Thesis

Explanation without Calibration: The Illusion of Understanding in AI-Supported Interview Preparation

Rafael Chrysanthou

UNIVERSITY OF CYPRUS

University of Cyprus
Department of Computer Science

May 2026

UNIVERSITY OF CYPRUS DEPARTMENT OF COMPUTER SCIENCE

Explanation without Calibration: The Illusion of Understanding in AI-Supported Interview Preparation

Rafael Chrysanthou

Supervisor

Dr. Vasos Vassiliou

The thesis was submitted in partial fulfillment of the requirements for the degree of Computer Science of the Department of Computer Science of the University of Cyprus.

May 2026

I declare that this dissertation and any accompanying software are my own original work, conducted in full compliance with the University of Cyprus Code of Conduct for Research (https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/), and with full accountability on my part for the accuracy, integrity, and validity of all content.

Acknowledgements

I would like to express my sincere gratitude to everyone who supported and guided me throughout the development of this thesis.

First and foremost, I would like to sincerely thank Dr. Argyris Constantinides for his continuous guidance, mentorship, and invaluable support throughout this project. His insight, feedback, and encouragement greatly influenced both the direction of this work and my academic growth.

I would also like to thank my supervisor, Dr. Vasos Vassiliou, for supporting the completion of the project within the academic framework. I sincerely appreciate his availability, and support throughout the development of this work.

Special thanks are also due to Dr. Marios Constantinides and Dr. Pooja Shikaripur Bheemasena Rao for their valuable advice, support, and contributions during different phases of the project.

Finally, I would like to thank my family for their continuous support, encouragement, and patience throughout this journey.

ABSTRACT

Artificial intelligence (AI) systems are increasingly used in interview preparation and other evaluative settings to provide automated feedback and performance assessment. However, while such systems may support reflection and learning, users may also misjudge their own performance after interacting with AI-generated feedback. This thesis investigates user calibration in AI-supported interview preparation, defined as the alignment between users' subjective self-assessments and objective performance evaluations. More specifically, the study examines how explanatory feedback and evaluative stress influence this alignment.

To investigate these effects, a web-based interview simulation platform was designed and implemented, integrating conversational interaction, automated evaluation, speech processing, and response-grounded explanatory feedback. A controlled between-subjects study (N = 87) examined the effects of explanatory feedback and evaluative stress on user calibration.

The findings show that evaluative stress significantly affected user calibration, leading participants to underestimate their interview performance more strongly under stressful conditions. In contrast, explanatory feedback alone did not significantly improve calibration, although it showed indications of partially attenuating the negative effects of stress. Additionally, calibration was more strongly associated with perceived usefulness and reliance on AI-generated feedback than with general trust in the system.

Overall, this thesis highlights calibration as an important dimension of human-AI interaction in evaluative contexts and contributes design implications for AI systems intended to support reflective self-assessment and feedback interpretation under varying cognitive conditions.

Keywords: Explainable AI, Human-AI Interaction, User Calibration, Interview Preparation, Evaluative Feedback

Chapter 1

Introduction

1.1 Introduction

AI systems increasingly shape how users prepare for high-stakes tasks such as job interviews. While these systems provide scalable and immediate feedback, they also introduce a critical challenge: users may misjudge how well they actually perform. In evaluative settings, inaccurate self-assessment can negatively affect learning, decision-making, and future performance, particularly when users rely on AI-generated feedback to evaluate their own abilities [1].

In this work, we examine this challenge through the lens of user calibration, defined as the alignment between users' subjective self-assessments and objective performance assessments [1, 2]. Prior research has shown that individuals frequently exhibit overconfidence or underconfidence, particularly in complex or uncertain tasks [3]. In AI-supported environments, this challenge becomes increasingly important, as users must interpret AI-generated evaluative feedback and incorporate it into their own judgments. Explainable AI (XAI) has been proposed to improve users' understanding of AI outputs and increase transparency in human–AI interaction [4]. However, prior work suggests that explanations do not always improve users' ability to independently evaluate AI-generated feedback and may instead increase confidence or perceived understanding without improving self-assessment accuracy [5, 6].

