Incorporating Respect into LLM-Based Academic Feedback: A BI-R Framework for Instructing Students after Q&A Sessions

Mayuko Aiba Grad. School of Engineering The University of Tokyo

Daisuke Saito Grad. School of Engineering The University of Tokyo

Nobuaki Minematsu Grad. School of Engineering The University of Tokyo

{aiba, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

In academic research, post-presentation Q&A sessions are crucial for deepening understanding and shaping research directions. Supervisors' comments are particularly valuable when they highlight perspectives that students have not yet fully considered. Such comments typically arise from careful reasoning within dialogue, yet large language models (LLMs) still struggle to reason precisely about dialogue context and communicative intentions.

Building on LLMs, this study proposes a feedback generation framework based on the Belief–Desire–Intention (BDI) model, which conceptualizes Q&A sessions as cognitive interactions between presenters and questioners. We further extend this framework into BI-R by introducing *Respect* as an explicit dimension, ensuring that generated feedback is not only accurate but also pedagogically constructive.

We evaluated the proposed frameworks (BDI and BI-R) through comparative experiments with master's students and field experiments with doctoral students during pre-defense presentations. Results showed that while the BDI prompt did not outperform the baseline, the BI-R prompt was particularly effective when students did not fully grasp the broader context or background of the questions. When comparing BDI and BI-R, the inclusion of Respect improved the tone and pedagogical appropriateness of feedback. These findings highlight the potential of the proposed framework as a supportive tool for training students and early-career researchers.

1 Introduction

Academic presentations typically consist of a formal talk that summarizes research findings, followed by an interactive Q&A session. For students, the Q&A enables deeper communication about their work, which can shape the future directions of their research and enhance the quality of their presentations.To maximize the impact of this process, it is essential for students to accurately interpret the questions they receive and respond appropriately. In many academic contexts, supervisors observe both the presentation and the Q&A session, often take note of the questions asked, and evaluate the adequacy of the presenter's responses. They then use these observations to offer concrete and actionable guidance. Supervisors' comments are especially valuable when questions introduce new perspectives that the student has not yet considered, or when the student's responses reveal gaps in understanding that need to be addressed in later revisions.

In this study, we aim to replicate this pedagogical process by using large language models (LLMs) to analyze Q&A transcripts and generate feedback for students, simulating the role of a supervisor.

In recent years, LLMs have shown remarkable progress across a wide range of natural language processing (NLP) tasks and have drawn growing attention in the domain of higher education. In the research process, students typically prepare manuscripts, submit them to conferences, practice their presentations, and anticipate potential Q&A. After the conference, they reflect on the questions and discussions they encountered and use these insights to guide the next steps of their research. Many of these activities are already supported by LLM-based tools. For example, Okgetheng and Takeuchi (2024) have demonstrated the use of LLMs for improving academic writing. Likewise, Aiba et al. (2024) developed a speech-based system that simulates Q&A sessions to help non-native English-speaking students prepare for international conferences.

Although previous studies have focused on supporting presentation preparation, such as automatically generating questions in advance, little research has addressed how to assist reflection and learning after the presentation, particularly through

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 288–301
February 26–March 1, 2026.

post-presentation Q&A sessions. Building on this line of research on LLM-based academic support, the present study investigates the analysis of actual post-presentation Q&A sessions and the generation of feedback using LLMs. Specifically, we aim to partially automate the feedback process that typically occurs after presentations, such as in mid-term progress reviews. It is common to observe students giving responses that are off the point during Q&A, often because they fail to fully grasp the questioner's intention or the background of the question. Our approach uses LLMs to infer the mental states of both presenters and questioners from Q&A transcripts, and to generate feedback that highlights unaddressed perspectives or gaps in understanding to guide future revisions. By leveraging LLMs, we aim to provide scalable feedback that can support many students across multiple domains, even when supervisors have limited time for detailed follow-up.

2 Related Research

2.1 LLM-based Support for Academic Contexts

Recent research has actively explored the use of LLMs in educational contexts (Shin and Lee, 2023; Hu et al., 2024; Zhang et al., 2024; Wu et al., 2024). In particular, these models have been applied to provide feedback and support students' independent practice (Allen and Mizumoto, 2024; Steiss et al., 2024; Zhong et al., 2024).

One line of research has focused on generating feedback on academic writing. For instance, Liang et al. (2024) investigated the use of GPT-4 to generate feedback for academic papers in the context of peer review and evaluated its quality. The study found that the overlap in feedback content between two human reviewers was comparable to that between a human reviewer and GPT-4, and that over 80% of researchers rated GPT-4's feedback as more useful than that of another human reviewer.

Another study, Aiba et al. (2024), developed a speech-based Q&A practice system that utilizes ChatGPT to support non-native English-speaking students preparing for international conferences. Multiple prompting strategies were tested to generate paper-specific questions using ChatGPT. This work suggests that LLM-based question generation can enable students to practice independently, a process that would conventionally require close collaboration with a supervisor who has a deep understanding of the paper's content.

2.2 The Belief–Desire–Intention Model for Dialogue Understanding

Originally developed in the fields of artificial intelligence and multi-agent systems, the Belief–Desire–Intention (BDI) model (Bratman, 1987) is widely used as a theoretical framework for describing the decision-making processes of intelligent agents. The model explains the behavior of intelligent agents in terms of three mental components:

Beliefs the agent's knowledge about the environment and other agents

Desires the goals or states of the world the agent wishes to achieve

Intentions the concrete plans and courses of action that the agent commits to in order to achieve its goals

By formalizing decision-making in this way, the framework accounts for how actions, including utterances, are selected by the agent.

