SocraticSeminar | Builders Blog | Jan 2025

"Improving Factuality and Reasoning in Language Models through Multiagent Debate" `(Du et al., 2024)` is focused on black-box approaches toward improving reasoning capabilities in LLMs. The core thesis of the paper draws upon the heuristic observation that reasoning among humans, even those who may be unreliable sources of information, becomes more refined through debate and discussion. The authors draw inspiration from Marvin Minsky's *Society of Mind (* *Minsky, Society of Mind.pdf* *),* which offers a cognitive theory on how intelligence can be composed of non-intelligent modules. While I haven't yet had the chance to read *Society of Mind* in full, I always take careful note when an author I respect (in this case, Josh Tenenbaum) cites a book that they found influential toward their work.

I found the paper to be straightforward overall and highly intuitive. I also happen to agree with the most frequent of the opposing comments in review, which is to say that the paper fails to provide concrete logic as to why debate may improve reasoning capabilities among LLM agents. Instead, it focuses on a wide variety of experimental conditions and reasoning tasks. Some of the conditions include summarizing the context of fellow agents in a round and/or self-reflection before submitting round answers, while some of the reasoning tasks tested were chess move prediction, grade school math questions, biographic facts, and arithmetic tasks.

The point to be proved felt somewhat intuitive, and thus the permutations of different experimental conditions began to feel repetitive. In my opinion the most significant finding identified is that a first round where all agents yield incorrect results may be followed by subsequent rounds where the set of agents converge to the correct answer. One of the other important critiques of the paper is that it is computationally expensive (for example, see `Table A9` Chess Reasoning task, where the multiagent debate is on average > 20x more expensive than a single agent). Personally, I feel that this is fine because the paper stresses multiple times over that their approach is orthogonal to other approaches toward improving reasoning. I would have loved, though, to see suggestions in a further work section as to how the debate process can be transferred to internal representations as opposed to postprocessing on a black box. I would assume that this is what most readers are left thinking about, as it is the natural solution to the problem of compute cost.

In general, I found the paper to be intuitive but somewhat "hacky", especially because it fails to identify how the interesting phenomenon of enhanced reasoning through debate might transfer to internal representations. (hacky because, hey, aside from the overkill of experimental cases this black box approach is not very difficult to implement, as seen below —>).

Below, I have attached my annotated copy from reading the paper:

IMPLEMENTATION

First, I will aim to implement the debate on a subset of the reasoning tasks in 3.1 and 3.2 (Arithmetic Task, GSM8K, BIG-bench Chess Move Validity, MMLU). Additionally, I reproduce the experimental conditions of self-reflection and context summarization, as well as a debate modification I have termed "Semantic RAFT", which I explain in more detail below.

I initially deployed an instance of `Llama-3.3-70B-Instruct` using GCP / Vertex AI as the LLM for my agents. Note, I had to increase my default quota from 2 to 8 `NVIDIA_L4` `g2-standard-96` in order to deploy the model. To be honest, I found the pricing to be BRUTAL (:...)), so I refunded and decided to use `Llama-3.3-70B` via `Llama API` instead.

I use `openai/gsm8k` and `cais/mmlu` from Hugging Face, and I design the arithmetic tasks from scratch. The Chess Validity task is not available via Hugging Face, so I pulled directly from the `BIG-bench` repo, focusing on real games of medium length, here:

`BIG-bench/bigbench/benchmark_tasks/chess_state_tracking/real_medium/task.json`

My experimental setup consists of the following conditions:

Task	[Arithmetic, GSM8K, Chess, MMLU]
# Rounds	[1, 5]
# Agents	[1, 5]
Semantic RAFT	[True, False]
# Samples	5
Reflection	False
Context Summarization	False

To be clear, "`# samples`" indicates I report accuracy values averaged over `# samples`. Additionally, I assess the following conditions on the Arithmetic Task only. These conditions are easily reproducible for each of the other tasks, and I evaluate them for just one task as a proof of concept (so as to avoid excessive API costs).

Task	Arithmetic
Reflection	True
Context Summarization	True
Context Summarization + Reflection	True
# Samples	10

SEMANTIC RAFT PROTOCOL

The protocol I have defined for Semantic RAFT is inspired both by the original `RAFT` leader election process as well as the idea of assigning roles to the debate agents from the `(Du et al., 2024)` paper.

