

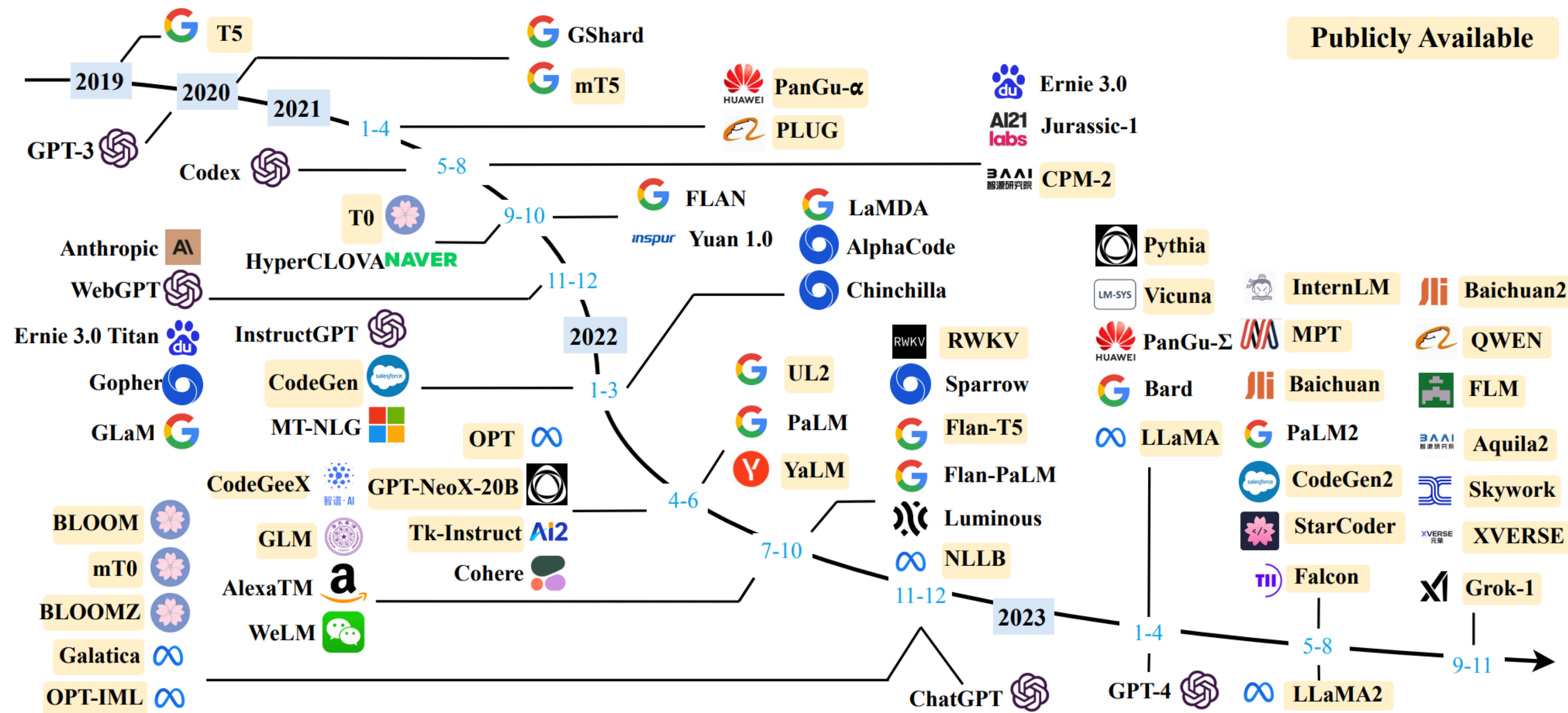
Does ChatGPT Know that It Does Not Know? Evaluating the Black-Box Calibration of ChatGPT COLING-2024

Youliang Yuan, Wenxuan Wang, Qingshuo Guo, Yiming Xiong, Chihao Shen, Pinjia He

School of Data Science, CUHK-Shenzhen

Background

- LLMs demonstrate impressive intelligence, driving the advancement of Artificial Intelligence
- A question: does ChatGPT know that it does not know?



Background

- Does ChatGPT itself know that it does not know under the black-box setting? For example
 - Give a reliable confidence for its answer
 - Correctly determine whether its answer is correct

Q: What is the remainder when 23 is divided by 4? ← Prompt
A: 3 ← Answer generated by GPT3 (greedy decoding)
Confidence: Medium ← Confidence generated by GPT3 (greedy decoding)

```
Question: Who was the first president of the United States?  
Proposed Answer: George Washington  
Is the proposed answer:  
  (A) True  
  (B) False  
The proposed answer is:
```

(Kadavath et al. 2022) Language Models (Mostly) Know What They Know.