In dialogue analysis, understanding an interlocutor's utterances often requires inferring the underlying beliefs, desires, and intentions that guide these utterances. This perspective highlights how dialogue involves more than surface-level content, as it reflects reasoning processes that must be reconstructed by the listener. Following prior work that applied the BDI model to reasoning and dialogue interpretation (Traum and Allen, 1994; Rao et al., 1995; Iida et al., 2024; Al Owayyed et al., 2025), we model each participant's mental state in terms of beliefs, desires, and intentions. Combining these mental components enables a more precise interpretation of the reasoning and motivations behind observed behaviors. It should be noted, however, that the mental state inferred by an observer does not necessarily match the actual mental state of the interlocutor, which can sometimes result in misunderstandings or off-point responses.

2.3 Reasoning and Understanding Challenges in LLMs and Structured Approaches to Overcome Them

Despite remarkable advances in recent years, LLMs still face notable challenges in achieving reasoning capabilities comparable to those of humans. For instance, even in relatively simple arithmetic word problems, they may fail when intermediate reasoning steps are omitted, leading to incorrect answers. They also tend to rely on brute-force strategies, such as exhaustive search, rather than exploiting structural properties of the problem that would allow more efficient solutions (e.g., binary search).

To mitigate these shortcomings, a growing body of research has explored a range of structured prompting techniques. A representative approach is Chain of Thought (CoT) prompting, which instructs the model to “think step by step” and has been shown to improve reasoning accuracy (Wei et al., 2022). Building on this idea, more advanced frameworks such as Tree of Thought (ToT) (Yao et al., 2023a), which encourages the branching and evaluation of multiple hypotheses, and ReAct (Yao et al., 2023b), which alternates between reasoning and acting (e.g., retrieval or calculation), have been proposed to enhance reasoning and decision-making.

Beyond reasoning tasks, in conversational contexts, LLMs often struggle to grasp speakers’ intentions, implicatures, or subtle contextual cues that extend beyond the literal meaning of utterances. To address these challenges, recent research has sought to strengthen conversational generative AI by incorporating mental state modeling. For instance, Iida et al. (2024) propose the Mental Model of Others (MMO), which represents beliefs, desires, and intentions, thereby enabling the model to grasp interlocutors’ implicatures in dialogue. They introduce two integration methods: LLM Embedded in Cognitive Model (LEC) and Cognitive Model Embedded in LLM (CEL). Among these, the LEC approach, in which the LLM operates as a module within a BDI-style cognitive architecture, effectively handles implicature-rich dialogues, outperforming conventional LLMs. These findings suggest that embedding BDI-style reasoning can enhance LLMs’ interpretive capabilities beyond surface-level content.

2.4 Aim of This Work

Building on these insights, this study employs a BDI-inspired framework to interpret ASR-based transcripts of post-presentation Q&A sessions using LLMs, with the aim of generating feedback that supports future research development and presentation improvement.

Within the broader research on structuring LLM reasoning through external guidance, our work specifically addresses the challenge of interpreting others’ reasoning processes in dialogue. By prompting LLMs to reconstruct the underlying beliefs, desires, and intentions behind each utterance, we aim to provide more supervisor-like feedback that is not only context-sensitive but also constructive and considerate of the presenter’s perspective.

We position this work as a case study illustrating how mental-state modeling can enhance feedback generation in educational contexts.

3 Proposed Method

3.1 Interpreting Q&A as Communication Through the BDI Model

In this study, Q&A interactions are analyzed through the lens of the BDI model, viewing them not merely as exchanges of information but as interactions between the mental states of presenters and questioners.

The questioner’s mental states can be decomposed as follows:

Belief The questioner’s understanding of the research content, shaped by listening to the presentation and integrating it with their own knowledge and values.

Desire The wish to confirm, clarify, or contribute something meaningful regarding the presentation.

Intention The aim to initiate discussion or provide new perspectives that help deepen the research.

Similarly, the presenter’s mental states can be described as:

Belief The presenter’s interpretation of the questioner’s thought process, situated within the context of their own research.

Desire The wish to respond clearly and accurately to the question.

Intention The aim to provide an answer that is grounded in the research and that helps develop the discussion further.

As illustrated in Fig.1, we define misalignment as a discrepancy between the questioner’s actual belief and the presenter’s inferred belief, as well as between their respective intentions. Such misalignments are often the source of inaccurate or uninformative responses. In our method, identifying and explicitly representing these discrepancies in BDI components may serve as the basis for generating precise and constructive feedback. In the present setting, we assume that the desires of the questioner and the presenter generally align, as both aim to deepen understanding of the research topic through accurate and constructive discussion. Therefore,

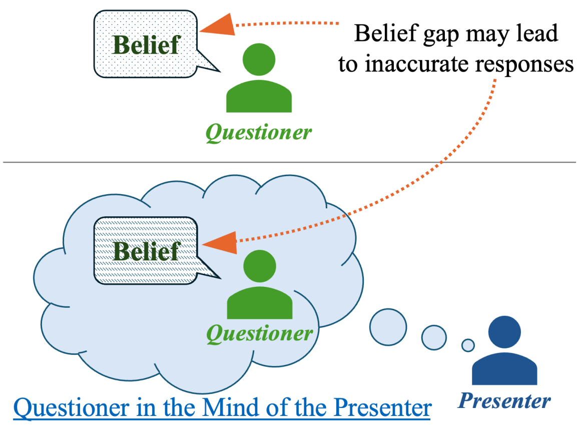

Figure 1: Misalignment between the questioner's actual belief and the questioner's belief inferred by the presenter often results in inaccurate or uninformative responses.

our analysis focuses primarily on inferring beliefs and intentions.

3.2 Prompt Design

To generate feedback from Q&A sessions using LLMs, we designed three types of prompts. All prompts take the presentation material and the Q&A log as input and generate feedback for each Q&A pair. For the BDI and BI-R prompts, we used a multi-turn, four-step procedure, where each step was a separate interaction and its output was fed into the next step. The flows of the prompts are summarized in Figure 2, and detailed prompt wording is provided in the Appendix A.

