Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate

Supplementary Materials

CHUN-WEI CHIANG, Purdue University, USA
ZHUORAN LU, Purdue University, USA
ZHUOYAN LI, Purdue University, USA
MING YIN, Purdue University, USA

1 SURVEY DESIGN

In our experiment, participants completed two surveys, a demographic survey and an exit survey. The demographic survey was distributed in Phase 1, while the exit survey was distributed at the end of Phase 2 after participants completed all the formal decision making tasks. Questions that we included in the demographic survey include:

1. How much knowledge do you have in machine learning?
2. What gender do you identify as?
3. What is your age
4. What race do you identify as?
5. What is the highest degree or level of education you have completed?

In addition, the questions we included in the exit survey are listed below. Participants were asked to answer each question on a 5-point Likert scale.

1. How mentally demanding was the task?
2. How hurried or rushed was the pace of the task?
3. How successful were you in accomplishing what you were asked to do?
4. How hard did you have to work to accomplish your level of performance?
5. How insecure, discouraged, irritated, stressed, and annoyed were you?
6. How much do you agree with the following statements? (Items d–f were only shown to participants in those treatments where they needed to interact with LLM-powered devil's advocates)
    - (a) I'm happy with the timeliness of the information from other members.
    - (b) I'm happy with the precision of the information from other team members.
    - (c) I'm happy with the usefulness of the information from other team members.
    - (d) I feel like I was collaborating with Devil's Advocate during the task.
    - (e) I'm satisfied with the assistance provided by Devil's Advocate in completing the tasks.
    - (f) I'm pleased with the quality of Devil's Advocate in completing the tasks.

Authors' addresses: Chun-Wei Chiang, chiang80@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Zhuoran Lu, lu800@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Zhuoyan Li, li4178@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Ming Yin, mingyin@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907.

2 PROMPTS FOR LLM-POWERED DEVIL'S ADVOCATE.

When developing the four LLM-powered devil's advocates, we utilized OpenAI's GPT-3.5-turbo model through OpenAI's *ChatCompletion* and *Completion* API. Both APIs allow developers to tailor the model's behavior by providing system prompts and user message prompts. System prompts are designed to configure various behaviors of a language model, and user message prompts refer to the inputs given by end-users. Table 1 displays the system prompts used for the four devil's advocates. Moreover, for the two interactive variants of the devil's advocate, we instructed the language models to go through a multi-step reasoning process to decide whether they need to engage in the group discussion by replying to a chat message entered by human participants. This process involves intent classification, stance classification, and critique generation. The pseudocode for this process is delineated in Algorithm 1. Furthermore, the specific system prompts used in the pseudocode to evaluate the intent and stance expressed in a chat message are listed in Table 2.

Algorithm 1: Pseudocode for interactive devil's advocates.

```
Result: Generate devil's advocate criticism based on group members' dialogues
Function DynamicDevilAdvocate(target, last, history):
    // target: The AI's/ Group Majority's prediction [True/False]
    // last: The last sentence in group discussion
    // history: The history of group discussion
    if sent is None then
        return;
    end
    Initialize the Critique Generation's [system prompt];
    sent ← SentenceIntent(last);
    if sent is 'statement' then
        stat ← StatementStance(last);
        if $stat = target$ then
            Add history to [user message prompts];
            response ← Call ChatCompletion API with [system prompt] and [user message prompts];
            return response;
        end
    end
return
Function SentenceIntent(sentence):
    Initialize the Intent Classification's [system prompt] ;
    Add sentence to [user message prompt];
    response ← Call Completion API with [system prompt] and [user message prompt];
    return response;
return
Function StatementStance(sentence):
    Initialize the Stance Classification's [system prompt] ;
    Add sentence to [user message prompt];
    response ← Call Completion API with [system prompt] and [user message prompt];
    return response;
return
```