(Lin et al. 2022) Teaching models to express their uncertainty in words..

Background

- To answer the previous question:
 - we propose three types of proxy confidence
 - evaluate the correlation intensity between these confidence and model accuracy

Proxy Type	Prompt	Output	Confidence
Quantitative	answer question and give your confidence (%): [query]	[ans], confidence is $c\%$	$c\%$
Qualitative	[hist]; is the answer true or false?	true / false	high / low
Consistent	<div><div>[query₁]; ... ; [query_n]</div><div>[hist]; please think again.</div></div>	<div><div>[ans₁]; ... ; [ans_n]</div><div>[ans_{new}]</div></div>	consistency

User: Answer the question and give your confidence (%): Please tell me 1+1=?

ChatGPT: 1+1=2, Confidence: 100%

Quantitative

User: Please tell me 1+1=?

ChatGPT: 1+1=2

User: Is the answer true or false?

ChatGPT: The answer 1+1=2 is true.

Qualitative

User: Please tell me 1+1=?

ChatGPT: 1+1=2

User: Please tell me 1+1=?

ChatGPT: 1+1=2

User: Please tell me 1+1=?

ChatGPT: 1+1=2

Consistent

User: Please tell me 1+1=?

ChatGPT: 1+1=2

User: Please think again.

ChatGPT: I apologize for the mistake. 1+1=11

Consistent

Experiment

- Dataset
 - TruthfulQA, MMLU, Modified Arithmetic, Analytic Entailment, Language Identification
- Metric
 - Expected Calibration Error (ECE), Pearson Correlation Coefficient (PCC)
 - In addition, we propose Monotonicity Score (MS) to quantify the degree to which the accuracy increases as the confidence increases

$$\text{MS} = \frac{1}{Z} \sum_{i=1}^{N-1} \frac{n_{i+1} + n_i}{2} \text{sign}(acc_{i+1} - acc_i),$$

where N is the number of the confidence intervals and acc_i the average accuracy for confidence interval i (large i refer high confidence).

