From Passive to Active Reasoning:
Can Large Language Models Ask the Right Questions under Incomplete Information?

Zhanke Zhou*, Xiao Feng*, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, Bo Han

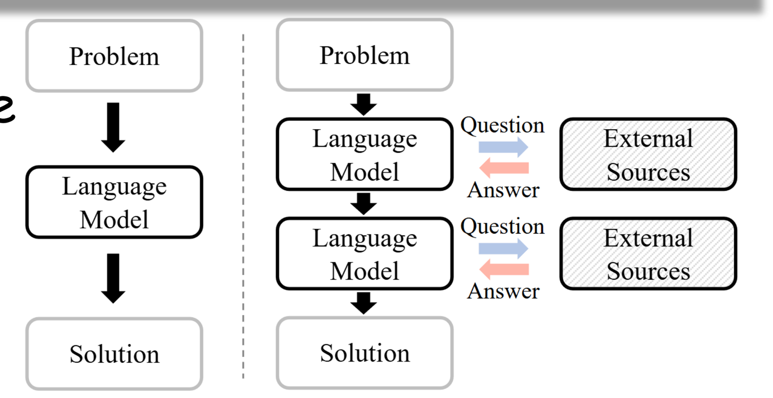
TMLR Trustworthy Machine Learning and Reasoning · 香港浸會大學 Hong Kong Baptist University · Mila · Université de Montréal · 上海交通大学 Shanghai Jiao Tong University · Stanford University · ICML International Conference On Machine Learning

Paper

Code

Problem: Active Reasoning

- **Passive Reasoning (PR): Full information** is provided for the model to derive the solution.
- **Active Reasoning (AR): Given incomplete information**, the model **interacts with external sources** for information



(a) Passive Reasoning (b) Active Reasoning

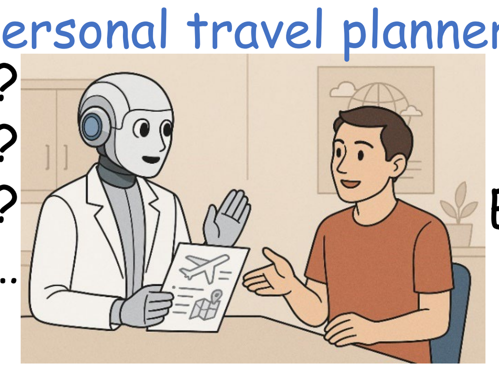
The Significance of Active Reasoning

Active reasoning is essential in real-world applications:

- Travel planners gather requirements to create travel plans
- Doctors ask targeted questions to collect critical information for accurate analysis and diagnoses

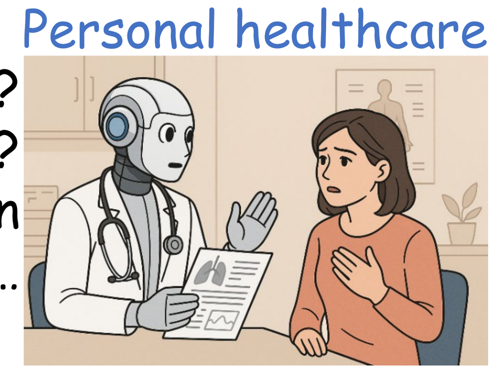
Personal travel planner

Budget?
Preference?
Time?
...