Beyond explainability, users do not interact with AI systems in a uniformly reflective manner. Prior work shows that users may over-rely on AI outputs through automation bias, while in other cases they may reject AI recommendations after observing errors, reflecting algorithmic aversion [7, 8]. These interaction patterns suggest that users may either defer to AI-generated feedback without sufficient critical evaluation or disregard it altogether depending on perceived

system reliability. Consequently, trust in AI systems may not necessarily correspond to improved evaluative judgment [9]. In evaluative contexts, such dynamics may distort how users assess their own performance, contributing to user miscalibration.

These challenges become particularly important stressful conditions. High-pressure situations such as interviews increase cognitive load and encourage more heuristic forms of decision-making [10], reducing reflective self-assessment and increasing reliance on external feedback. As a result, evaluative stress may influence not only task performance, but also how users interpret AI-generated feedback and assess their own performance, potentially increasing discrepancies between subjective and objective performance assessments [11, 12].

To investigate these challenges, this thesis studies AI-supported interview preparation as a high-stakes human-AI interaction setting in which users interpret system feedback, reflect on their performance, and assess their own interview abilities under different cognitive conditions. As illustrated in Figure 1, the study followed a controlled experimental methodology investigating user calibration under different feedback and stress conditions. The figure summarizes the overall study workflow, the four experimental conditions, the research questions guiding the investigation, and the main findings related to calibration, stress, trust, and perceived usefulness. Within this framework, a web-based interview simulation system was developed to systematically manipulate two factors: (i) the presence of explanatory feedback and (ii) the presence of evaluative stress during the interaction.

Although the system supported context-sensitive conversational behavior, interview flow, stage progression, and evaluation procedures remained backend-controlled to preserve experimental consistency across participants and conditions.

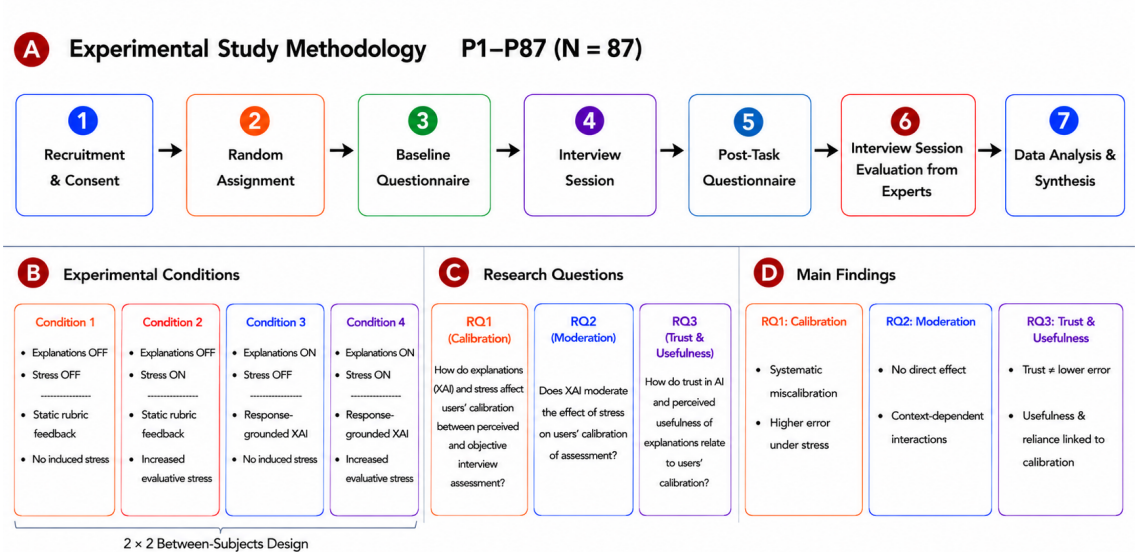

Figure 1: Overview of the study workflow, experimental conditions, research questions, and key findings.

