NEURAL INFORMATION PROCESSING SYSTEMS

Project Homepage
https://statqa.github.io/

Are Large Language Models Good Statisticians?

Yizhang ZHU[1], Shiyin DU[1], Boyan LI[1], Yuyu LUO[1,2]✉, Nan TANG[1,2]

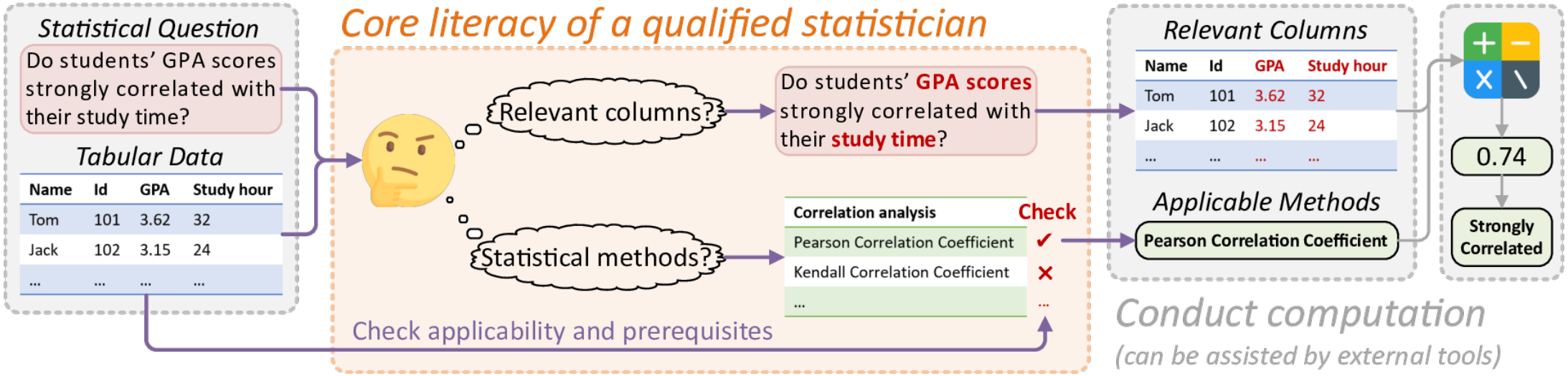
[1]The Hong Kong University of Science and Technology (Guangzhou),
[2]The Hong Kong University of Science and Technology

✉ Corresponding author: Yuyu Luo (yuyuluo@hkust-gz.edu.cn)

38th Conference on Neural Information Processing Systems

Motivation

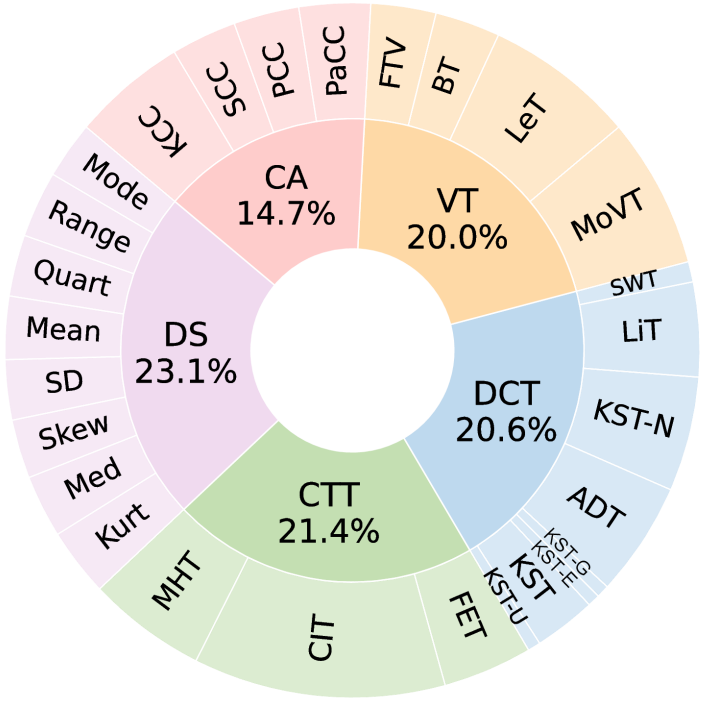
Statistical Analysis



Do LLMs truly understand such "statistical literacy"?

