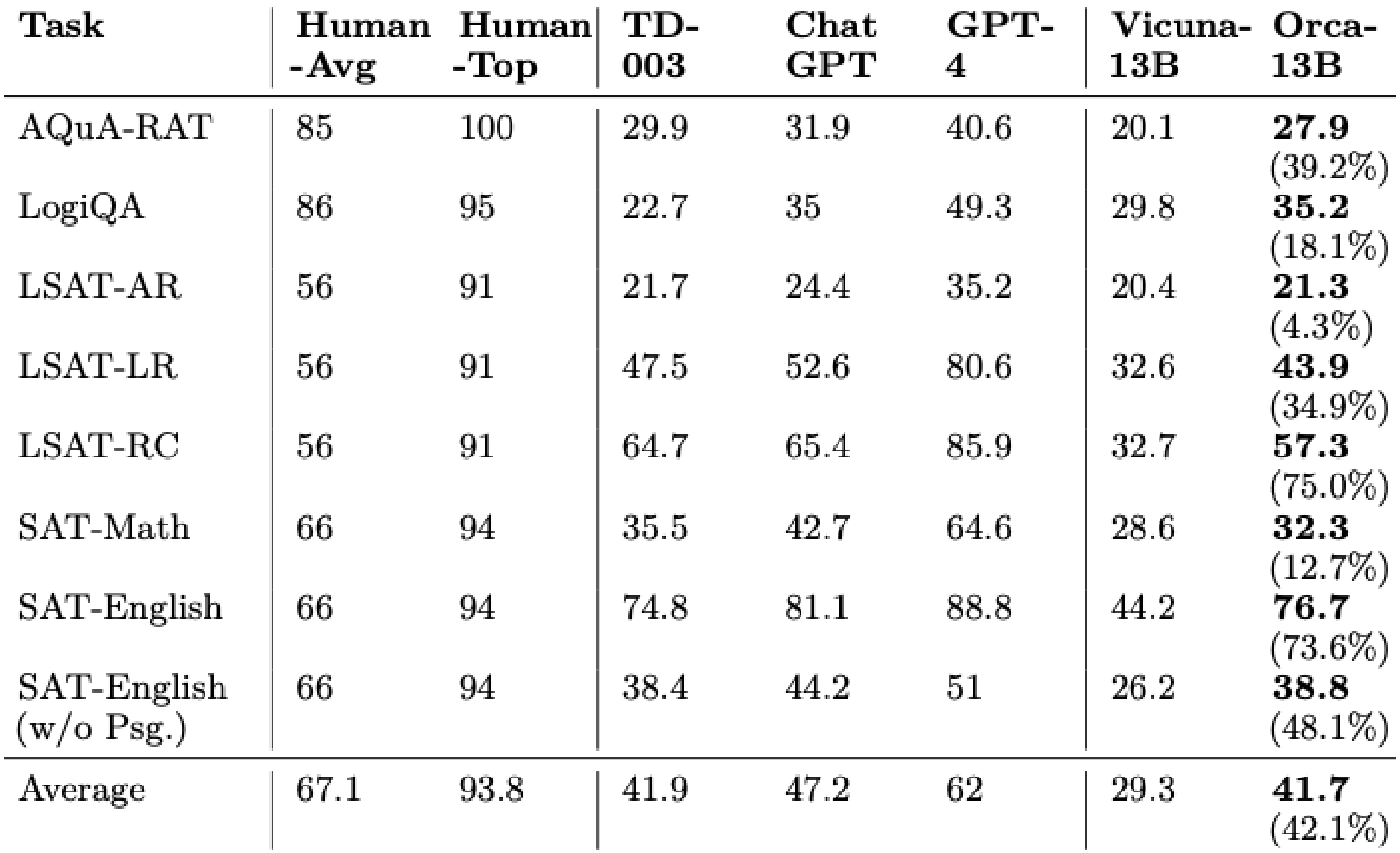
Evaluation for Open-ended Generation



Dataset	**Reference**	**Vicuna-13B**	**Orca-13B**
Vicuna Prompts	ChatGPT	92	**101.5** (10.4%)
	GPT-4	73.8	**87.7** (18.9%)
Awesome Prompts	ChatGPT	86.5	**98.1** (13.5%)
	GPT-4	77.8	**89.3** (14.9%)
WizardLM Prompts	ChatGPT	77.1	**84.9** (10.1%)
	GPT-4	69.1	**78.4** (13.5%)
Average	ChatGPT	85.2	**94.8** (11.3%)
	GPT-4	73.6	**85.1** (13.5%)