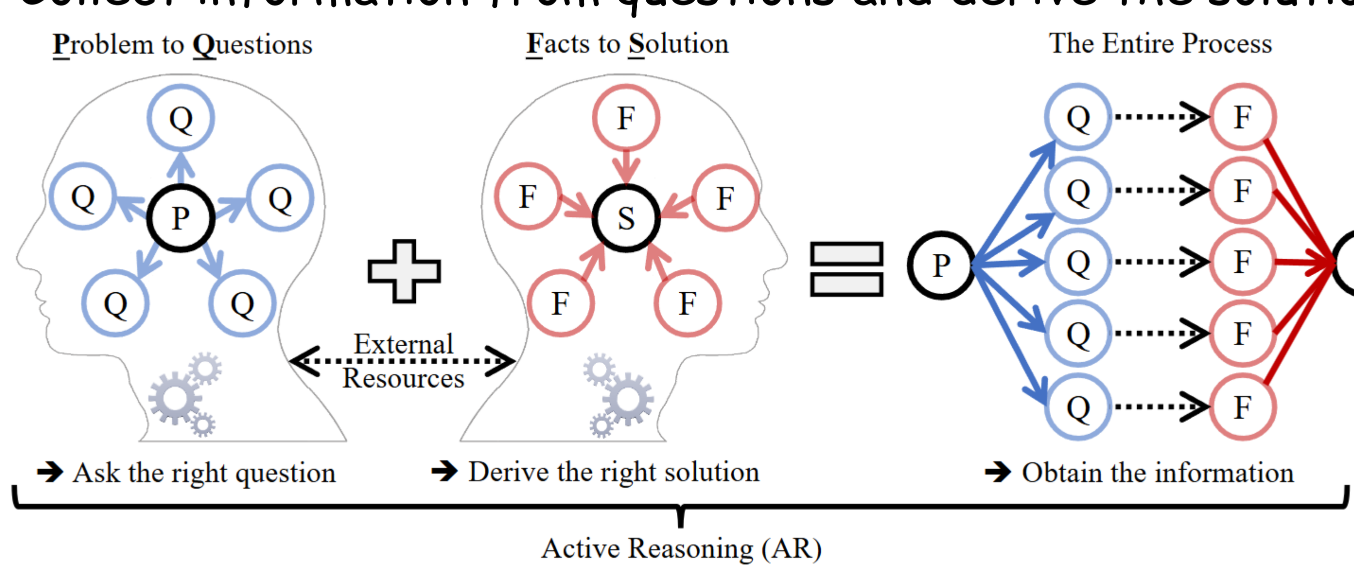
Personal healthcare

Symptoms?
History?
Examination
...



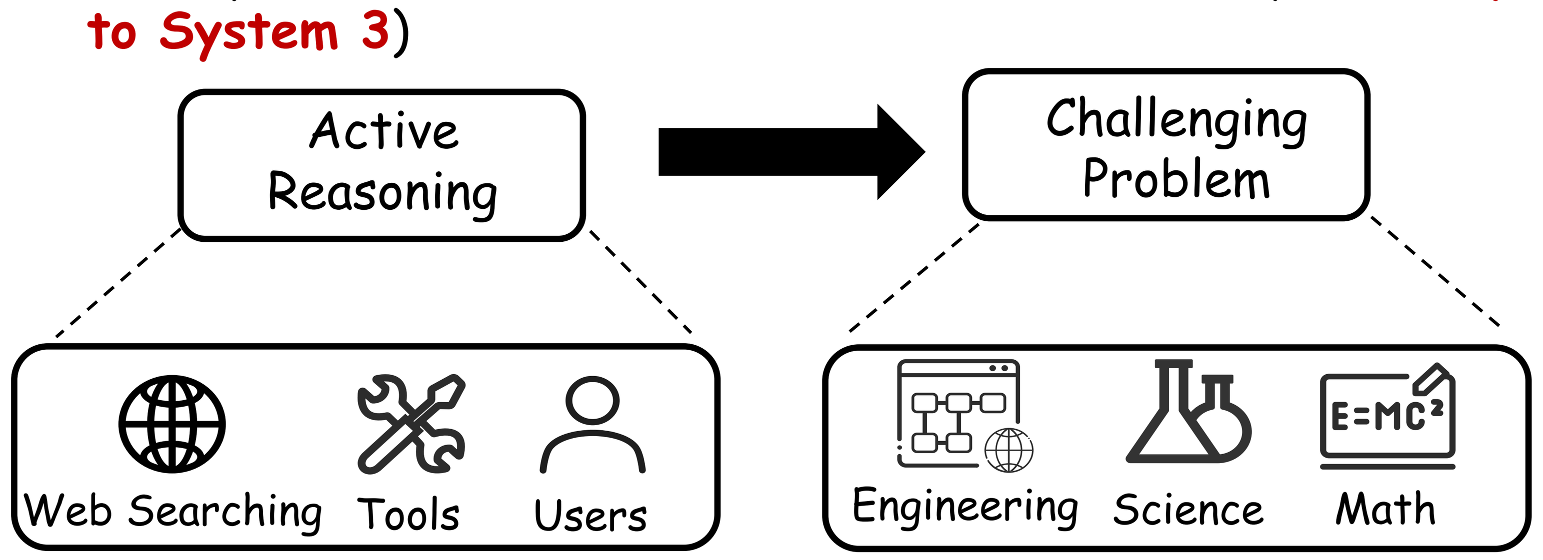
The core requirements of Active Reasoning