Note a few important differences between this and the actual `RAFT` leader election protocol:

- There are no “heartbeats” sent from leader to follower nodes because this protocol does not care for failure of nodes
- It is impossible to receive a vote request from a previous term
- No election timeouts
- Candidates do not instantiate elections independent of the other nodes and because of this they need not vote for themselves

In this sense, please note explicitly that the `Semantic RAFT` protocol is related to `RAFT` only in spirit. Below, I've included an outline of the protocol I implemented. Function names listed are common sense, and they implement exactly what they are named for.

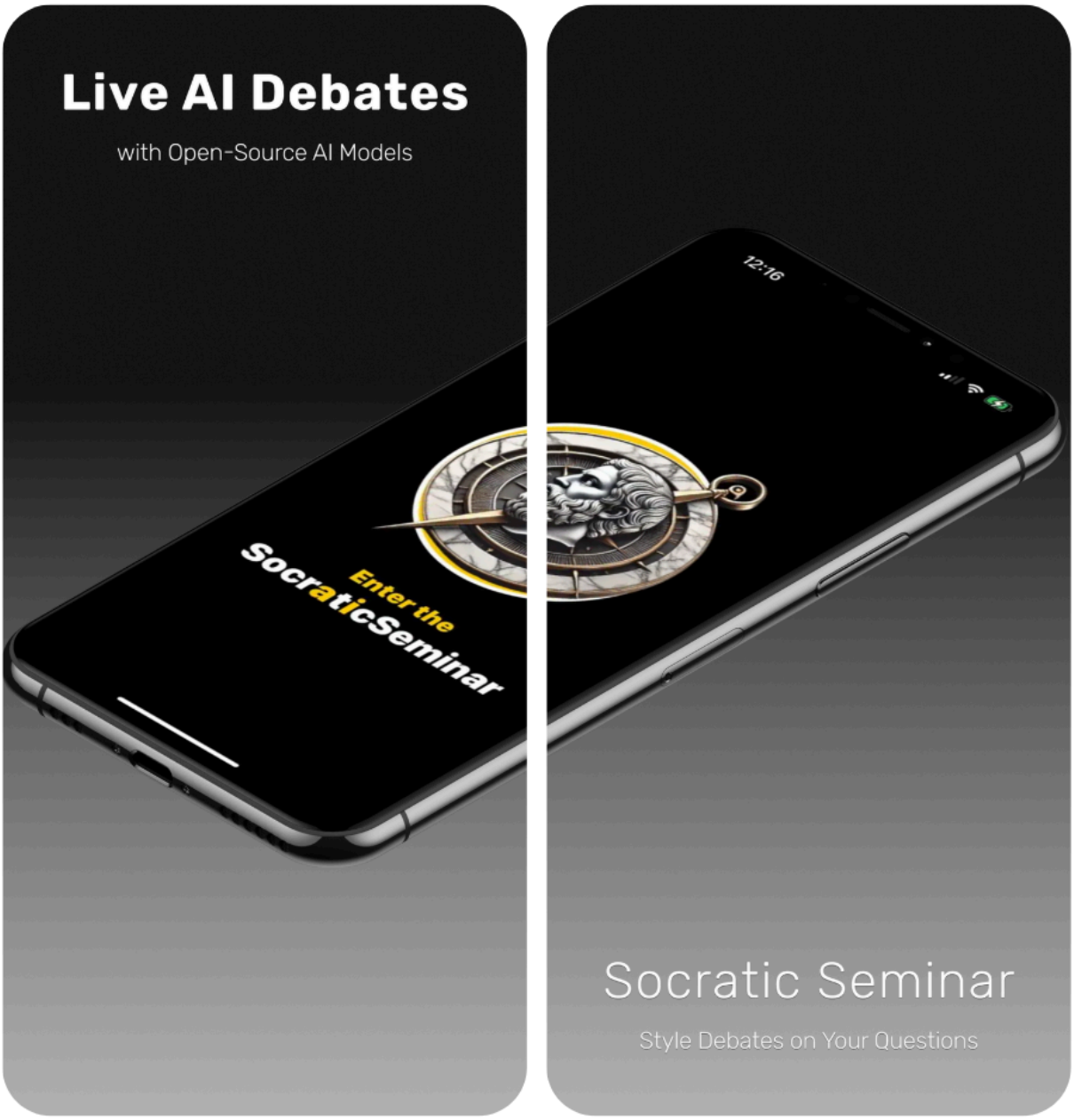
Algorithm 1 Semantic RAFT Debate Protocol