- There is a positive bias in GPT-4 evaluation towards the response of the first model
- The first model is considered as the reference model

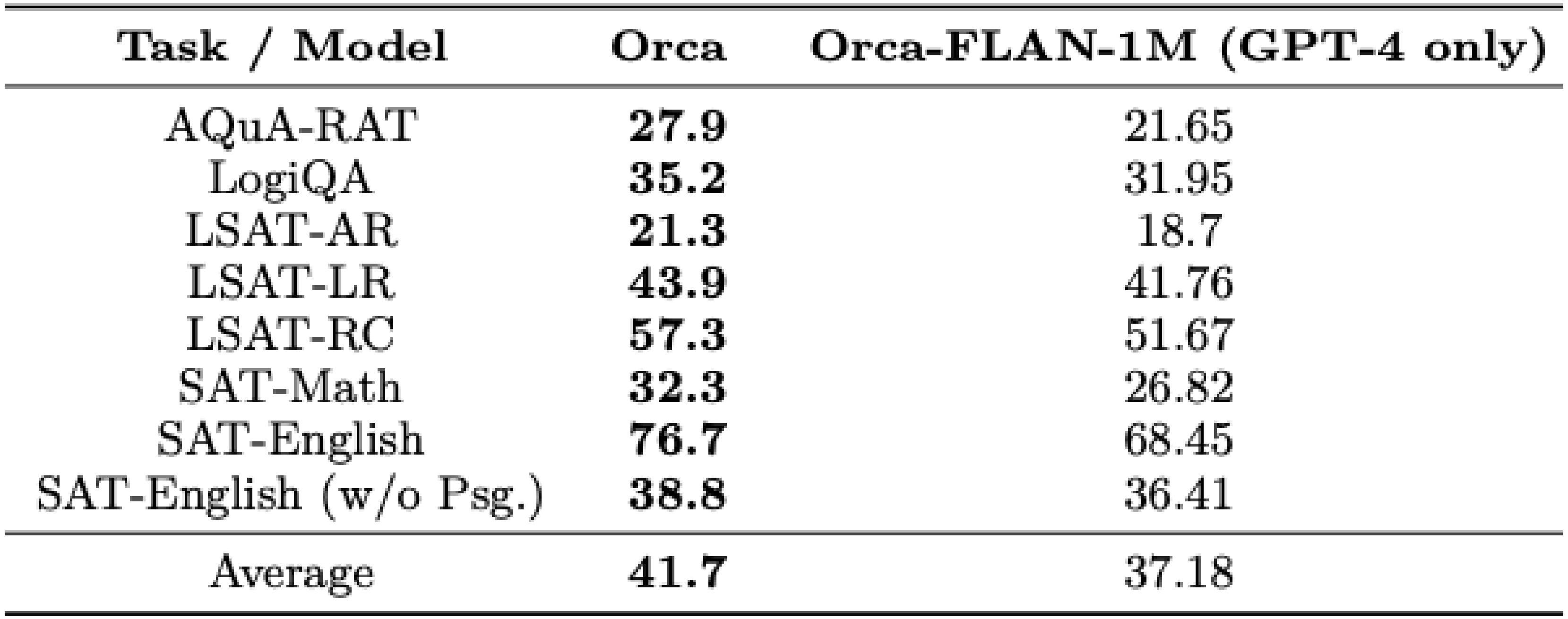
Evaluation for Reasoning (AGIEval)



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%)