- **Target important missing details** and **form questions**
- Collect information from questions and derive the solution



From Laurens Van der Maaten's keynote at CVPR 2025:

- **System 3 is thinking together. Interactive and collaborative.** Finds others with complementary skills or experience to **solve more complex tasks**
- AR depicts the interactive and collaborative ability (**the key to System 3**)



Dataset: AR-Bench

AR-Bench (**A**ctive **R**easoning **Bench**mark) contains **6040** questions and **3** tasks, covering **commonsense**, **logic**, and **symbolic** reasoning tasks

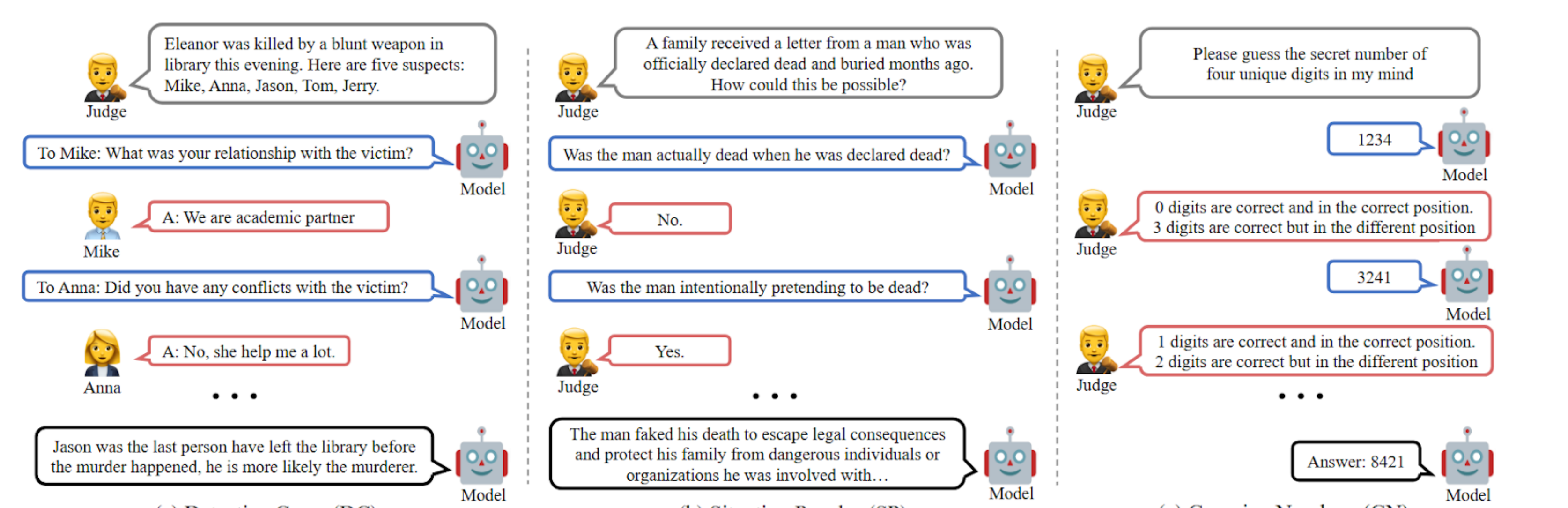
The AR-Bench covers the **three tasks:**

- **Detective Cases (DC):** Interrogation between a detective and 5 suspects
- **Situation Puzzles (SP):** the puzzle to reveal the truth from a mystery
- **Guessing Numbers (GN):** the game to uncover a 4-unique-digits number

Dataset statistics for the three tasks in AR-Bench:

Task	DC	SP	GN
Size (train/test)	400/100	400/100	4940/100
Avg. problem tokens	564.06	178.53	176.00
Interaction feedback	Narrative	Yes/No	Info. about correct digits
Answer space	5		5040
Metric	Accuracy	F1 score	Exact match

