Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making

Shuai Ma
The Hong Kong University of Science and Technology
Hong Kong, China
shuai.ma@connect.ust.hk

Qiaoyi Chen
The Hong Kong University of Science and Technology
Hong Kong, China
qchench@connect.ust.hk

Xinru Wang
Purdue University
West Lafayette, Indiana, USA
xinruw@purdue.edu

Chengbo Zheng
The Hong Kong University of Science and Technology
Hong Kong, China
cb.zheng@connect.ust.hk

Zhenhui Peng
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong Province, China
pengzhh29@mail.sysu.edu.cn

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

Xiaojuan Ma
The Hong Kong University of Science and Technology
Hong Kong, China
mxj@cse.ust.hk

Abstract

Traditional AI-assisted decision-making systems often provide fixed recommendations that users must either accept or reject entirely, limiting meaningful interaction—especially in cases of disagreement. To address this, we introduce *Human-AI Deliberation*, an approach inspired by human deliberation theories that enables dimension-level opinion elicitation, iterative decision updates, and structured discussions between humans and AI. At the core of this approach is *Deliberative AI*, an assistant powered by large language models (LLMs) that facilitates flexible, conversational interactions and precise information exchange with domain-specific models. Through a mixed-methods user study, we found that *Deliberative AI* outperforms traditional explainable AI (XAI) systems by fostering appropriate human reliance and improving task performance. By analyzing participant perceptions, user experience, and open-ended feedback, we highlight key findings, discuss potential concerns, and explore the broader applicability of this approach for future AI-assisted decision-making systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713423

Keywords

AI-Assisted Decision-making, Human-AI Collaboration, Deliberation, Appropriate Reliance, Large Language Models

ACM Reference Format:
Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3713423

1 Introduction

With remarkable technological advancements, AI has been increasingly used to support people in making decisions in various domains, including criminal justice [30, 33], admissions [23, 135], financial investment [51], and medical diagnosis [19, 75], among others. Concerns surrounding AI's accuracy, safety, ethics, and accountability [12, 19, 75] have led to the widespread adoption of the *AI-assisted decision-making* paradigm in real-world applications [8, 16, 127, 137]. In this paradigm, AI performs an assistive role by providing a recommendation, while human decision-makers can choose to accept or reject it in their final decision [69].

Research in recent years, however, identified two challenges within the existing AI-assisted decision-making paradigm. **[Challenge 1]** First, a battery of empirical studies found that people rarely trigger analytical thinking when directly presented with AI's suggestions [11, 15, 105]. As a result, people frequently over-rely on the AI's incorrect recommendations (over-reliance) or mistakenly ignore AI's correct suggestions (under-reliance) [16, 85, 127]. Although some solutions have been proposed, such as displaying AI explanations [8] and forcing people to think more effortfully

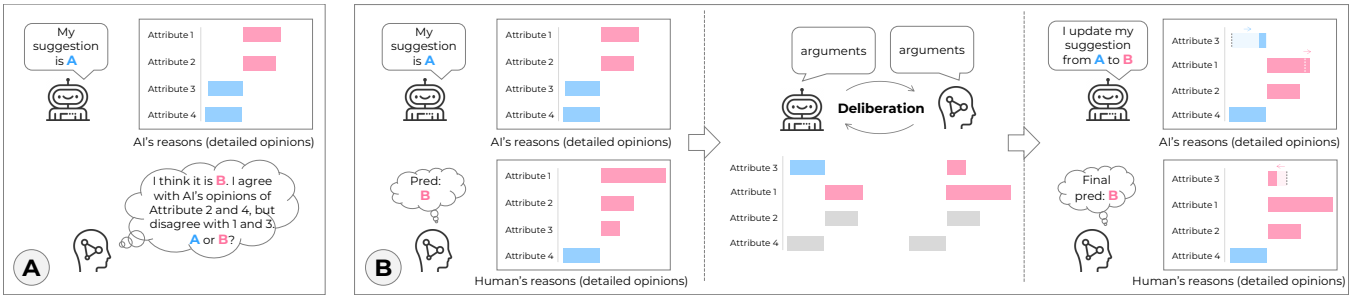

Figure 1: An illustration of *Human-AI Deliberation*. (A) In traditional AI-assisted decision-making, when humans disagree with AI's suggestions (and only find parts of AI's reasons applaudable), it is difficult for humans to decide whether and how much to adopt AI's suggestion. (B) In our proposed *Human-AI Deliberation*, we provide opportunities for the human and the AI model to deliberate on conflicting opinions by discussing related evidence and arguments. Then, AI and humans can update their thoughts (when find it necessary) and reach final predictions.

[16], the results are mixed at best [68, 103]. **[Challenge 2]** Second, instead of full consensus or complete divergence, human and AI decision rationales often exhibit partial alignment [113, 128]. While they may concur on certain aspects, differences may persist on others [95]. However, in current AI-assisted decision-making systems, AI always provides a fixed recommendation regardless of human thoughts and humans can only accept or reject AI's recommendation *as a whole* [69], with limited support for resolving conflicts or engaging in a meaningful exchange of ideas with the AI system [95]. For example, as shown in Figure 1 (a), when the human decision-maker's prediction is inconsistent with the AI model's recommendation and the human only partially agrees with the AI's reasoning (e.g., explanation), existing AI-assisted decision-making interfaces do not support any communication between humans and AI regarding conflicting opinions. This limitation may impede the effective utilization of both human and AI knowledge, hindering collaborative and complementary human-AI team performance.

Deliberation, characterized by thoughtful and reasoned discussion, plays a pivotal role in facilitating constructive discourse and consensus-building across various contexts [3, 115]. Deliberation proves effective in facilitating diverse human decision-making tasks, including deliberative politics [13, 52, 119], clinical diagnosis [60, 104, 107], criminal justice [27, 122], among others. It offers individuals an opportunity to rigorously evaluate different perspectives, including their own, which can potentially address Challenge 1 in AI-assisted decision-making. Moreover, deliberation allows participants to refine their viewpoints through informed discussions about opinion discrepancies [48, 49, 102]. Such a structured process may also enable humans and AI to engage in detailed discussions, potentially mitigating Challenge 2. Despite the potential benefits of deliberation, how to design mechanisms to facilitate deliberative conversation between humans and AI and how deliberations influence AI-assisted decision-making remain to be explored.

In this paper, building on established guidelines for enhancing discourse quality and identifying common ground in human deliberation [3, 4, 115], as well as the weight-of-evidence approach in decision-making [2, 10, 130], we propose a novel solution: *Human-AI Deliberation* for AI-assisted decision-making (Figure 1 (b)). Instead of presenting a fixed AI suggestion for humans to accept or reject, our approach encourages humans to externalize their thoughts, enables an interactive deliberation process between humans and AI around the conflicting points of their opinions and rationales, and fosters dynamic, fine-grained updates of humans and AI's decisions. The key component of this approach is *Deliberative AI*, which has the ability to locate viewpoint dissimilarities, stimulate comprehensive deliberation with human decision-makers, and make necessary changes, even compromises, in its own suggestion as the constructive discussion unfolds. To design such an AI assistant, we propose to integrate the strength of domain-specific models (for reliable assistant information generation) and Large Language Models (LLMs, for interactivity and conversation capability). We elaborate on the architecture design of *Human-AI Deliberation* and *Deliberative AI* in Section 3 and demonstrate how to instantiate the architecture in an illustrative task in Section 4.

Since the primary purpose of deliberation is to resolve conflicts between human and AI perspectives, we intentionally selected task cases with notable human-AI disagreements for our user study, drawing from insights in our pilot study. As human-AI deliberation is designed to both resolve conflicts and mitigate inappropriate human reliance on AI, we investigate its impact on task performance and human reliance on AI. Additionally, since deliberation explicitly highlights these conflicts, we are particularly interested in examining its effect on human perceptions of AI and the overall decision-making experience. Specifically, using our proposed concept of *Human-AI Deliberation* as a research probe, we aim to explore the following research questions.

- **RQ1**: How will *Human-AI Deliberation* affect task performance and humans' reliance on AI suggestions?
- **RQ2**: How will *Human-AI Deliberation* affect humans' perceptions of the AI partner and their user experience?
- **RQ3**: How will humans perceive the effectiveness of the proposed *Human-AI Deliberation* and what can be improved for future *Human-AI Deliberation* design?

To answer these questions, we conducted an exploratory study using a graduate admissions task. We recruited participants with

graduate admissions experience (at least once admitted to a graduate program) on Prolific and asked them to predict an applicant's chance of getting an offer based on the applicant's profile. We compared the proposed *Deliberative AI* with traditional *explainable AI (XAI)* and *human alone* baselines. Our experimental results revealed that *Human-AI Deliberation* has the potential to enhance decision accuracy and promote appropriate reliance on AI recommendations compared to traditional XAI assistants. We conclude by discussing key implications and addressing generalizability concerns based on our design and study findings.

In summary, we make three contributions:

- We propose a novel architecture, *Deliberative AI*, designed to enable *Human-AI Deliberation*—a collaborative process where humans and AI deliberate together to resolve conflicting perspectives in decision-making tasks—by seamlessly integrating domain-specific models with large language models (LLMs).
- We demonstrate the instantiation of the *Deliberative AI* in an illustrative task (college graduate admission), including the implementation of different components and interface design.
- We conduct an exploratory study to gain an initial empirical understanding of how *Human-AI Deliberation* might impact the decision-making process and how humans would perceive this novel AI assistance. Additionally, we demonstrate its potential to improve decision accuracy and promote appropriate human reliance on AI.

2 Related Work

2.1 AI-Assisted Decision-Making: Objectives and Challenges

Artificial Intelligence (AI) is increasingly used in decision-making across various domains [26, 29, 88, 90]. However, AI's real-world applications are not infallible, still far from 100% accuracy [44, 86, 109]. This is especially concerning in high-stakes domains like medicine and criminal justice, leading to ethical and legal complexities [12, 19, 75]. To address this, the prevalent paradigm of *AI-assisted decision-making* has emerged, drawing substantial attention in the Human-Computer Interaction (HCI) and AI communities [8, 16, 127, 137]. In this paradigm, AI takes on an assistant role, offering recommendations for human decision-makers to accept or reject in their final decisions [69].

Research in AI-assisted decision-making spans a range of objectives, including enhancing team performance [85, 137], promoting decision fairness [24, 126], improving efficacy and efficiency [23], fostering understanding of AI [23, 127], building trust and appropriate reliance on AI [55, 108], and enriching subjective user experiences [79, 86]. Among these, a key goal is achieving complementary performance—where the collaborative outcomes of human-AI teams exceed what either humans or AI could achieve independently [8]. Despite its importance, recent empirical studies highlight persistent difficulties in reaching this goal [8, 105, 137], driven by two primary challenges.

One challenge is the underutilization of human and AI domain knowledge [5]. Some researchers aim to leverage the complementary aspects of human and AI intelligence by training AI to complement human knowledge [7, 132]. Moreover, existing AI-assistant interfaces do not efficiently harness the knowledge of both parties [117]. AI contributes its knowledge to humans by providing recommendations with AI explanations serving as a means of representing its detailed reasoning [69]. These explanations could facilitate the collaborative synthesis of human and AI intelligence, allowing them to combine insights into different features for final decisions. However, when conflicting views arise, current interfaces provide limited support for the communication and exchange of human and AI knowledge [79]. To address this, we propose *Human-AI Deliberation* to resolve conflicts through natural discussions.

The second challenge concerns human reliance on AI suggestions [8, 85, 87, 89, 137]. Achieving complementary performance relies on human decision-makers' ability to judiciously determine when to consider AI recommendations and when to be skeptical [16, 105, 137]. Both over-reliance, where individuals trust AI excessively [74, 100], and under-trust, where individuals fail to utilize AI when necessary [74], can lead to adverse outcomes. Successful decision-making requires individuals to decide whether and how to rely on AI recommendations on a case-by-case basis [6–8, 120, 137]. Current approaches present AI performance indicators, explanations, outputs, and confidence levels to assist humans in making informed decisions. However, existing research has found that when people are provided with a recommendation and passively look at it, they often lack analytical thinking, leading to over-reliance or under-reliance on AI systems [15, 42, 64, 79]. *Human-AI Deliberation*, as proposed in this paper, encourages a careful evaluation of AI rationales through discussions of conflicts in human and AI opinions. By engaging humans in the deliberation process, it promotes a more comprehensive understanding of AI insights, reducing the potential for both under-reliance and over-reliance.

2.2 The Role of Deliberation in Human Decision Making

The meaning of deliberation is *"the act of thinking about or discussing something and deciding carefully"* [91]. It involves considering all relevant individuals as moral agents who must justify their viewpoints and listen to others' reasons [50]. Rather than seeking consensus, the process aims to enhance individual perspectives by incorporating others' viewpoints, thus increasing decision maturity and wisdom [50]. The origin of group deliberation can be traced back to public deliberation or deliberative democracy, where citizens convene to discuss policies with potential implications for their lives [112]. Recent studies on online deliberation have showcased its ability to enhance the accuracy of crowd-working tasks [22, 32], improve perceptions of procedural justice [38], and facilitate consensus-building among participants [76, 107, 122, 134]. Furthermore, deliberation proves effective in facilitating various decision-making tasks, including clinical diagnosis [60, 104, 107], criminal justice [27, 122], and more.

