

# Overcoming Majority Errors in Multi-Agent Debate

Meitong Liu, Jieyi Zhao, Wangjia Zhan, Maojie Xu, Ian Jiang

# Motivation & Research Questions

- Q1. How is the effectiveness of multi-agent debate on tasks where majority voting fails?
- Q2. How is the effectiveness of multi-agent debate on these difficult tasks as the number of agents or debate rounds grows?
- Q3. Does the number and ratio of correct and incorrect responses affect the chance of an agent flipping its opinion?

# Evaluation on questions with majority errors

- Setups
  - Model: GPT-3.5-Turbo as in [1]
  - Dataset: GSM8K, graduate level mathematical reasoning tasks

*Q1. How is the effectiveness of multi-agent debate on tasks where **majority voting fails**?*

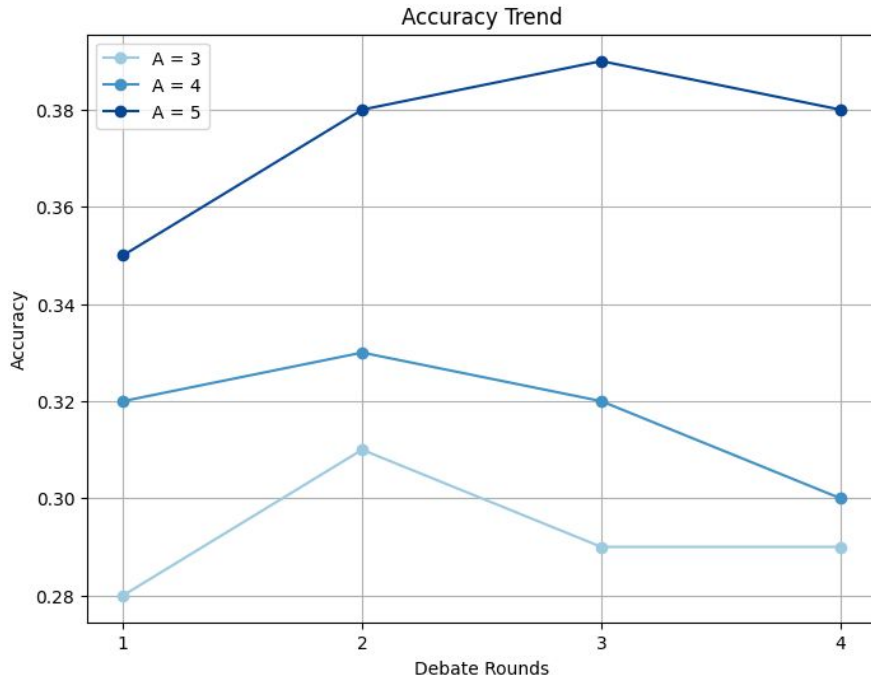
- We obtained a subset of **225 questions** from the test set where  $A = 3$  agents fail to reach a correct consensus through majority voting.