Input: n: number of agents, Q: debate question
$\mathcal{A} \leftarrow \{a_1, ..., a_n\}$ {set of agents}
$\mathcal{R} \leftarrow$ generateRoles(Q, n) {domain-specific roles}
assign($\mathcal{A}$, $\mathcal{R}$)
$votes \leftarrow \emptyset$
for agent $i = 1, \ldots, n$ **do**
  $v_i \leftarrow$ vote(a_i, $\mathcal{R}$)
  $votes \leftarrow votes \cup v_i$
end for
$leader \leftarrow$ getMajorityVote($votes$)
while debate not complete **do**
  $responses \leftarrow$ debateRound($\mathcal{A}$, $leader$)
  $analysis \leftarrow$ evaluateResponses($responses$)
  $votes \leftarrow \emptyset$
  for agent $i = 1, \ldots, n$ **do**
    $v_i \leftarrow$ vote(a_i, $\mathcal{R}$, $responses$)
    $votes \leftarrow votes \cup v_i$
  end for
  $leader \leftarrow$ getMajorityVote($votes$)
end while

RESULTS - APP

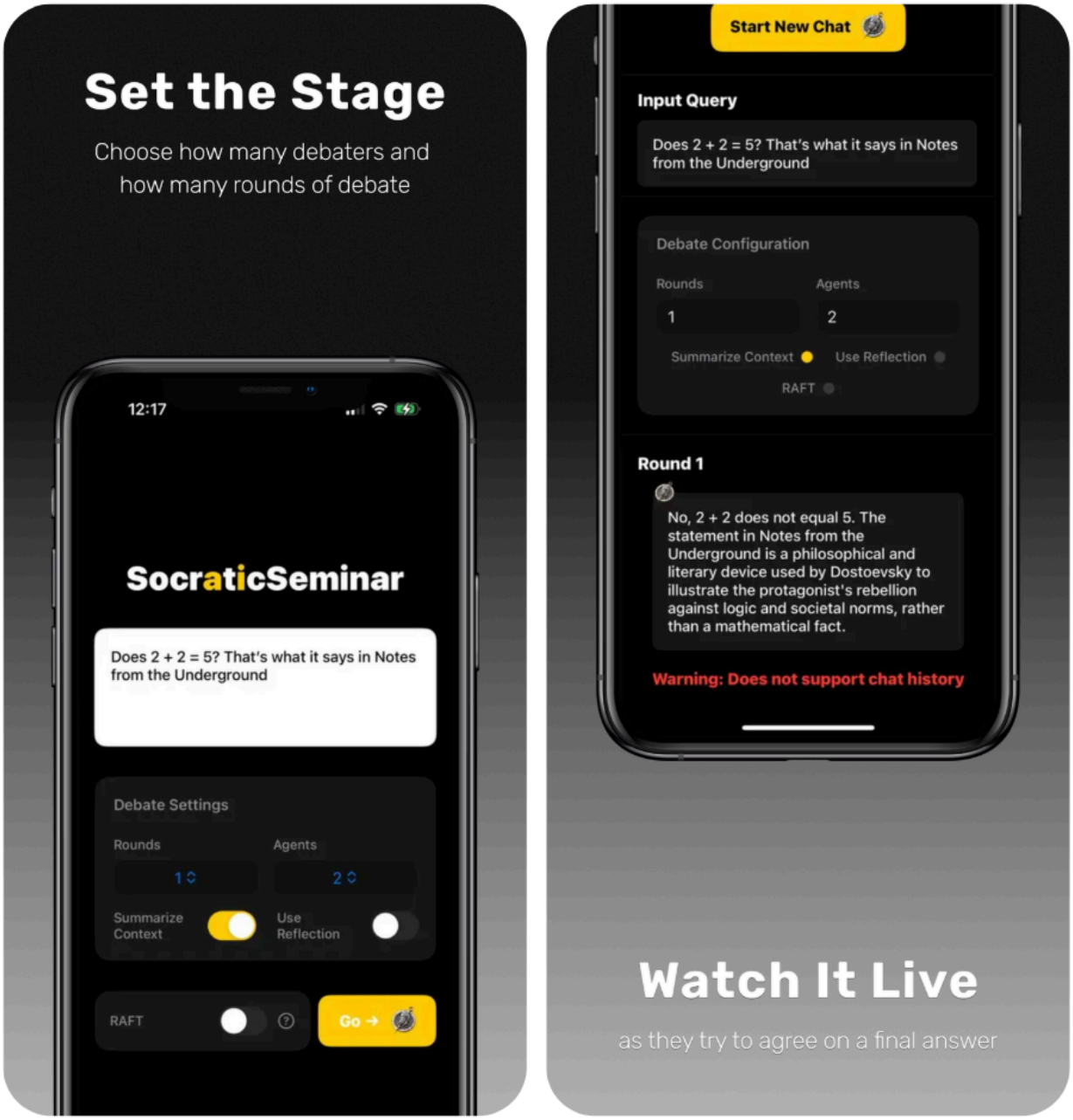
I wrapped a general-query version of my experimental LLM debate setup into a native iOS app. The Socratic Seminar app is now available on the iOS App Store at the link below.