Effective decision-making is crucial across various domains, and deliberation offers significant advantages. It improves decision quality and problem-solving by enabling comprehensive analysis and evaluation of options, fostering a deeper understanding of issues and outcomes [9, 71]. Deliberative decisions are often wiser and more effective due to their basis in thorough analysis [28, 70]. Additionally, deliberation promotes participation and collaboration,

encouraging stakeholder engagement and facilitating communication, which aids in resolving complex issues and ensuring decision acceptance [40, 97, 133]. Finally, it helps mitigate decision biases and enhance fairness by enabling objective analysis and reducing emotional influences [57, 66, 118].

Despite the significance of deliberation in decision-making, there is a dearth of research on its integration into AI-assisted decision-making processes. To address this gap, drawing upon theories and practices in deliberation [13, 52, 81, 116, 119], we propose *Human-AI Deliberation* to facilitate human reflection and discussion on conflicting human-AI opinions. Based on this approach, we aim to move a first step towards designing a *Deliberative AI* and investigating its effects on decision processes and outcomes through an exploratory empirical study.

2.3 Existing Studies on Deliberation in AI-Assisted Decision Making

Deliberation enhances decision-making by integrating diverse perspectives, improving solution quality, and fostering critical thinking [50]. It involves analytical reflection and active discussion[50], both of which have been explored separately in research on AI-assisted decision making.

To stimulate analytical thinking, different interventions have been designed to encourage deeper engagement with System 2 thinking [62], such as "cognitive forcing" techniques that prompt more deliberation. Examples include asking individuals to make independent predictions before receiving AI input [16, 101, 105] or using "slow algorithms" to reduce reliance on AI. Additionally, providing AI explanations without concrete recommendations [42] and using AI-framed questioning [25] have been shown to enhance critical thinking. However, these approaches may lead to under-reliance and may not fully address differences between human and AI perspectives.

Some studies have explored human-AI dialogues in cooperative games [37, 65], but these were not tailored to decision-making tasks. Recent work has begun integrating discussions into AI-assisted decision-making. For example, Zheng et al. [139] included AI in group decisions for student essay evaluations, though these efforts often rely on Wizard of Oz setups rather than purpose-built AI systems. Similarly, Chiang et al. [24] studied collaboration between AI and two humans in recidivism risk assessment but limited the AI's role to providing suggestions without active discussion. Other efforts, such as Zhang et al. [135], used AI models as tools to facilitate deliberation among organizations but did not address direct deliberation between humans and AI. Perhaps the most relevant work is by Slack et al. [114], who explored dialogue-based AI explanations to handle follow-up user questions and improve understanding. However, the key difference is that we focus on deliberation design and propose *Deliberative AI* which can not only "explain to users" but also actively engage users in the deliberative discussions by "asking or challenging" the users, aiming to promote people's critical thinking.

Previous work has touched upon the idea of having AI serve deliberation among humans, but to our knowledge, no research has directly facilitated deliberation between humans and AI. Our work takes an initial step toward designing and evaluating human-AI deliberation in decision-making contexts, offering valuable insights into integrating deliberation into AI-assisted decision-making processes.

3 Deliberative AI for Human-AI Deliberation

This section introduces deliberation into the decision-making process, focusing on weighing the evidence. Building on the Weight of Evidence (WoE) framework, we propose a *Human-AI Deliberation* architecture and present the design considerations and architecture of *Deliberative AI*, an AI assistant capable of engaging in deliberation with humans.

3.1 Integrating Deliberation into Decision-Making

Decision-making involves selecting the best choice from a range of alternatives to achieve a desired outcome [35]. The process is typically summarized in seven steps[1], with *weighing the evidence* being a critical step due to its direct influence on subsequent decision outcomes [83, 121]. We propose conducting *Human-AI Deliberation* during this phase because disagreements often arise between human perspectives and AI suggestions in this phase and these human-AI disagreements can lead to conflicting opinions and divergent outcomes in later steps [69].

To structure the deliberation process, we break down decision-making problems and human-AI thoughts into four components (Figure 2 (a)): (1) **Decision**: The overall choice to be made for a specific problem; (2) **Dimension**: An aspect considered when forming the overall decision. In tabular datasets commonly used in decision-making [45, 127], dimensions often correspond to attributes or sets of related attributes, such as academic excellence or research ability in a graduate admission task; (3) **Opinion** on a dimension: The assessment of a dimension's impact on the overall decision (e.g., academic excellence contributes +50% to admission probability); (4) **Evidence**: The foundation for forming an opinion. For humans, evidence may include facts, heuristics, or personal experiences [110, 111]. For AI, evidence is often rooted in the information encoded in its training data.

We propose deliberation at the dimension level, focusing on how each dimension supports or opposes the overall decision. Both humans and AI must substantiate their opinions with evidence and assess its credibility and probative value [10]. This process generates the Weight of Evidence (WoE) [20], which quantifies evidence significance/importance relative to alternatives [48, 49, 102]. WoE is widely used in decision-making for its intuitive meaning and practical implementation [49, 106]. In the rest of this paper, we will use "opinion" and "WoE" interchangeably.

3.2 An Architecture of Human-AI Deliberation

Based on the WoE-centered decision-making approach, we propose *Human-AI Deliberation*, an architecture to stimulate deliberative processes between humans and AI (Figure 2). This architecture comprises the following essential activities:

[1]Seven-step decision making: Step 1: Identify the problem, Step 2: Collect information, Step 3: Identify the alternatives, Step 4: Weigh the evidence, Step 5: Choose from the alternatives, Step 6: Implement action, and Step 7: Evaluate the results.

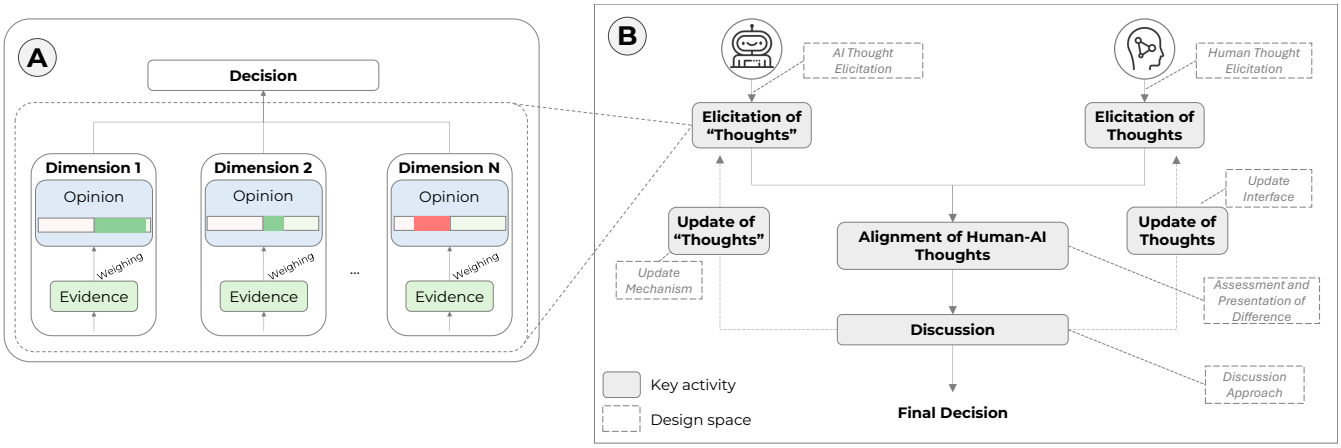

Figure 2: The architecture of *Human-AI Deliberation*. (A) Illustrates the Weight of Evidence (WoE) concept in decision-making, showcasing how decision-makers assess evidence across dimensions to shape opinions and arrive at a final decision. (B) Presents the Architecture for *Human-AI Deliberation*, with key activities (shown in grey boxes) and potential design space (shown in dashed-line boxes).

- **Elicitation of Thoughts**: Human and AI start with articulating their dimension-level perspectives on the decision problem. While AI presenting its "thoughts" (e.g., in the form of feature importance explanation) is rather common in AI-assisted decision-making [79], this activity also encourages individuals to clarify their ideas and examine their reasoning, which prompts analytical thinking in human [16, 92]. Two aspects of this activity require careful design. First, AI thought elicitation demands a good balance between human information needs and interpretability [1, 96]. Second, human thought elicitation, while encouraging thoughtful reasoning [16, 92], can impose a potential workload. It thus demands suitable, friendly interface designs.
- **Alignment of Human-AI Thoughts**: As human's and AI's viewpoints and the process they form those viewpoints may diverge [58, 63, 94], this activity is tasked with establishing a common language for the two parties to compare their WoE and determine the extent of discrepancy. Proper assessment and presentation of human-AI WoE differences can help effectively navigate humans' attention and efforts in the subsequent activities [14].
- **Discussion**: This activity fosters constructive discussions where humans and AI substantiate their opinions, clarify evidence choices, and explain weight assignments. It promotes critical thinking, reduces biases, and highlights differences between parties [56, 72, 92]. With a broad design space, it must be tailored to specific decision tasks, considering factors like content, style, leadership (who initiates and leads the conversation), and duration. A potential solution is adapting human-human discussion [56, 72, 92] to the human-AI discussion contexts.
- **Update of Thoughts**: In-depth discussions may expose potential flaws and conflicts in the original decisions as humans and AI are both imperfect [8]. This activity provides an opportunity for them to reflect on the gaps in thinking [92] and revise their thoughts accordingly. For AI, this means designing appropriate mechanisms to interactively update its recommendations. For humans, the interface should possess the flexibility for them to change their WoE.

In summary, the proposed *Human-AI Deliberation* architecture includes four interlinked activities and requires appropriate designs for both the decision-making interface and the AI (as shown in the dashed box in Figure 2 (b)). Since interface design is task-dependent, we focus on the design of *Deliberative AI* in the next subsection.

3.3 Deliberative AI: Design Considerations and Overall Structure

Based on deliberative theories [13, 52, 119] and practices [81, 116], the Discourse Quality Index (DQI) [115] and its improved versions [4, 17] provide a comprehensive framework for assessing human deliberation. We adapt DQI to AI-assisted decision-making and derive the following design considerations (DCs):

- **DC 1. Participation equality**: *Deliberative AI* should ensure that both parties possess equal voice [21] and share similar opportunities to offer opinions and reasons as well as to participate in discussions.
- **DC 2. Justification rationality**: *Deliberative AI* should adeptly provide rational justifications for its stances during interactions and encourage humans to do the same.
- **DC 3. Constructive updates**: Rather than rigidly adhering to its initial opinions or blindly leaning towards others, *Deliberative AI* should aim to facilitate compromise, reconciliation, or consensus as deliberation evolves. It should help both sides to think carefully and rationally and update their WoE in a timely manner.
- **DC 4. Interactivity**: *Deliberative AI* should be able to understand human intentions and dynamically generate appropriate responses based on human's questions, arguments, and statements.
- **DC 5. Respect and agreement**: *Deliberative AI* must ensure polite discourse and respect for other participants, especially during discussions. Even if it disagrees with humans on some

aspects, *Deliberative AI* should show respect and understanding, creating a positive environment for continued engagement.

To fulfill these considerations, we integrate Large Language Models (LLMs) and domain-specific models (DS models) to build *Deliberative AI*. DS models are responsible for the initial generation and subsequent refinement of AI's WoE. DS models' predictive power and domain knowledge offer reliable (instead of potential hallucination) information for deliberation activities. LLMs, on the other hand, bridge the interactions between humans and DS models with their conversation abilities. Overall, the architecture of *Deliberative AI* (illustrated in Figure 3) comprises three layers: **Communication** layer, **Control** layer, and **Knowledge** layer.

- The **Communication** layer, empowered by LLMs, incorporates three components:
  - *Intention Analyzer* (for DC 4) understands human intent and argument evidence, facilitating cross-referencing with the Knowledge layer through the Control layer.
  - *Deliberation Facilitator* (for DC 2&5) encourages careful thinking and rational justifications while maintaining respectful deliberation.
  - *Argument Evaluator* (for DC 2&3) assesses human justification rationality, which can be used to further prompt humans' reasoning and update AI opinions.
- The **Control** layer, encompassing four components, oversees:
  - *Dialogue Controller* (for DC 1) manages the deliberative discussion process (e.g., when to elicit thoughts, when to update opinions, when to move on to the next dimension, etc.)
  - *Regulator* (for DC 2) guides and constrains LLM output with domain-specific model insights and training data.
  - *Knowledge Extractor* (for DC 2) extracts data insights from domain-specific models and training data based on human intent analysis.
  - *Opinion Update Controller* (for DC 3) adjusts AI viewpoints based on human-AI dynamics (e.g., the strength of justifications, uncertainty behind AI's opinions, etc.).
- The **Knowledge** layer comprises a domain-specific model and training data, providing both domain-specific knowledge and data-derived insights.

In the next section, we will provide a detailed description of how we implemented *Human-AI Deliberation* and *Deliberative AI* architecture in the context of a graduate admissions task.

