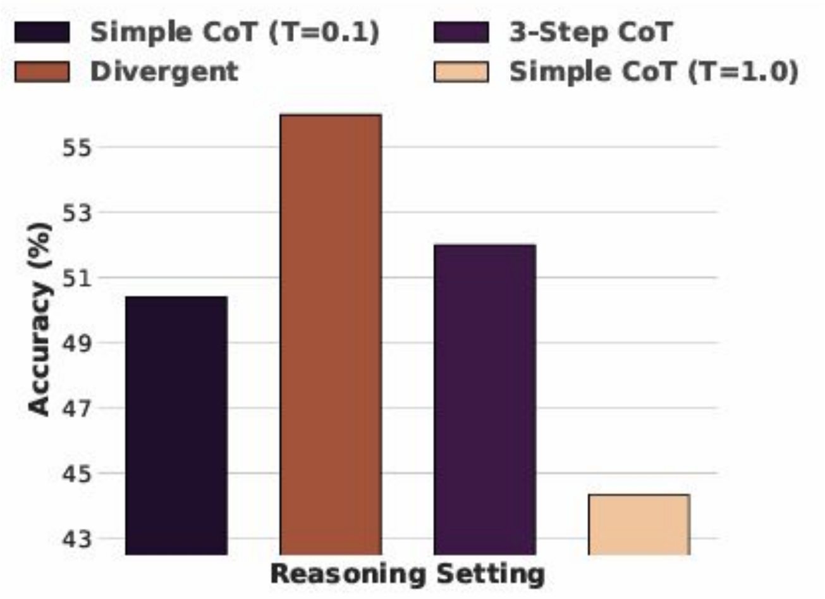
AdaReasoner: Adaptive Reasoning Enables More Flexible Thinking

Xiangqi Wang[1]* Yue Huang[1]* Yanbo Wang[2] Xiaonan Luo[1] Kehan Guo[1]
Yujun Zhou[1] Xiangliang Zhang[1]†
[1] University of Notre Dame [2]MBZUAI
{xwang76, yhuang37, xluo6, kguo2, yzhou25, xzhang33}@nd.edu
yanbo.wang@mbzuai.ac.ae

A lightweight policy module selects LLM generation settings conditioned on question nuisances.

- Introduced a few-shot adaptive controller that tunes temperature and reasoning prompt based on task features.
- Trained via RL with a factorized action space and Boltzmann exploration for optimal flexibility.
- Achieved theory-backed fast convergence with a sublinear policy gap.
- Demonstrated consistent gains across six LLMs on diverse reasoning suites, improving Out-of-Distribution (OOD) robustness.