https://apps.apple.com/us/app/socratic-seminar/id6740616041



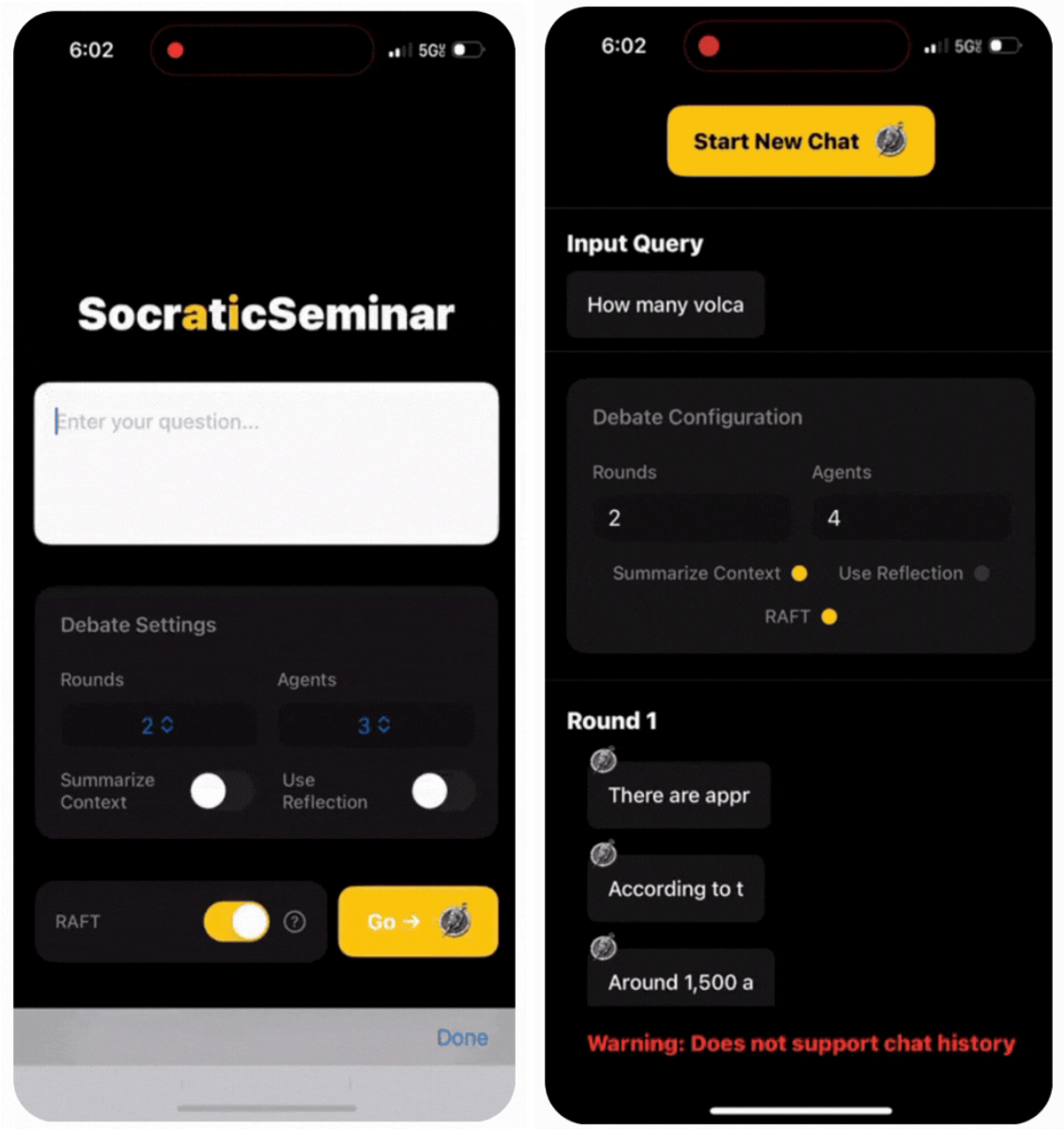
[App Store Screenshots 1 & 2]