4 Instantiating the Architecture: Graduate Admission Prediction

4.1 Task, Dataset and AI Model

We choose to use graduate admission as an illustrative task to demonstrate how to instantiate the proposed *Human-AI Deliberation* architecture. In this task, participants decide on admitting or rejecting applicants to a U.S. university based on their profiles. We chose this task for two key reasons. First, this task is widely used in AI-assisted decision-making research [23, 34, 135, 138], with real-world universities employing AI algorithms for decision consistency and workload reduction [99, 129]. Second, graduate admission often involves deliberation among committee members [135], making it ideal for studying the effects of our proposed *Human-AI Deliberation*.

The task utilizes a synthesized dataset [23] that simulates profiles of applicants at a U.S. public university based on publicly available aggregate statistics and distributions[2]. The dataset comprises 100 student applications' profiles, featuring attributes considered by admission committees in actual scenarios, e.g., *GRE Verbal*, *GRE Quant*, *GRE Writing*, *GPA*, *Statement of Purpose Strength*, *Diversity Statement Strength*, *Country*, *Major*, *Applicant's Undergraduate Institution Rank*, and *Recommendation Letter Strength*. The dataset also includes a decision label for each case: strong reject, weak reject, weak accept, or strong accept.

To build a domain-specific model (DS-model) that can generate suggestions, we trained a multi-category linear model using a 70% random split of the dataset as in [23]. We employed a linear regression model as a decision classifier, discretizing the predicted responses into one of the four decision labels. Consistent with common practices [34], we further binarized the original labels, mapping strong/weak reject to "reject" and strong/weak accept to "accept" as the ground truth for assessing AI model's and participants' prediction accuracy. The trained model achieved an 80% accuracy on the remaining 30% test set. The task samples used in the study were selected from the test set.

4.2 Human-AI Thought Representation and Alignment

In our *Human-AI Deliberation* architecture, the initial step involves both humans and AI externalizing their thoughts. We employ feature contribution [78, 79] to represent their weight of evidence (WoE) along each dimension [2]. Feature contribution is represented by contribution scores indicating the positive or negative influence of each feature x_i on the final prediction y. In graduate admission tasks, we treat each attribute in an applicant's profile as a dimension, and feature contribution requires humans and AI to assess the influence of each dimension on the final decision. It provides a common ground for AI and humans to express, compare, and initiate discussions.

For the human side, Weight of Evidence (WoE) represents humans' perceived influence of an attribute on the overall likelihood of an applicant being admitted. On the AI side, we utilized SHAP (SHapley Additive exPlanations) [82], a widely-used posthoc explainable AI algorithm, to assess feature contribution/importance. SHAP values indicate both the direction and strength of a feature's impact on predictions. SHAP offers two key advantages: First, it captures feature interactions (e.g., how multiple features jointly influence the final decision), aligning with human decision-making processes. Second, its additive nature mirrors how humans combine evidence for or against options [10]. In essence, SHAP allows feature contribution/importance to be linearly aggregate to match the model's actual prediction no matter the AI model is linear or a complex non-linear neural network.

However, when applied to *Human-AI Deliberation*, two limitations of SHAP values become evident: (1) **Interpretability of Raw**

[2]Due to privacy issues, public graduate admission data sets are all synthesized. We acknowledged that it may deviate from the real-world setting.

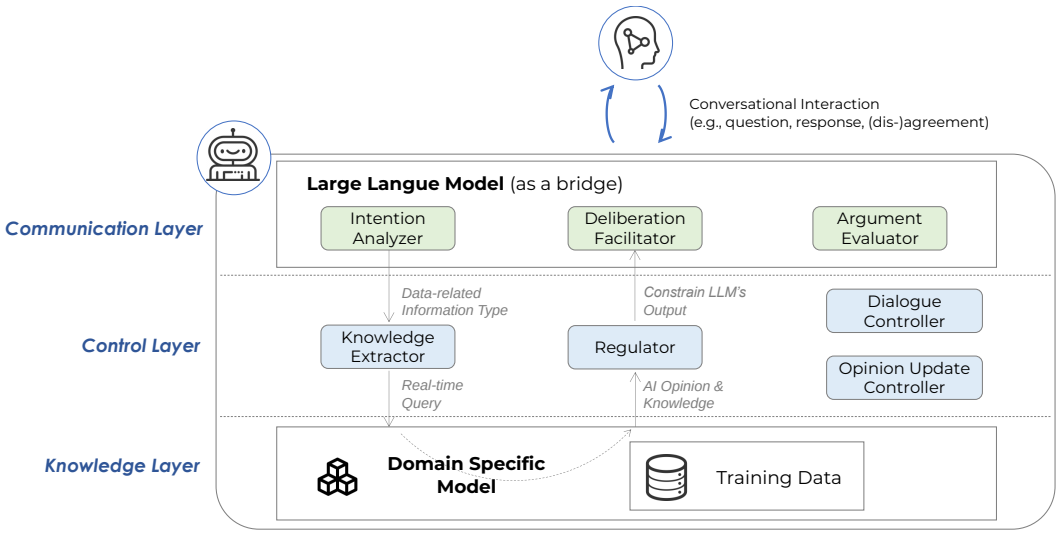

Figure 3: The architecture of *Deliberative AI* which integrates a domain-specific model and a Large Language Model, enabling the AI to engage in natural communication with humans while also harnessing domain knowledge derived from the specialized model.

SHAP Values: Raw SHAP values (e.g., 2.15) can be difficult for non-experts to interpret directly. To address this, we converted SHAP values into probabilities by using a regression model, mapping the four-category label range to a 0-100% scale. This conversion allows SHAP values to be directly interpreted as probabilities. (2) **Explanation of Feature Importance**: SHAP values indicate the importance of features but do not explain why a feature is important. To address this, we propose generating "meta-explanations" using the *Knowledge Extractor* (Sec. 3.3) to extract relevant evidence (e.g., data patterns) from the training data for each dimension. This approach aims to enhance the transparency of AI during the deliberation process.

Next, we describe how we implemented each component of *Deliberative AI*.

4.3 Implementation of Deliberative AI

4.3.1 ***I. Communication Layer.*** This layer serves as a vital bridge between humans and the DS-model, facilitating effective communication by comprehending human inputs and crafting relevant responses. Note that all components in this layer are based on our designed prompts, which enable LLMs to play different roles, rather than incorporating additional predictive models.

I-1. Intention Analyzer. We harnessed the language capabilities of LLMs[3] to identify human intentions and targeted dimensions in discussion. To formulate effective prompts for intention analysis, we conducted a pilot study to gather common dialogues around graduate admission decisions, including questions, arguments, critiques, and challenges. In the pilot study, we developed a preliminary version *Deliberative AI* (with basic deliberative discussion capability) to carry out conversations with 30 participants from Prolific[4]. Each participant expressed their opinions on various dimensions of applicant profiles and engaged in discussions with the AI. We collected 226 human deliberative statements. To extract various intentions from these statements, two authors performed qualitative coding using thematic analysis [59], and the results are summarized in Table 1. We then iteratively refined LLM prompts based on the collected data and built an "Intention Analyzer" with a 96% accuracy in identifying themes of participant statements. Specific prompts are available in the supplementary materials.

I-2. Deliberation Facilitator. This component addresses DC2 (Justification Rationality) and DC5 (Respect and Agreement) by designing corresponding LLM prompts. In particular, we instruct LLM to (1) Demonstrate a nuanced understanding of the human's statement; (2) analyze the specific content of the person's statement; and (3) provide a thoughtful and critical response. For detailed prompts, please refer to the supplementary materials.

I-3. Argument Evaluator. The main function of this component is to assess the strength of a person's statement/argument, which later informs updates to AI opinions. Drawing from established theories in human argumentation evaluation [53, 123, 124], we devised a comprehensive scoring mechanism with nine key items: Clarity, Relevance, Evidence, Logic, Consistency, Counterarguments, Depth, Credibility, and Alignment. These criteria are integrated into a prompt, guiding the LLM to evaluate human statements. We then average and scale the scores to obtain the overall human argument strength S_{Human} (from 0 to 1; 0: weakest, 1: strongest). We conducted a pilot study to evaluate the reliability of the LLM in scoring human arguments, using data collected from our Intention Analyzer pilot study (Sec. 4.3.1 I-1). Two authors independently scored the arguments using predefined schemas, resolving disagreements to reach consensus. We then calculated Cohen's Kappa (κ) to measure agreement between the LLM and human scores. The resulting κ value of 0.78 indicates substantial agreement, suggesting

[3]During our experiment (conducted in August 2023), we used the GPT-3.5 model. For consistency, we will refer to it as "LLM" throughout this paper.
[4]www.prolific.co

Table 1: Qualitative analysis of the sentiment/intention category of participants' statements (arguments, justifications, questions, critiques, etc.) in the deliberative discussion.

Themes	Definitions and Examples	#Participants
Distribution/Level of an attribute's values	**Participants evaluate how attribute values are distributed among the pool of applicants.** "*3.16 isn't a bad GPA - it's only slightly below average, sure, but it's still fairly good*" (P2)	35 (15%)
Overall importance of an attribute	**Participants consider or challenge the overall importance of an attribute on the admission decision.** "*Diversity is extremely important to the institution as a whole so the students highly rated diversity statement would highly influence their admittance.*" (P33)	24 (10%)
Contribution of an attribute	**Participants directly express their opinion on an attribute's contribution or challenge the contribution given by the AI but without evidence.** "*I know Applicant Undergraduate School Ranking has a significant impact on the chance of admission. But why is medium rank not good?*" (P1)	47 (20%)
Contrastive evaluation	**Participants compare an attribute's current value with other values (often using the average) to judge an attribute's impact.** "*I am surprised you ranked the applicant's GPA on a negative scale. 3.26 is not that much lower than the 3.5 of the last applicant.*" (P10)	41 (18%)
Holistic review of multiple attributes	**Participants evaluate how different attributes interact, taking into account the influence of certain attribute values on the strength of others.** "*The engineering major is incredibly difficult and any GPA above a 3.5 is considered successful.*" (P3) "*I said 2% positive influence because this individual went to a top rank school, which I assume is harder academically than some lower ranked schools.*" (P22)	23 (10%)
Data-irrelevant questions/arguments	**Participants give data-irrelevant statements based on their heuristics, past experiences, personal beliefs, etc.** "*Statement of purpose is the only part of the application process where the applicant gets to show us who they really are in their own words - not just a score or some data value. I ranked these higher for this reason.*" (P5)	77 (34%)

that the LLM provides reliable annotations with minimal disagreement in the context of our graduate admission task. Additional details, including scoring schemas and prompts, can be found in the supplementary materials.

In summary, the communication layer enables general interactions with humans. To integrate it with the DS-model's predictions and knowledge, a control layer is required to connect the two.

4.3.2 ***II. Control Layer.*** This layer manages the querying and extraction of specific DS-model opinions and knowledge while controlling the entire conversation flow.

II-1. Dialogue/Discussion Controller. This component serves as the control center for the discussion process, orchestrating a structured deliberation flow as shown in Figure 4. It unfolds as follows: [Thought Elicitation] Participants express their WoE on each dimension; AI responds with its perspectives. [Discussion] AI highlights commonalities and discrepancies, inviting participants to provide justifications or question differing viewpoints. AI responds with critical insights. All three components of the Communication layer (*Intention Analyzer*, *Deliberation Facilitator*, and *Argument Evaluator*) play vital roles in this phase. After one round of discussion, AI offers input options for participants to update, maintain, or continue the discussion. AI proceeds based on participants' choices. If they wish to move to the next dimension, AI summarizes any pending dimensions, highlighting differences. Participants can choose to explore untouched dimensions, revisit previous discussions, or skip this round. Participants have the flexibility to initiate dialogues on any dimension at any time, using quick input options or free text. They can refine their views on the decision interface independently of AI opinion updates.

II-2. Knowledge Extractor. Based on the attributes/dimensions and intent types identified by the "Intention Analyzer" (see Table 1), we developed a series of query functions to extract relevant data knowledge from the DS-Model. These functions help pull evidence for the LLM to generate responses in deliberative discussions appropriately. We established a mapping between the recognized intent type and the query function and called different query functions based on the recognized intent type. In practice, for a decision-making task, an existing dataset is typically available for training the AI model. This same dataset can be used to extract data patterns based on task-specific features, such as the percentile of a single feature value or the combination of multiple features. We recommend that researchers interested in applying this approach to other tasks first identify the data patterns that users are likely to find relevant, and then design the corresponding extraction functions. Below is a brief overview of the designed functions corresponding to different human intent types. Please refer to the supplementary materials for detailed codes and examples.

- *Distribution/Level of an attribute's value*:

  - Function `get_distribution(attr_val)` calculates attribute value percentiles within the applicant pool, along with contextual comparisons (with minimum, maximum, quartiles, mean, and median).
- *Overall attribute importance*:
  - Function `get_global_feature_importance(attr)` returns global importance.
  - Function `get_correlation(attr)` provides Pearson correlation between the selected attribute and the admission chance.
  - Function `get_influence_on_admission_chance(attr)` calculates admission chance changes for varying attribute values.
- *Contribution of an attribute*:
  - Function `get_current_value_influence(attr)` calculates admission chance differences when the value of a selected attribute is randomized.
- *Contrastive Evaluation*:
  - Function `get_contrastive(attr, contrast)` computes admission chance differences compared to a contrastive value.
- *Holistic review of multiple attributes*:
  - Function `get_holistic_analysis(attr, fix_attr)` evaluates the impacts of attribute(s) in specific scenarios, e.g., the impact of the GPA percentile in [top-ranked universities].