Experiment: Quantitative Confidence

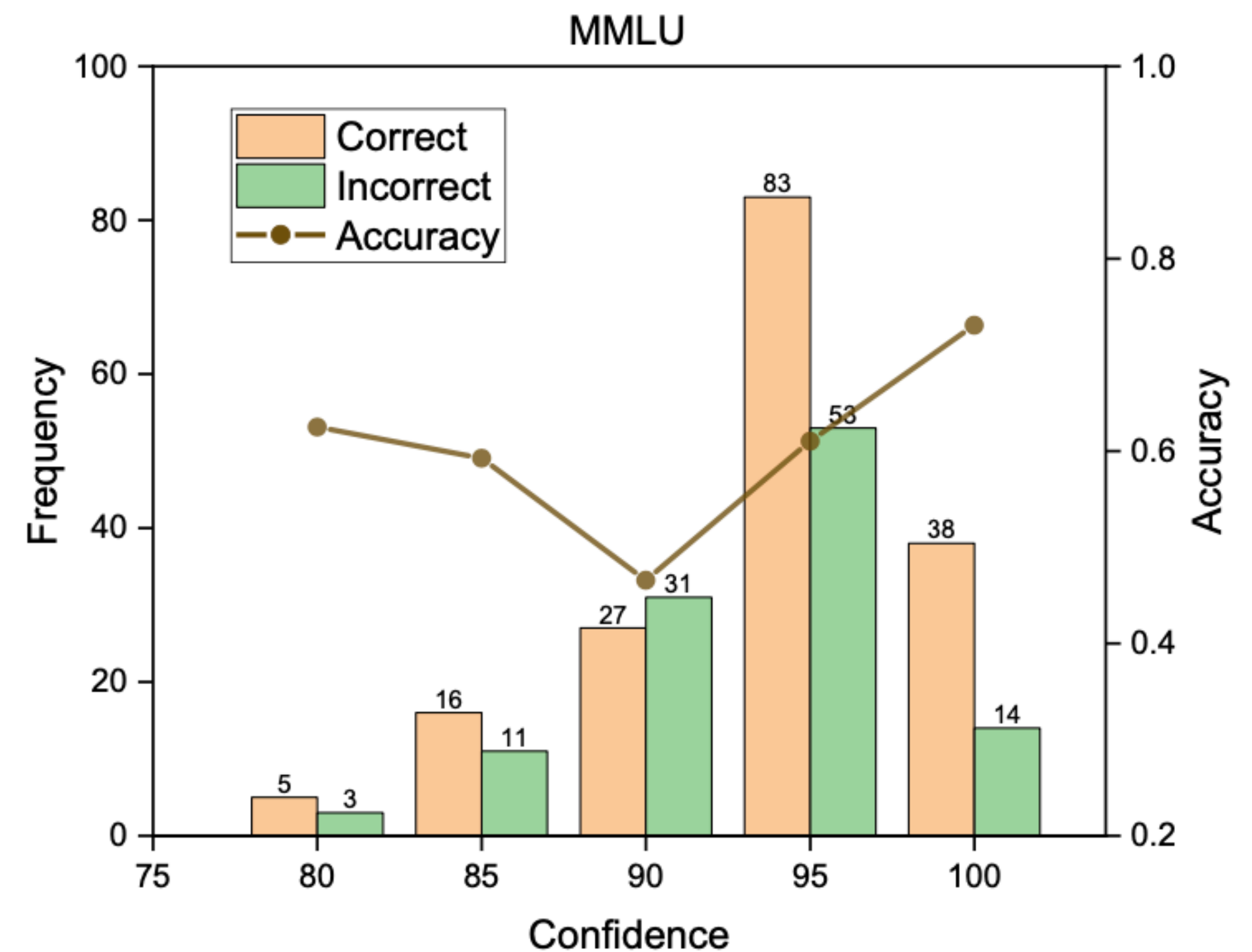
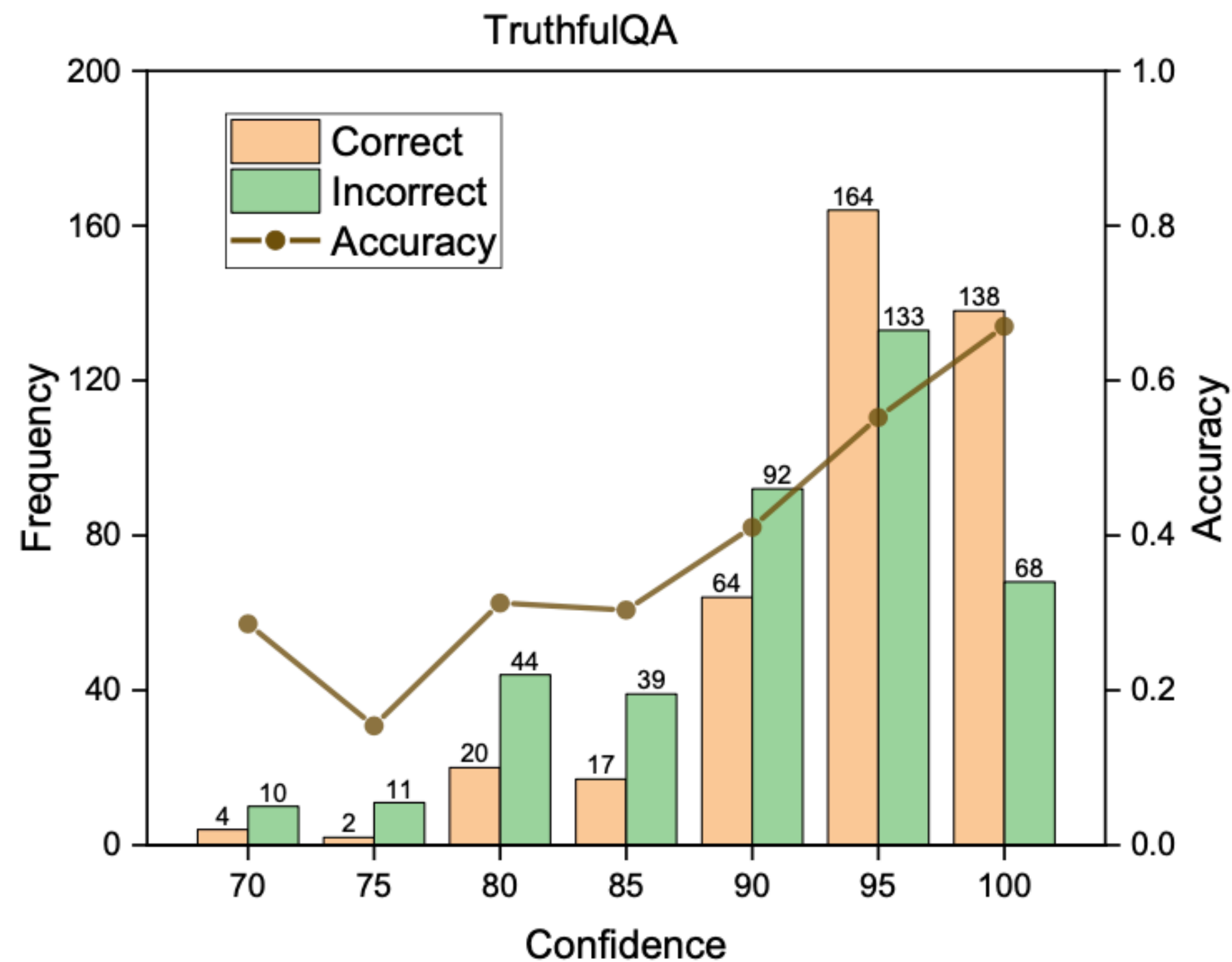
- Evaluate ChatGPT's ability to determine the accuracy of its answers based on quantitative confidence