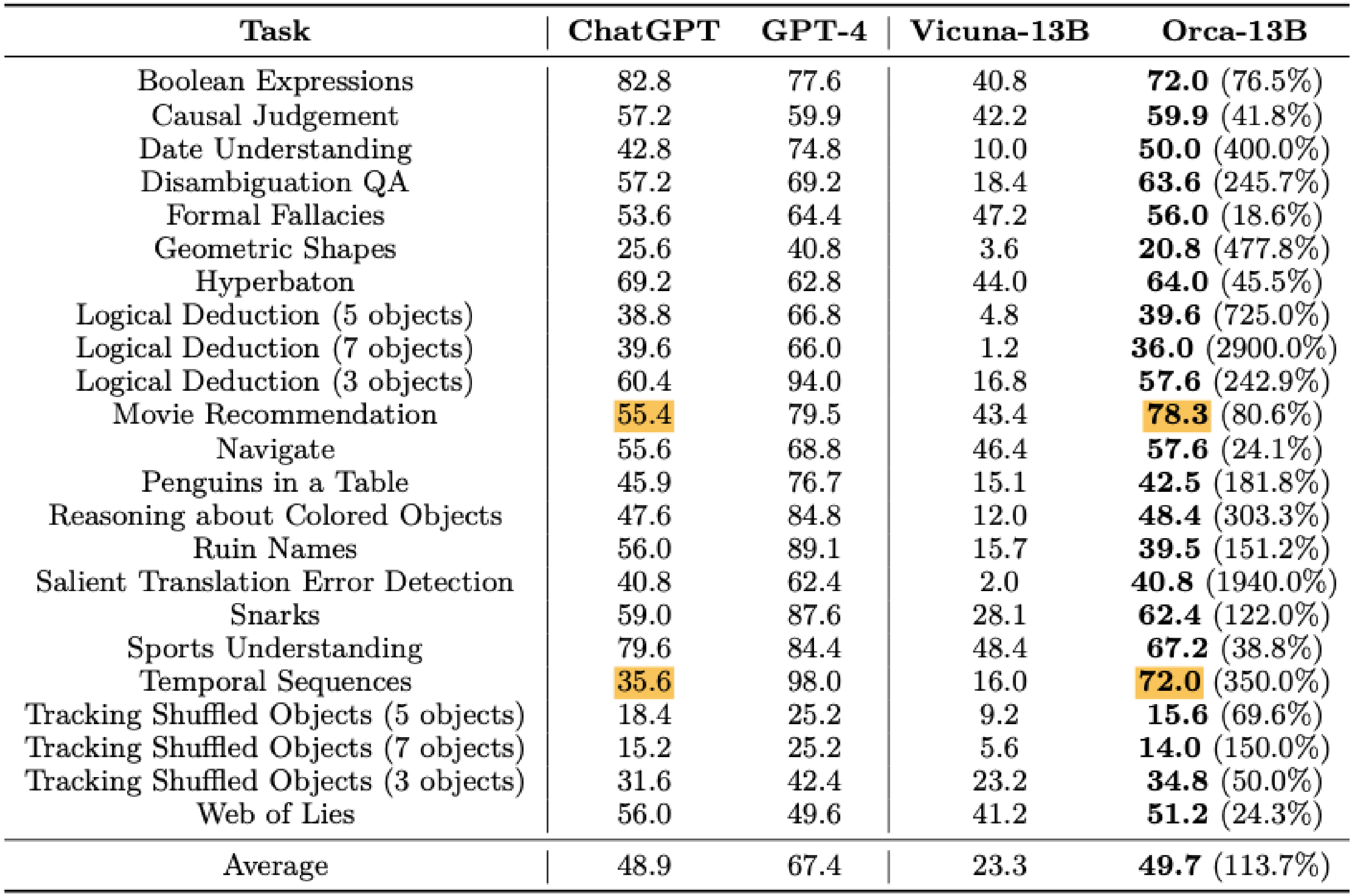
Evaluation for Reasoning (AGIEval)