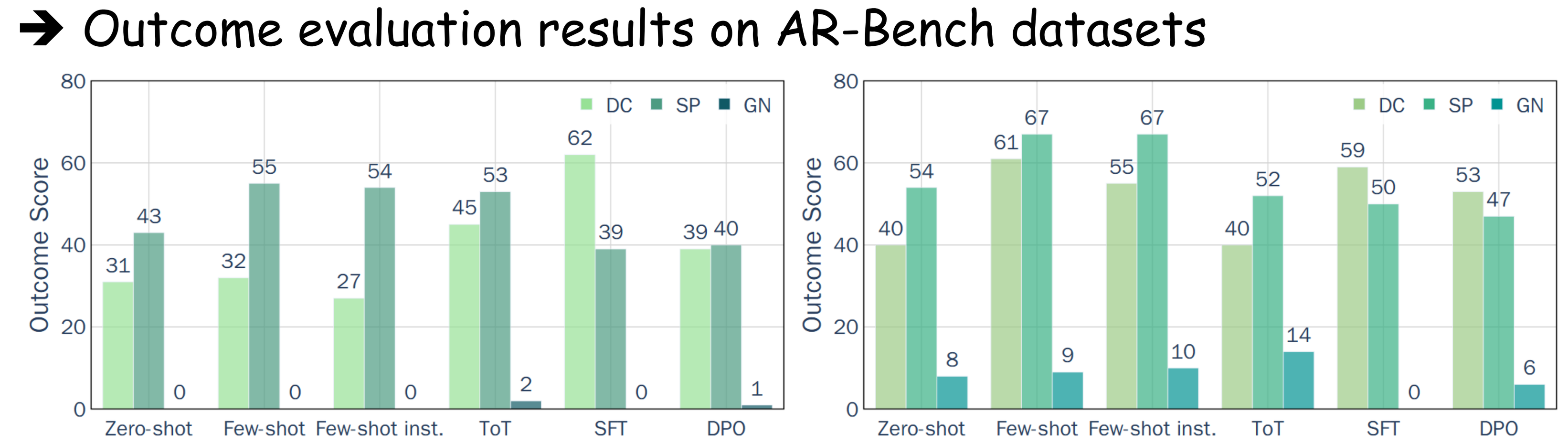
Examples of three tasks:



(a) Detective Cases (DC) (b) Situation Puzzles (SP) (c) Guessing Numbers (GN)

Experiments

➔ Outcome evaluation results on AR-Bench datasets



(a) Llama-3.1-8B (b) Llama-3.1-70B

Figure 4: The evaluation results of outcome scores for Llama-3.1-8B and Llama-3.1-70B on the AR-Bench across various methods. The outcome scores represent accuracy, F1 score, and exact match rate for tasks DC, SP, and GN, respectively.

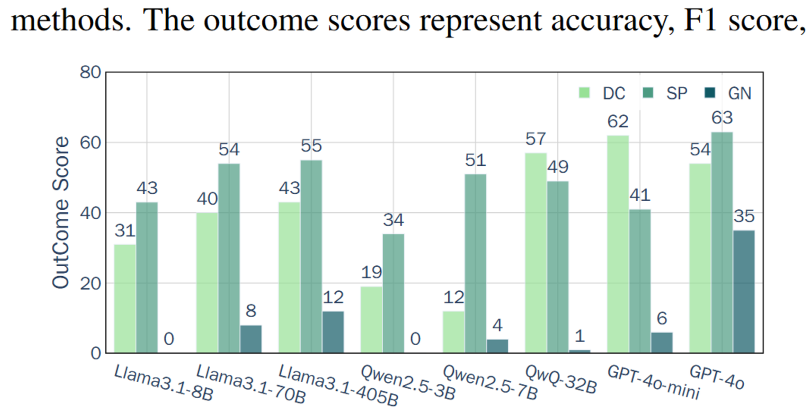


Figure 5: Reasoning accuracy on the AR-Bench with different language models. We set zero-shot as the default setting.

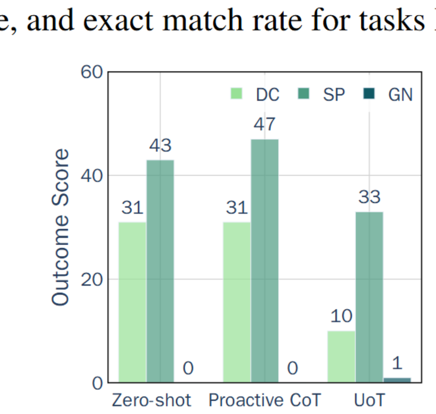


Figure 6: Compare advanced methods using Llama-3.1-8B.