1.2 Aims and Objectives

The main aim of this thesis is to investigate how explanatory feedback and evaluative stress influence users' ability to accurately assess their performance in AI-supported interview preparation. More specifically, the work examines calibration as a core dimension of human-AI interaction, focusing on how users interpret and utilize AI-generated evaluative feedback under conditions with and without evaluative stress.

To achieve this goal, the thesis pursues the following objectives:

- Design and implement an AI-supported interview preparation system integrating conversational interaction, automated evaluation, and explanatory feedback.
- Investigate how explanatory feedback influences users' calibration between subjective and objective performance assessments.
- Examine how evaluative stress affects users' self-assessment accuracy and interaction with AI-generated feedback.
- Analyze the relationships between calibration, trust, perceived usefulness, and reliance on AI feedback.
- Derive design implications for evaluative AI systems that support accurate and reflective self-assessment.

1.3 Research Questions

To structure the investigation, this thesis addresses the following research questions:

- **RQ1:** How do explanatory feedback and stress influence users' calibration between subjective and objective interview performance assessments?
- **RQ2:** To what extent does explanatory feedback moderate the effect of stress on users' calibration of interview performance assessments?
- **RQ3:** How do user perceptions (trust, perceived usefulness) and reliance on AI feedback relate to users' calibration of performance assessment?

1.4 Contribution

This thesis makes several contributions to Human-Computer Interaction (HCI) and AI-supported evaluative systems research.

First, the work introduces a controlled AI-supported interview preparation framework integrating conversational interaction, backend-guided interview orchestration, automated evaluation,

speech processing, and response-grounded explanatory feedback.

Second, the study provides empirical evidence that evaluative stress significantly impairs user calibration in AI-supported interview preparation, highlighting calibration as an important outcome of human-AI interaction.

Third, the findings demonstrate that explanatory feedback does not uniformly improve calibration, but instead exhibits context-dependent effects under stress conditions.

Finally, the thesis shows that calibration is associated more strongly with perceived usefulness and reliance on AI feedback than with general trust in the system, emphasizing the need to move beyond trust-centered evaluations of AI interaction.

1.5 Structure of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2 presents the theoretical background and related work on explainable AI, calibration, trust and reliance in AI systems, human–AI interaction, and AI-supported interview preparation.

Chapter 3 presents the architecture and interaction workflow of the proposed AI-supported interview preparation system. It describes the main system components, interview orchestration process, evaluation pipeline, stress manipulation, and the generation and presentation of explanatory feedback.

Chapter 4 presents the design and implementation of the platform, including the database architecture, interaction workflow, evaluation components, and experimental condition integration.

Chapter 5 presents the evaluation study and experimental findings. It describes the study design, participants, procedure, outcome measures, data analysis, and results regarding calibration, stress, explanatory feedback, trust, and reliance on AI feedback.

Finally, Chapter 6 summarizes the main findings of the thesis, discusses limitations, and outlines directions for future research.

1.6 Chapter Summary

This chapter introduced the research problem addressed in this thesis, focusing on user calibration in AI-supported interview preparation. It discussed how explanatory feedback, evaluative stress, trust, and reliance may influence users' self-assessment after interacting with AI-generated feedback. The chapter also presented the aims, research questions, contributions, and overall structure of the thesis.

Chapter 2

Related Work

2.1 Introduction

This chapter reviews prior research relevant to AI-supported interview preparation and user calibration in evaluative human-AI interaction settings. Existing literature suggests that AI-generated feedback can influence not only user performance, but also how users interpret, evaluate, and reflect on their own abilities [8, 9, 13]. At the same time, prior work highlights the importance of explainability, trust, reliance, and cognitive factors in shaping users' interactions with AI systems [4, 12, 14].

To better situate the present work, the chapter is organized into three broader research areas. First, we review AI-supported evaluative and educational systems, with a focus on interview preparation and automated feedback generation. Second, we examine research on explainable AI and human understanding, focusing on how explanations influence user interpretation and decision-making. Third, we discuss calibration, trust, reliance, and stress in human-AI interaction, emphasizing factors that affect users' ability to accurately assess their performance.