The backend server is deployed for `us-central1 (Iowa)`, so the app may experience delays if you are using this internationally. Please let me know if this is a problem, and I can deploy an instance closer to where you are.



[App Store Screenshots 3 & 4]

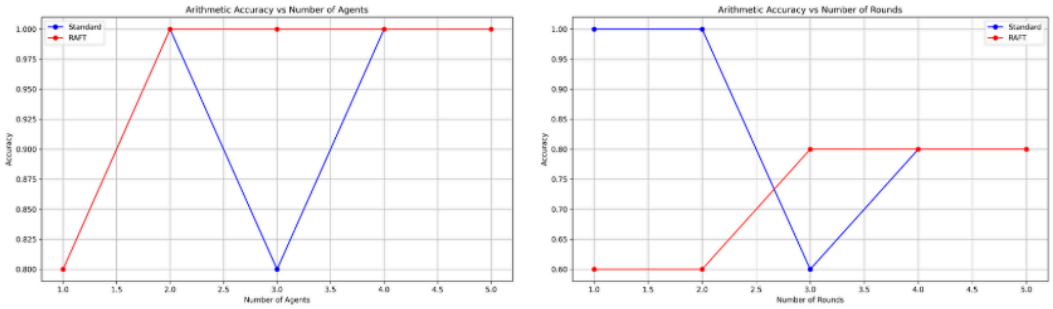
Below, I've included gifs that display the two central user experience flows:



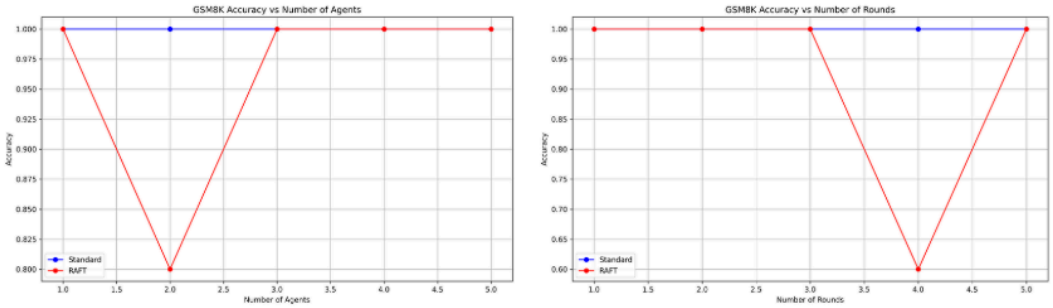
The app allows users to set parameters of the debate indicated in the `(Du et al., 2024)` paper such as context summarization of other agents' answers, self-reflection, the number of agents per round, and the total number of rounds. Additionally, it provides an optional toggle for my Semantic RAFT implementation. Rather than sample queries in the form of the curated tasks defined in the paper's experimental setup, the Socratic Seminar app takes as input a general user query as is done in any typical LLM interface. Notably, and unlike any typical LLM interface, I did not design the app to support chat history because I felt this was outside of the scope of the paper's experimental setup.