II-3. Regulator. The primary objective of this component is to harness the expertise of the DS-Model to regulate the responses generated by LLM. This approach makes certain that LLM's responses always align with the DS-Model's knowledge and decisions. To achieve this goal, we created consistency-ensuring prompts based on three key elements: (1) the findings extracted by the *Knowledge Extractor*, (2) the overarching decisions made by the DS-Model, and (3) the DS-Model's viewpoint on the current attribute under discussion.

II-4. Opinion Update Controller: We updated the AI's opinions by taking into consideration: (1) the current opinions of both the human (O_{Human}) and the AI (O_{AI}) on the discussed attribute, (2) the strength of the human's argument (S_{Human}, see *Argument Evaluator*), and (3) the AI's uncertainty (U_{AI}) measured and calibrated via Uncertainty Quantification 360 toolbox [47] (the uncertainty ranges from 0 to 1: the closer to 1, the more uncertain AI's prediction is). We propose the following formula to update the AI's opinions ($\hat{O}_{AI}$) on an attribute based on these factors, inspired by result aggregation in crowd intelligence [43, 84]. It is important to clarify that the term "update" here does not refer to modifications to the domain-specific model itself, such as retraining or fine-tuning. Instead, it pertains to the adjustment of the AI's expressed viewpoints regarding the current task case. Notably, LLM is not directly involved in updating the domain-specific model's viewpoints. Rather, LLM generates a strength score for the human's arguments, with the update being performed using Eq. 1.

$$\hat{O}_{AI} = \frac{1 - U_{AI}}{1 - U_{AI} + S_{Human}} \cdot O_{AI} + \frac{S_{Human}}{1 - U_{AI} + S_{Human}} \cdot O_{Human}, \quad (1)$$

4.3.3 ***III. Knowledge Layer***. This layer comprises the DS-Model and the training dataset. The DS-Model provides overall predictions and opinions on each dimension, while the training dataset supplies essential information (e.g., data distributions and patterns) for the *Knowledge Extractor* to perform real-time calculations and queries.

Overall, in this architecture, LLM is used for language understanding and generation. The opinions and evidence used by LLM are retrieved in real time from the DS-Model and training data through our logic code (like retrieval augmented generation [77]). In this way, the LLM is used in a controllable and responsible manner, minimizing the potential hallucination. We provide an example of how data is processed in Deliberative AI in the Appendix.

4.4 Interface Design

The interface for the graduate admission task is structured into three main regions:

- **Profile Region** (Figure 5 (A)) displays the applicant's profile, providing a table with the current value and possible range of each attribute. Users can access attribute definitions and basic data distribution statistics (minimum, maximum, average, and median values) by hovering over pink circular markers.
- **Opinion and Prediction Region** (Figure 5 (B)) is dedicated to thought elicitation by both users and the AI.
  - The upper part displays aggregate predictions from both humans and AI. This includes a legend (Figure 5-1) and two slide bars (Figure 5-2 and -3) representing AI's and the user's overall predictions, respectively. Each slide bar shows three line indicators: a white line representing the average admission probability of all applicants, a green line showing the initial predictions made by humans/AI, and a yellow line denoting the updated prediction by humans/AI (only shown after an update is made).
  - The bottom part of this region enables both humans and AI to express opinions on each decision dimension (i.e., applicant attribute). A simplified profile in the middle (Figure 5-5) reduces attention shifts. Status indicators show if a dimension has been discussed (green), is being discussed (orange), or is yet to be discussed (gray). Separate "concrete opinion" sliders (Figure 5-4 and Figure 5-6) allow AI and humans to share dimension-level opinions.
  - Each dimension's slide bar starts in a central position (0% contribution). Users can drag the slider any time to the right to increase the weight on an attribute toward a positive "admit" decision or to the left to reduce its contribution. Alternatively, users can directly input contribution values in a box below the slider.

  Slide bars in Figure 5-2 and Figure 5-4 are interconnected, so as those in Figure 5-3 and Figure 5-6. Values within the "overall prediction bar" reflect the cumulative values from the "concrete opinion bars." Any changes in the dimension-level bars immediately update the overall prediction. AI's and the user's opinions are displayed side by side for easy comparison. Note that at the beginning of each case, users have to complete their opinion inputs and click the [Submit Opinion] button to see AI's initial (overall and concrete) suggestions.
- **Discussion Region** (Figure 5 (C)) is where all deliberative dialogues take place. Users can type out their opinion arguments, questions, disagreements with AI, responses to AI queries, and more. Importantly, changes made in the Opinion Region are seamlessly integrated by AI and reflected in ongoing discussions, and

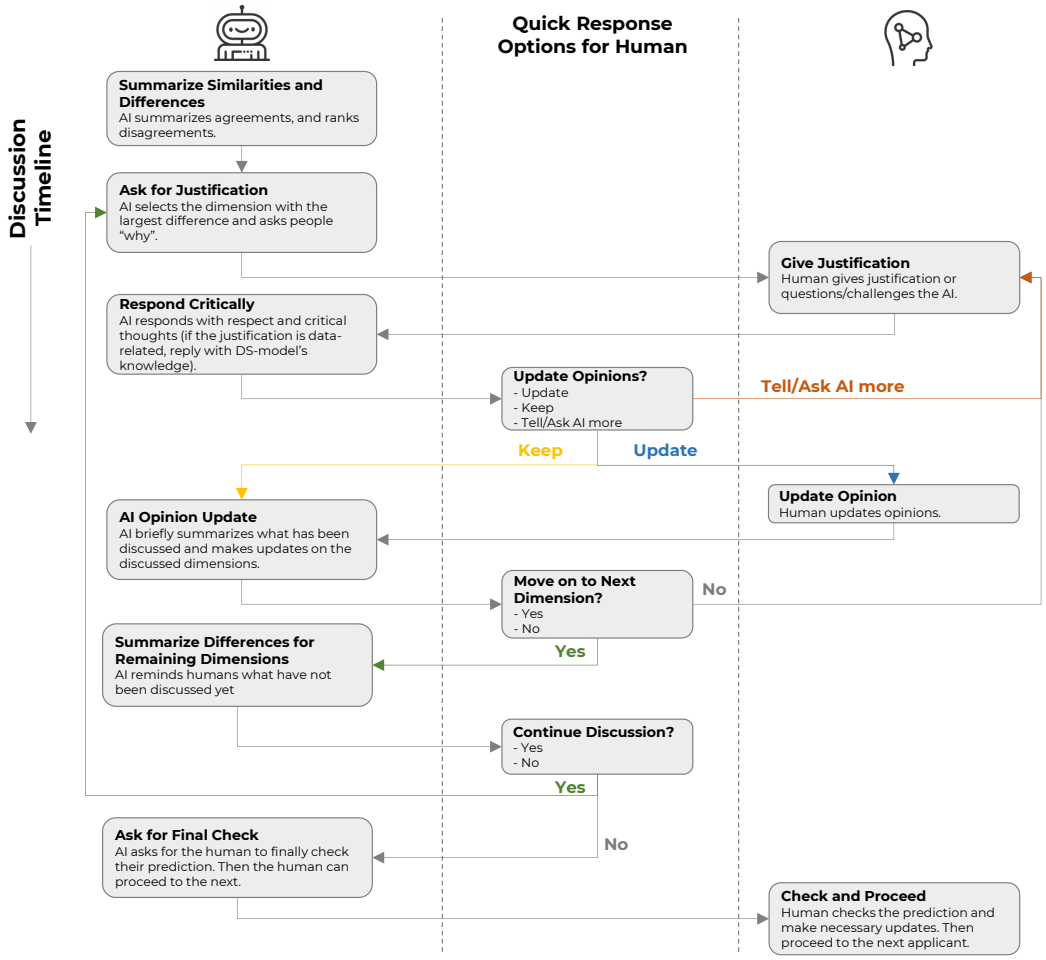

Figure 4: The conversation flow for the deliberative discussion.

conversely, any viewpoint changes mentioned in the dialogue are instantly updated in the Opinion Region.

5 Exploratory User Study

To gain an initial understanding of the impact of the *Human-AI Deliberation*, we conducted a mixed-methods study within the context of graduate admissions. This study is termed *exploratory* because our objective was not to assess the effectiveness of *Human-AI Deliberation* specifically for graduate admissions. Rather, we used graduate admissions as an illustrative task to explore the potential effects of *Human-AI Deliberation* on AI-assisted decision-making.

5.1 Task Setup

We used the graduate admission task as our testbed. To ensure manageable study durations and prevent participant fatigue, we selected five task cases based on the length of the pilot study. These cases included one for the tutorial and four for the main tasks. To investigate human-AI deliberation, we selected cases from a pilot study where human and AI opinions conflicted, often involving ambiguous applicant profiles near the admission borderline. As a result, predicting the admission outcome for these cases was challenging for both humans and AI, leading to performance levels between 50% and 60% for both. However, this does not imply that the AI model used in our study is of unrealistically low performance; rather, we focused on ambiguous task cases that require conflict resolution.

5.2 Conditions

We compared the proposed *Deliberative AI* with the traditional explainable AI assistant (*XAI*) and *Human Alone*.

- **Deliberative AI (DAI)**: Participants share their thoughts on various dimensions before viewing AI recommendations. We present AI's "thoughts" on each dimension afterward. After comparing conflicting viewpoints, we offer a dialogue interface for participants and AI to discuss any of the perspectives, as shown in Figure 5.
- **Explainable AI (XAI)**: After individuals provide their predictions, they receive AI recommendations (along with feature contribution-based explanations) and then make their final judgments (see Figure 10 in Appendix).

Figure 5: The interface of *Deliberative AI*. The interface contains three parts. The top part (A) is the applicant's profile. The bottom left part (B) is the region for humans and AI to indicate (and update) their opinions. The bottom right part (C) is the discussion region where humans and AI can discuss conflicting opinions. (All the dashed lines are only for illustration)

- **Human Alone**: Participants need to make predictions independently without any AI assistance.

5.3 Procedure

With our institutional IRB approval, we conducted a between-subjects study. After obtaining consent, we had participants complete a background questionnaire to gather demographic data and assess their AI expertise. We then introduced the study, explained the task, workflow, and AI's functions, including its ability to update opinions, without delving into the specifics of the adjustment mechanism. This approach reflects real-world scenarios, where non-technical users focus more on functionality than technical details. Participants then engaged in an interactive tutorial, practiced with one example task, and received distribution and summary statistics for each attribute of the applicant's profile. After the tutorial, we asked qualification questions to check participants' understanding of the task, allowing only those who answered all questions correctly to proceed to the main task. In the main task, participants worked on four graduate admission task cases, which were presented randomly. Finally, we collected participants' perceptions, experiences, and feedback on the AI system and the discussion process in the exit survey.

5.4 Participants

We first conducted a power analysis to determine the required sample size for using G*Power [39] with a default effect size f=0.25 (indicating a moderate effect), a significance threshold α=0.05, and a statistical power $1-\beta$=0.8. This resulted in a required total sample size of 159 participants for the three conditions. After obtaining institutional IRB approval, we recruited a total of 174 participants from Prolific[4]. To ensure high-quality responses, participants had to meet specific criteria: (1) residing in the United States; (2) having been admitted to a US graduate program before (as the task involved predicting graduate admission in a US university); (3) having at least a 99% approval rate with at least 1000 previous submissions; (4) using English as their first language; and (5) using a desktop computer for the experiment. After filtering based on attention-check questions, we obtained 153 valid responses (*Deliberative AI*: 48, *XAI*: 51, *Human Alone*: 54). Among the final participants, 84 self-identified as male, 67 as female, and 2 as others. The age distribution was as follows: 23 participants aged 24–29, 42 aged 30–39, 33 aged 40–49, 30 aged 50–59, and 25 aged over 59. Regarding education, 125 participants held an MA/MSc/MPhil degree, and 28 held a Doctorate (PhD or equivalent). Participants also had diverse levels of AI knowledge: 9 reported having no knowledge, 86 were familiar with basic AI concepts, 50 had experience using AI algorithms, and 8 identified as AI experts. Participants in the *Deliberative AI* condition received bonuses based on the actual study length. To motivate high-quality work, participants received a $0.50 bonus if their overall accuracy exceeded 75%. On average, participants earned about $12 per hour.

5.5 Measurement

To answer the aforementioned research questions, we comprehensively measured the effects of human-AI deliberation across four dimensions: *task performance*, *reliance*, *perceptions of AI*, and *user experience*.