- How can we evaluate LLMs' performance in more complex and specialized statistical testing tasks?
- How capable are current LLMs in this field, and how can we improve their performance?
- How do humans perform compared to LLMs, and what are the differences in their performance?

Statistics of StatQA



- DS (Task): Descriptive Statistics (2690)
- CTT (Task): Contingency Table Test (2491)
- DCT (Task): Distr. Compliance Test (2400)
- VT (Task): Variance Test (2330)
- CA (Task): Correlation Analysis (1712)
- Mode: Mode
- Range: Range
- Quart: Quartile
- Mean: Mean
- SD: Standard Deviation
- Skew: Skewness
- Med: Median
- Kurt: Kurtosis
- MHT: Mantel-Haenszel Test
- CIT: Chi-square Independence Test
- FET: Fisher Exact Test
- KST-U: Kolmogorov-Smirnov Test for Uni. distr.
- KST: Kolmogorov-Smirnov Test
- KST-G: Kolmogorov-Smirnov Test for Gam. distr.
- KST-E: Kolmogorov-Smirnov Test for Exp. distr.
- ADT: Anderson-Darling Test
- KST-N: Kolmogorov-Smirnov Test for Normality
- LiT: Lilliefors Test
- SWT: Shapiro-Wilk Test of Normality
- MoVT: Mood Variance Test
- LeT: Levene Test
- BT: Bartlett Test
- FTV: F-Test for Variance
- PaCC: Partial Correlation Coefficient
- PCC: Pearson Correlation Coefficient
- SCC: Spearman Correlation Coefficient
- KCC: Kendall Correlation Coefficient

Table 1: Statistics of StatQA

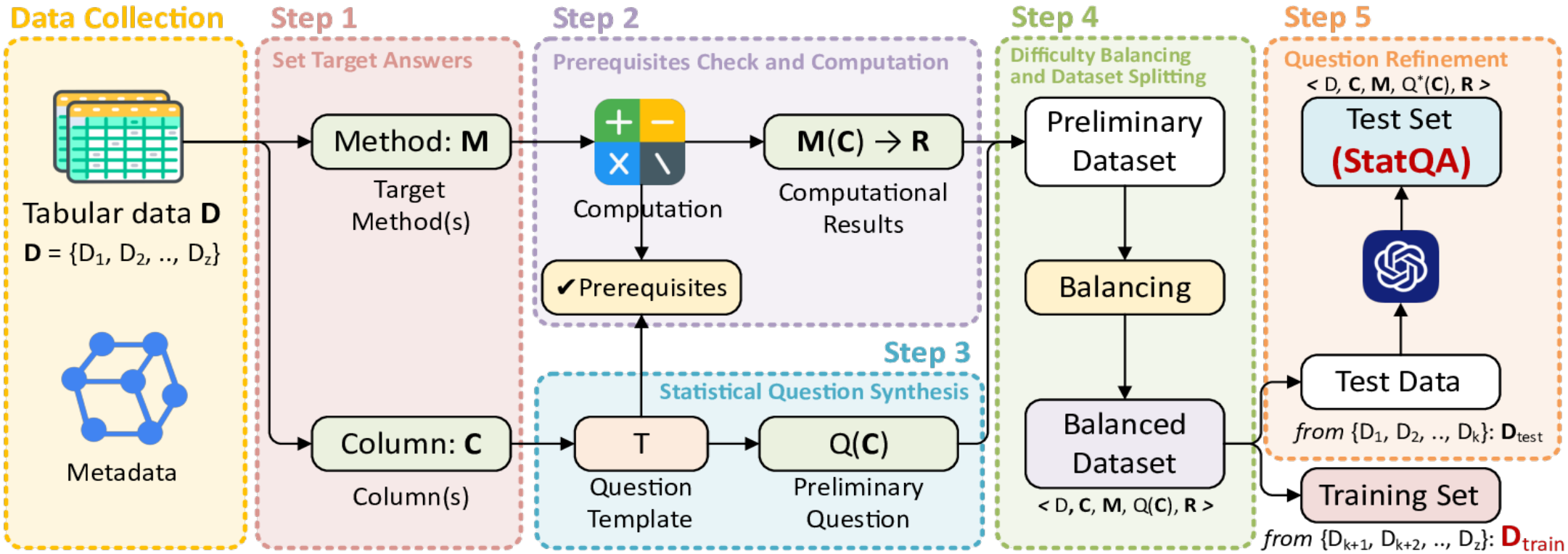
Item	Tabular Data		Question Length (Chars)			Difficulty		#-Examples	
	Avg #-Rows	Avg #-Cols	Max	Min	Avg	Easy	Hard	StatQA	mini-StatQA
Stats	6,228	14	346	21	113	7,401	4,222	11,623	1,163