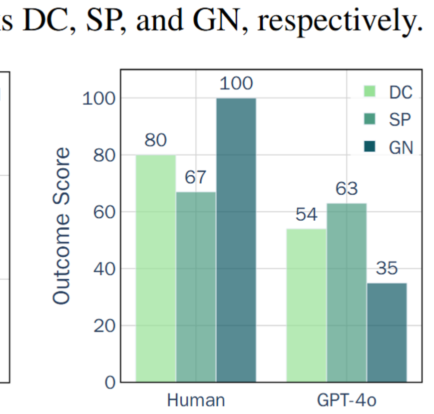


Figure 7: Compare zero-shot GPT-4o with human eval.

※ Key Observations:

1. **AR-Bench demonstrates challenges across all models and methods**
2. Existing active reasoning methods fail in AR-Bench
3. Human baselines significantly surpass cutting-edge language models

➔ Process evaluation results on AR-Bench datasets

Figure 8: The process score across three tasks, evaluating Llama-3.1-8B (a) and Llama-3.1-70B (b) with different methods.

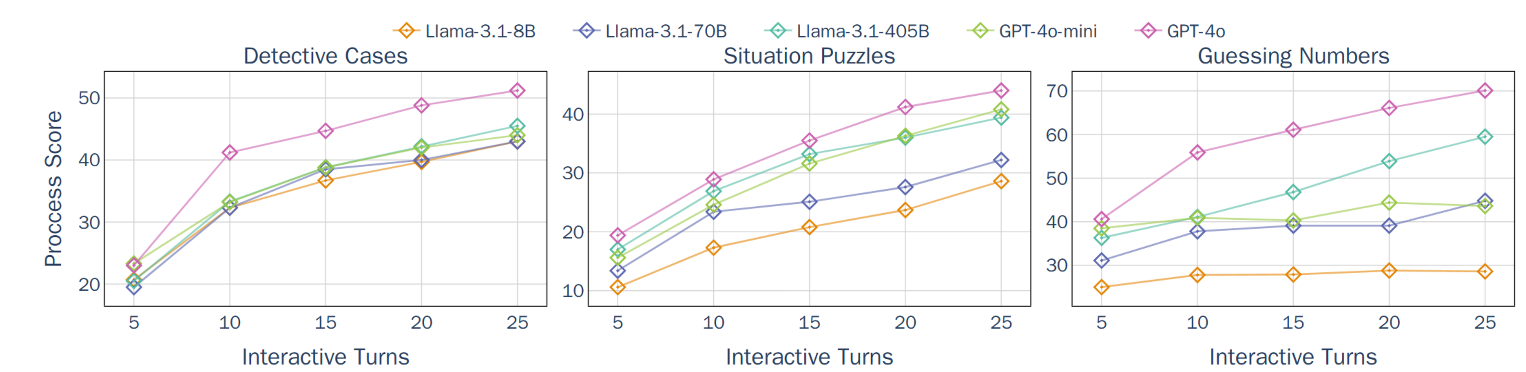
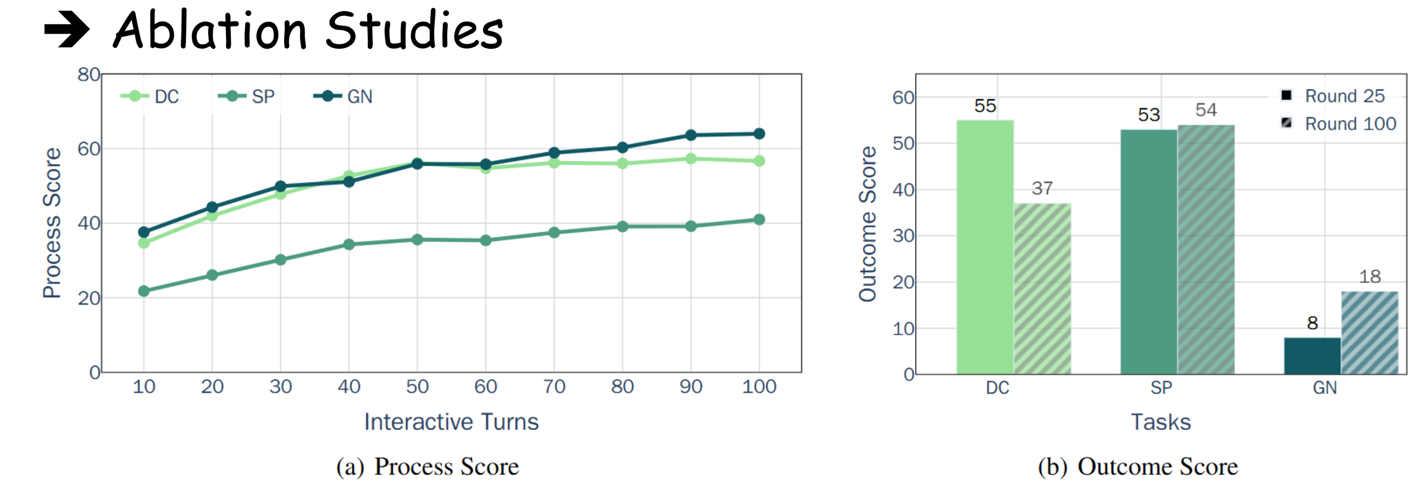