- **Task Performance**. We evaluated decision-making accuracy using *Decision Accuracy* metrics [8, 137].
- **Reliance**. Participants' reliance on AI suggestions was assessed through the *Agreement Fraction* [55, 137] and *Switch Fraction* [55, 137]. Additionally, the appropriateness of reliance was measured using the *Over-reliance Ratio* [100, 125, 127] and *Under-reliance Ratio* [100, 125, 127].
- **Perceptions of AI**. Participants' perceptions of AI were measured using 7-point Likert scales for *Helpfulness* [15, 18, 73], *Trustworthiness* [16, 45], and *Understanding* [127] (1: Strongly disagree; 7: Strongly agree).
- **User Experience**. Participants' *Decision Confidence* was evaluated using established measures [93]. Given that deliberation requires additional effort, we also assessed *Mental Demand* [16, 45, 54, 68], *Effort* [54], *Complexity* [16], and *Satisfaction* [16, 45], all measured on 7-point Likert scales.

To gain a deeper understanding of participants' perceptions of both *Deliberative AI* and the deliberative decision-making process, we also gathered open-ended feedback in the exit survey. These questions explored participants' perceptions of the usefulness of deliberating with AI, the AI's updates, and their expectations for system improvements. A detailed overview of the metrics and questions is provided in Table 2 in the Appendix.

5.6 Analysis Methods

We conducted mixed-methods analyses to examine the aforementioned metrics. For quantitative analysis, we first performed normality tests (Shapiro-Wilk) and found that the data did not fit the normality assumption. Therefore we ran the non-parameter tests. Specifically, to compare *Deliberative AI* and *XAI* (such as humans' reliance on AI, and their perceptions of AI), we run Mann-Whitney U test. To compare all three conditions (such as task performance, and user experience), we employed Kruskal–Wallis tests with Bonferroni post-hoc correction and we reported adjusted p-values.

For qualitative analysis, two authors independently coded participants' open-ended feedback and conversation logs using an inductive thematic analysis approach [59]. The final themes emerged through discussions and harmonization over several iterations. We also identified representative examples from the source texts for demonstration in this paper.

6 Results

In this section, we report our exploratory findings regarding the three research questions: (**RQ1**) how *Human-AI Deliberation* affects task performance and human reliance (and reliance appropriateness) on AI, (**RQ2**) how *Human-AI Deliberation* affects human perceptions and task experience, and (**RQ3**) how humans perceive the effectiveness of *Human-AI Deliberation* and what should be improved.

6.1 RQ1: How will *Human-AI Deliberation* affect task performance and humans' reliance (and its appropriateness) on AI suggestions?

6.1.1 Decision Accuracy. As shown in Figure 6, participants in the *Deliberative AI* condition demonstrated significantly higher decision accuracy (M=0.598, SD=0.169) compared to those in the *XAI* condition (M=0.524, SD=0.16, p <0.05). This finding suggests that in scenarios where tasks are challenging for both humans and AI—where the individual performance of humans and AI is relatively low—traditional Explainable AI (XAI) may not improve performance and could even have negative effects. In contrast, *Human-AI Deliberation* shows potential to yield positive outcomes, even when the performance of the underlying AI models in these difficult task cases is suboptimal.

6.1.2 Reliance. We measured participants' objective reliance by agreement fraction and switch fraction. As shown in Figure 7 (a), participants agreed significantly less with AI's suggestions in *Deliberative AI* (M=0.57, SD=0.24) than in *XAI* (M=0.68, SD=0.27, p <0.05), and switched significantly less to AI's predictions in *Deliberative AI* (M=0.23, SD=0.35) than in *XAI* (M=0.51, SD=0.41, p <0.001). Combined with participants' open-ended feedback (Sec. 6.3), this may be because people invest more in independent thinking in the process of deliberation with AI and realize the problematic aspects of AI's perspective.

6.1.3 Appropriateness of Reliance. We further measured the appropriateness of participants' reliance on AI's suggestion by under-reliance and over-reliance (Figure 7 (b)). Results show that there is no significant difference between *Deliberative AI* (M=0.32, SD=0.28) and *XAI* (M=0.29, SD=0.30) in terms of under-reliance. While significantly less over-reliance was observed in *Deliberative AI* (M=0.47, SD=0.31) than in *XAI* (M=0.65, SD=0.33, p <0.001), which means that participants had more appropriate reliance on AI when collaborating with our proposed *Deliberative AI*. This result aligns with existing research on cognitive forcing functions [16], where high cognitive effort reduces over-reliance on AI. However, unlike in [16], where cognitive effort significantly increased under-reliance (i.e., humans under high cognitive effort may blindly ignore AI's suggestions), our findings reveal no such adverse effect. This indicates that our tool mitigates over-reliance not merely by increasing cognitive effort but by fostering meaningful deliberation.

6.2 RQ2: How will *Human-AI Deliberation* affect humans' perceptions of the AI and user experience?

We measured the effects of different AI conditions on participants' perceptions and user experience via a 7-point Likert scale (1: strongly disagree, 7: strongly agree).

6.2.1 Perceptions of AI. Figure 8 (a) shows participants' perceptions of the AI model. There were no significant differences in *perceived helpfulness* and *understanding* between *Deliberative AI* and *XAI*. However, participants reported significantly less trust in *Deliberative AI* (M=4.47, SD=1.68) compared to *XAI* (M=5.52, SD=1.27, p <0.01), aligning with their reliance behaviors (Sec. 6.1.2). This difference may be attributed to participants identifying more AI flaws through deliberation than by solely observing AI's explanations, supported by conversation logs analysis (see Sec. 6.3 for details).

6.2.2 User experience. First, we want to see participants' decision confidence. As indicated in Figure 8 (b), participants in *XAI* reported significantly higher confidence (M=6, SD=1.10) in their predictions than those in *Human Alone* (M=5.59, SD=1.12, p <0.05). However, from Figure 6 (a) we found that the final accuracy of participants in *XAI* is even lower than those in *Human Alone*. This indicates that the traditional *XAI* might lead to humans' ***illusionary confidence***, which could prevent humans from making optimal decisions.

Given that the *Deliberative AI* requires participants to externalize thoughts at a dimension level and engage in deliberative discussions on conflicting opinions, it's crucial to explore how these activities influence the user experience. Results showed no significant difference among the three conditions concerning *Mental Demand*, *Effort*, and *Perceived System Complexity*. However, we find participants reported significantly lower *Satisfaction* in *Deliberative AI* than in *Human Alone*. This decrease in user experience may be due to the AI exposing more conflicts for humans to resolve. As P3 noted, "*It's annoying because I have to try to find evidence to prove that my point of view is correct.*" This result suggests that there is a trade-off between encouraging users' deliberative thinking and optimizing their user experience, in line with findings in previous studies [16]. Future work should focus on finding a balance between the benefits of deliberation and maintaining a positive user experience.

6.3 RQ3: How will humans perceive the effectiveness of the proposed *Human-AI Deliberation* and what can be improved for future *Human-AI Deliberation* design?

In addition to quantitative measures, we aimed to gain a deeper understanding of participants' perceptions of the helpfulness of the proposed *Human-AI Deliberation* and the *Deliberative AI* feature, particularly the opinion-updating aspect. We also sought insights to inform future design improvements for *Human-AI Deliberation*. To achieve this, we analyzed participants' open-ended feedback, supported by their conversation logs. Our key findings are summarized in Figure 9. Below, we present key insights, with themes highlighted in bold.

6.3.1 Perceptions of Deliberative AI and the Discussion Process. Participants offered both positive and critical feedback on *Deliberative AI* and the deliberation process. Of the 48 participants who experienced with *Deliberative AI* condition, 43 felt that **deliberation helped them make more informed decisions**. Specifically, AI-assisted discussions enabled participants to "**identify AI's limitations**" (21/48), "**consider different perspectives**" (10/48), and "**reflect on and correct their own mistakes**" (15/48). For example, P1 (Male, 32) pointed out the AI's overreliance on GPA scores while underestimating the role of recommendation letters: *"The AI relies too much on GPA scores but undervalues recommendation letters. It didn't provide convincing justifications, so I couldn't rely on its opinion for these factors."* P19 (Male, 42) noted how the deliberation process encouraged self-reflection: *"The AI made me question what I believed to be sufficient reasoning, especially given the data."* These

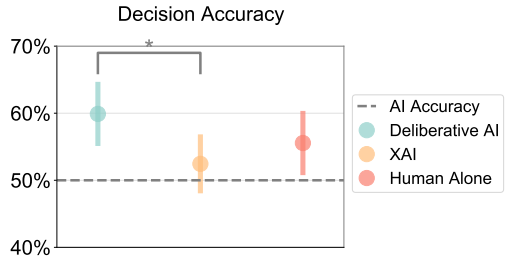

Figure 6: Task performance in different conditions. It is important to note that we intentionally selected ambiguous task cases that are prone to conflicts between humans and AI and are challenging for both. As a result, the accuracy of both humans and AI individually is relatively low. The error bars represent 95% confidence interval. (*: p <0.05)

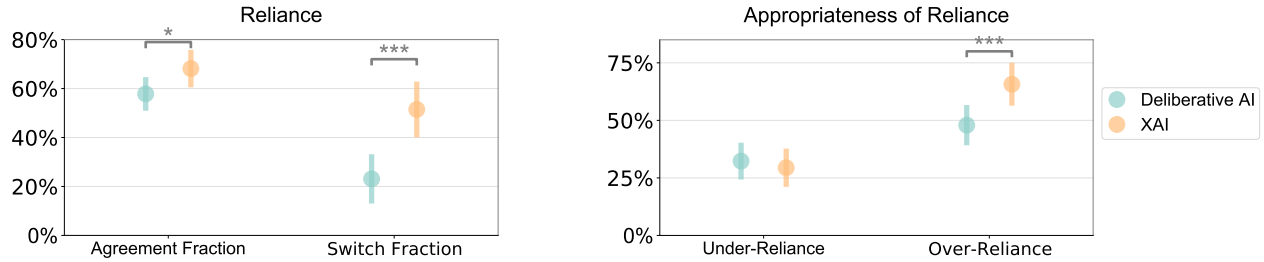

Figure 7: Participants' reliance and the appropriateness of their reliance. (A) Participants' reliance on AI's suggestions was measured by agreement fraction and switch fraction. (B) The appropriateness of participants' reliance on AI's suggestions, including under-reliance (the ratio where participants did not use a correct AI suggestion) and over-reliance (the ratio where participants used an incorrect AI suggestion). The error bars represent 95% confidence interval. (*: p <0.05, ***: p <0.001)

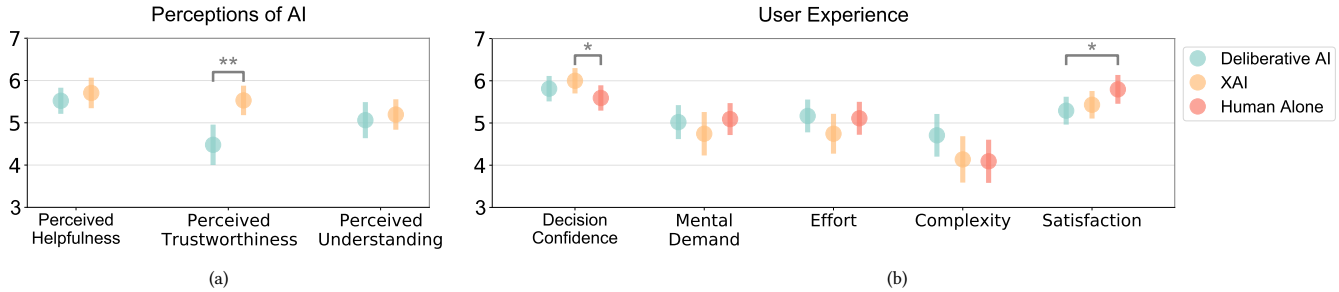

Figure 8: Participants' perceptions and task experience. (A) Participants' perceptions of different AI assistant. (B) Effects on user experience. The error bars represent 95% confidence interval. (*: p <0.05, **: p <0.01)

reflections are consistent with our analysis of participants' conversation data with *Deliberative AI*, where participants expressed doubts to the AI in 32% of dialogues, acknowledged its arguments in 17% of dialogues, and engaged in self-reflection and correction in 15% of dialogues.