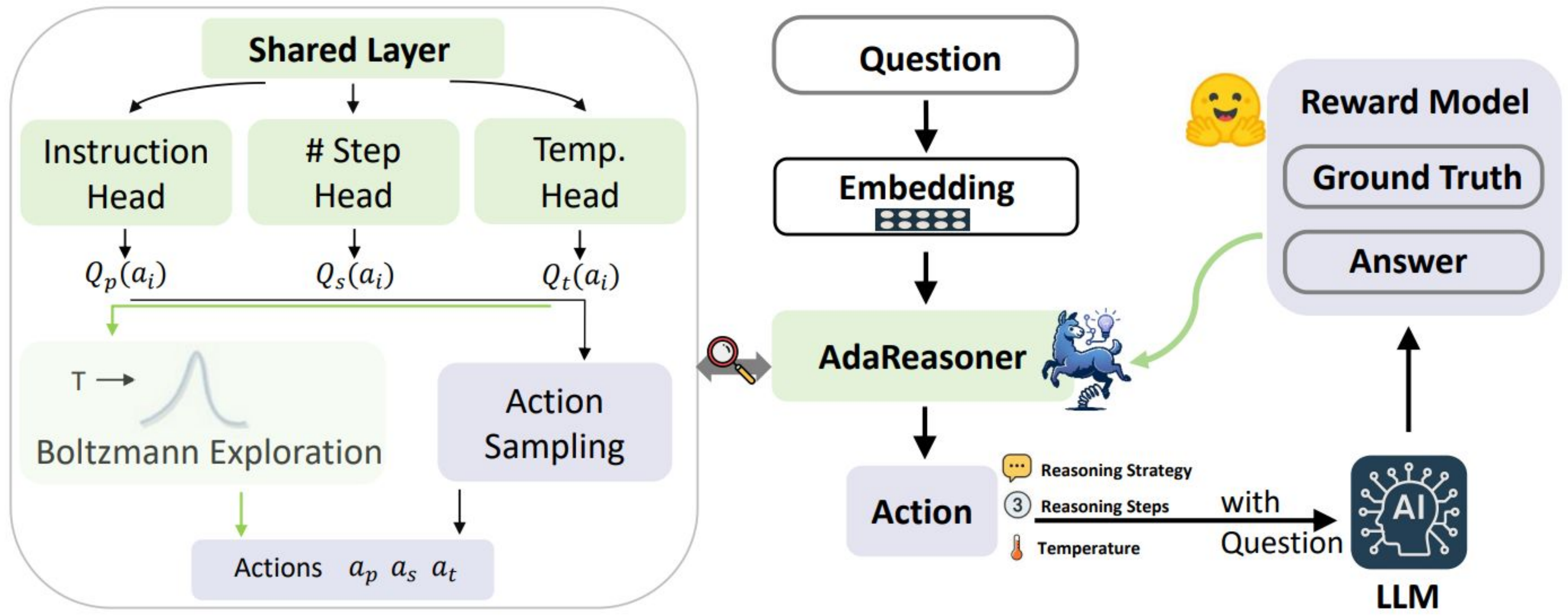
Pipeline diagram of adareasoner

A: Dual Network Design for Factorized Action Space

AdaReasoner factorizes instruction, temperature, and step count into separate heads anchored by one shared encoder. This keeps all decisions aligned in the same feature space, reducing drift to avoid sub-optimality

B. REINFORCE++ Algorithm with Proxy Pre-trained Reward Model

AdaReasoner uses a lightweight policy-network REINFORCE++ algorithm, where actions are sampled and updated via policy gradients. A DeBERTa-based proxy reward model provides a scalar score in [0,1], giving dense and consistent feedback that stabilizes learning



Configuration Action Space of AdaReasoner

Action Space	Expression
Number of Steps	$\mathcal{A}_s = \{x \mid x \in \mathbb{Z}, 3 \leq x \leq 10\}$
Temperature	$\mathcal{A}_t = \{0.0 + 0.1k \mid k \in \mathbb{Z}, 0 \leq k \leq 10\}$
Reasoning Instructions	$\mathcal{A}_p = \{\text{base + variation}\}$

Base Instruction
Break down your reasoning into clear, sequential steps.
Systematically structure your analysis, elaborating on each step with thorough detail.
Examine the logical connections between concepts and articulate each step in depth.
Consider multiple perspectives and explore alternative viewpoints comprehensively.
Apply creative reasoning to unearth unconventional insights and challenge standard assumptions.
Adopt a detailed and rigorous approach, balancing specific details with overarching themes.
Reflect on your assumptions and refine your argument through critical self-questioning and validation.
Explain your reasoning step-by-step in a clear, accessible manner for all audiences.
Include a systematic self-check and verification of your reasoning process to ensure consistency.
Conclude by summarizing your key points and re-evaluating your final answer for completeness.

Variation Instruction
Thoroughly analyze all possible interpretations for comprehensive understanding.
Decompose the problem into smaller, logical components for clarity and precision.
Cross-reference reasoning with similar examples or prior cases for validation.
Review and verify each step to ensure no key detail is overlooked.
Challenge conventional thinking while maintaining logical soundness.
Ensure every premise is clearly understood and meticulously applied.
Pay close attention to minor details that might otherwise be neglected.
Use simple, straightforward language to guarantee clarity and accessibility.
Perform a detailed self-audit to detect and correct inconsistencies.
Validate conclusions by aligning them with established principles or empirical data.

Table of divide of action space (generation temperature, generation step in prompt, and reasoning psychology).