RESULTS - DATA

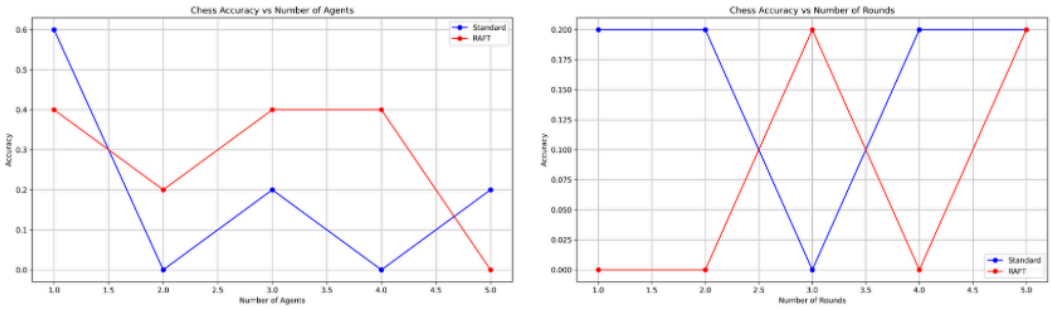
Below, accuracy is reported for each task averaged over 5 samples for both debate protocols.



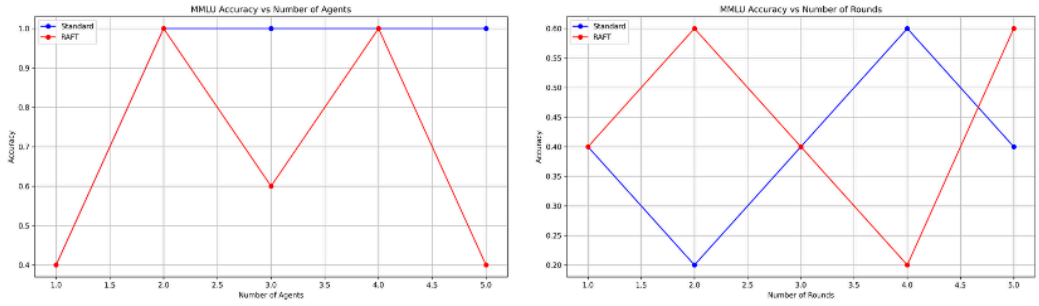
[Performance on Arithmetic Task Averaged over 5 Samples]



[Performance on GSM8K Task Averaged over 5 Samples]

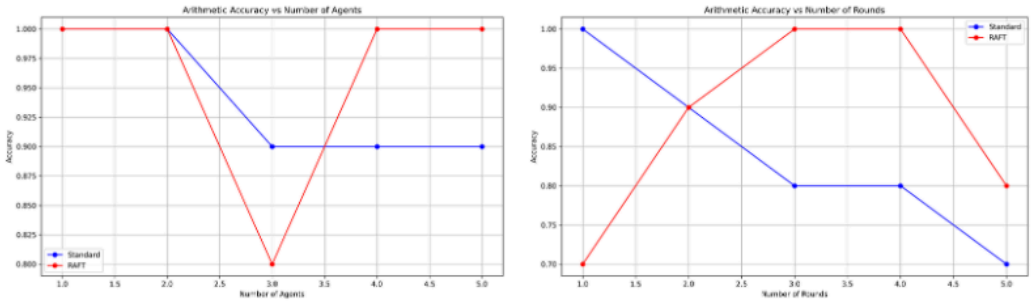


[Performance on Chess Task Averaged over 5 Samples]

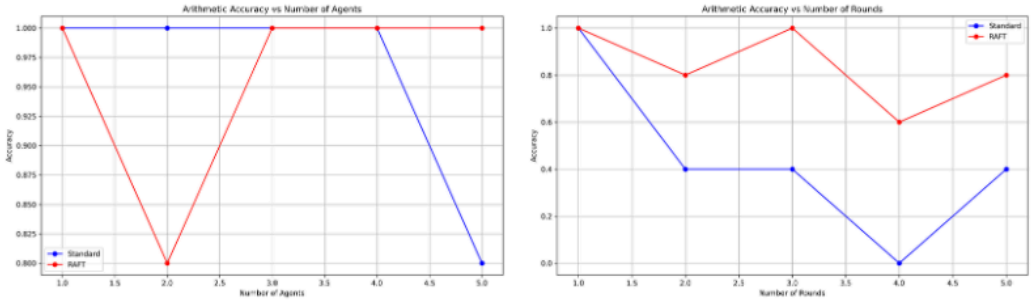


[Performance on MMLU Task Averaged over 5 Samples]