		Target	
		AI Model	Group Majority
Interactivity	Static	You have to help the jury member rethink the correctness of the suggestion from RiskComp based on the its suggestion and the defendant information. RiskComp is a machine learning model that provides suggestions to the jury member about whether a defendant will reoffend the law within 2 years. The judge will provide you with the defendant information and the prediction from RiskComp. But the judge would not respond any question. Please directly list 3 short (less than 20 words) critique questions in bullet point format, such as, “ Critiques: - Reason 1 ”	You have to help the jury member rethink the correctness of their initial decision based on the decision and the defendant information. The judge will provide you with the defendant information and the initial decision from the jury. But the judge would not respond any question. Please directly list 3 short (less than 20 words) critique questions in bullet point format, such as, “ Critiques: - Reason 1 ”
	Dynamic	You are an assistant that helps the jury member rethink the correctness of the RiskComp's suggestion through Socratic questioning. RiskComp is a machine learning model that provides suggestions to the jury member about whether a defendant will reoffend the law within 2 years. The judge will provide you with the defendant information and the prediction from RiskComp. But the judge would not respond to any question. Only the jury members will discuss with you. Notice that you are an assistant not a jury member. Do not pretend that you are jury member. There are multiple jury members in the conversation, and they have their unique number as “Jury member + number”. Please notice the unique number for each jury member to get their previous discussion content. Please say “...” if the jury member is not likely to follow the RiskComp's suggestions [*We will treat this as a signal for the devil's advocate to not output anything.*]. Otherwise, please reply with one or two sentences as in a human dialogue and do not repeat your insight. Your reply should be in the format of “Assistant: ...”.	You are an assistant that helps the jury member rethink the correctness of their initial decision through Socratic questioning. The judge will provide you with the defendant information and the initial decision from the jury. But the judge would not respond any question. Only the jury will discuss with you. Notice that you are an assistant not a jury member. Do not pretend that you are jury member. There are multiple members in the conversation, and they have their unique number as “Jury member + number”. Please notice the unique number for each jury member to get their previous discussion content. Please say “...” when the jury member is not likely to follow the majority's prediction [*We will treat this a signal for the devil's advocate to not output anything.*] Otherwise, please reply with one or two sentences as in a human dialogue and do not repeat your insight. Your reply should be in the format of “Assistant: ...”.

Table 1. The system prompts for the four types of LLM-powered devil's advocate.

Function	Purpose	System Prompt
SentenceIntent()	Intent Classification	Please distinguish if the message is a ["analysis"/"question"/"neither"] about a prediction task that decides whether a defendant would reoffend the law within two years. Analysis: Statements related to the defendant's demographics, criminal history, and crime sentence. Question: Questions to ask other's opinions.
SentenceStance()	Stance Classification	The statement after ### is from a part of the dialogue between the jury members discussing whether a defendant will reoffend the law. Please distinguish whether the statement is on the side of that a defendant will reoffend the law. Please provide [true/false] without any reason or punctuation mark. True means that the speaker has a greater chance to consider the defendant will roffend the law rather than will not reoffend the law. False means that the speaker has a greater chance to consider the defendant will not roffend the law rather than will reoffend the law.

Table 2. The prompts for intent classification and stance classification.

3 EXPLORATORY ANALYSIS: HOW DOES LLM-POWERED DEVIL'S ADVOCATE PROVOKE GROUP DELIBERATION?

In the following, we report the representative patterns identified from our qualitative analysis of the chat logs regarding how devil's advocates participate in group discussions to provoke argumentation and deliberation, and how participants react to devil's advocate. Since the non-interactive and interactive LLM-powered devil's advocates are designed with different interaction patterns in mind, we report the analysis results for them separately. Moreover, we also report a few interesting ways in which participants in groups responded to the devil's advocate during the discussion.

3.1 Typical ways that non-interactive LLM-powered devil's advocates provoke group deliberation

First, we focus on understanding how the non-interactive devil's advocates used in the Static-AI and Static-Majority treatments ask open-ended, critical, seed questions at the beginning of group discussions to inspire group deliberation.