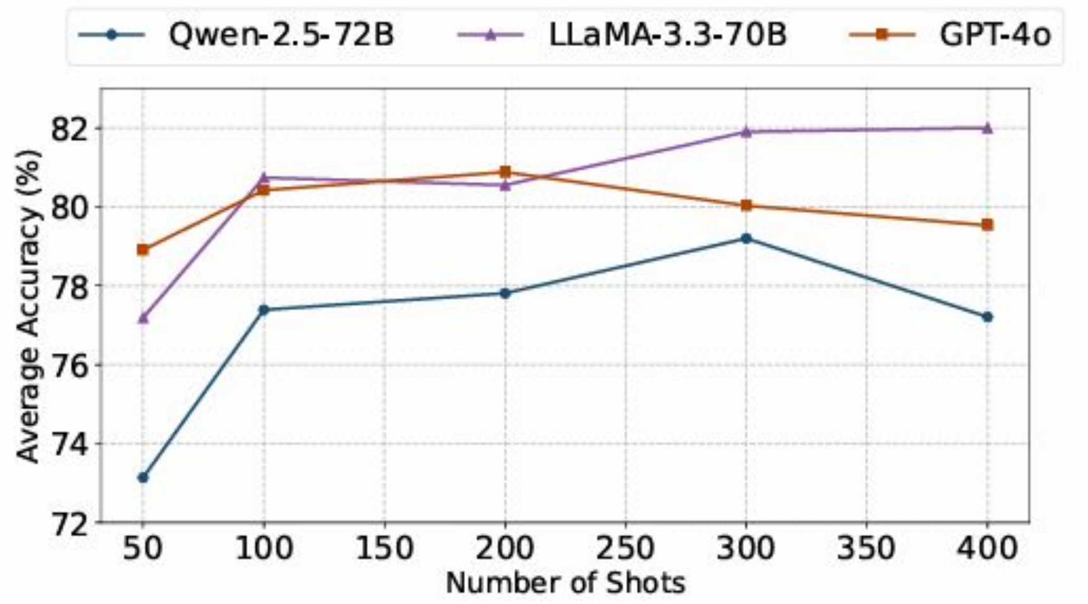
We evaluate AdaReasoner on four reasoning datasets—Metaphor, TruthfulQA, MMLU (Math), and LogiQA—using accuracy under the LLM-as-Judge protocol. Across six LLMs, we benchmark against key reasoning baselines including CoT, Think-Short, ToT, Best-of-N, Auto-CoT, and In-context CoT.

Table 1: Performance of various reasoning methods across multiple datasets for different LLM models (accuracy in %). The highest score for each dataset and the average in each model group is highlighted in **bold** and underlined.

Model	Reason Method	Dataset (%)				Average
		Metaphor	TruthfulQA	MMLU	LogiQA	
GPT-4o	CoT	50.40	78.40	76.04	70.00	68.71
	Think Short	61.00	64.81	68.52	70.81	66.28
	ToT	48.25	74.29	86.11	73.90	70.91
	Best-of-N	52.60	79.41	83.41	72.37	71.95
	Auto-CoT	62.33	**83.09**	72.15	71.71	72.32
	In-context CoT	53.98	77.04	83.63	80.04	74.42
	AdaReasoner	**71.56**	81.30	**86.49**	**82.31**	**80.42**
Llama-3.3-70B-Ins.	CoT	51.56	75.77	83.33	75.56	71.56
	Think Short	59.56	75.77	81.61	73.78	72.68
	ToT	60.89	75.33	86.24	83.56	76.51
	Best-of-N	52.89	77.09	**89.69**	76.00	73.92
	Auto-CoT	45.33	78.85	81.82	76.00	70.50
	In-context CoT	52.71	82.45	84.57	75.59	73.60
	AdaReasoner	**66.11**	**83.09**	87.77	**85.00**	**80.74**
Qwen-2.5-72B-Ins.	CoT	60.18	79.36	73.89	78.26	72.92
	Think Short	71.24	80.28	64.16	75.22	72.73
	ToT	62.26	77.50	66.57	79.51	71.46
	Best-of-N	59.73	78.44	76.11	78.26	73.14
	Auto-CoT	65.93	83.49	76.11	79.13	76.17
	In-context CoT	**73.39**	78.94	71.93	74.83	74.77
	AdaReasoner	65.19	**83.82**	**80.14**	**80.79**	**77.49**
Claude-3.5-sonnet	CoT	62.13	86.13	85.00	80.43	78.42
	Think Short	**67.71**	83.43	78.95	77.95	77.01
	ToT	59.45	85.12	86.43	81.98	78.25
	Best-of-N	41.41	83.43	81.87	78.95	71.42
	Auto-CoT	65.04	84.86	88.50	78.70	79.28
	In-context CoT	55.81	**88.60**	79.23	79.53	75.79
	AdaReasoner	65.77	86.17	**89.21**	**84.55**	**81.43**
Deepseek-R1	CoT	54.35	83.34	96.13	81.82	78.91
	Think Short	67.71	80.00	95.55	77.71	80.24
	ToT	63.33	86.16	**98.70**	83.22	82.85
	Best-of-N	54.55	85.51	94.37	87.01	80.36
	Auto-CoT	61.04	82.61	97.70	80.52	80.47
	In-context CoT	50.06	84.21	96.15	84.25	78.67
	AdaReasoner	**72.00**	**88.17**	96.33	**88.60**	**86.28**
GPT-o3-mini	CoT	45.10	84.00	95.71	83.87	77.17
	Think Short	57.14	80.00	93.21	67.74	74.52
	ToT	53.85	84.91	98.18	80.00	79.24
	Best-of-N	56.99	82.10	93.55	84.22	79.22
	Auto-CoT	51.00	**86.79**	**97.78**	76.14	77.92
	In-context CoT	53.00	82.25	95.56	77.19	77.00
	AdaReasoner	**67.29**	86.45	96.13	**87.67**	**84.39**