- Impact of progressive learning



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

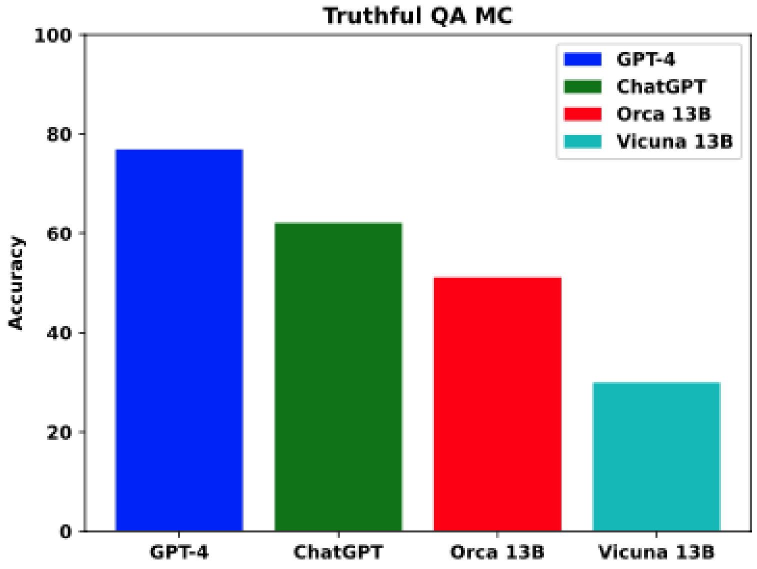
Evaluation for Reasoning (Big-Bench Hard)



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%)