# Evaluation on questions with majority errors

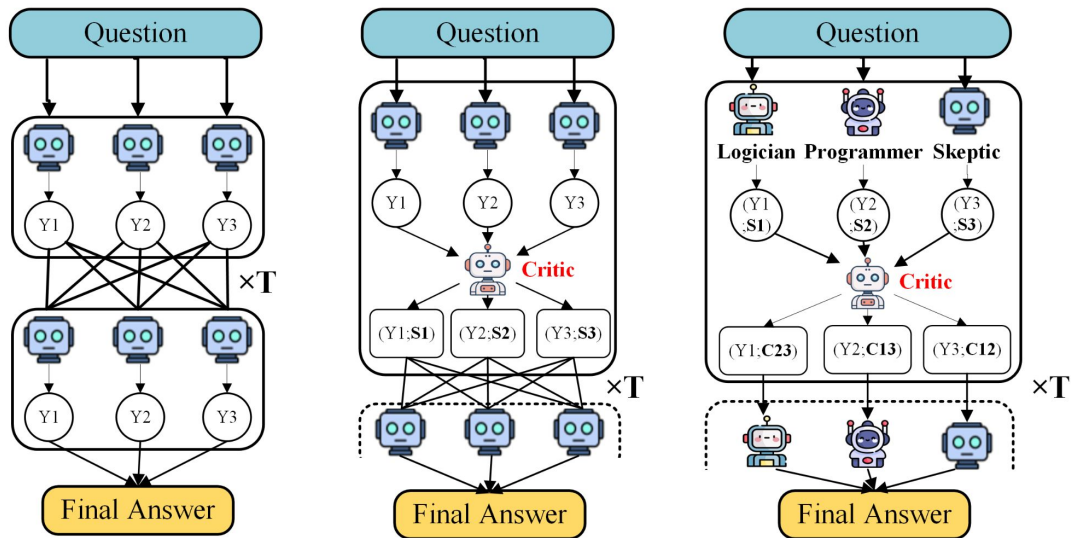
Q2. How is the effectiveness of multi-agent debate on these difficult tasks as **the number of agents (A)** or **debate rounds (R)** grows?

- As A grows, accuracy increases.
- However, accuracy fluctuates as R grows.

↳ **Fluctuation** as the debate progresses indicates that vanilla MAD is not effective handling majority errors.



# MAD with Confidence score



(a) Multi-Agent Debate    (b) MAD w/ confidence score (ours)    (c) MAD w/ confidence score & specialization (ours)

## Workflow

Y: output(reasoning+answer), S: confidence score, C: comment

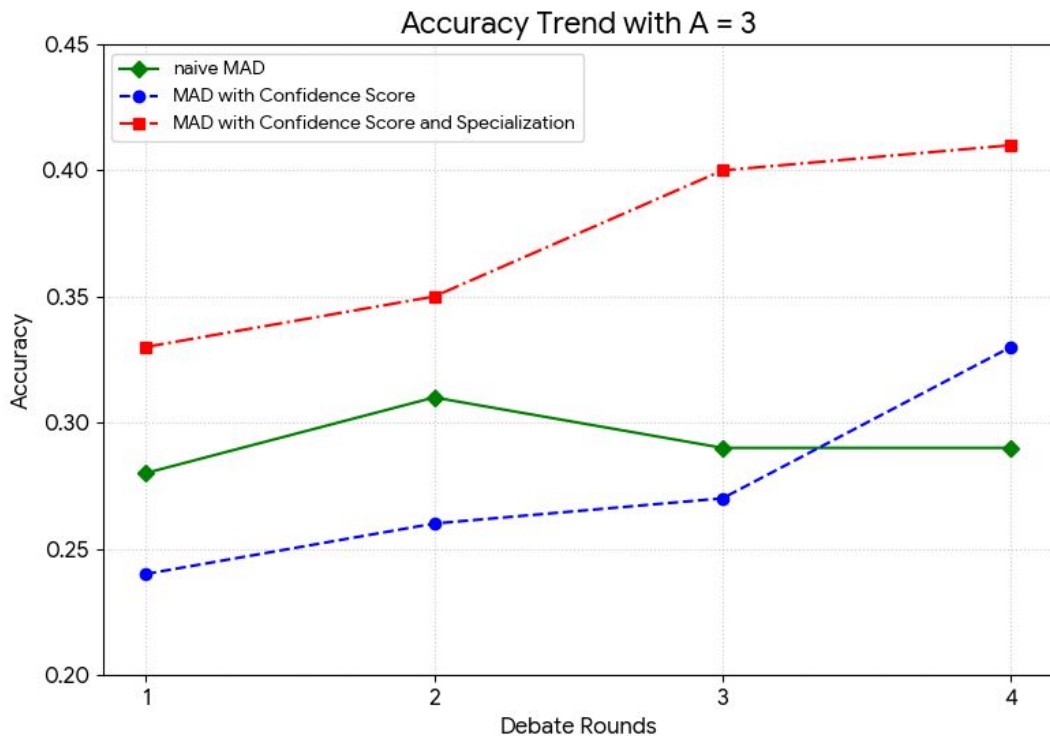
Naive MAD's **weakness**: (Lack of diverse perspectives; Easy Causing hallucinations)

**Framework(b)**: Introduce a critic model to assign confidence scores (1 = low, 10 = high). Debate quality is highly enhanced.