Dataset	UnderConf	OverConf	ECE	Dataset	PCC	MS	Avg.Con	Acc
TruQA	0.10	41.95	42.05	TruQA	95.16	78.91	92.49	50.49
LangId	0.00	52.70	52.70	LangId	74.97	63.73	89.18	36.48
MMLU	0.11	33.39	33.50	MMLU	54.47	52.44	93.36	60.07
AnaEnt	0.29	36.14	36.43	AnaEnt	87.95	35.51	96.42	55.65
ModAr	0.75	50.12	50.87	ModAr	-33.45	-86.96	96.73	47.28
Average	0.25	42.86	43.11	Average	55.82	28.73	93.64	49.99

- Conclusion:
 - ChatGPT demonstrates a significant tendency towards overconfidence
 - ChatGPT appears to know it is uncertainty but not always.
 - Hight positive correlation (PCC, MS) with four datasets

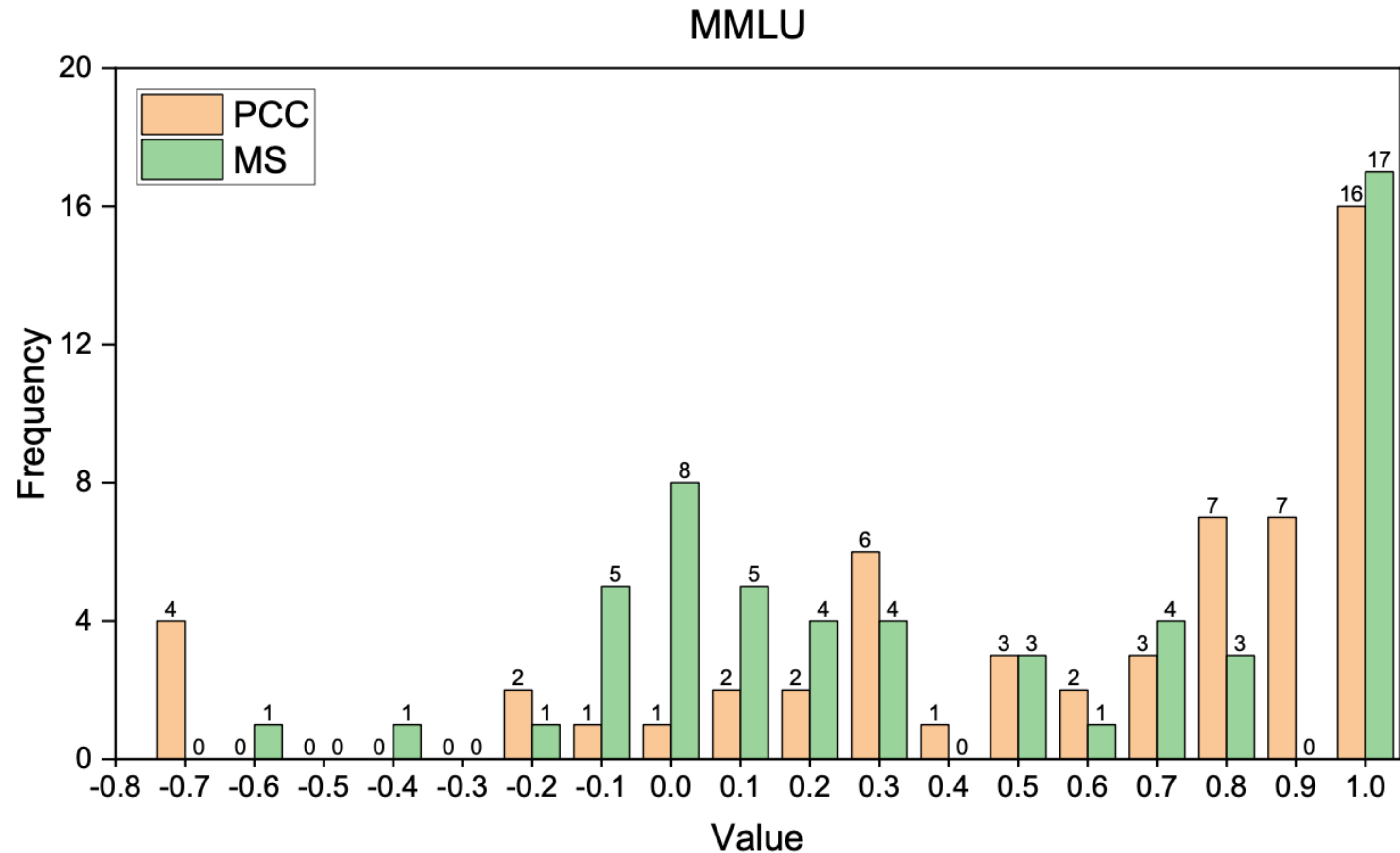
Experiment: Quantitative Confidence

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on quantitative confidence