3.2.1 Baseline Prompt

This serves as the baseline method in our experiments. Given the presentation material and the Q&A log, the model is simply instructed to generate feedback directly, without additional reasoning steps.

3.2.2 BDI Prompt

This prompt incorporates reasoning steps based on the BDI model to analyze the mental states of both questioner and presenter:

1. Infer the questioner's belief and intention for each question.
2. Infer the presenter's belief and intention for each response, and identify points of alignment and discrepancy with the inferred questioner's mental state.
3. Generate feedback based on the analysis from step 2.

3.2.3 BI-R Prompt

This extends the BDI Prompt by incorporating pedagogical considerations into the feedback generation process.

While the BDI model accounts for rational decision-making, generating feedback for supporting students during academic Q&A sessions also requires sensitivity to educational and interpersonal factors. To address this, we introduce *Respect* not as a mental-state component, but as a guiding principle for feedback generation. This ensures that the output is not only logically accurate, but also pedagogically appropriate to the presenter's level of understanding and expressed in a respectful manner. The steps for the BI-R prompt are as follows:

1. Infer the questioner's belief and intention for each question.
2. Infer the presenter's belief and intention for each response, and identify points of alignment and discrepancy with the inferred questioner's mental state.
3. Generate feedback that reflects the analysis from step 2, with explicit guidance to **respect** the presenter by ensuring that the feedback is appropriate in tone, content, and amount.

3.3 Experimental Evaluation

In Sections 4 and 5, we describe two experiments conducted to evaluate the effectiveness of the proposed method. The evaluation consisted of (i) a comparative experiment with master's students, in which two prompting strategies were compared pairwise (Section 4), and (ii) both field and comparative evaluations with three doctoral students, in which all three prompting strategies were directly compared during pre-defense dissertations (Section 5).

4 Evaluation with Master's Students

4.1 Data Collection

We collected data from six master's students at the University of Tokyo's Graduate School of Engineering during their pre-defense presentations. In these presentations, the students presented the progress of their own research after preparing an eight-page mid-term review paper that accompanied their presentation. Each presentation was conducted online via Zoom and consisted of a 20-minute presentation followed by a 15-minute Q&A session.

The following materials were obtained: (i) an eight-page mid-term review paper, (ii) presentation

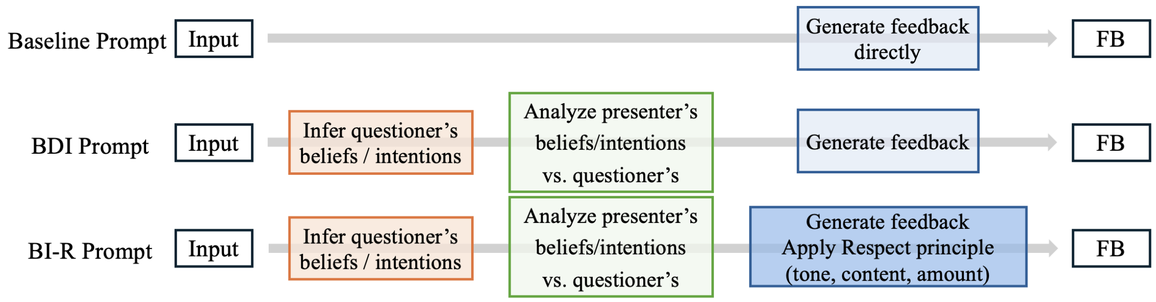

Figure 2: Overview of the prompt designs. The Baseline prompt directly generates feedback. The BDI prompt first infers the questioner's beliefs and intentions, then analyzes the presenter's beliefs and intentions in relation to those of the questioner before generating feedback. The BI-R prompt further incorporates explicit guidance emphasizing respect when generating feedback.

slides, and (iii) transcripts of the presentations and Q&A sessions. The transcripts were first generated using Zoom's automated captioning feature. Subsequently, they were corrected for minor recognition errors and typographical inconsistencies with the assistance of ChatGPT (OpenAI, 2023), taking into account the context of the research presentation. Trivial responses consisting only of short acknowledgments (e.g., "Thank you," "That's right") were excluded from analysis.

4.2 Experimental Procedure

Feedback was generated using GPT-4o with the three prompting strategies described in Section 3.2: Baseline, BDI, and BI-R. For each Q&A pair, two strategies were directly compared in a controlled evaluation. Because the Q&A sessions varied in length across presentations, and only questions posed in Japanese were analyzed, the number of evaluated Q&A pairs differed by participant.

Two comparison settings were conducted: Comparison I (Baseline vs. BDI; $N = 5$ participants) and Comparison II (Baseline vs. BI-R; $N = 6$ participants). A direct comparison between the BDI and BI-R prompts is presented in Section 5.

The participants (P1-P6) were second-year master's students from the same university where the data were collected. Of these, five students (P1–P5) took part in both Comparison I and Comparison II.

For each comparison, participants received an evaluation sheet containing the transcript of their own Q&A session and the feedback generated by two different prompts. In the evaluation sheets, the paired prompts were anonymized and presented simply as Prompt 1 and Prompt 2.

All presenters and questioners were native Japanese speakers, and the feedback was also provided in Japanese.

4.3 Evaluation Criteria

In this study, we do not evaluate the correctness of intermediate BDI/BI-R inferences themselves, but assess the practical value of the framework through the usefulness and acceptability of the final feedback. We employed two evaluation metrics: usefulness and acceptability.

Usefulness refers to the extent to which the feedback provided balanced and sufficient information to improve the student's research, presentation, and subsequent Q&A performance. It was evaluated on a 5-point Likert scale (5 = Very useful, 4 = Useful, 3 = Neutral, 2 = Not useful, 1 = Very unhelpful).