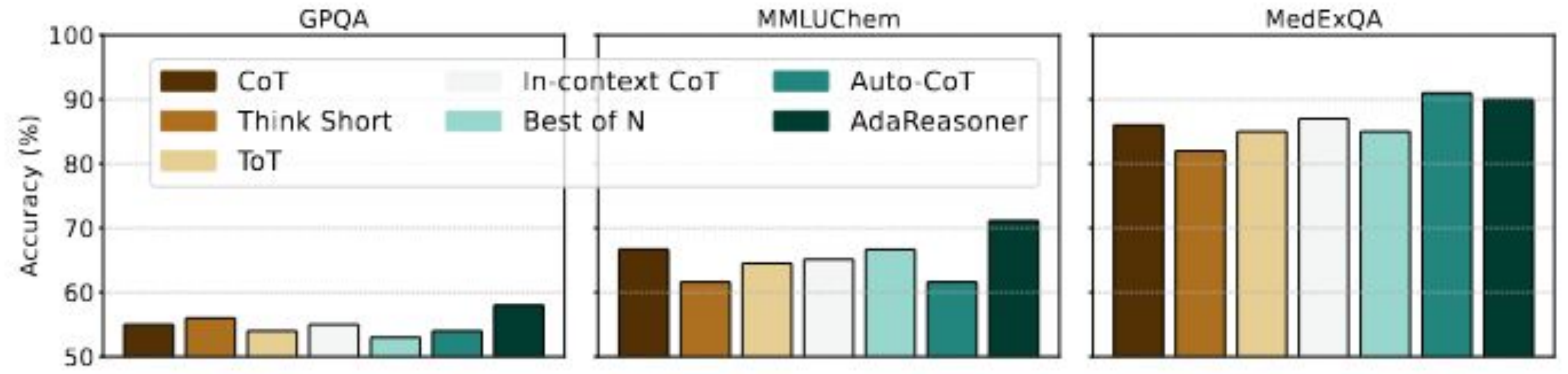
Ablation	Metaphor	TruthfulQA	MMLU	LogiQA	Average
Random Action	55.92	76.15	80.32	76.81	72.30
AdaReasoner (a_t)	62.91	80.00	77.71	75.67	74.07
AdaReasoner (a_s)	68.11	74.29	82.11	74.44	74.74
AdaReasoner (a_p)	70.66	78.31	84.50	81.01	78.62
AdaReasoner (Random Seed)	53.17	70.55	79.13	73.90	69.19
w/ Bandit Adapter	68.30	76.11	80.00	79.13	75.89
w/ Perturbed Reward	70.83	79.26	85.07	77.89	78.26
w/ [-0.5, 0.5] Reward	56.66	76.15	79.04	77.63	72.37
w/ Qwen Adapter	65.76	73.80	69.69	80.00	72.31
Adareasoner (Close-perturb)	66.05	79.39	85.18	80.03	77.66
Adareasoner (Distant-perturb)	57.69	71.77	81.42	74.96	71.46
Adareasoner (Emsemble)	65.73	79.54	84.71	80.04	77.50
AdaReasoner	**71.56**	**81.30**	**86.49**	**82.31**	**80.42**

Ablation Studies of different action and parameter



Few-shot training performance.