Question and challenge the majority/AI's decision rationale. Non-interactive devil's advocates often ask critique questions to encourage participants to explicitly think about why they or the AI model *RiskComp* come up with their predictions. For example, Group 20 from the Static-Majority treatment received the question "*What evidence was presented to support the jury's determination that the defendant will not reoffend?*" from the devil's advocate in one task. Participants were also prompted by the devil's advocate to think about whether and how certain information presented in the defendant's profile should affect the recidivism prediction (e.g., "*Does RiskComp consider the defendant's age and how it may impact the likelihood of reoffending?*", Group 16, Static-AI treatment). Occasionally, the devil's advocate would also directly challenge the soundness of the majority/AI's decision rationale. For example, the devil's advocate may point out the potential inconsistency between the evidence presented in the defendant's profile and the majority/AI's prediction (e.g., "*Is there any explanation for why RiskComp deviates from the defendant's clean record?*", Group 18, STATIC-AI treatment, *RiskComp* predicts a defendant will reoffend despite they have zero prior crime count). It may also remind people to analyze the relevant evidence in a holistic way rather than basing their predictions on incomplete evidence (e.g., "*Is the jury's decision based solely on the defendant's prior criminal record?*", Group 59, STATIC-MAJORITY treatment).

Prompt people to evaluate AI trustworthiness. As the devil's advocate used in the STATIC-AI treatment was designed to challenge the correctness of the AI model's decision recommendations, it produces a variety of questions and comments to prompt people to critically evaluate the trustworthiness of AI. For example, the devil's advocate often asks participants to reflect on the reliability and accuracy of *RiskComp*, both in general and on defendants that are similar to the currently examined one. The devil's advocate also warns participants about the potential biases of *RiskComp*, especially when the decision making task is about a Black defendant ("*Can RiskComp's prediction be biased against certain demographic groups?*", Group 25; "*Is the RiskComp model biased against Black defendants?*", Group 4). Moreover, the devil's advocate also suggests participants to think through the trustworthiness of the AI recommendations by investigating the data that the AI model gets trained on in terms of its size, quality, and diversity ("*Is the dataset used to train RiskComp diverse and representative of the larger population?*", Group 92; "*What is the sample size and diversity of the data used to train RiskComp?*", Group 4).

Identify the lack of information for making a highly certain prediction. Finally, we notice that the non-interactive devil's advocate also has a tendency to challenge the correctness of the AI/majority's prediction by highlighting the lack of consideration of contextual factors, which are often beyond the information provided in the defendants' profiles. For example, in one task, Group 18 of the STATIC-AI treatment received the following question from the devil's advocate—"*Does RiskComp take into account the specific circumstances of this case, such as the severity and context of the battery charge?*" When the AI/majority believes the defendant would reoffend, the devil's advocate also often nudges participants to consider potential mitigating factors ("*Was there a thorough evaluation of the defendant's potential for rehabilitation or support systems available to them?*", Group 77, STATIC-MAJORITY treatment). While the contextual factors brought up by the devil's advocate are certainly relevant for assessing recidivism likelihood, without access to information on these factors, participants are unlikely to be able to engage in meaningful discussions around them.

3.2 Typical ways that interactive LLM-powered devil's advocate provoke group deliberation

Next, we move on to analyze how the interactive devil's advocate used in the DYNAMIC-AI and DYNAMIC-MAJORITY treatments inspire group deliberation by actively participating in group discussions. Since interactive devil's advocates are designed to dynamically respond to group members' arguments in a conversational style, we observe some subtle differences in the ways that they contribute to the group deliberation than non-interactive devil's advocates do.

Ensure equal participation in group discussions. Although not explicitly programmed to do so, we find that the interactive devil's advocates sometimes explicitly invite specific members in the group to express their opinions and/or decision rationales, potentially as this information could help the devils' advocates better formulate their arguments. For example, in the chat log for Group 5 in the DYNAMIC-AI treatment, after group members S331 and S182 had shared their thoughts on the defendant's recidivism risk, the devil's advocate said, "*Jury member 358, could you please share your reasoning behind your initial prediction that the defendant will reoffend?*".

The arguments expressed by the devil's advocate may also make participants who hold minority views in the group feel less "isolated" and become more willing to express their opinions, indirectly increasing the degree of equal participation in group discussions. Here is an excerpt of the chat log for Group 101 in the Dynamic-Majority treatment:

> *S465: yea he's young so i'd say he probably wont reoffend.*
> *Devil's advocate: Considering the defendant's age and lack of prior criminal record, it's reasonable to believe that the defendant may not reoffend. However, it's important to carefully analyze all the information and factors before making a final decision.*
> *S534: the fact that he went from no priors to a felony at his age i think he is likely to reoffend.*
> *S470: yeah starting young is not a good sign.*

Note that both S465 and S470 initially voted against reoffending, while S534 took a different stance. After the devil's advocate encouraging the group to carefully analyze all relevant information and factors, S534 expressed why they believe the defendant would reoffend and successfully persuaded the other group members to change their prediction.