Acceptability refers to how naturally the feedback could be received without feeling inconsistent with the presenter's perspective. Presenters compared the two types of feedback and selected which one they felt was more natural and easier to accept.

4.4 Results

Tables 1, 2, and 3 present the evaluation outcomes. Table 1 summarizes the mean usefulness scores, the preference rates for BDI and BI-R over the baseline, and the results of statistical tests across the two comparisons. Table 2 corresponds to Comparison I (Baseline vs. BDI), and Table 3 corresponds to Comparison II (Baseline vs. BI-R). Both tables report participant-level results, showing individual variability in preferences and ratings.

4.4.1 Usefulness

As shown in Table 1, in Comparison I, the mean usefulness score on the 5-point scale was 3.57 for the baseline and 3.32 for the BDI prompt. A paired-samples t-test revealed no significant difference between the two conditions, $t(26) = -0.17$, $p = .866$.

In Comparison II, the mean usefulness score was 3.50 for the baseline and 3.93 for the BI-R prompt.

Table 1: Summary of usefulness and preference results across prompting strategies.

Comparison	Usefulness (Base)	Usefulness (BDI/BI-R)	Preference Rate	Statistical Test
Baseline vs. BDI	3.57	3.32	.568	$t(26) =$ n.s., $p > .05$
Baseline vs. BI-R	3.50	3.93	.748	$W = 1,\ p < .05$

Preference rate is proportion of Q&A pairs where participants preferred BDI/BI-R feedback in terms of acceptability. Wilcoxon signed-rank test compares the per-participant preference rates shown in Tables 2 and 3 against chance (0.5).

Table 2: Participant-level acceptability results for Comparison I (Baseline vs. BDI).

Participant	Baseline	BDI	Preference Rate (BDI)
P1	1	1	.500
P2	0	3	1.000
P3	5	2	.286
P4	4	5	.556
P5	3	3	.500
Overall	13	14	**.568**

Table 3: Participant-level acceptability results for Comparison II (Baseline vs. BI-R).

Participant	Baseline	BI-R	Preference Rate (BI-R)
P1	0	2	1.000
P2	0	3	1.000
P3	4	3	.429
P4	2	7	.778
P5	2	4	.667
P6	5	8	.615
Overall	13	27	**.748**

This difference was statistically significant, $t(39) = 2.42$, $p = .020$, indicating that BI-R feedback was rated as more useful overall.

4.4.2 Acceptability

For Comparison I, the mean preference rate for BDI-generated feedback was 56.8% (Table 2), showing no statistically reliable advantage over the baseline (Table 1).

For Comparison II, the mean preference rate for BI-R-generated feedback was 74.8% (Table 3), which was significantly greater than chance ($p < .05$; Table 1).

4.5 Discussion

The evaluation results offer several insights into how BDI-inspired reasoning can structure LLM-based feedback. In summary, while the BDI prompt alone did not yield significant improvements in usefulness or acceptability compared to the baseline, the BI-R prompt produced feedback that was both more useful and more readily accepted by the participants. This pattern highlights the importance of balancing logical accuracy with pedagogical and interpersonal appropriateness in the generation of feedback within academic Q&A contexts.

The limited impact of the BDI prompt suggests that merely introducing belief and intention inference is not sufficient to generate feedback perceived as helpful by learners. Participants often found the BDI feedback overly detailed or overly critical, with some describing it as "too harsh." This tendency aligns with Shute (2008)'s pedagogical observation that feedback perceived as overly detailed or critical can exceed the learner's cognitive tolerance, underscoring the need for calibration.

In contrast, the BI-R prompt explicitly guided the model to produce feedback that integrated both analytical rigor and interpersonal sensitivity. This adjustment enabled the generation of feedback that participants were more willing to accept. The results resonate with Hattie and Timperley (2007)'s framework, which emphasizes that constructive, positively framed feedback enhances learning motivation more effectively than purely corrective remarks.

In this study, we extend the application of the BDI model to the context of academic post-presentation Q&A sessions. Unlike purely rational decision-making, this setting requires educational and social considerations. Therefore, we argue that incorporating Respect as an additional dimension is essential for generating feedback that is not only analytically sound but also pedagogically and interpersonally appropriate.

Although most participants favored the BI-R prompt, one participant (P3) consistently preferred the baseline feedback in both Comparison I and Comparison II, commenting that overly long responses were less appealing and that brevity was preferable when the content was equivalent. This divergence highlights the need to consider individual preferences regarding feedback length. Similar mechanisms are already employed in LLM training, where models occasionally ask users to choose between alternative responses to collect preference data.

Taken together, these findings suggest that the quality of feedback in academic Q&A contexts should be shaped not only by logical soundness but also by pedagogical sensitivity and social awareness. Explicitly instructing models to respect the presenter appears to help balance tone, detail, and length, making the feedback more acceptable to students. This indicates a promising direction for designing LLM-based feedback systems that support learning by combining structured reasoning with social and educational sensitivity.

5 Evaluation with Doctoral Students

5.1 Data Collection with Doctoral Students

We collected data from three third-year doctoral students (P7–P9) during their pre-defense reviews. In these reviews, each student presented the progress of their doctoral research. The reviews were conducted face-to-face in a university conference room and consisted of a one-hour presentation followed by a one-hour Q&A session.

The following materials were obtained: (i) presentation slides, (ii) transcripts of the presentations and Q&A sessions, and (iii) research papers authored by the presenters that were directly related to the presentation content. The transcripts were generated using the conference room microphone connected to Zoom's automated captioning feature. Subsequently, they were corrected for minor recognition errors and typographical inconsistencies with the assistance of ChatGPT, taking into account the context of the research presentation. As described in Section 4.1, trivial responses consisting only of short acknowledgments (e.g., "Thank you," "That's right") were excluded from analysis. All presenters and questioners were native Japanese speakers, and the feedback was also provided in Japanese, as in Section 4.