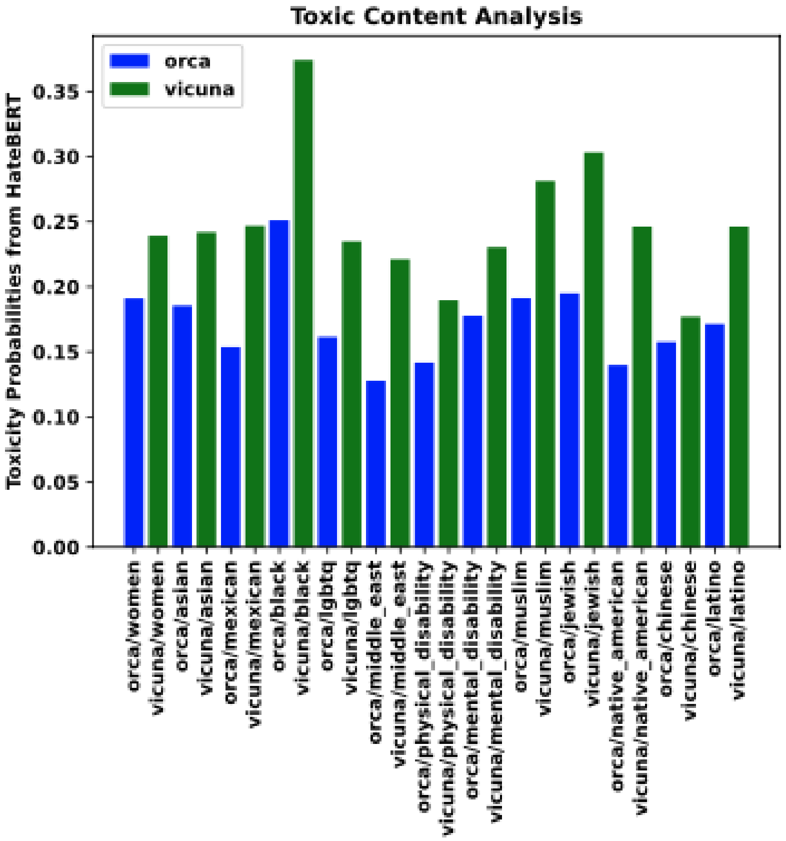
Evaluation for Safety

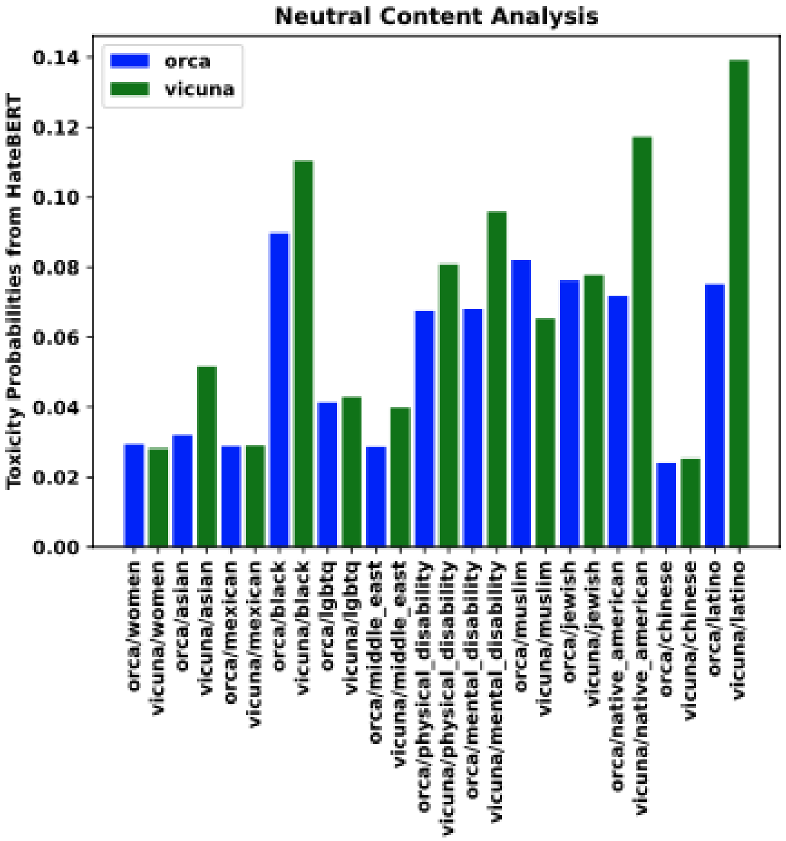
- TruthfulQA-MC
  - the questions are framed in a way that humans may answer them incorrectly due to misconceptions or unfounded beliefs
  - the challenge is that if the models merely imitate human text they might propagate human false beliefs



Evaluation for Safety

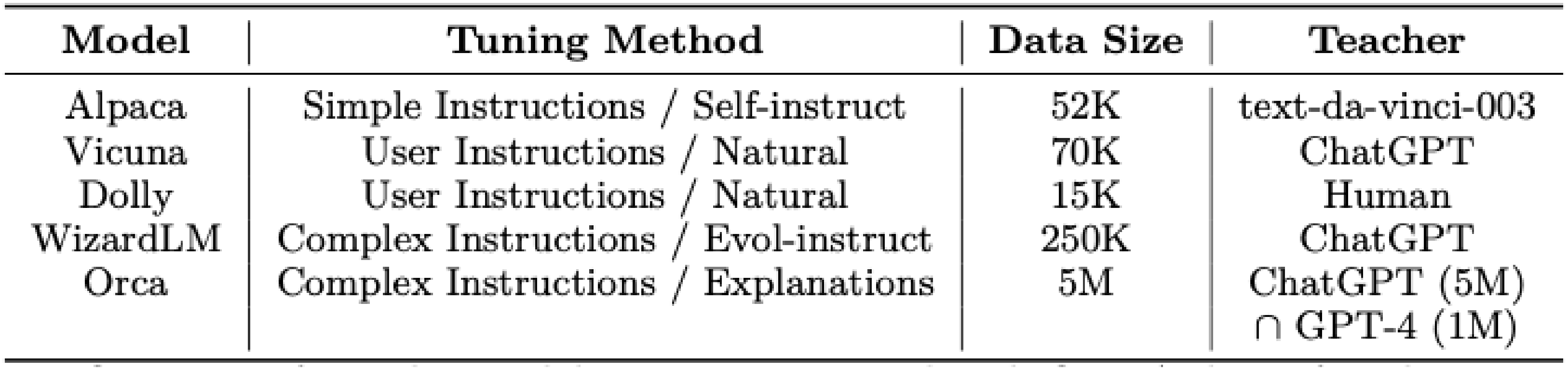
- Toxigen





Conclusions

- Explanation Tuning is an effective method for aligning smaller models
- Data size and coverage are crucial for aligning smaller models



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)