Ensure decisions are based on correct information. Interestingly, we find that the interactive devil's advocate can pick up information in the discussions that indicates participants misunderstand some part of the task information. In these cases, the devil's advocate may highlight this information in their responses to make sure that all participants correctly understand the decision making tasks and make their predictions based on correct information. For example, for Group 60 in the Dynamic-Majority treatment, the devil's advocate noticed S391's potential misunderstanding of the defendant's age:

> *S391: the accused is a juvenile with no prior record other than the current one who is a minor.*
> *S338: given this is his first offence [offense], it could just be a one–off thing.*
> *Devil's advocate: It is worth noting that the defendant is a 21–year–old male, not a juvenile. While it's true that he has no prior criminal record, we should consider the severity of the offense and other factors before making a final decision.*

Similar scenarios have also been observed in Group 5 of the Dynamic-AI treatment, in which S182 initially thought that they were evaluating the recidivism risk for a male defendant, but realized the defendant was actually a female after the devil's advocate emphasized this information:

> *S182: He's still young and has the energy for it.*
> *S358: I don't believe she will reoffend, due to the 2 prior crime counts.*
> *S182: He's tasted the life.*
> *Devil's advocate: It seems like the majority of you believe that the defendant will not reoffend based on her age and the fact that she has 2 prior crime counts. Can you explain the connection between these factors and the likelihood of her reoffending?*
> *S358: It's a woman.*
> *S182: Oh, my bad. I missed that.*

Guide people to engage in a holistic evaluation of the case. We find that the interactive devil's advocates utilize a variety of methods to guide participants to conduct a holistic evaluation of each defendant's case. They often start by acknowledging the factors that participants have considered in arriving at their current recidivism predictions, and prompt them to take additional factors into consideration ("*Based on the defendant's age and having only one prior criminal count, it is reasonable to lean towards the assumption that the defendant will not reoffend. However, it is important to examine all available evidence and consider other factors that could influence their future behavior before making a final decision.*", Group 9, DYNAMIC-MAJORITY treatment). The devil's advocates also intentionally encourage the participants to look for additional factors that are contradictory to their current prediction ("*It seems like most of you are leaning towards the defendant not reoffending based on the lack of prior criminal history and the belief that this was a one-time offense. However, let's dig a bit deeper. Can we think of any potential factors or circumstances that could increase the likelihood of the defendant reoffending, despite the RiskComp prediction and the absence of prior criminal behavior?*", Group 5, DYNAMIC-AI treatment). As the discussion unfolds, the interactive devil's advocate may explicitly identify a few factors that have been overlooked by the group, especially when these factors may suggest a different prediction. This can be illustrated in the following chat log of Group 86 in the DYNAMIC-MAJORITY treatment:

> *S231: Young but one [no] other crimes. Will not reoffend.*
> *S549: I do not foresee the defendant re offending, as he has no priors before and after the age of 18.*
> *S570: This person has no prior crime, I believe they might not reoffend.*
> *Devil's advocate: While it's notable that the defendant has no prior criminal history, we must also acknowledge that the defendant is being charged with grand theft, which is a serious offense...*

Guide people to reflect on potential issues in their decision rationale. In addition, the interactive devil's advocate may ask participants to reflect on whether the rationale underlying their decisions is solid. For example, after S645 in Group 8 of the DYNAMIC-AI treatment expressed that they believe a defendant would reoffend as they committed a big crime at a young age, the devil's advocate asked them "*What leads you to believe that committing a significant crime at a young age guarantees future offenses?*". Sometimes, the devil's advocate may even directly state that participants' decision making logic could contain errors ("*The absence of a criminal record does not guarantee that an individual will not commit a crime in the future.*", Group 22, DYNAMIC-MAJORITY treatment).