**Framework(c)**: Introduce two significant agents

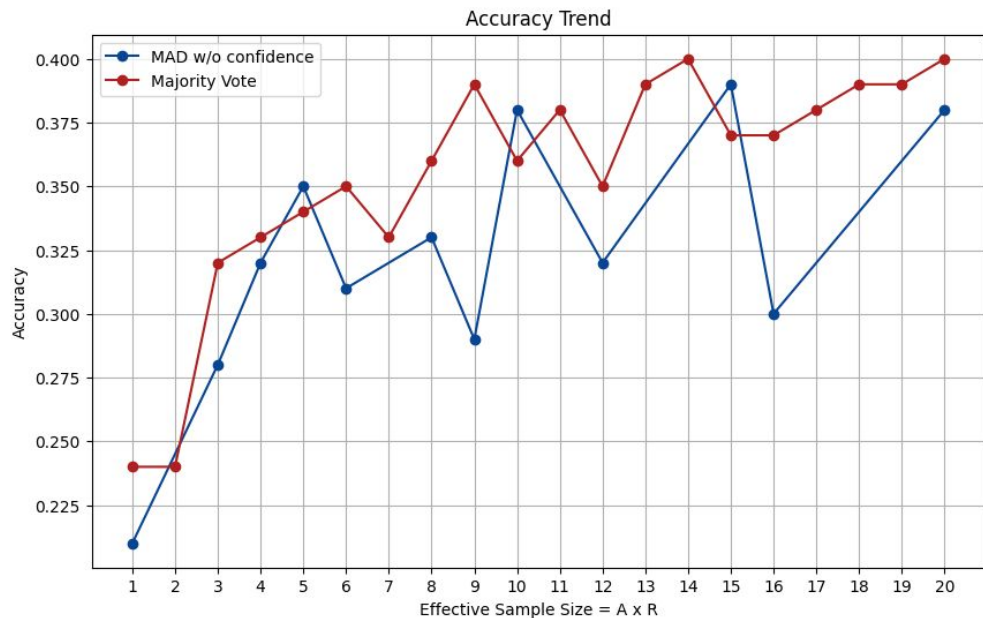
1. Logician Forcing CoT
2. Programmer Python format
  - When correct:
    - Clear logic -> High persuasiveness
  - When incorrect:
    - Immediate exposure of logical flaws -> Enables rapid Critic optimization.

# Results for Confidence score



1. Our method exhibits a steady, linear improvement, avoiding the fluctuations in the naive MAD.
2. Both variants outperform naive MAD; specialization delivers the strongest overall gains.

# Is repeated sampling sufficient?



Is the performance gain in MAD due to increased sampling times?