5.2 Experimental Procedure

Feedback was generated using GPT-4o with the three prompting strategies described in Section 3.2.

For each Q&A pair, the feedback outputs from all three prompting strategies were presented side-by-side to each participant for evaluation. The students rated each feedback item on a 5-point usefulness scale and selected which feedback felt most acceptable, based on the evaluation criteria introduced in Section 4.3.

Table 4: Preference rates and mean usefulness (5-point scale) for doctoral students; each cell shows "rate, mean," with the values in parentheses after each participant indicating the number of Q&A pairs. Means are computed across participants (unweighted).

Participant	Baseline	BDI	BI-R
P7 (n=12)	.583, 4.55	.333, 4.25	.083, 4.25
P8 (n=7)	.143, 4.14	.429, 4.29	.429, 4.43
P9 (n=27)	.407, 4.48	.074, 3.81	.519, 4.52
Mean	**.378**, 4.39	.279, 4.12	.344, **4.40**

5.3 Results from Doctoral Students

Table 4 summarizes both the number of feedback items judged as most acceptable and the corresponding mean usefulness ratings for each doctoral participant.

Across all participants, the BI-R prompt showed the highest mean usefulness rating (4.40), closely followed by the Baseline (4.39). However, the Baseline achieved a slightly higher overall preference rate (.378 vs. .344). At the individual level, two participants (P8 and P9) rated the BI-R feedback highest, whereas one participant (P7) preferred the Baseline feedback.

5.4 Discussion

As shown in the results, doctoral students exhibited diverse preferences: some favored concise feedback generated with the Baseline prompt, while others valued the more respectful tone of BI-R-based feedback.

Compared with the sessions described in Section 4, the doctoral pre-defense reviews examined in this study included a larger proportion of clarification questions. Consequently, the Baseline prompt may have been preferred, as it provided concise, surface-level remarks based directly on what was explicitly said in the exchange, whereas the BDI/BI-R prompts emphasized deeper inference and explanation of the questioner's implicit intentions.

Furthermore, doctoral students are generally capable of inferring questioners' intentions on their own. Thus, the more elaborate interpretive reasoning generated by the BDI and BI-R prompts, though conceptually rich, may have appeared redundant or overly explicit to these advanced students.

When comparing the BDI and BI-R prompts, the inclusion of *Respect* in the latter appeared to improve both tone and pedagogical alignment of the generated feedback. A paired-samples *t*-test confirmed that the BI-R feedback was rated significantly higher in usefulness than the BDI feedback ($p = .005$). Specifically, the BI-R feedback was often perceived as more supportive and considerate, suggesting that explicit guidance toward respectful expression can improve perceived usefulness without compromising clarity. Overall, while deep mental-state reasoning may not always be necessary depending on the question type or student characteristics, feedback generation incorporating social sensitivity through *Respect* appears capable of producing feedback that is both educationally effective and contextually appropriate.

5.5 Comments for Feedback made with LLMs from Doctoral Students

All doctoral students rated feedback as useful or very useful. They commented that the feedback accurately captured the essence of the Q&A exchanges, provided valuable perspectives for improving the persuasiveness of answers, and helped them reflect on overlooked aspects of their research.

At the same time, several limitations were identified. Participants pointed out that the quality of transcription strongly influenced the accuracy of the feedback, noting that when presentations included many technical terms, the inaccuracy of Zoom's automatic captions often reduced precision. They also suggested incorporating non-verbal cues such as facial expressions and tone of voice into the interpretive process. In real academic discussions, some presenters adjust their responses by observing the interlocutor's facial expressions and tone to judge whether their explanations are on the right track. Enabling LLMs to capture and reflect such non-verbal signals could therefore make their feedback more adaptive and human-like.

Additionally, the supervisor of two doctoral participants reviewed the generated feedback and offered the following comments: "It's very convenient not to have to take notes manually, and the feedback itself is quite accurate. In the doctoral presentations, most students were able to respond appropriately to the questions, but the examples of clearer or more effective answers suggested by the system were particularly helpful. However, for undergraduate or master's students, such feedback might create pressure. They might feel they have to deliver answers at this level."

6 Summary and Future Directions

This study highlights both the promise and the challenges of using LLMs to generate feedback in academic Q&A sessions. While the BDI prompt alone showed no consistent advantage over the baseline, the BI-R prompt yielded clear improvements in both usefulness and acceptability, depending on the question type or student characteristics such as expertise and experience level. This indicates that incorporating an explicit dimension of Respect is not only ethically and stylistically valuable but also has measurable effects on how feedback is received by students.

At the same time, the nuanced and context-dependent nature of Respect calls for deliberate design choices and adaptive prompting strategies to ensure feedback that is both pedagogically sensitive and socially attuned.

Several avenues remain open for extending this work. First, the current study involved small-scale evaluations with native Japanese master's and doctoral students in engineering. Larger studies across diverse disciplines, cultures, and language backgrounds are needed to validate the generalizability of the BI-R framework. In future work, objective measures such as improvements in subsequent presentations or third-party assessments should be incorporated to complement the subjective, perception-based evaluations adopted in this study for practical deployment in educational settings.

Second, the method relied solely on textual transcripts generated by a general-purpose ASR system. Incorporating multimodal inputs such as prosody, gestures, or facial expressions could provide richer cues for mental-state inference and yield more contextually appropriate feedback.

Finally, ethical issues warrant attention. Automated feedback systems may produce inaccurate or overly directive guidance, potentially affecting students' self-assessment. Future work should establish safeguards and ensure alignment with pedagogical intent so that such systems complement rather than replace human supervision.

References

Mayuko Aiba, Daisuke Saito, and Nobuaki Minematsu. 2024. A ChatGPT-based Oral Q&A Practice System for First-time Student Participants in International Conferences. In *Proc. Interspeech 2024*, pages 5202–5203.