Figure 9: The process score of different models across three tasks in AR-Bench. All models are in a zero-shot setting.

※ Key Observations:

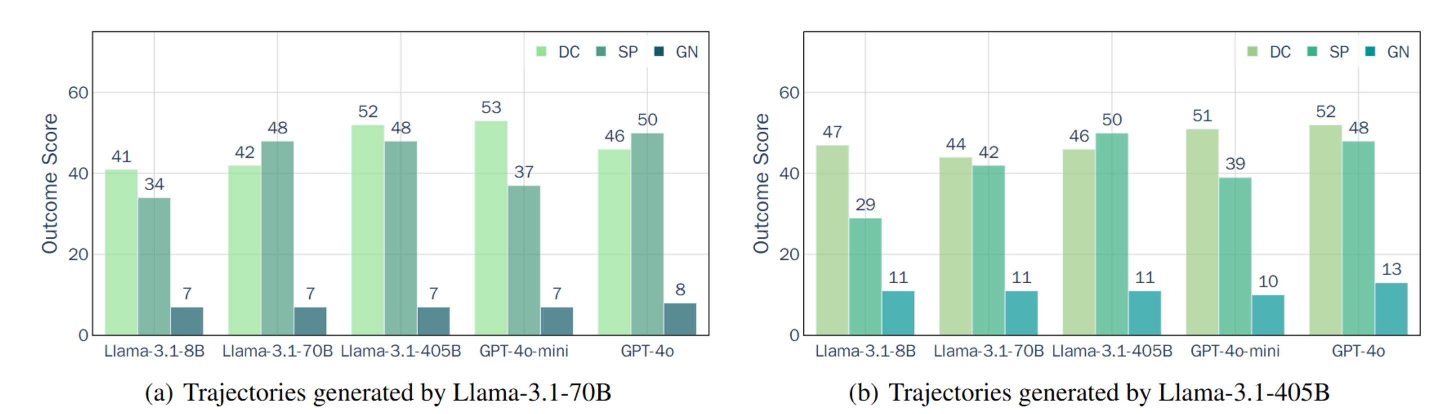
1. **LLMs struggle to consistently propose good questions**
2. The unreliable verifier limits the performance of ToT
3. The reliability of verifiers varies, strong in GN but weaker in SP
4. Underperforming LLMs ask low-quality questions
5. Larger models can retrieve more useful information

➔ Ablation Studies



(a) Process Score (b) Outcome Score

Figure 10: We present the results of scaling up the interaction rounds from 25 to 100 across three tasks using the Llama-3.1-70B model. The results include a comparison between the final outcomes and those in Fig. 5, and the process scores.



(a) Trajectories generated by Llama-3.1-70B (b) Trajectories generated by Llama-3.1-405B

Figure 11: The outcome scores of reasoning given the generated question-answering traces. We employ various models to make predictions in the traces generated by Llama-3.1-70B (a) and Llama-3.1-405B (b) to evaluate to what extent the question-answering history affects these models to draw the final conclusion.

※ Key Observations:

1. Larger models demonstrate robustness to insufficient information to derive more correct conclusions
2. More question-asking turns cannot directly indicate more accurate conclusions in AR-Bench