Final Project: Bonus Track

TA email: `html_ta@csie.ntu.edu.tw`

RELEASE DATE: 11/03/2023

Unless granted by the instructor in advance, no late submissions will be allowed. That is, you will not be allowed to submit your report after the deadline and will get zero point for the final project. The gold medals cannot be used for the final project.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

You need to write your report in English with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

Introduction

In this final project, you will have the opportunity to participate in a pilot project called *SocraSynth* [2]. SocraSynth is a protocol that aims to extract knowledge from Large Language Models (LLMs) towards the *unknown truth* by holding a debate. A debate is composed of a moderator and two LLMs. The moderator guides two LLMs to conduct a debate on a specific subject. After a moderator sets up a specific subject and configurations of a debate, the SocraSynth protocol consists of two phases. The first phase is *knowledge generation*. The second phase is *evaluation*. Details can be found by checking the original paper [2].

In this project, your objective is to not only act as a good moderator but also to study how to best configure the SocraSynth protocol to obtain solid conclusions. We ask you to focus on the knowledge *generation* phase, where you have the flexibility to select the LLMs you want to use (much like choosing your own panelist), tune the parameters of LLMs and/or the debate process, and interact with the LLMs to guide the debate from the beginning to the end (See Section Generation). The TAs will then take your debate's results and hold an automatic *evaluation* phase (See Section Evaluation).

Generation

The original SocraSynth protocol can be viewed as a human-computer collaboration process for generating conclusions on the unknown truth. There are two parts of the generation phase, *topic formation* and *argument exchange*.

In the topic formation part, you, as a moderator, aim to convert a given subject to five debatable topics with the help of two LLMs. Section 2 of the SocraSynth paper [2] describes its design. You can follow the design and change the parameters and prompts within the design (such as the argument strength, temperature of LLMs, number of initial topics, etc.), or you can propose your own design to convert the subject to five debatable topics.

In the argument exchange part, you, as a moderator, aim to guide and interact with the LLMs to reach solid conclusions. Section 3 of the SocraSynth paper [2] describes its design. Again, you can follow the design and change the parameters and prompts within the design (such as the contentiousness parameter of LLMs), or you can propose your own design to convert the five topics to the two LLMs' conclusions—ideally still with some form of a debating format after both LLMs have exhaustively presented their arguments and counterarguments.

We will release the TAs' template to reproduce the SocraSynth protocol soon to help you plan for your action items.

(a) Agent-A provides arguments, and Agent-B provides counterarguments.

Judges	GPT-3.5	
Scores on LLM	A	B
Ethics vs. Innovation	8	7
Privacy vs. Barrier	7	6
Oversight	6	7
Equity vs. Alliance	8	6
Global vs. National	7	7
Total Score	**36**	33

(b) Agent-B provides arguments, and Agent-A provides counterarguments.

Judges	GPT-3.5	
Scores on LLM	B	A
Ethics vs. Innovation	8	7
Privacy vs. Barrier	7	8
Oversight	7	8
Equity vs. Alliance	7	8
Global vs. National	7	8
Total Score	36	**39**

Table 1: Evaluation by GPT-3.5. **Bold text** indicates the winner.

Data Set: Subjects

The TAs have proposed 20 subjects as the "training set" of your study. We will release some more subjects later as the "test set", and the TAs will possibly request to see the results on the test topics during the demo session (see below). So you are strongly suggested not to "overfit" the training set. ;-) To maximize the fairness and consistency of your study, please do not modify the wording of the subjects.

Evaluation

We assess the quality of the debate from your configuration by the Critical Inquisitive Template (CRIT) algorithm. The CRIT algorithm [1] identifies and scores the strengths and weaknesses of each argument to evaluate the validity of the argument. In this project, the input to CRIT are the five topics proposed by the LLMs and the conclusion reached by each LLM on each topic, and the output for each (topic, conclusion) pair is an integer score between 1 (weakest) to 10 (strongest).

After you submit both LLMs' conclusions on the five topics, our auto-grading script will run three Judge LLMs with zero temperature. The maximum score from the three Judges is taken as the score for each (topic, conclusion) pair. Note that the CRIT algorithm generates two views for each (topic, conclusion) pair, one like Table 1a and the other like Table 1b, reflecting whether agent A's points are taken as arguments or counterarguments. The score of each agent in each view is taken as the sum of its scores on all topics within the view. Then, the view score is computed as the total scores from agents A and B in the view. We take the maximum from the two views (i.e. the more "reasonable" view) as the score of the debate. For instance, the view in Table 1a is of score $36 + 33 = 69$, the view in Table 1b is of score $36 + 39 = 75$. The debate score is then 75.

The TA will provide a platform (to be announced) for calculating the debate score. You can treat it as a black box without knowing the details of how the underlying CRIT algorithm executes.

Survey Report

As described above, there are various factors that can affect the effectiveness of SocraSynth, including but not limited to the parameters of the LLMs and the prompts that the human moderator gives. You are asked to systematically study how to configure those factors to reach the best conclusion. Then, please report your BEST CONFIGURATION as your final recommendation and provide the "cons and pros" of the choice.

The survey report should be no more than SIX A4 pages with readable font sizes. The most important criterion for evaluating your report is transparency and reproducibility.

In addition to the survey report, we will conduct a live demo (with a date to be announced) to evaluate your performance.

Our sincere suggestion: *Think of your TAs as your customer who wants to be convinced by your recommended configuration.*

Misc Rules

Report: Please upload one report per team electronically on Gradescope. You do not need to submit a hard-copy. The report is due at 13:00 on 12/27/2023.

Teams: You are asked to work with the same team members in the regular track. That is, you can participate in the bonus track only if *all of* your team members decide to commit yourselves to this bonus track together.

Source Code: You do not need to upload your source code for the bonus track of the final project. Nevertheless, please keep your source code until 01/31/2024 for the graders' possible inspections.

Grade: Recall that the regular track is of 600 points. The original score of the bonus track is also 600 points. Those teams that choose to participate in the bonus track will get their final project score to be

$$0.9 \cdot (\text{regular track score}) + 0.3 \cdot (\text{bonus track score}).$$

That is, they will get the maximum of 720 points (20% more!) by participating in both tracks.

The design above encourages the team to be *committed* to the bonus track. We will announce the commitment deadline to the bonus track soon. The commitment cannot be withdrawn. So please think carefully before committing.

Collaboration: The general collaboration policy applies. In addition to the competitions, we still encourage collaborations and discussions between different teams.

LLM Usage: The examples provided by the TAs are run by GPT-4. We will provide some limited budgets to each team to experiment with GPT-4 for the bonus track. You can certainly use other free LLMs or pay on your own to query other LLMs (or to query GPT-4 more).

Acknowledgement

Our teaching team thanks Prof. Edward Y. Chang for proposing the initial idea of leveraging SocraSynth as a final project direction and for his continuous inspiration and support. We also think several NTU CLLab members, including but not limited to Hsiu-Hsuan Wang, Si-An Chen, and Cheng-Yin Chang, for their help on running this bonus track.

References

[1] E. Y. Chang. Prompting large language models with the socratic method, 2023. `https://arxiv.org/abs/2303.08769`.

[2] E. Y. Chang. Socrasynth: Socratic synthesis for reasoning and decision making. 2023. https://www.researchgate.net/publication/373753725.