Mohammed Al Owayyed, Adarsh Denga, and Willem-Paul Brinkman. 2025. Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–10.

Todd J Allen and Atsushi Mizumoto. 2024. ChatGPT Over My Friends: Japanese English-as-a-Foreign-Language Learners' Preferences for Editing and Proofreading Strategies. *RELC Journal*, page 00336882241262533.

Michael Bratman. 1987. Intention, Plans, and Practical Reason.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of educational research*, 77(1):81–112.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. *IEEE Transactions on Learning Technologies*, 17:1445–1459.

Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. 2024. Integrating Large Language Model and Mental Model of Others: Studies on Dialogue Communication Based on Implicature. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pages 260–269.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*, 1(8):AIoa2400196.

Boago Okgetheng and Koichi Takeuchi. 2024. Modeling Score Estimation for Japanese Essays with Generative Pre-trained Transformers. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 64–73.

OpenAI. 2023. GPT-4 Technical Report. `https://arxiv.org/abs/2303.08774`.

Anand S Rao, Michael P Georgeff, et al. 1995. BDI Agents: From Theory to Practice. In *Icmas*, volume 95, pages 312–319.

Dongkwang Shin and Jang Ho Lee. 2023. Can ChatGPT make reading comprehension testing items on par with human experts?

Valerie J Shute. 2008. Focus on Formative Feedback. *Review of educational research*, 78(1):153–189.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91:101894.

David R. Traum and James F. Allen. 1994. Discourse Obligations in Dialogue Processing. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 1–8, USA. Association for Computational Linguistics.

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*, 35:24824–24837.

Minglin Wu, Jing Xu, Xixin Wu, and Helen Meng. 2024. Prompting Large Language Models with Mispronunciation Detection and Diagnosis Abilities. In *Proc. Interspeech 2024*, pages 2990–2994.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating Classroom Education with LLM-Empowered Agents. *arXiv preprint arXiv:2406.19226*.

Huihang Zhong, Yanlu Xie, and ZiJin Yao. 2024. Leveraging Large Language Models to Refine Automatic Feedback Generation at Articulatory Level in Computer Aided Pronunciation Training. In *Proc. Interspeech 2024*, pages 2600–2604.

A Prompt Designs

This appendix provides the exact prompt instructions used in the experiments for each condition: Baseline, BDI, and BI-R.

Each prompt was implemented in a multi-turn setting, where the model first received the presentation materials (slides and manuscript) and then the Q&A transcript for feedback generation. Although the original experiments were conducted in Japanese, the English versions of the prompts are presented here for clarity and reproducibility.

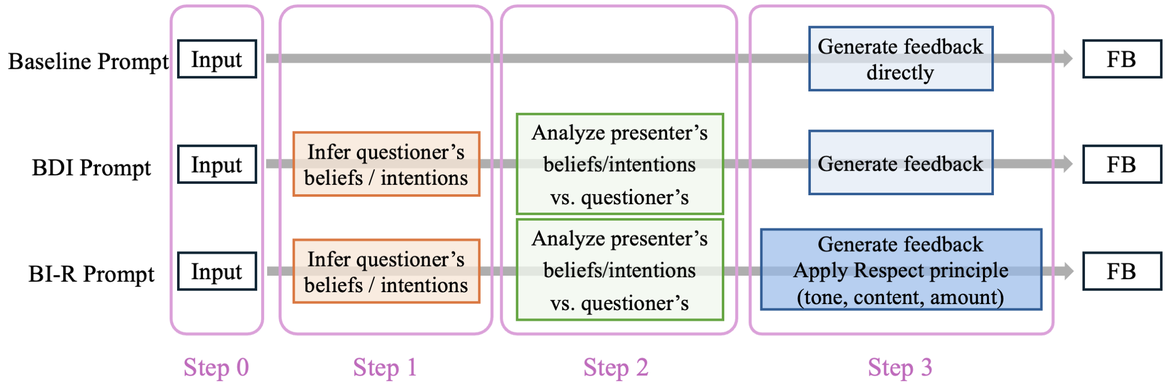

Figure 3: Overview of the prompt designs, corresponding to the stepwise descriptions (Steps 0–3) used in the main text. The boxed regions explicitly indicate which operations are performed at each step.

Figure 3 illustrates the overall prompt structure, where each boxed region corresponds to the stepwise descriptions (Steps 0–3) used throughout the paper.

A.1 Baseline Prompt

Step 0: Context Input

You are given the manuscript and slides used in a doctoral pre-defense presentation. Later, you will be asked to evaluate the Q&A session that followed the presentation. First, carefully read and understand the contents of the provided materials.
(Uploaded presentation materials here.)

Step 3: Feedback Generation

You are given a transcript of questions and answers exchanged during the Q&A session. Based on this log, generate feedback for each Q&A pair on the presenter's responses.
(Q&A transcript here.)

A.2 BDI Prompt

Step 0: Context Input

Humans have the following internal representations:

- **Beliefs:** A set of information representing their understanding of the world.
- **Desires:** The goals or states they wish to achieve.
- **Intentions:** The concrete plans or strategies they adopt to take action toward their goals.

You are given the manuscript and slides used in a doctoral pre-defense presentation. Later, you will be asked to evaluate the Q&A session that followed the presentation. First, carefully read and understand the contents of the provided materials.
(Uploaded presentation materials here.)

Step 1: Mental State Analysis (Questioner)

Next, analyze the following transcript of the Q&A exchange step by step. Do the following for each question: (1) infer the questioner's intention and report it; (2) based on that intention, infer the questioner's beliefs and report them.
(Q&A transcript here.)

Step 2: Mental State Analysis (Presenter)

Next, for each question, output the presenter's intention behind their response. Then, analyze and describe any misalignments or points of alignment between the questioner's and presenter's inferred beliefs and intentions.

