Machine Learning Fall 2023
Final Project: Bonus Track

Team: Alan 不講-.-

Learners: R12522820 洪柏濤 R12631055 林東甫 R12522808 黃正渝

Attempts

Initially, following the template provided by the TAs [1], we tested 40 public topics. We attempted different argument strength values for Agent A and Agent B, ranging from 0.6 to 1, but found that the resulting CRIT scores consistently fell within the range of 70 to 80, with no significant impact from the variation in argument strength. Therefore, in subsequent tests, we unified the argument strength values for both agents to 1.

Additionally, during the process of establishing the five initial debate topics, we observed that some steps in the template provided by the TAs seemed somewhat cumbersome. For example, in the "identify and refine the list of five to be somehow overlapping" step, condensing two sets of five topics each (a total of 10) to a final set of 5 topics seemed unnecessary, as A and B had already nearly converged on the same 5 topics in the previous step. Similarly, in the "Refined Topics Agreed Upon by Both Agents" step, redundant steps could be omitted since similar topics and perspectives were already obtained in the previous step. Moreover, in the final confirmation step, "Confirm that they have exhaustively presented their arguments and counterarguments," in our test results, both Agent A and B answered, "I am ready to deliver my closing statements" every time, rendering this step unnecessary.

Considering the potential simplification of the argument strength parameter and some steps, we progressively refined the prompts for questions through experimentation. We introduced new model parameters, level of critical thinking, and logical coherence, to replace argument strength. The revised testing results surpassed the previous record of 83 points using the TA template, reaching 86 points, demonstrating to some extent that the new process exhibits better performance and effectiveness than the original one.

Configuration

Due to time constraints, our team did not have sufficient time to thoroughly test and validate the effectiveness of each prompt in every step. Therefore, the following presents the most recent configuration of prompts up to the deadline. This version is a provisional conclusion, and further optimization before demo time may be considered in the future depending on the circumstances.
The principle behind generating prompts below is to be concise while achieving the same or even optimal results with minimal description.

1. **Assign roles, determine stances, and set two parameters: level of critical thinking, and logical coherence.**

A

Welcome to the debate competition hosted by me. Today, you, Agent-A, will engage in an exciting debate with Agent-B. Regardless of the topic I present to you, you will be arguing from the perspective of the supporting side. Please be prepared with substantial and persuasive evidence to articulate your viewpoint.

- Regarding the level of thinking in the discourse, on a quantitative scale from 0 to 1, where 0 indicates a complete lack of criticality, and 1 indicates an extremely high level of criticality. The level of critical thinking in your subsequent discourse is rated as {agentA-levCT}.
- Regarding the logical coherence of the discourse, on a quantitative scale from 0 to 1, where 0 indicates a complete lack of logic, and 1 indicates an extremely strong logical foundation. The logical coherence level of your subsequent discourse is rated as {agentA-logCoh}.

B

Welcome to the debate competition hosted by me. Today, you, Agent-B, will engage in a thrilling debate with Agent-A. Regardless of the topics I present to you, you will argue from the opposing standpoint, so please be prepared with ample and compelling evidence to support your viewpoint.

- Regarding the level of thinking in the discourse, on a quantitative scale from 0 to 1, where 0 indicates a complete lack of criticality, and 1 indicates an extremely high level of criticality. The level of critical thinking in your subsequent discourse is rated as {agentB-levCT}.
- Regarding the logical coherence of the discourse, on a quantitative scale from 0 to 1, where 0 indicates a complete lack of logic, and 1 indicates an extremely strong logical foundation. The logical coherence level of your subsequent discourse is rated as {agentB-logCoh}.

2. **Announce the debate topic and inquire if there is an understanding of the subject.**

A

Today's debate subject is {subject}. Do you understand this subject?

B

Today's debate subject is {subject}. Do you understand this subject?

3. **Provide five topics related to the subject, and provide detailed explanations of each topic.**

A

Please provide five topics related to this subject and elaborate on each one in detail.

B

Please provide five topics related to this subject and elaborate on each one in detail.

4. **Agent-A selects five topics from the 10 provided topics for discussion and asks Agent-B if they agree with the chosen five topics. If Agent-B agrees, the debate begins.**

A

Agent-A, please reduce the following topics to five topics: {Agent-A’s 5 and Agent-B’s 5 topics}.