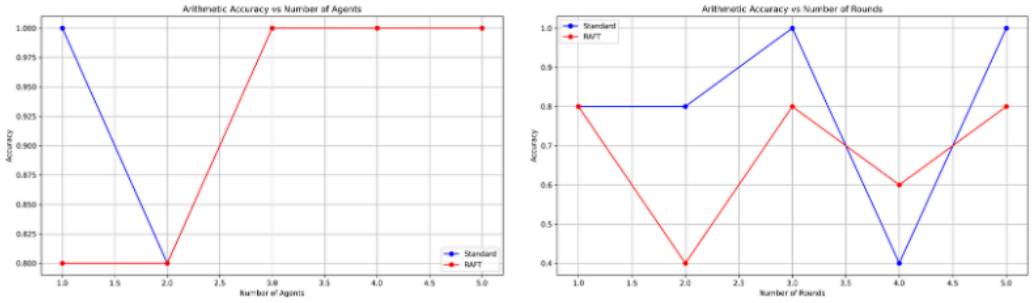
Additionally, accuracy is reported for the Arithmetic Task after changing one experimental condition in each case and holding all others constant.



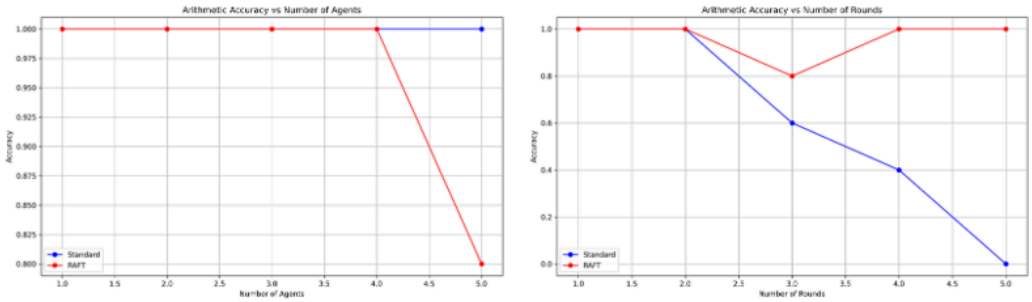
[Performance on Arithmetic Task]
[10 Samples, reflection=false, summarize_context=false]



[Performance on Arithmetic Task with Summarization]
[5 Samples, reflection=false, summarize_context=true]



[Performance on Arithmetic Task with Reflection]
[5 Samples, reflection=true, summarize_context=false]



[Performance on Arithmetic Task with Reflection + Summarization]
[5 Samples, reflection=true, summarize_context=true]

Averaging over all samples and experimental conditions for `# rounds`, `# agents`, `reflection`, and `summarize_context` for the Arithmetic Task only, the following composite accuracy values are achieved.

Task	Arithmetic
Standard	0.850
Semantic RAFT	0.884

While this cannot be used to make any definitive claims on the effectiveness of the Semantic RAFT debate setup, heuristically it makes sense that this style of debate modification should outperform a standard debate protocol. In fact, it touches on the nature of debate itself. Debate hardly counts as debate if all parties are in agreement, and enforcing roles in one way or another is a way to stimulate a more intense debate.

It is not necessarily true that opposing viewpoints are inherently better than conforming viewpoints, but rather that confrontation with opposition forces one to strengthen their own position and therein pushes them closer to the chance of landing on the truth. Among humans at least.

NOTE

As far as further work is concerned for this section, the FIRST change I would love to make is to compare performance across each of the 4 benchmarks for `Llama 3.3 70B` vs. `DeepSeek-R1-Distill-Qwen-32B`. I have a strong feeling that this distilled model will still outperform my results for `Llama 3.3 70B`.

Also, as a point of curiosity, below is a snapshot of Socratic Seminar's performance on the App Store as of 1/24/25. While it is a trivially small amount, still cool to see! Shoutout to my two users in Eswatini and Sweden.