Step 3: Feedback Generation

Finally, based on the above analysis, generate clear, constructive feedback for each Q&A pair on the presenter's responses.

A.3 BI-R Prompt

Step 0: Context Input

Humans have the following internal representations:

- **Beliefs:** A set of information representing their understanding of the world.
- **Desires:** The goals or states they wish to achieve.
- **Intentions:** The concrete plans or strategies they adopt to take action toward their goals.

In two-party communication such as Q&A sessions, mutual understanding of each other's beliefs and intentions constitutes a key aspect of "high-quality communication."

You are given the manuscript and slides used in a doctoral pre-defense presentation. Later, you will be asked to evaluate the Q&A session that followed the presentation. First, carefully read and understand the contents of the provided materials. (Uploaded presentation materials here.)

Step 1: Mental State Analysis (Questioner)

Next, analyze the following transcript of the Q&A exchange step by step. Do the following for each question: (1) infer the questioner's intention and report it; (2) based on that intention, infer the questioner's beliefs and report them.
(Q&A transcript here.)

Step 2: Mental State Analysis (Presenter)

Next, for each question, output the presenter's intention behind their response. Then, analyze and describe any misalignments or points of alignment between the questioner's and presenter's inferred beliefs and intentions.

Step 3: Respectful Feedback Generation

Finally, based on the above analysis, generate clear and constructive feedback for each Q&A pair to help the presenter achieve higher-quality communication. If a response is already of sufficient quality, feedback is not required. Adjust the **amount and content** of feedback according to the degree of misalignment identified. When discussing the questioner's mental state, avoid overly assertive expressions. Ensure that the feedback demonstrates **respect toward the presenter**, maintaining an appropriate tone and level of detail so that the feedback feels considerate and acceptable.

B Illustrative Example of Intermediate Inference and Resulting Feedback

To illustrate that our framework performs intermediate, stepwise inference rather than merely attaching labels, we provide one example Q&A exchange together with the corresponding outputs, including intermediate inference results and the resulting feedback, which were produced by the BI-R prompt described in Section 3.2. This illustrative example was generated using GPT-5.2, whereas the quantitative experiments reported in the main paper were conducted using GPT-4o.

The example is drawn from a pilot presentation on our preliminary study, which used the BDI framework to categorize question types. At that stage, the prompt design explicitly included a step that modeled the presenter's beliefs by incorporating additional contextual information written from a supervisor's perspective. Nevertheless, the underlying theoretical basis (mental-state modeling in dialogue) is consistent with the present paper.

It is provided solely for illustrative purposes and is not included in the quantitative results reported in this paper.

The original conversations were conducted in Japanese and are translated into English for presentation in this paper.

B.1 Example Q&A Excerpt (translated from Japanese)

Questioner: I found it interesting that linking internal mental states made the question content more convincing. One question: you seem to input beliefs and related information, but how much data do you input and in what format? More generally, what kinds of information, at what granularity, contribute to convincing outputs? Do you have any analysis or observations on that?
Presenter: Thank you for the question. From the perspective of automating part of a supervisor's role, we considered how much context should be provided. We prepared an appendix that summarizes the actual prompts. In the BDI prompt, we first explain the BDI model and the terminology such as "belief." As context, we input the student's manuscript, approximately eight pages in our setting, and the question text. We then infer the questioner's intention and beliefs, followed by the presenter's beliefs. For modeling the presenter's beliefs, in addition to the presentation materials, we also input a short summary of the research themes of the presenter's lab, taken from the lab website, as it reflects the supervisor-written description of current research focus. We compare the inferred beliefs of the questioner and presenter and output the final classification result and its rationale; for "developmental" questions, we also output possible future research directions.

B.2 Intermediate and Final Outputs

Below we show representative outputs for the intermediate inference steps as well as the resulting feedback. The content is lightly edited for readability and translated into English, but the structure (the questioner's intentions and beliefs $\rightarrow$ the presenter's intentions and beliefs $\rightarrow$ alignment/misalignment $\rightarrow$ feedback generation) follows the actual prompting procedure.

B.2.1 Step 1: Inferred Questioner Intention

The questioner's intention can be summarized into the following three main points.

1. From the perspective of practical deployment, the questioner seeks clarification on how information such as beliefs is designed and provided in the BDI prompt, specifically regarding the amount of information required and the format in which it is supplied.
2. While acknowledging the reported improvement in perceived plausibility, the questioner assumes that this effect is likely dependent on the type, granularity, and quantity of the input information. Accordingly, the questioner aims to elicit insights into which levels of granularity are particularly effective.
3. As a broader research concern related to reproducibility, generalizability, and applicability, the questioner also seeks to assess the methodological validity of the input design, including whether it may be arbitrary or rely on excessively rich contextual information.

In essence, the question reflects an interest in how mental-state information is operationalized in practice—namely, what kinds of information are incorporated, in what form and quantity, and why such design choices are justified.

Step 1: Inferred Questioner Beliefs

1. The estimation and presentation of mental states such as beliefs and intentions are highly sensitive to how input information is structured; vague or underspecified inputs are likely to result in unstable or unconvincing outputs.
2. The observed increase in plausibility is assumed to be attributable, at least in part, to the amount, granularity, and format of the contextual information provided. Without explicit discussion of these factors, it is difficult to evaluate which aspects of the method are truly effective.
3. For real-world deployment and extension to more complex settings (e.g., noisy ASR outputs, data from other labs, or different research domains), input design should not be treated as "the more, the better," but rather as a trade-off between cost (e.g., data collection and preprocessing) and effectiveness (e.g., plausibility and validity of the feedback).
4. Consequently, from a research standpoint, there is a need to systematically organize and articulate design principles regarding what information should be provided, in what format, and at what level of detail to achieve desirable outcomes.
5. Conversely, without such clarification, there remains a concern that the reported improvements may stem from overly enriched inputs and may not generalize beyond the current experimental setting, motivating the questioner's request for further justification.