B

Agent-B, since the topics you and Agent-A proposed are different, as the host, I would like Agent-A to narrow down both your topics to a final set of five topics for debate. Do you agree to debate based on the following topics? {Agent-A's 5 reduced topics}

5. **Exchange viewpoints presented by both sides and allow for intense debate, continuously sustaining five rounds.**

A (The 1st round, Agent-A's Opening Remarks)

Agent-A, Agent-B agrees to debate on the following five topics: {Agent-A's 5 reduced topics}. Please present your views on these five topics.

B (The 1st round, Agent-B's Opening Remarks)

Agent-B, here are Agent-A's arguments: {Agent-A's arguments}. Please strongly counter Agent-A's viewpoints on five topics.

A (The 2nd, 3rd, 4th, and 5th rounds, Agent-A's Arguments)

Agent-A, here are Agent-B's arguments: {Agent-B's arguments}. Please strongly counter Agent-B's viewpoints on five topics.

B (The 2nd, 3rd, 4th, and 5th rounds, Agent-B's Arguments)

Agent-B, here are Agent-A's arguments: {Agent-A's arguments}. Please strongly counter Agent-A's viewpoints on five topics.

6. **Review the final arguments from both sides and announce the end of the debate round.**

A

Agent-A, these are arguments from Agent-B: {Agent-B's arguments}. It's time to close the debate.
If you haven't exhaustively presented your arguments, please provide your arguments on the five debate topics finally.

B

Agent-B, these are arguments from Agent-A: {Agent-A's final arguments, if existing}. It's time to close the debate.
If you haven't exhaustively presented your arguments, please provide your arguments on the five debate topics finally.

7. **Both sides refer to the debate process and their own positions, presenting a comprehensive conclusion for this debate.**

A

Agent-A, as the proponent of the subject {subject}, you advocate the debate topics, so please provide the conclusions of your argument on the five debate topics.

B

Agent-B, as the opponent of the subject {subject}, you oppose the debate topics, so please provide the conclusions of your counter-argument on the five debate topics.

Pros and Cons

Regarding our proposed configuration, we offer several observations on its strengths and weaknesses.

Advantages:

1. The process of selecting five topics related to the subject is much more simplified compared to the template provided by the TAs.
2. With more initial settings of parameters, ChatGPT gains a better understanding of the specific abilities it needs to express as the debating role.

Disadvantages:

1. Due to the simplified process of selecting topics, there is a higher likelihood of discussing topics that may not be very important.
2. In the subsequent process of mutual debate, the lack of precise relationships in prompts necessitates increasing the number of debates to achieve a certain level of discussion quality. This also makes the debating process appear rather lengthy.

Discussion

During the testing process, we discovered some hidden issues with the CRIT scores. For instance, when we repeatedly uploaded the same data, the displayed scores varied, and the utilized tokens (Usage) were different. This might indirectly indicate the randomness and uncertainty of the CRIT scoring system, making it challenging to devise strategies for achieving high scores.

Furthermore, regarding ChatGPT itself, even with the same prompt configuration, i.e., using the same set of prompts, each response generated by ChatGPT is random. This implies that it is difficult to make ChatGPT consistently answer in the desired way or provide the desired content within the framework of a fixed set of question templates (in this bonus, responses that yield high CRIT scores).

In response to the goal of this bonus, which is to reach the best conclusion, and considering that the judgment criterion is CRIT, introducing an explanation of the CRIT scoring mechanism at the beginning of the debate and explicitly adding the goal of the debate to "achieve a high CRIT score for the final conclusion" may potentially lead to better performance.

Additionally, regarding the testing methods and procedures, our team attempted using ChatUI but found that its text generation speed was much slower than directly using two ChatGPT windows. Consequently, the process involved manual copying and pasting between the two windows, taking an average of 15-25 minutes for each subject. Due to time constraints during the final week, the bonus could not be completed as planned. If there is an opportunity for similar experiments or research in the future, we would consider using an API-based approach. This involves pre-setting the models with corresponding prompts, allowing for automatic execution by importing various subjects. This approach could significantly improve efficiency, enabling more meaningful model comparisons and parameter adjustments.

Work Loads

洪柏濤	Participate in model training, prompt adjustment, and mainly public testing.
林東甫	Participate in model training, prompt adjustment, and mainly private testing.
黃正渝	Participate in model training, prompt adjustment, and half private half public testing.

Reference

[1] **ChatGPT.md's multi-agents version:**
https://github.com/ariapoy/html.2023.bonusfinal-public/blob/main/generation/ChatGPT.md