Conventional Benchmark Construction

- Collect suitable dataset *D*
- Formulate question *Q*
- Manually annotate answer *A*

→ High-quality
Time-consuming
Limited extensibility

StatQA Construction: Reversed Pipeline



Experimental Protocols

LLM Experimental Settings

Models:

- Open-source LLMs: LLaMA-2-7b/13b-chat-hf, LLaMA-3-8b, LLaMA-3-8b-Instruct
- Proprietary LLMs: ChatGPT, GPT-4, GPT-4o

Strategies:

- Few-shot, CoT, domain-knowledge prompting, fine-tuning (for open-source LLMs)

Human Experimental Settings

Grouping:

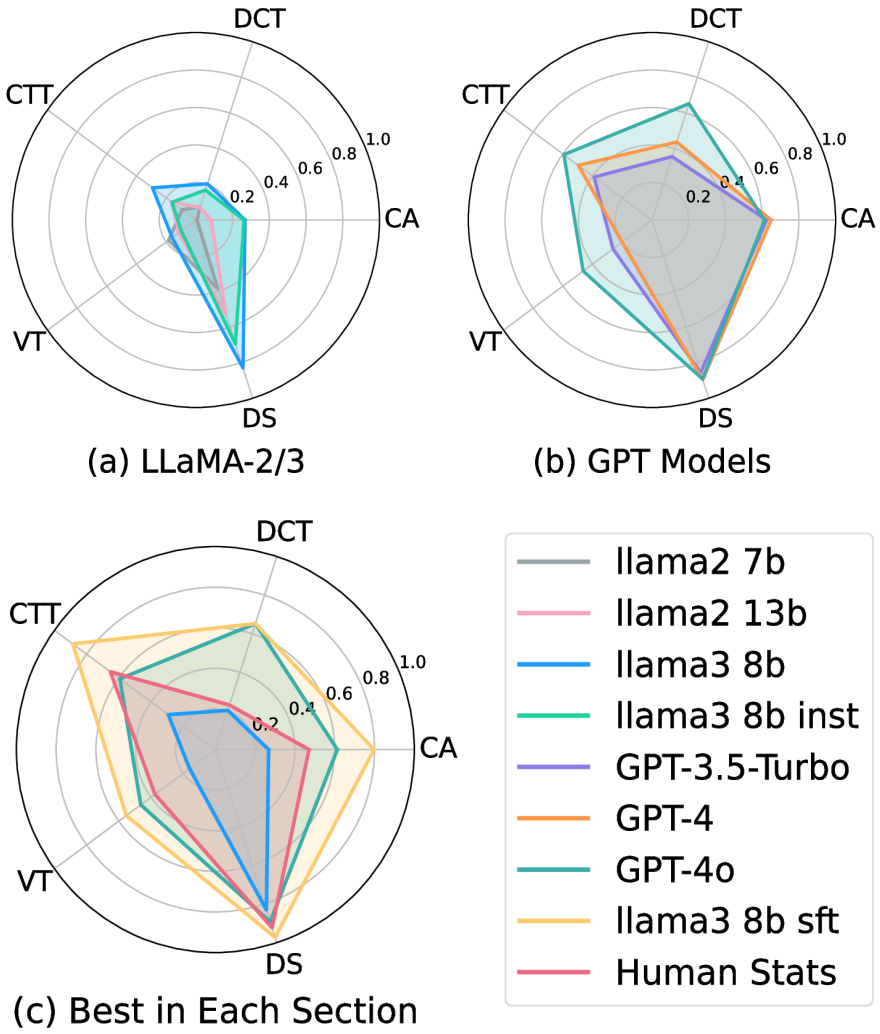
- Non-Stat: 3 STEM PG-students without a statistical background
- Stat: 3 PG-students in statistics major

Mode:

- Closed-book, Open-book

Experimental Results

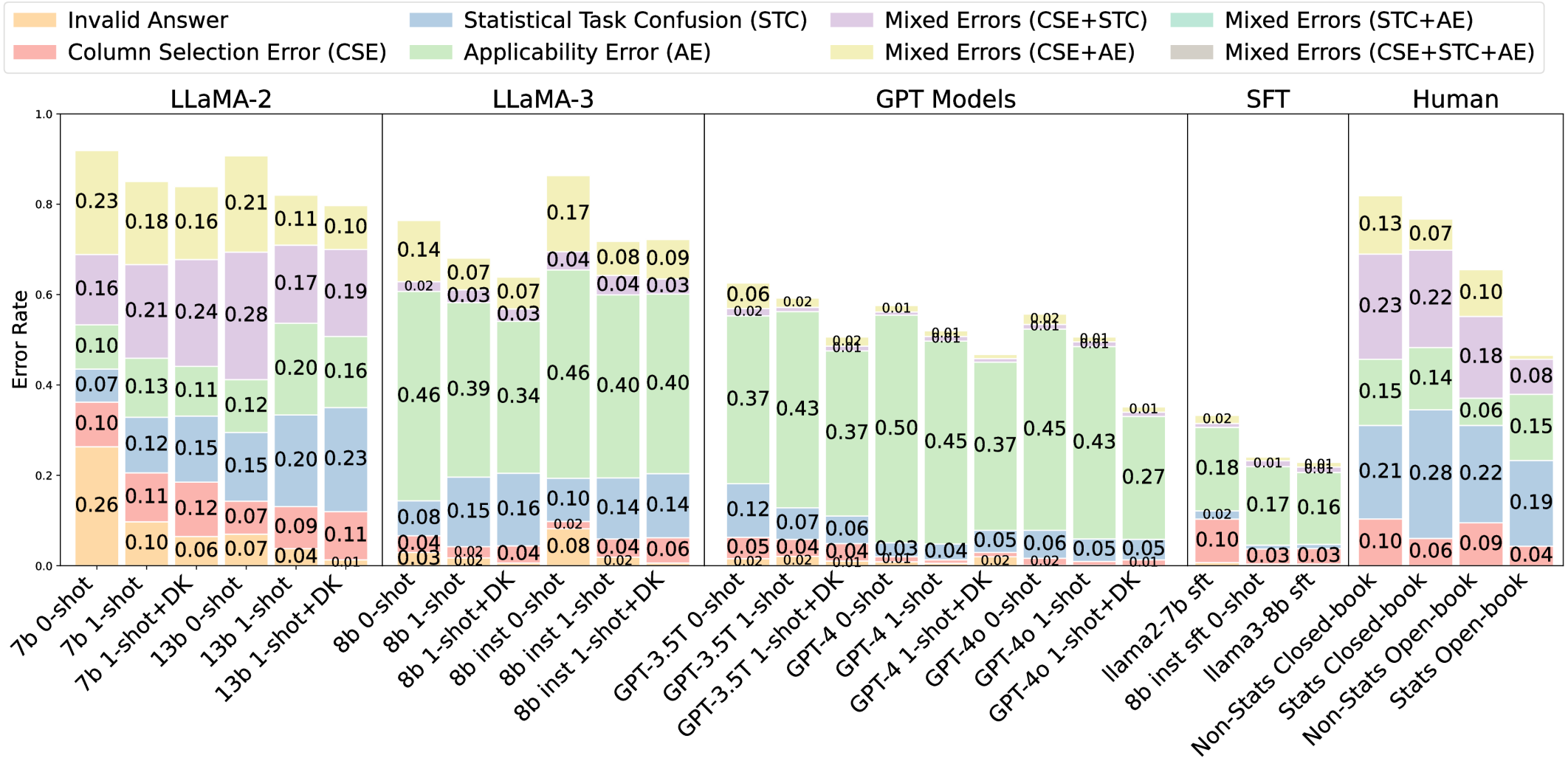
Model	Strategy	Overall	CA	CTT	DCT	VT	DS
Open-source LLMs: LLaMA-2/3							
LLaMA-2 7B	0-shot	8.08	1.79	1.17	2.12	6.97	25.48
	1-shot	14.96	0.60	6.25	5.93	19.26	37.07
	0-shot-CoT	6.36	1.19	0.78	2.12	5.74	19.69
	1-shot-CoT	14.45	1.79	4.30	8.48	19.67	33.21
	1-shot+DK	16.08	0.60	7.42	9.32	18.44	38.61
LLaMA-2 13B	0-shot	9.29	1.79	0.39	8.48	3.28	29.34
	1-shot	17.97	9.52	5.47	9.32	2.05	58.69
	0-shot-CoT	9.03	2.38	0.00	9.32	2.87	27.80
	1-shot-CoT	17.63	6.55	9.38	9.75	0.41	56.37
	1-shot+DK	20.29	8.33	7.03	16.53	11.48	52.90
LLaMA-3 8B	0-shot	23.56	1.19	0.00	16.53	16.39	74.52
	1-shot	31.90	17.86	8.20	18.64	25.41	82.63
	0-shot-CoT	22.01	1.19	0.39	15.68	13.93	70.27
	1-shot-CoT	32.24	14.29	5.86	19.92	**29.10**	**84.17**
	1-shot+DK	**36.11**	**26.79**	20.31	**29.24**	15.98	83.01
LLaMA-3 8B Instruct	0-shot	13.67	10.12	13.28	5.09	1.23	35.91
	1-shot	28.20	**26.79**	12.11	13.56	9.84	75.68
	0-shot-CoT	11.61	10.71	14.84	6.78	0.00	24.32
	1-shot-CoT	28.29	26.19	16.80	16.10	9.84	69.50
	1-shot+DK	27.77	19.64	**22.27**	20.34	8.20	63.71
Proprietary LLMs: GPT-3.5-Turbo, GPT-4 and GPT-4o							
GPT-3.5-Turbo	0-shot	37.40	47.02	31.25	26.27	12.71	70.66
	1-shot	40.76	53.57	12.50	27.54	26.23	86.10
	0-shot-CoT	38.17	45.24	33.59	25.85	13.93	72.20
	1-shot-CoT	39.64	51.79	10.94	26.70	26.23	84.56
	1-shot+DK	49.36	62.50	35.55	38.98	26.23	85.71
GPT-4	0-shot	42.39	66.67	20.70	45.76	2.46	82.63
	1-shot	47.98	67.86	26.56	44.07	14.75	91.12
	0-shot-CoT	43.34	67.86	23.44	46.19	1.64	83.78
	1-shot-CoT	47.46	67.26	30.08	41.95	11.07	91.12
	1-shot+DK	53.22	64.88	43.75	49.58	20.08	89.56
GPT-4o	0-shot	44.23	62.50	19.53	25.00	31.56	86.49