Moreover, 18 participants found that **discussing with the AI introduced "new knowledge, insights, and perspectives"**. For instance, P5 (Male, 45) commented: *"The AI provided information I didn't know, like percentiles and how similar stats influenced past decisions, which I found extremely helpful."*

Nine participants said that **deliberation helped them recognize biases**. For example P35 (Female, 29) mentioned: *"It made me realize the AI had inherent biases, which prompted me to pause and reflect."* This aligns with findings from the participants' conversation data with AI, where many participants (15/48) identified biases in the AI's reasoning. For example, the AI exhibited bias by assigning more importance to U.S. applicants while downplaying those from Asia. P3 (Male, 42) questioned: *"Why does their country matter? Penalizing someone for their nationality seems biased."*

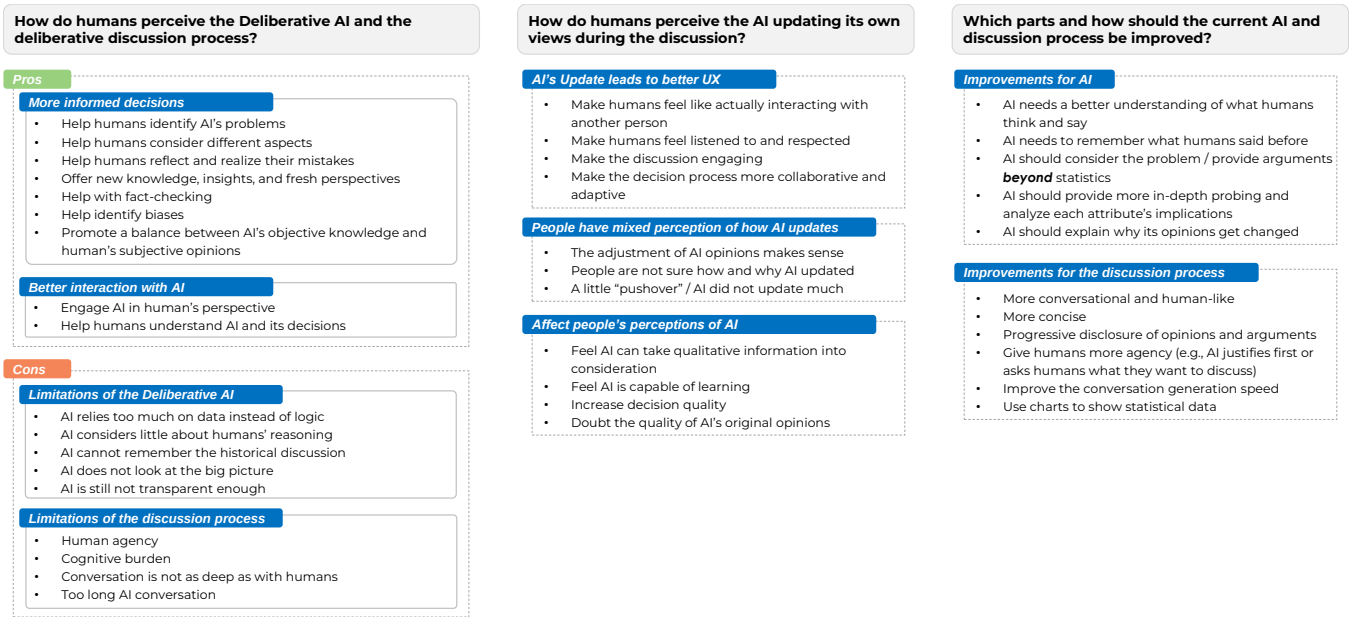

Figure 9: The main results of our thematic analysis of the open-ended questions.

Furthermore, 15 participants appreciated that **deliberation balanced the AI's objective data with human subjective judgment**." For instance, P2 (Male, 42) said: *"It provides an opportunity to consider multiple perspectives and statistical data, resulting in more balanced decision-making."*

Despite the benefits, participants also identified limitations in the deliberative process. Five participants felt that their **sense of agency was reduced as the AI prompted them to think, rather than passively waiting for their input**. Four participants found the **discussion is mentally demanding** and criticized the AI's verbose responses, reflecting the results on user experience (see Sec. 6.2.2).

Regarding the limitations of *Deliberative AI*, two main concerns emerged: that "**Deliberative AI relied too heavily on data over logic**" (15 participants) and "**failed to consider human reasoning**" (10 participants). This feedback aligns with our analysis, where participants integrated personal experiences and logic into their decisions, while the AI lacked this depth. For example, participants considered factors such as: *"Business programs are highly competitive, requiring a higher GPA."* (P17, Female, 60) and *"A strong SOP reflects a deep understanding of the project and strong commitment."* (P32, Male, 28)

Additional concerns included the AI's insufficient ability to faithfully remember previous discussions (five participants), failure to consider the broader context (three participants), and lack of transparency (three participants).

6.3.2 Perceptions of Deliberative AI*'s Updates.* Participants generally appreciated the dynamic updating feature. 25 out of 48 participants felt that **the AI updates improved the user experience**. They described the experience as more interactive, comparing it to "*interacting with a real person*" (6 participants), where their opinions felt "*heard and respected*" (19 participants). Moreover, these updates made the discussion "*more engaging*" (8 participants) and the decision-making process "*more collaborative*" (5 participants).

However, perceptions of AI's updates varied. While 11 participants felt the AI adjusted its views too frequently (e.g., "*the AI updated too much*"), 13 participants expressed uncertainty about "*why and how the AI was updating*." For example, P2 (Male, 42) noted: *"I'm not sure if the AI really changed its opinion based on what I said or if it was just programmed to do so in response to my input."* Interestingly, **some participants thought the AI was too quick to change its stance, while others felt it wasn't flexible enough**. These updates also affected participants' perceptions of the AI itself. Eight participants felt the AI could "*incorporate qualitative information*," and five believed the AI was "*learning from human knowledge during the discussion*," which could "*enhance the quality of decision-making*" (3 participants). However, three participants found the lack of transparency in the update process led them to "*doubt the reliability of the AI's initial opinions*."

Although the AI's dynamic updates mimic human deliberation, making these updates more meaningful requires considering diverse user perceptions and designing the feature more thoughtfully. Specifically, the future design of AI opinion updates should draw insights from social science [79], addressing user preferences regarding the frequency and magnitude of updates. Recall that in the introduction before the experiment, we did not disclose the specific mechanism of AI opinion updates to the users, so this process was not transparent to them. Enhancing transparency and users' understanding of AI update could alleviate their concerns.

6.3.3 Opportunities for Future Improvements of Deliberative AI. Participants provided several suggestions for improving *Deliberative AI* in future versions. Fifteen participants emphasized **the need for AI to develop a deeper understanding of human**

thoughts and arguments, calling for more nuanced and context-aware communication. Eight participants stressed the importance of AI remembering previous interactions to enable more personalized and coherent conversations, avoiding repetition and improving the AI's responsiveness to specific perspectives.

Twelve participants suggested that **AI should move beyond statistical reasoning** to offer broader, more holistic solutions, while 10 participants advocated for deeper analysis and consideration of broader implications in each context. Transparency was also highlighted as crucial, with eight participants asking for clearer explanations about the AI's changing opinions to foster trust and understanding.

Regarding the design of human-AI discussions, eight participants preferred a more conversational and human-like interaction, while 11 favored concise AI responses. Two participants recommended gradually disclosing the AI's arguments, and three wanted more user control over steering the discussion. Additional suggestions included improving dialogue generation speed and enhancing the display of visual information beyond just text.

7 Discussion

In this paper, we introduce the *Human-AI Deliberation* approach to address two key challenges in AI-assisted decision-making: insufficient analytical engagement with AI recommendations and limited support for resolving human-AI disagreements. Traditional interfaces often restrict users to accepting or rejecting AI suggestions as a whole, limiting nuanced understanding and collaboration. Our exploratory assessment demonstrates that *Human-AI Deliberation* improves decision accuracy and fosters appropriate reliance on AI compared to conventional explainable AI systems. In this section, we discuss key implications, generalizability, limitations, and directions for future work.

7.1 Deliberation as a New Paradigm Complementing Existing (X)AI Assistance

The *Human-AI Deliberation* approach introduces a conflict-driven discussion model that complements traditional AI assistance. *Deliberative AI* builds on explainable AI (XAI) principles, challenging human perspectives while respecting their agency as decision-makers. It enhances AI assistants by fostering deeper analytical thinking, encouraging users to form independent opinions before seeing AI suggestions (as in the Cognitive Forcing Function approach [16]) and stimulating critical thinking through AI-generated questions (similar to [25]). Additionally, *Human-AI Deliberation* involves users in active discussions, improving communication and transparency with AI. This engagement helps lead to more informed and nuanced decisions.

Application Value of Deliberative AI. Deliberative AI is well-suited for critical decision-making tasks, such as investment or hiring, where decision quality outweighs the need for speed or convenience. While engaging with Deliberative AI may demand more time and effort, decision-makers appreciate its ability to support well-considered outcomes. Additionally, Deliberative AI fosters user reflection by encouraging individuals to examine biases in their reasoning and evaluate differences between their perspectives and the AI's suggestions. Participants frequently emphasized this in their qualitative feedback, highlighting its value even in subjective decision contexts. Beyond enhancing objective accuracy, Deliberative AI promotes thoughtful consideration and introspection, which are vital for decisions requiring nuanced judgment.

7.2 Reducing Human Over-Reliance by Exposing AI Mistakes

Human-AI Deliberation significantly reduces participants' tendency to over-rely on inaccurate AI suggestions. This approach utilizes cognitive forcing theory, which encourages forming independent opinions before viewing AI recommendations. This helps counteract *anchoring bias*—the undue influence of initial AI predictions on subsequent judgments—and fosters more analytical, System 2 thinking [62]. While our *Explainable AI (XAI)* baseline also promotes independent opinion formation, *Deliberative AI* proves more effective by involving deeper deliberative discussions, especially when facing conflicting viewpoints.

In the *Deliberative AI* architecture, participants become more aware of AI's limitations through engagement with conflicting opinions, reducing over-reliance. Notably, 31% of conversations involve participants questioning AI's logic, reflecting a critical evaluation of its insights. This approach effectively minimizes over-reliance without increasing under-reliance, as participants adjust their judgments based on deliberative dialogue. We recommend that AI-assisted decision-making systems should emphasize transparency by **explicitly** highlighting potential AI errors rather than merely suggesting that "AI may make errors".

7.3 Human-AI Conflict Resolution: Key to Enhancing Collaboration

We propose that addressing conflicts in decision-making is more beneficial than merely seeking consensus. Our approach, *Human-AI Deliberation*, prioritizes resolving disagreements between humans and AI—an often overlooked aspect in current AI-assisted decision-making. Conflicts serve as a lens to uncover underlying flaws and biases, making them essential to effective human-AI collaboration. This conflict-centered methodology offers several advantages: it enhances decision-making accuracy by mitigating over-reliance on AI, fosters deeper introspection, and facilitates reconciliation between differing human perspectives and AI-generated recommendations. Additionally, addressing conflicts allows for the identification of biases in AI interpretations, thereby promoting fairness in decision-making [57, 66].

However, prioritizing conflict resolution can influence user experience. Our experimental results show that while *Deliberative AI* demonstrates no significant differences from *XAI* and *Human Alone* in terms of mental demand, effort, or perceived complexity, it introduces a more complex decision-making process, resulting in lower user satisfaction. We identify three primary sources of the system's burden and propose solutions to mitigate them:

- **Articulating Opinions Across Dimensions:** Currently, users must articulate their views comprehensively across multiple dimensions. Future interfaces could streamline this by letting users focus on dimensions they consider most critical or provide approximate opinions in natural language. AI could then identify

the direction and intensity of human opinions from languages, allowing users to refine these interpretations as needed.

- **Dialogue Effort:** The AI currently initiates discussions on dimensions with significant disagreements. Future designs could let users take a more active role or adopt a bidirectional approach to enhance autonomy. Alternatively, AI could prompt users to reflect on specific dimensions and acknowledge the completion of reflection with simple actions, such as clicking an "OK" button, reducing dialogue input.
- **Emotional Challenges in Conflict Resolution:** Resolving conflicts with AI can be difficult and reduce satisfaction with *Deliberative AI*. To improve this, interfaces could reframe conflicts as opportunities for reflection and feedback, using Constructive Conflict Theory [67] to emphasize that AI and humans share the same goals. Drawing from Positive Psychology [98], AI could demonstrate active listening, show appreciation for user viewpoints, and foster a supportive atmosphere to reduce dissatisfaction and encourage deeper engagement.

7.4 Obstacles to Discussion: Humans and AI Think Differently

Humans Use Heuristics and Logic, While AI Relies on Data: Integrating LLMs and DS-Models enhances *Deliberative AI* for dynamic human-AI discussions. However, AI's data-centric approach often clashes with human decision-making, which relies on personal understanding, logic, heuristics, and creativity [31, 36, 46, 110, 111]. We propose using DS-Models to guide LLMs on data-related discussions, while LLMs handle non-data matters autonomously. Despite this, engaging users with strong subjective perspectives remains challenging [80].

Existing Explanation Methods Fall Short: Current XAI methods, such as local feature importance [79], inadequately explain specific feature contributions. Our approach, which provides data-related insights like distribution and value comparisons, aims to support feature-level discussions. Yet, challenges persist: data patterns may not match AI explanations precisely, and users often rely on subjective reasoning [31, 46]. As Miller et al. note, "probabilities are not as important as causal links" [94]. Thus, data-driven insights alone may not align with users' cognitive processes or replicate human-like conversations.

Future human-AI deliberation designs should address these challenges by aligning with human intuition and cognitive processes, and by crafting AI explanations that facilitate more effective discussions.

7.5 Ethical Concerns in Deliberative AI Design

Responsible Use of LLMs in Assisting Human Decision-Making: LLMs are evolving rapidly but are prone to "hallucinations," where they generate plausible but incorrect information [61]. Relying solely on LLM-generated opinions without appropriate fine-tuning is irresponsible. Even with fine-tuning, LLMs may produce unpredictable responses. We advocate for a responsible approach that integrates LLMs with DS-Models, where DS-Models guide LLM responses, positioning LLMs primarily as intermediaries between users and DS-Models. Although this approach limits LLM flexibility, it enhances security and control. Nonetheless, even with DS-Model guidance, inaccuracies can still occur. Researchers should exercise caution and transparency, informing users about the potential for errors and the limitations of LLM-generated information.