Experiment: Quantitative Confidence

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on quantitative confidence



- The distribution of 57 tasks in the MMLU across various PCC and MS intervals.

Experiment: Qualitative Confidence

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on qualitative confidence

Table 4: The "True" response proportion across correct, incorrect, and overall subsets.

Dataset	Correct	Incorrect	Overall
TruQA	97.24	97.98	97.54
MMLU	98.85	99.07	92.76
ModAr	90.37	90.00	90.18
AnaEnt	100.00	100.00	100.00
LangId	91.75	99.33	96.34
Average	94.97	97.83	95.36

User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

User: Is the answer true or false?

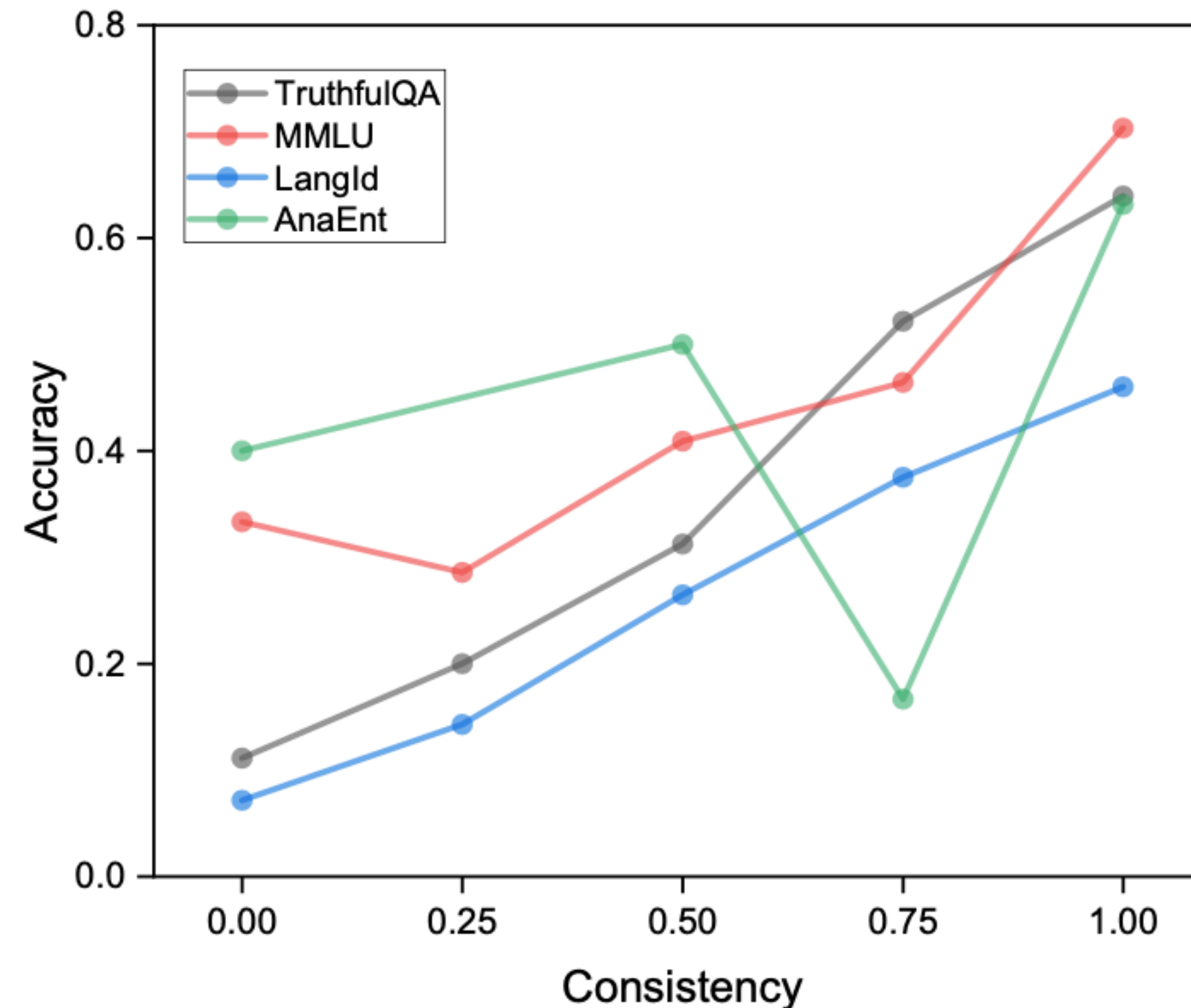
ChatGPT: The answer $1+1=2$ is true.