Step 2a: Inferred Presenter Intention

The presenter's intention underlying the response can be summarized as follows.

1. In response to the questioner's abstract concerns regarding the amount, format, and granularity of input information, the presenter aims to provide concrete examples drawn from the actual experimental setup in order to make the design choices more tangible and convincing.
2. Rather than offering a purely theoretical or generalized explanation, the presenter seeks to ground the response in a fact-based account of how the prompts were designed and implemented in the current study, thereby demonstrating the feasibility and practical realism of the approach.
3. To address the implicit concern that the model's performance may rely on arbitrarily enriched inputs, the presenter emphasizes a clear design rationale, namely the intention to approximate the role typically played by an academic advisor.
4. By explaining the structure of the BDI prompt and the types of input information in a stepwise manner—such as the manuscript, the question text, and lab-level research summaries—the presenter aims to convey that the method is not a black box but follows a transparent and interpretable design.
5. At the same time, the presenter deliberately avoids an in-depth discussion of granularity optimization or generalized design guidelines, instead clarifying the current scope of the study and delineating which aspects are left for future investigation.

Step 2a: Inferred Presenter Beliefs

1. Feedback perceived as convincing does not necessarily require the accurate recovery of interlocutors' true mental states; rather, it emerges from presenting these states in a structured and plausible manner that supports coherent interpretation.
2. In actual supervisory practice, advisors typically evaluate students' work based on a limited but high-quality set of information sources, such as the full manuscript, the presentation content,

and an understanding of the lab's research direction.

3. Accordingly, contextual information at the level of lab website research summaries is considered sufficiently appropriate for approximating the presenter's beliefs within the proposed framework.
4. The effectiveness of incorporating beliefs and intentions as inputs depends more on the use of a structured framework (e.g., the BDI model) and on the reliability of information sources (e.g., texts authored by advisors) than on fine-grained decomposition or quantitative parameterization.
5. The primary objective of the current study is not to identify an optimal level of input granularity, but to demonstrate that explicitly modeling mental states in a BDI-inspired manner can improve the quality of generated feedback.
6. Therefore, the optimization of input quantity, format, and granularity is reasonably positioned as future work, and a comprehensive analysis of these factors is beyond the intended scope of the present study.

Step 2b: Alignment/Misalignment

Based on the inferred beliefs and intentions of both interlocutors, we analyze points of alignment and misalignment that characterize the exchange.

Alignment in Beliefs. Both the questioner and the presenter share the view that explicitly modeling mental states in a BDI-inspired manner is a meaningful and effective approach.

- The questioner considers it reasonable and intellectually appealing that linking beliefs and intentions contributes to more convincing question classification and feedback.
- The presenter believes that structuring mental states using the BDI framework enables the generation of more plausible and interpretable feedback.

Both parties also agree that the plausibility of the output depends on how contextual information is designed.

- The questioner emphasizes that plausibility is influenced by what information is provided, in what quantity, and at what level of granularity.
- The presenter emphasizes the importance of providing contextual information at a level comparable to that typically used by academic advisors.

Misalignment in Beliefs. The core divergence lies in expectations regarding the appropriate research scope at the current stage.

- The questioner expects some degree of organization regarding which levels of granularity are effective and how far the findings can be generalized, even at the present stage.
- The presenter prioritizes demonstrating the effectiveness of introducing a BDI-style structure itself, rather than optimizing or comparing input granularity in detail.

Another point of misalignment concerns the required strictness of belief inference.

- The questioner assumes that vague belief inference makes it difficult to explain why the output appears convincing.
- The presenter assumes that belief inference does not need to be strictly accurate, as long as the inferred mental states are structured in a plausible and coherent manner.

Alignment in Intentions. Despite these differences, both interlocutors share a constructive, improvement-oriented stance.

- The questioner seeks to clarify future research directions and design principles rather than to reject the approach.
- The presenter aims to explain the current design rationale and ensure transparency, rather than responding defensively.

Misalignment in Intentions. A final divergence emerges in the type of outcome each party expects from the exchange.

- The questioner aims to elicit generalized design insights and implications for future evaluation criteria.
- The presenter focuses on explaining what has been implemented in the current study and on clarifying the scope of the present contribution.

Overall, while the exchange is cooperative and grounded in shared methodological assumptions, these differences in expectations regarding scope and outcomes give rise to subtle but important misalignments.

Step 3: Feedback

Overall assessment (shared premises). As a starting point, this response can be positively evaluated in that it

- does not substantially misinterpret the questioner's intent,
- provides concrete explanations of the actual prompt design and input information, and
- offers a sincere and high-quality response that reduces the black-box nature of the research.

In this respect, the response successfully fulfills its role in the Q&A exchange.

Room for improvement (feedback based on misalignment of intentions). At the same time, the questioner may have been interested not only in *what was done in this study*, but also in broader design implications, such as

- which levels of input granularity are likely to be effective, and
- how such design choices could be organized or generalized in future work.

Addressing these points more explicitly could have aligned the response more closely with the questioner's underlying concerns.

Concrete suggestion. For example, the response could have been strengthened by adding a remark such as: "In this study, we prioritized approximating the role of a supervisor and therefore provided relatively rich contextual information, such as the full manuscript and summaries of the lab's research themes. However, it remains an open question which levels of granularity are truly essential, and whether similar levels of plausibility can be maintained with simplified inputs. We plan to investigate these issues through comparative experiments that progressively reduce the amount of input information in future work." By doing so,

- the current design choices would be justified,
- while the questioner's interests would be explicitly acknowledged as future research directions.

Such framing would further reinforce the impression that the response carefully digests the intent of the question and engages with it in a reflective and forward-looking manner.