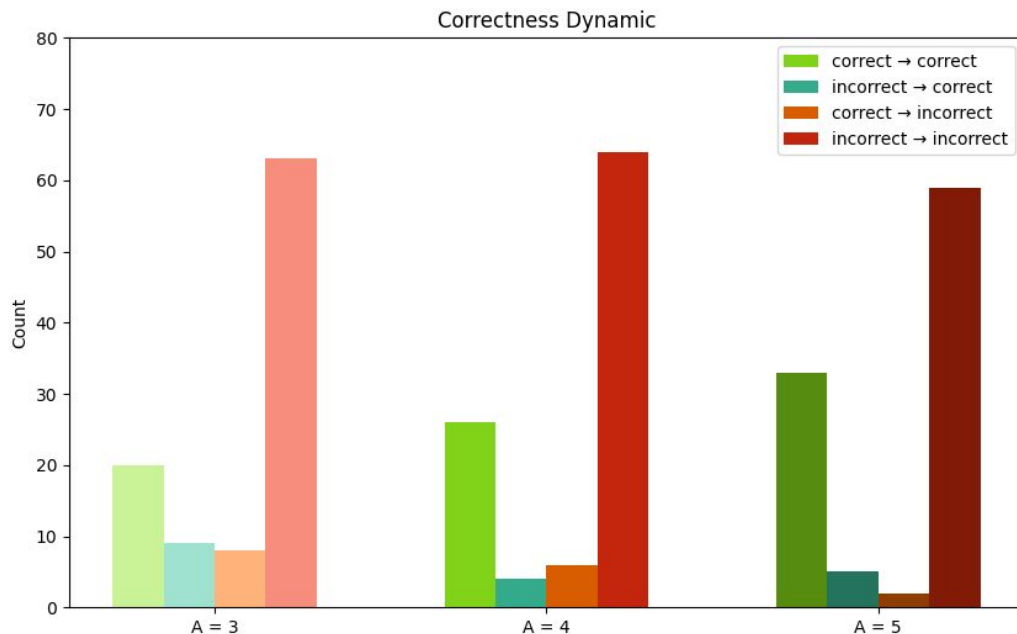
Effective Sample Size =  $A \times R$



We found that under **the same computation budget**, i.e., the same effective sample size, majority voting among repeatedly sampled responses **outperforms** MAD.

# Is repeated sampling sufficient? A closer look.

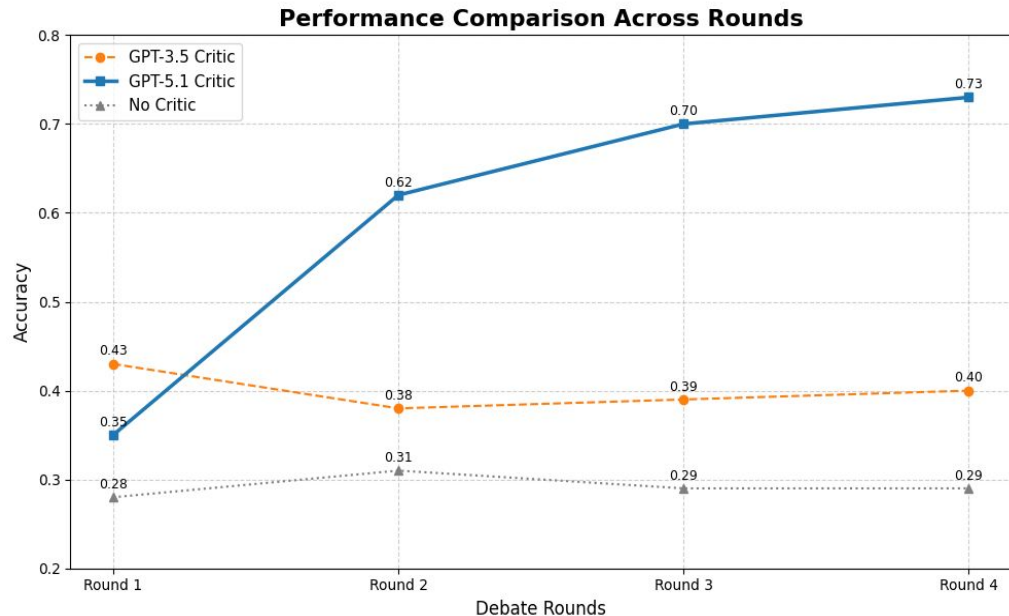
The short answer is YES!



# Potential for MAD

- **Observation:**
  - Repeated sampling + majority vote often outperforms MAD.
  - Supervision and debate with agents of same level cannot really help
- **Solution:** Use MAD to bridge the gap by seeking high-quality supervision instead of relying solely on agents of same level.

# Potential for MAD: High Quality



We can observe that with the supervision of a stronger model, the performance of debate grows stably. This basically means that if we provide some valuable insight or feedback during the discussion, weak agents can realize their mistakes and correct accordingly.

Thanks for your listening!