Qualitative

- Conclusion:
 - ChatGPT always think its answer is correct.

Experiment: Repetition Consistency

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on repetition consistency



User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

Consistent

- Conclusion:
 - Higher consistency means higher accuracy.

Experiment: Think Again Consistency

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on think again consistency
- Flip means the model change its answer after thinking again

Dataset	Acc (Flip)	Acc (Non-flip)	Flip Rate
TruQA	47.21	95.12	82.77
MMLU	55.64	96.00	91.13
AnaEnt	62.32	100.00	98.57
LangId	44.29	100.00	85.37
Average	48.41	87.64	79.48

User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

User: Please think again.

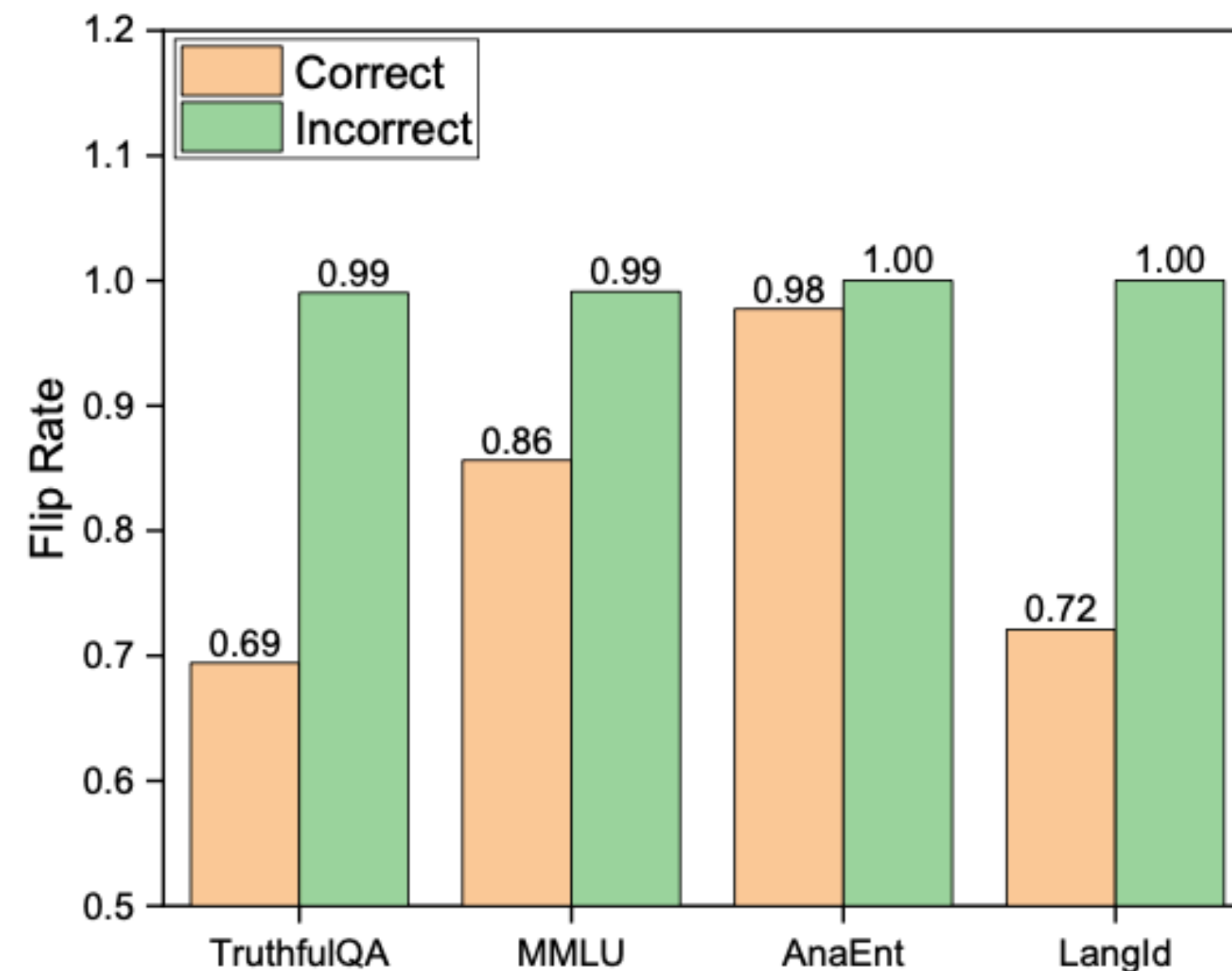
ChatGPT: I apologize for the mistake. $1+1=11$