Ethical Issues with AI Opinion Updates: Research indicates users value AI's responsiveness to their arguments [139]. Our AI opinion update mechanism, which adjusts AI stances based on user input and prediction uncertainty, aims to reflect this need. While users appreciate feeling heard, ethical concerns arise, particularly about accountability. A key issue is who should be responsible if the AI, initially correct, updates to an incorrect prediction after discussion. Additionally, the AI's adaptability may create the impression of learning and progress, even if the underlying model remains unchanged, potentially leading to unrealistic expectations. It is crucial to design AI systems with transparent updating mechanisms to ensure users understand how updates are made and manage their expectations effectively.

7.6 On the Generalizability of Human-AI Deliberation

Task Suitability: *Human-AI Deliberation* is less suited for repetitive, low-stakes tasks like content moderation [68] but is more appropriate for high-stakes, complex tasks such as healthcare [75], finance [51], and criminal justice [24].

Discussion Effort: Discussing every feature, as in our study, may be impractical for tasks with many attributes. Grouping features into broader categories could streamline discussions.

Opinion Representation and Alignment: We used the Weight of Evidence method to quantify opinions, but this may not be intuitive for all users. Future designs could infer opinions from natural language, emotional intensity, or comparisons, or simplify input with rankings or pairwise comparisons [41].

Data Type Applicability: While designed for tabular data, the architecture can extend to text tasks with LLM's capability in dealing with textual data. Adapting it to image data may require advances in vision-language models [136].

Using LLM as a Communication Bridge: In Deliberative AI, the LLM acts as a deliberation facilitator, intention analyzer, and argument evaluator, relying mainly on its language understanding capabilities with minimal reliance on its reasoning abilities. Through two pilot studies, we validated the LLM's effectiveness in these roles. However, challenges remain regarding its reasoning capabilities [131]. Designers should leverage the LLM's strengths in communication and avoid overburdening it with complex reasoning tasks.

7.7 Limitations and Future Work

The study design has several limitations. **First**, the college admissions task used for illustration does not fully capture real-world admissions processes, which typically involve in-depth discussions about a student's materials, such as her/his statement of purpose (SOP), background, and overall fit for the department. However, our used public dataset quantifies the "strength" of the SOP and recommendation letter as a scale value, which inevitably oversimplifies these nuanced evaluations. Additionally, while participants had relevant experience, they were not actual admissions committee members, leading to an expertise gap. Furthermore, most

participants had experience applying to master's programs, with only a few familiar with PhD admissions. Given the distinct evaluation criteria for these two application types, our findings may not generalize to PhD admissions. Future research should assess the approach's effectiveness in real-world admissions contexts. **Second**, graduate admissions is a subjective task that lacks a definitive ground truth. In our study, since the dataset labels were provided by a professional admissions committee, we used decision accuracy and over/under-reliance as objective metrics to assess decision quality "to some extent". The dataset also abstracts subjective elements (e.g., SOP and recommendation letter) into numerical strength values, making the task relatively more objective. Nonetheless, future work should explore the effects of Deliberative AI in the context of more objective tasks. **Third**, the number of decision tasks in the study was limited. During the pilot, using 8-10 tasks led to participant fatigue and a drop in engagement after completing 3-4 tasks. To maintain engagement, we reduced the task count to four, which restricts the generalizability of our results. Future studies should conduct long-term evaluations to collect more deliberative decision data. **Fourth**, the independent decision accuracy of both humans and AI was relatively low (50-60%) in our selected task cases, as we intentionally chose tasks prone to conflicts that require discussion. The ambiguity in these cases led to suboptimal performance from both. However, our study did not address less-ambiguous cases, where differing but firm opinions may still arise. We believe Deliberative AI can still help reduce errors in such cases by prompting reflection on biases or overlooked perspectives, especially when humans' confidence is high. Further research is needed to explore different task cases for a more comprehensive understanding of human-AI deliberation.

8 Conclusion

In this paper, we introduce *Human-AI Deliberation*, as a new paradigm of AI assistance for decision-making. *Human-AI Deliberation* encourages the externalization of thoughts, facilitates interactive deliberation between humans and AI, and allows for dynamic updates of decisions. To enable the deliberation, we present a novel AI assistant called *Deliberative AI*, which can identify differences in viewpoints, engage in comprehensive deliberation, and adapt its suggestions during discussions. We apply this architecture to an illustrative task (graduate admissions decisions) and conduct an exploratory study to assess its potential impact on decision-making processes, outcomes, user perceptions, and experiences. Results indicate the potential of *Deliberative AI* to improve decision accuracy and promote more appropriate human reliance on AI. Additionally, we analyze participants' open-ended feedback to gain deeper insights into how users use and perceive *Deliberative AI*, uncovering areas for improvement. With the key insights and implications derived from our study, we aim for this work to serve as an exploratory step toward establishing a new paradigm of AI assistance that enhances decision-making.

Acknowledgments

This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16207923.

References

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] David Alvarez-Melis, Harmanpreet Kaur, HD III, Hanna M Wallach, and Jennifer Wortman Vaughan. 2021. A Human-Centered Interpretability Framework Based on Weight of Evidence. *arXiv* (2021).

[3] André Bächtiger and John Parkinson. 2019. *Mapping and measuring deliberation: Towards a new deliberative quality*. Oxford University Press.

[4] André Bächtiger, Susumu Shikano, Seraina Pedrini, and Mirjam Ryser. 2009. Measuring deliberation 2.0: standards, discourse types, and sequenzialization. In *ECPR General Conference*. Potsdam, 5–12.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2020. Optimizing ai for teamwork. *arXiv preprint arXiv:2004.13102* (2020).

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[9] Jason Barabas. 2004. How deliberation affects policy opinions. *American political science review* 98, 4 (2004), 687–701.

[10] Donald J Baumann, John D Fluke, Len Dalgleish, and Homer Kern. 2014. The decision-making ecology. *From evidence to outcomes in child welfare: An international reader* (2014), 24–40.

[11] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.

[12] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[13] Laura W Black, Stephanie Burkhalter, et al. 2010. Methods for analyzing and measuring group deliberation. In *Sourcebook for political communication research*. Routledge, 345–367.

[14] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[15] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[17] André Bächtiger, Marlène Gerber, and Eléonore Fournier-Tombs. 2022. 83Discourse Quality Index. In *Research Methods in Deliberative Democracy*. Oxford University Press. doi:10.1093/oso/9780192848925.003.0006 arXiv:https://academic.oup.com/book/0/chapter/378695331/chapter-pdf/49943298/oso-9780192848925-chapter-6.pdf

[18] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[19] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[20] Nancy Cartwright and Jacob Stegenga. 2011. A theory of evidence for evidence-based policy. (2011).

[21] Simone Chambers. 2005. Measuring publicity's effect: Reconciling empirical research and normative theory. *Acta Politica* 40 (2005), 255–266.

[22] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[23] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[24] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[26] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[27] Dennis J Devine, Laura D Clayton, Benjamin B Dunford, Rasmy Seying, and Jennifer Pryce. 2001. Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, public policy, and law* 7, 3 (2001), 622.

[28] Ap Dijksterhuis, Maarten W Bos, Loran F Nordgren, and Rick B Van Baaren. 2006. On making the right choice: The deliberation-without-attention effect. *Science* 311, 5763 (2006), 1005–1007.

[29] Steven E Dilsizian and Eliot L Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16, 1 (2014), 1–8.

[30] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[31] Charles A Doswell. 2004. Weather forecasting by humans—Heuristics and decision making. *Weather and Forecasting* 19, 6 (2004), 1115–1126.

[32] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 32–41.

[33] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[34] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[35] Franz Eisenfuhr. 2011. Decision making.

[36] Jonathan St BT Evans. 2002. Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin* 128, 6 (2002), 978.

[37] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074.

[38] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[39] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[40] James S Fishkin. 2018. *Democracy when the people are thinking: Revitalizing our politics through public deliberation*. Oxford University Press.

[41] Johannes Fürnkranz and Eyke Hüllermeier. 2010. Preference learning and ranking by pairwise comparison. In *Preference learning*. Springer, 65–82.

[42] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[43] Francis Galton. 1907. Vox populi. *Nature* 75, 1949 (1907), 450–451.

[44] Jing Gao, Feng Tian, Junjun Fan, Dakuo Wang, Xiangmin Fan, Yicheng Zhu, Shuai Ma, Jin Huang, and Hongan Wang. 2018. Implicit detection of motor impairment in Parkinson's disease from everyday smartphone interactions. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[45] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[46] Johny Ghattas, Pnina Soffer, and Mor Peleg. 2014. Improving business process decision making based on past experience. *Decision Support Systems* 59 (2014), 93–107.

[47] Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. 2021. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410* (2021).

[48] Isidore Jacob Good. 1950. Probability and the Weighing of Evidence. (1950).

[49] David Gough. 2007. Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research papers in education* 22, 2 (2007), 213–228.

[50] Diego Gracia. 2003. Ethical case deliberation and decision making. *Medicine, Health Care and Philosophy* 6 (2003), 227–233.

[51] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[52] Jürgen Habermas. 2005. Concluding comments on empirical approaches to deliberative politics. *Acta politica* 40 (2005), 384–392.

[53] Robert Harris. 1997. Evaluating Internet research sources. *Virtual salt* 17, 1 (1997), 1–17.

[54] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[55] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[56] Randy Y Hirokawa. 1985. Discussion procedures and decision-making performance: A test of a functional perspective. *Human Communication Research* 12, 2 (1985), 203–224.

[57] Guy Hochman, Shahar Ayal, and Dan Ariely. 2015. Fairness requires deliberation: The primacy of economic over social considerations. *Frontiers in psychology* 6 (2015), 747.

[58] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[59] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[60] Giulia Inguaggiato, Suzanne Metselaar, Bert Molewijk, and Guy Widdershoven. 2019. How moral case deliberation supports good clinical decision making. *AMA journal of ethics* 21, 10 (2019), 913–919.

[61] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[62] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[63] Robert A Kaufman and David Kirsh. 2022. Cognitive Differences in Human and AI Explanation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[64] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[65] János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. 2022. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications* 13, 1 (2022), 7214.

[66] Geoffrey P Kramer, Norbert L Kerr, and John S Carroll. 1990. Pretrial publicity, judicial remedies, and jury bias. *Law and human behavior* 14, 5 (1990), 409–438.

[67] Louis Kriesberg. 2007. *Constructive conflicts: From escalation to resolution*. Rowman & Littlefield.

[68] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[69] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[70] Hélène Landemore. 2012. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press.

[71] Hélène Landemore and Scott E Page. 2015. Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, philosophy & economics* 14, 3 (2015), 229–254.

[72] James R Larson, Pennie G Foster-Fishman, and Christopher B Keys. 1994. Discussion of shared and unshared information in decision-making groups. *Journal of personality and social psychology* 67, 3 (1994), 446.

[73] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76.

[74] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[75] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical

Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[76] Sung-Chul Lee, Jaeyoon Song, Eun-Young Ko, Seongho Park, Jihee Kim, and Juho Kim. 2020. Solutionchat: Real-time moderator support for chat-based structured discussion. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[77] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[78] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[79] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
[80] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* (2023).
[81] Christopher Lord and Dionysia Tamvaki. 2013. The politics of justification? Applying the 'Discourse Quality Index'to the study of the European Parliament. *European Political Science Review* 5, 1 (2013), 27–54.
[82] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[83] Fred C Lunenburg. 2010. The decision making process.. In *National Forum of Educational Administration & Supervision Journal*, Vol. 27.
[84] Aidan Lyon and Eric Pacuit. 2013. The wisdom of crowds: Methods of human judgement aggregation. In *Handbook of human computation*. Springer, 599–614.
[85] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
[86] Shuai Ma, Mingfei Sun, and Xiaojuan Ma. 2022. Modeling Adaptive Expression of Robot Learning Engagement and Exploring its Effects on Human Teachers. *ACM Transactions on Computer-Human Interaction* (2022).
[87] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. " Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2403.09552* (2024).
[88] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. 2019. SmartEye: assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
[89] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. *arXiv preprint arXiv:2403.01791* (2024).
[90] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *CHI Conference on Human Factors in Computing Systems*. 1–25.
[91] Merriam-Webster. [n. d.]. Deliberation Definition. https://www.merriam-webster.com/dictionary/deliberation.
[92] Katherine L Milkman, Dolly Chugh, and Max H Bazerman. 2009. How can decision making be improved? *Perspectives on psychological science* 4, 4 (2009), 379–383.
[93] Deborah J Miller, Elliot S Spengler, and Paul M Spengler. 2015. A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology* 62, 4 (2015), 553.
[94] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
[95] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 333–342.
[96] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.
[97] Tina Nabatchi and Matt Leighninger. 2015. *Public participation for 21st century democracy*. John Wiley & Sons.
[98] Masi Noor, Rupert Brown, Roberto Gonzalez, Jorge Manzi, and Christopher Alan Lewis. 2008. On positive psychological outcomes: What helps groups with a history of conflict to forgive and reconcile with each other? *Personality and Social Psychology Bulletin* 34, 6 (2008), 819–832.
[99] DJ Pangburn. 2019. Schools are using software to help pick who gets in. What could go wrong. *Fast Company* 17 (2019).
[100] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
[101] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.
[102] Charles Sanders Peirce. 2014. *Illustrations of the Logic of Science*. Open Court.
[103] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
[104] Anne Preisz. 2019. Fast and slow thinking; and the problem of conflating clinical reasoning and ethical deliberation in acute decision-making. *Journal of paediatrics and child health* 55, 6 (2019), 621–624.
[105] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.
[106] Thomas L Saaty. 2008. Decision making with the analytic hierarchy process. *International journal of services sciences* 1, 1 (2008), 83–98.
[107] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[108] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.
[109] Chuhan Shi, Yicheng Hu, Shenan Wang, Shuai Ma, Chengbo Zheng, Xiaojuan Ma, and Qiong Luo. 2023. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[110] Herbert A Simon. 1990. Bounded rationality. *Utility and probability* (1990), 15–18.
[111] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.
[112] Robert L Simon. 2008. *The Blackwell guide to social and political philosophy*. John Wiley & Sons.
[113] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[114] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations. (2022).
[115] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics* 1 (2003), 21–48.
[116] Jürg Steiner, André Bächtiger, Markus Spörndli, and Marco R Steenbergen. 2005. Deliberative politics in action. Analysing parliamentary discourse. (2005).
[117] Mark Steyvers and Aakriti Kumar. 2022. Three Challenges for AI-Assisted Decision-Making. (2022).
[118] Philip E Tetlock. 2017. Expert political judgment. In *Expert Political Judgment*. Princeton University Press.
[119] Dennis F Thompson. 2008. Deliberative democratic theory and empirical political science. *Annu. Rev. Polit. Sci.* 11 (2008), 497–520.
[120] Amy Turner, Meena Kaushik, Mu-Ti Huang, and Srikar Varanasi. 2022. Calibrating trust in AI-assisted decision making.
[121] Dartmouth U Mass. [n. d.]. 7 STEPS TO EFFECTIVE DECISION MAKING. https://www.umassd.edu/fycm/decision-making/process/.
[122] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.
[123] Frans H Van Eemeren and A Francisca Sn Henkemans. 2016. *Argumentation: Analysis and evaluation*. Taylor & Francis.
[124] Frans H Van Eemeren, A Francisca Sn Henkemans, and Rob Grootendorst. 2002. *Argumentation: Analysis, evaluation, presentation*. Routledge.
[125] Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*. Springer, 251–262.
[126] Xinru Wang, Chen Liang, and Ming Yin. [n. d.]. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets. ([n. d.]).
[127] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[128] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[129] Austin Waters and Risto Miikkulainen. 2014. Grade: Machine learning support for graduate admissions. *Ai Magazine* 35, 1 (2014), 64–64.
[130] Douglas L Weed. 2005. Weight of evidence: a review of concept and methods. *Risk Analysis: An International Journal* 25, 6 (2005), 1545–1557.
[131] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[132] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
[133] Magdalena E Wojcieszak, Young Min Baek, and Michael X Delli Carpini. 2010. Deliberative and participatory democracy? Ideological strength and the processes leading from deliberation to political engagement. *International Journal of Public Opinion Research* 22, 2 (2010), 154–180.
[134] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[135] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
[136] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
[137] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
[138] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. 2023. Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics* (2023).
[139] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

Appendices

A Baseline XAI Interface

Figure 10 shows the baseline XAI interface.

B An Example Dataflow of Deliberative AI

Figure 11 provides the details of a conversation between a human and a *Deliberative AI* discussing how an applicant's GPA affects admissions chances. Here's a step-by-step breakdown:

(1) The user inputs GPA-related arguments in the dialogue interface.
(2) The system packages the user's input as a prompt for the *Intention Analyzer* in the Communication Layer.
(3) The *Intention Analyzer* recognizes attributes and intentions and saves in JSON format, then forwards it to the *Knowledge Extractor* in the Control Layer.
(4) The *Knowledge Extractor* generates a query function and fetches statistical results from the DS-Model and training data.
(5) The statistical results are transmitted to the *Regulator*.
(6) *Regulator* crafts a constraint prompt ensuring consistency between the LLM's output and the DS-Model's prediction, feeding it to the LLM-based *Deliberation Facilitator*.
(7) The *Deliberation Facilitator* generates responses to the user's initial arguments.

C Detailed Metrics and Questions

Table 2 shows the detailed metrics and questions used in our measurement.

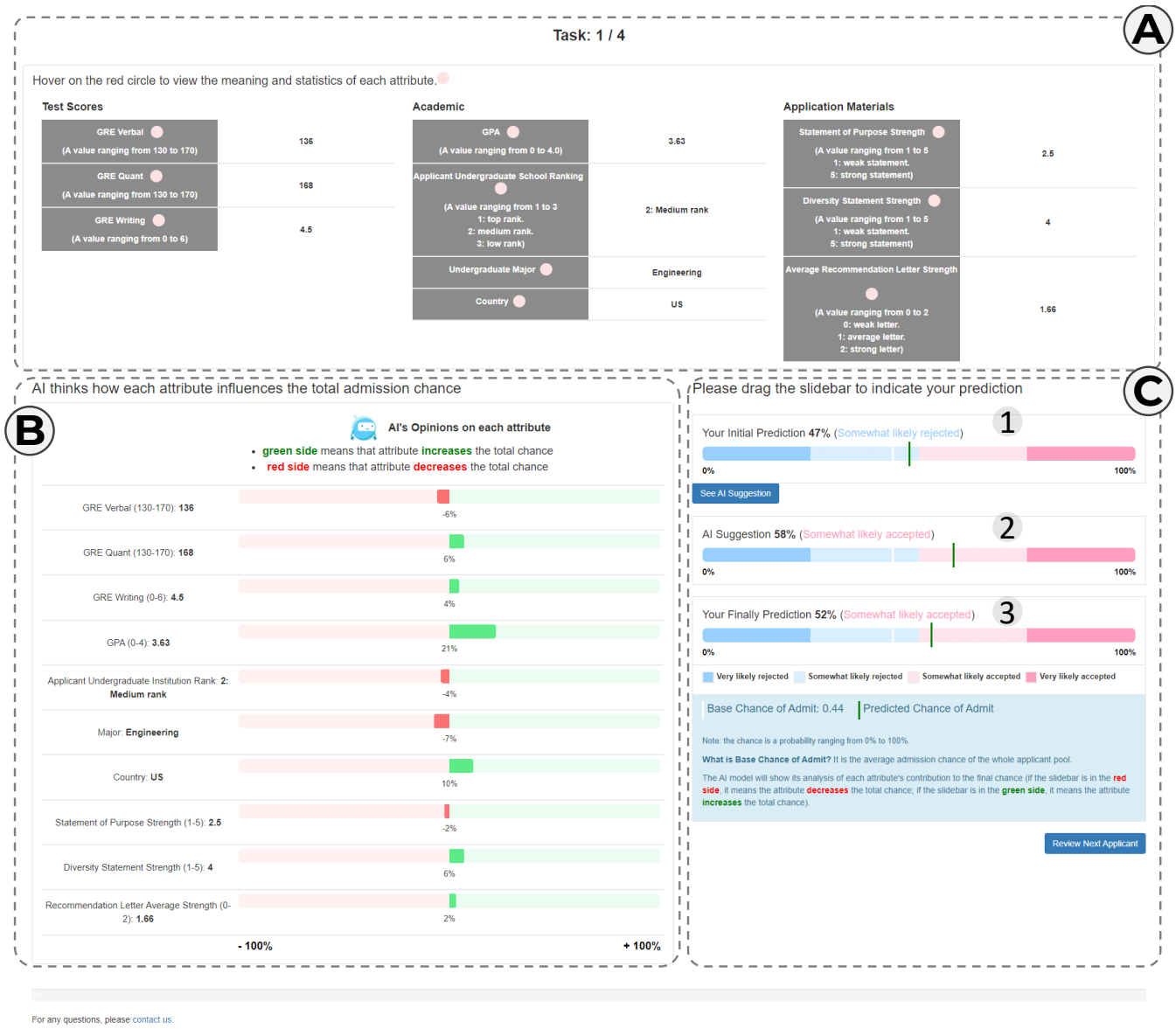

Figure 10: The baseline XAI (traditional explainable AI) interface in our user study. The interface contains three parts. The top (A) is the applicant's profile. The bottom left part (B) shows AI's feature contribution explanation. The bottom right part (C) is for humans to (1) indicate their initial predictions, (2) see AI's suggestions, and (3) indicate their final decisions. Note that AI's suggestions and explanations are only shown after humans make their initial predictions. (All the dashed lines are only for illustration)

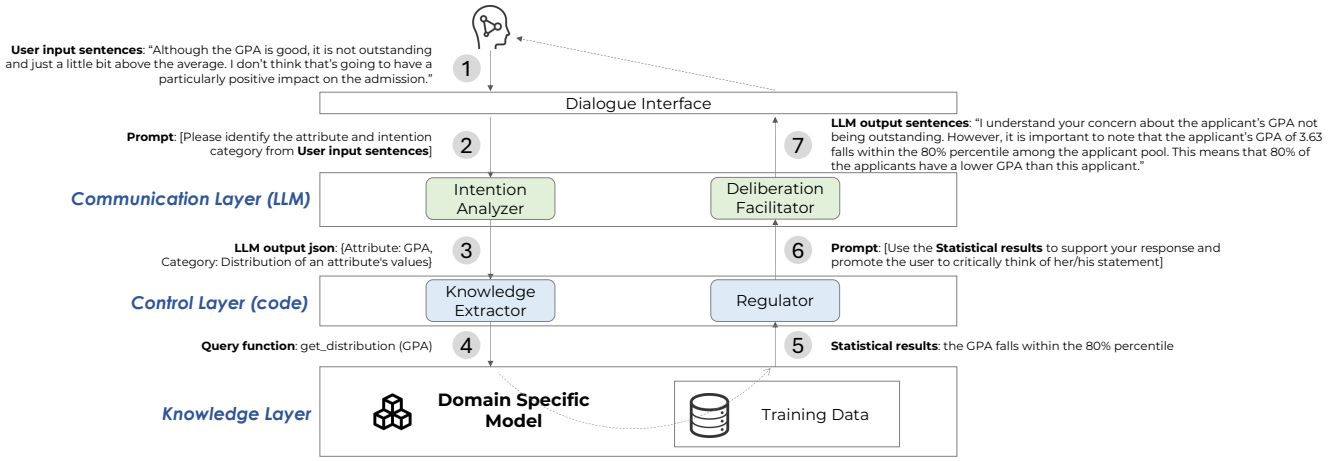

Figure 11: An illustration of how *Deliberative AI* processes humans' inputs and how it generates outputs. The prompts used are simplified in this figure for illustration purposes. (for the complete prompts, please check our supplementary materials)

Table 2: Measurements used in our user study. We collected participants' objective decision data, subjective questionnaire data, and qualitative open-ended feedback.

Aspect	Metrics	Detailed Meaning and Questions
Objective Measures		
Performance	Decision Accuracy	Accuracy of participants' final predictions.
Reliance	Agreement Fraction	Percentage of tasks where participants' final prediction agreed with AI's prediction. $\frac{\text{Number of final decisions same as the AI suggestion}}{\text{Total number of decisions}}$
	Switch Fraction	Percentage of tasks where AI's prediction was used when initial disagreement existed. $\frac{\text{Number of decisions user switched to agree with the AI model}}{\text{Total number of decisions with initial disagreement}}$
	Over-reliance Ratio	Fraction of tasks where participants used an incorrect AI prediction. $\frac{\text{Number of incorrect human final decisions with incorrect AI suggestions}}{\text{Total number of incorrect AI suggestions}}$
	Under-reliance Ratio	Fraction of tasks where participants did not use a correct AI prediction. $\frac{\text{Number of incorrect human final decisions with correct AI suggestions}}{\text{Total number of correct AI suggestions}}$
Subjective Measures		
Perceptions of AI	Helpfulness	*"I think the AI model's assistance is helpful/useful for me to make good decisions."*
	Trustworthiness	*"The AI model can be trusted to provide reliable decision support."*
	Understanding	*"I understand how the AI model works to predict an applicant's chance of being admitted."*
User Experience	Decision Confidence	*"I feel confident in the decisions I made."*
	Mental Demand	*"The decision-making process is mentally demanding."*
	Effort	*"I have to work hard (mentally and physically) to accomplish my level of performance."*
	Complexity	*"The decision-making process and the interaction with AI models are complex."*
	Satisfaction	*"I am satisfied with the AI model's assistance and the decision-making process."*
Open-ended Feedback	Perception of helpfulness	*"Do you think the discussion with AI is (or not) helpful? Could you tell us the reasons why you think the discussion is helpful (or not helpful)?"*
	Perception of AI update	*"What do you think of the AI updating its own views during the discussion?"*
	Potential Improvement	*"To make a better discussion, which parts do you think the current AI needs to be improved, and how should it be improved?"*