	1-shot	49.36	**69.05**	26.56	30.93	34.43	89.97
	0-shot-CoT	44.71	63.10	20.70	24.58	32.38	86.49
	1-shot-CoT	48.67	67.86	25.78	28.81	32.79	**91.89**
	1-shot+DK	**64.83**	61.31	**65.23**	**59.32**	**46.31**	89.19
Fine-tuned LLMs							
SFT LLaMA-2 7B	0-shot	66.72	69.05	35.94	83.48	54.51	91.89
SFT LLaMA-3 8B	0-shot	**77.13**	**79.76**	65.23	**88.56**	55.33	**97.30**
SFT LLaMA-3 8B Instruct	0-shot	75.92	69.64	**68.75**	85.17	**57.38**	96.14
Human experiments (On subset of mini-StatQA)							
Human (Non-Stats)	Closed-book	18.10	5.88	3.85	8.70	0.00	65.39
	Open-book	34.48	**52.94**	0.00	30.44	8.33	84.62
Human (Stats)	Closed-book	23.28	29.41	0.00	17.39	0.00	69.23
	Open-book	**53.45**	47.06	**23.08**	**65.22**	**37.50**	**92.31**



Radar Charts for Best Results of Each Model in Sub-tasks

Error Analysis

Distribution of Error Categories Across Experiments.



Contributions

- StatQA:
  - Benchmark for statistical analysis tasks, particularly focusing on the applicability assessment
  - Introduce an automated pipeline to construct StatQA
- Systematic Evaluation:
  - Extensive evaluations on widely used LLMs
  - Several strategies, including in-context learning, domain-specific prompts and fine-tuning
- Comparative Study:
  - Group-based human experiments and comparatively analyze differences between humans and LLMs
  - Comparative error analysis, highlighting distinct strengths, revealing potential for collaboration
- New Empirical Findings and Research Opportunities:
  - We summarize six key findings and discuss research opportunities in this field

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

Thank you!

For more details and findings, welcome to refer to our project homepage:

https://statqa.github.io/