Consistent

- Conclusion:
 - Generally, ChatGPT tends to change its answer (avg 79.48%)
 - Higher consistency (Non-flip) means higher accuracy

Experiment: Think Again Consistency

- Evaluate ChatGPT's ability to determine the accuracy of its answers based on think again consistency
- Flip means the model change its answer after thinking again



User: Please tell me $1+1=?$

ChatGPT: $1+1=2$

User: Please think again.

ChatGPT: I apologize for the mistake. $1+1=11$

Consistent

- Conclusion:
 - Generally, ChatGPT tends to change its answer (avg 79.48%)
 - Higher consistency (Non-flip) means higher accuracy
 - If its answer is wrong, it will be changed with a extremely high probability

Discussion: negative correlation in ModAr

- Why the model shows high confidence for incorrect answer in ModAr?
- We first split the ModAr dataset into two parts: normal and overriding

ModAr	Normal	Overriding
Query	<p>In the following lines, the symbol -> represents a simple mathematical operation.</p> <p>355 + 927 -> 1282 694 + 206 -> 900 767 + 234 -> 1001 441 + 354 -> 795 514 + 207 -> 721 708 + 852 -></p>	<p>In the following lines, the symbol -> represents a simple mathematical operation.</p> <p>980 + 605 -> 1586 660 + 623 -> 1284 566 + 848 -> 1415 935 + 466 -> 1402 627 + 142 -> 770 200 + 179 -></p>
Response	<p>ANSWER: 1560. The operation 708 + 852 results in 1560. This is obtained by adding the two numbers together. Percentage of confidence: 100%</p>	<p>ANSWER: 379 . The last mathematical operation was 200 + 179, which equals 379. My confidence in this answer is 100%. (ground-truth: 380)</p>

Discussion: negative correlation in ModAr

- Why the model shows high confidence for incorrect answer in ModAr?

The number of correct/incorrect predictions on normal data and overriding data.

Right / Wrong	Conf=90	Conf=95	Conf=100
Normal	11 / 0	29 / 0	91 / 0
Overriding	0 / 4	0 / 21	0 / 113
with CoT	0 / 2	0 / 6	0 / 120
with Expl	0 / 3	2 / 4	20 / 109
ModAr	11 / 4	29 / 21	91 / 113

- Conclusion:
 - The model tends to be excessively confident in its answers when it possesses strong prior knowledge, as it achieve almost 100% accuracy on normal data while 0% on overriding data