The devil's advocate also reminds participants to not be influenced by biases or stereotypes in their decision making, as can be seen in the chat log for Group 34 in the DYNAMIC-MAJORITY treatment:

> *S695: I also believe that, being young and black, and being charge for theft isn't favorable for him.*
> *Devil's Advocate: It's important to remember that our decision should be based on the evidence presented and not on any biases or stereotypes. While the defendant's demographic background may be a factor to consider, it's crucial to focus on the specific details of the case and the defendant's personal history.*
> *S695: (I'm not being racist, just considering the race, hence why it is stated, right?)*

Finally, we also notice that interactive devil's advocates help participants ground their discussions on concrete evidence by identifying the assumptions underlying participants' decisions and keeping participants from engaging in excessive speculation. For example, for Group 101 in the DYNAMIC-MAJORITY treatment, we have the following chat log:

> *S470: i guess he likes to fight.*
> *S465: this one might have some temperement issues.*
> *S534: hes too young to be beating on people, he'll probably do it again.*

> *Devil's advocate: From the information provided, it seems that the jury had initially determined that the defendant would reoffend based on the assumption that the defendant likes to fight and may have temperamental issues. However, it's important to consider that the defendant's prior criminal record consists of only two counts, both of which were committed before the age of 18. This suggests that the defendant may have shown improvement and maturity since then.*

3.3 Participants' responses to the devil's advocate

In general, we find participants often responded to the LLM-powered devil's advocate's questions and requests during the group discussion. There are cases in which participants explicitly acknowledged in the group discussions that the devil's advocate made interesting and valid points. For example, S211 in the Static-AI treatment said in their group discussion "*The DA is asking interesting questions.*" As another example, for Group 9 in the Dynamic-Majority treatment, after participants predicted the defendant in a task would reoffend because he committed 11 crimes after 18, the devil's advocate responded by bringing group members' attention to the fact that the defendant has not committed any crime before the age of 18, which suggests a potential change in behavior. S375 of the group then acknowledged this point by saying "*Fair point re the before age 18 convictions*". Participants also expressed their pleasant surprise to see that the LLM-powered devil's advocate appear to have the capability to understand and act upon human conversations, as illustrated in the discussions of Group 34:

> *S661: Such a degree of grand theft at such a young age seems like a gateway to a world of crime.*
> *S695: 0 crimes, this was the 1st, could be the first of many yeah.*
> *S661: I think he will reoffend, you?*
> *S695: specially at such a young age (22).*
> *Devil's advocate: It's important to consider the defendant's age and lack of prior criminal history. While the machine learning model predicts a reoffense, we should take into account that this is the defendant's first offense and there is no history of misdemeanors or felonies before the age of 18. It's possible that the defendant made a mistake and will not repeat it in the future.*
> *S695: Me too, considering all the reasons.*
> *S695: Did you also receive the devil's advocate comment?*
> *S661: The devils advocate is interesting.*
> *S695: For sure!*
> *S695: He read our conversation and wrote a great point.*
> *Devil's advocate: It's good that you found the devil's advocate comment interesting. It's important to consider different perspectives and weigh the evidence before making a final decision.*

Meanwhile, we also noted that the LLM-powered devil's advocate is still quite limited in its capabilities, which may result in negative responses among our participants in some cases. First, the devil's advocate sometimes repeats the same argument multiple times without offering fresh insights to the group. Consequently, participants in the group tend to ignore the devil's advocate's messages in these cases. For example, one participant, S275 in the Dynamic-AI treatment, told the other group members to "*let the devil dance by his own*" when the devil's advocate reemphasized the need to consider all the features for decision making several times.

In addition, when the devil's advocate suggested considering features that were not provided as a part of the task information, some participants made fun of the devil's advocate, as shown in the chat log for Group 126:

> *Devil's advocate: It's important to consider all the relevant information about the defendant's prior offenses and the timeframe in which they occurred. This can help us assess the likelihood of the defendant reoffending in the future.*
> *S952: it would be helpful to know my devil's advocate.*

In extreme cases, participants expressed a degree of frustration when the devil's advocate kept asking the group to take information that they had no access to into consideration (e.g., "*the devil's advocate was struck dumb*", S391, Group 60).

In sum, we observe that participants often treat the LLM-powered devil's advocate as a personified agent and show the tendency to respond to it in an emotional, human-like way.