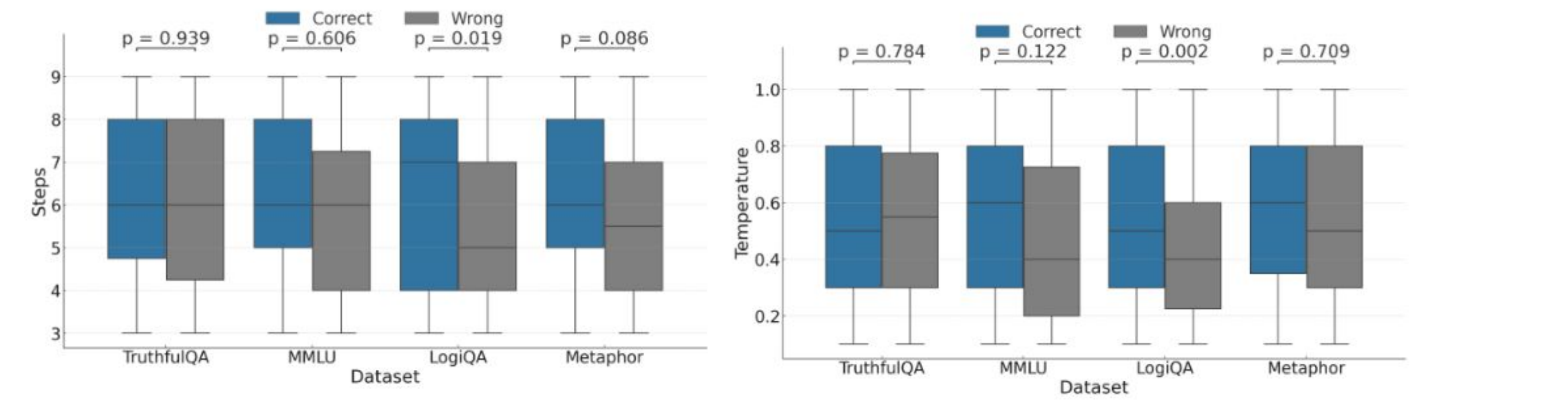
Adareasoner is few-shot efficient



Out of Distribution Robustness

Model	CoT	Think Short	ToT	Best-of-N	Auto-CoT	In-context CoT	AdaReasoner
GPT-4o	98.83	99.17	99.17	99.17	99.50	99.00	99.00
Llama-3.3-70B-Ins.	99.50	100.00	99.17	99.33	99.00	98.00	100.00
Qwen-2.5-72B-Ins.	98.83	98.83	99.50	98.83	99.33	99.17	99.33
Claude-3.5-Sonnet	99.33	99.00	99.50	99.50	99.50	100.00	99.33
DeepSeek-R1	99.33	99.17	99.00	98.83	99.17	98.00	100.00
GPT-o3-mini	100.00	100.00	99.00	100.00	100.00	99.00	99.50

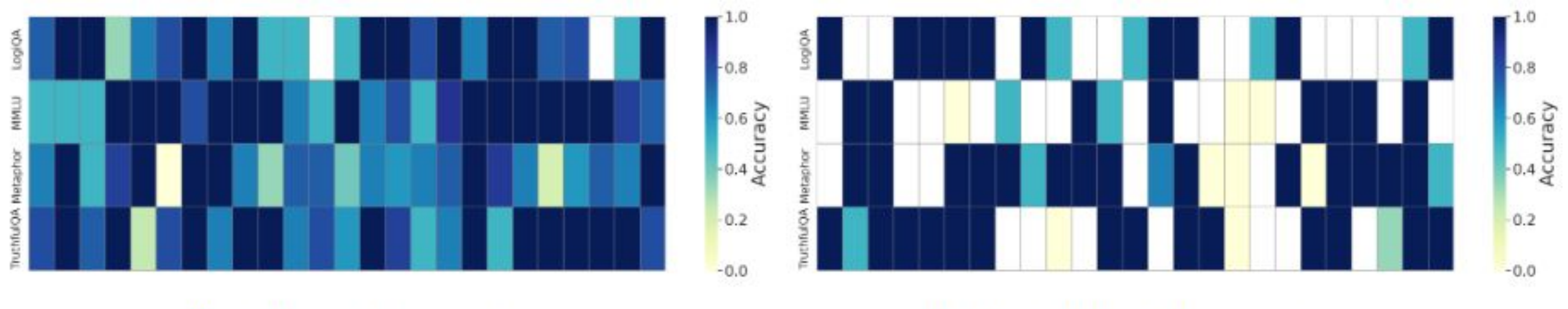
LLM-as-Judge Reliability Result Table



(a) Steps a_s

(b) Temperature a_t

Generation pattern of steps and temperature among different dataset types



(a) Top-25 most frequent a_p

(b) Top-25 least frequent a_p

Heatmap of correctness of most/least frequently used psychology method

Table 6: Action Statistics across Datasets

Configuration Action	Metaphor	TruthfulQA	MMLU	LogiQA
# Steps a_s	5.86 ± 0.57	6.04 ± 1.44	6.54 ± 0.71	6.14 ± 1.02
Temperature a_t	0.542 ± 0.110	0.629 ± 0.281	0.572 ± 0.155	0.538 ± 0.209

Statistics Table of Actions Mean among different datasets

https://mine-lab-nd.github.io/project/adareasoner.html