Discussion

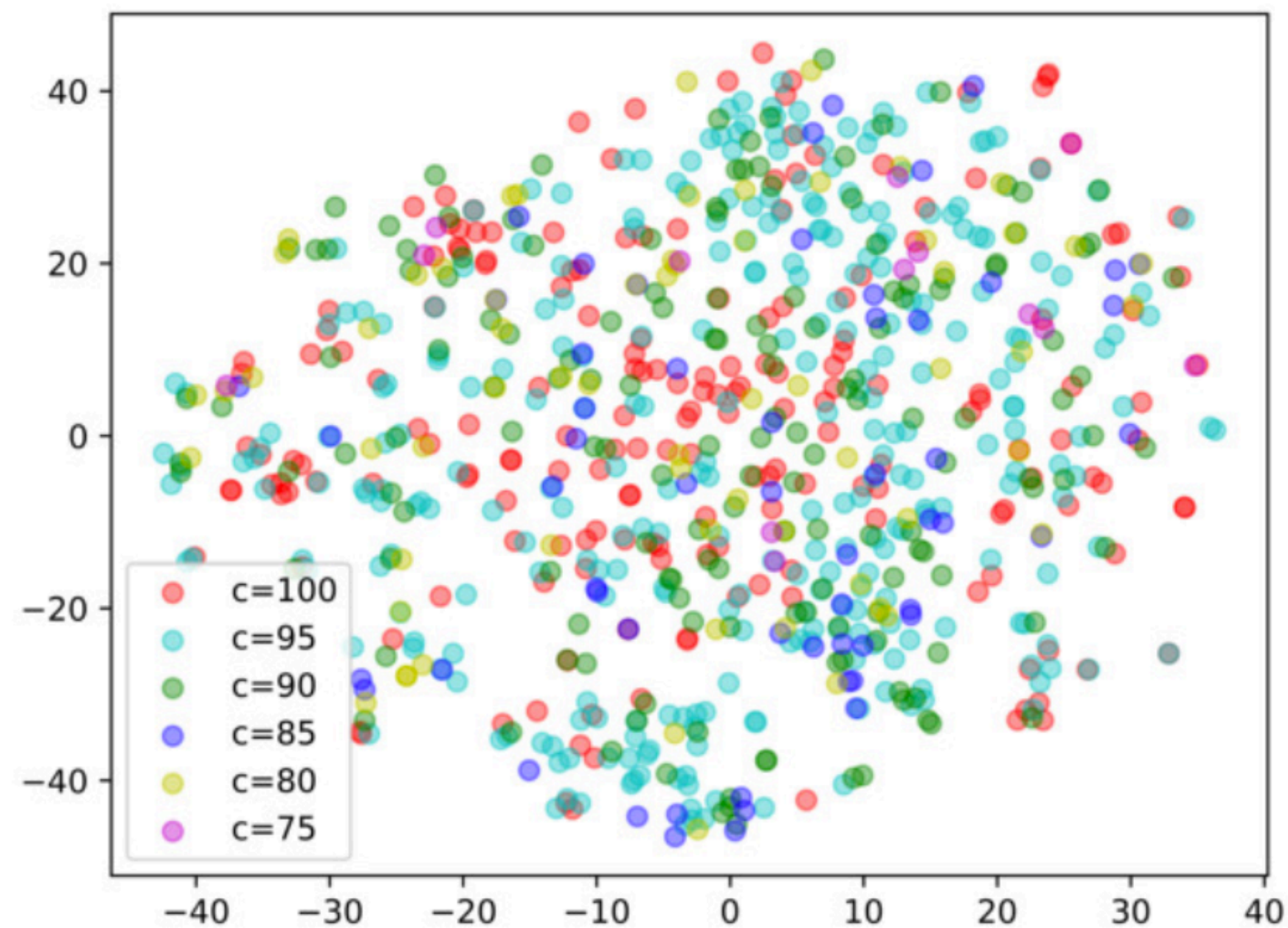
- Does ChatGPT solely utilize some simple heuristics (e.g. options number) in questions to generate confidence?

Heuristic / Group	Conf=100	Conf=95	Conf=90	Conf=85	Conf=80
Option Number	4.67	5.21	5.28	5.66	5.27
Perplexity	10.72	8.57	9.11	7.92	8.71
Length (TruQA)	274	338	320	355	318
Length (H.a)	155	193	171	204	167
Length (P.m)	-	733	790	731	705

- Conclusion:
 - The calibration capability can not be explained by heuristics we provide.

Discussion

- Does ChatGPT solely utilize some simple heuristics (e.g. options number) in questions to generate confidence?



Task & Difficulty (Conf/Acc)	Elementary	High School	College	Professional
Biology	-	94.75/81.82	94.25/71.00	-
Chemistry	-	93.47/58.16	93.35/46.39	-
Computer Science	-	94.10/66.00	93.93/41.84	-
Mathematics	96.60/45.00	95.73/34.38	94.90/35.35	-
Medicine	-	-	92.95/65.00	90.20/73.74
Physics	-	93.60/36.00	93.43/43.43	-
Psychology	-	92.95/83.00	-	91.17/68.37
Average	96.60/45.00	94.10/59.89	93.80/50.50	90.69/71.06

- Conclusion:
 - The calibration capability can not be explained by heuristics we provide.

Discussion

- Can ChatGPT's capability be attributed to techniques in the training process?

Model	PCC	MS	Avg.Con	Acc
Alpaca(7B)	-16.37	15.89	82.32	18.89
Davinci(175B)	-20.69	25.10	87.97	28.98
ChatGPT	95.16	78.91	92.49	50.49

- Conclusion:
 - This capability may be implicitly learned in reinforcement learning

Summary

- We examined ChatGPT's black-box calibration capability and proposed three types of proxy confidence: quantitative confidence, qualitative confidence, and answer consistency
- Although the quantitative confidence metric showed a strong positive correlation with accuracy in some datasets, this was not the case for all datasets
- Our analysis indicates that ChatGPT does not learn this ability from simple heuristics but may learn it implicitly through reinforcement learning