2.2 AI-supported Evaluative and Educational Systems

2.2.1 Structured Evaluation and Interview Preparation

Employment interviews are widely used to evaluate candidates, with prior work emphasizing the importance of structured formats, standardized questions, and consistent evaluation criteria for improving reliability and validity [15, 16]. Structured interviews are generally associated with reduced evaluator bias, improved comparability across candidates, and more reliable assessment outcomes compared to less standardized interview formats [16,17]. In these settings, predefined rubrics and scoring schemes are commonly used to support consistency and comparability across candidates and evaluators [18, 19].

Recent advances in artificial intelligence have introduced AI-supported systems for interview preparation and evaluation, enabling scalable simulation of interview scenarios and automated feedback generation [20,21]. Early conversational agents such as MACH demonstrated how AI systems can support interview training by providing users with real-time evaluative feedback on their responses [13]. Since then, AI-assisted interview platforms have increasingly incorporated automated scoring, feedback generation, and conversational interaction mechanisms [22].

These systems aim to provide scalable, accessible interview training while reducing the need for human supervision and evaluation. In addition to supporting repeated practice opportunities, AI-assisted interview platforms may encourage reflection through immediate evaluative feedback and conversational guidance [13,23]. However, the increasing use of automated evaluation mechanisms has also raised concerns regarding transparency, consistency, fairness, and the extent to which users critically engage with AI-generated assessments [5, 9, 21].

More recent conversational AI systems have additionally explored adaptive interaction behaviors, including dynamic follow-up questioning, conversational personalization, and context-sensitive interviewer behavior [13, 23, 24]. While adaptive interaction may improve conversational realism and user engagement [23], excessive variability in interviewer behavior may affect comparability across sessions and introduce inconsistencies in evaluative outcomes. These challenges highlight the need to balance adaptive conversational realism with standardized evaluation procedures in AI-supported interview environments.

2.2.2 AI-generated Feedback and User Reflection

Despite the scalability and accessibility benefits offered by AI-supported evaluative systems, concerns remain regarding the validity, fairness, and potential biases of automated evaluation mechanisms [21]. In many cases, AI-generated feedback is reduced to numerical scores, rubric-based evaluations, or brief textual suggestions, potentially oversimplifying complex human per-

formance and limiting deeper reflection [4, 9]. Although such forms of feedback may improve efficiency and scalability [20], they may provide limited support for deeper reflection and nuanced self-evaluation in evaluative settings.

Prior work further suggests that users may rely on automated outputs without critically engaging with them, particularly when feedback is perceived as authoritative or objective [9]. In such cases, users may accept automated evaluations with limited critical scrutiny, potentially reinforcing automation bias and reducing independent judgment [7, 8]. Recent research has additionally highlighted that AI-generated evaluations may influence not only users' task performance, but also their confidence, trust, and perceptions of their own abilities [5, 25].

Collectively, these findings suggest that AI-supported evaluative systems may influence not only task performance, but also how users form judgments about their own abilities and performance. These effects may become particularly important in evaluative contexts such as interview preparation, where users must simultaneously interpret automated evaluations and assess their own abilities [1, 2]. However, limited attention has been given to how different forms of AI-generated evaluative feedback, including explanatory feedback, influence user calibration in interview preparation contexts.

2.3 Explainability and Human Understanding in AI Systems

2.3.1 Explainable AI and Interpretability

Explainable AI seeks to make AI outputs more understandable by providing users with explanations for system decisions and recommendations [26, 27]. These explanations are commonly presented through justifications, feature-based feedback, evidence highlighting, example-based explanations, or structured reasoning intended to help users interpret AI outputs and understand how particular conclusions are reached [28, 29]. From a human-centered perspective, explanations are often designed to support users' sensemaking processes, improve transparency, and help users form more accurate mental models of AI system behavior [27, 30].

Prior work further suggests that interpretability plays an important role in shaping how users perceive, understand, and interact with AI systems [4, 26]. In many human-AI interaction settings, explanations are intended to help users engage more critically with AI-generated outputs and support informed decision-making [6, 30]. As a result, explanations are frequently associated with improved user understanding, interaction quality, and perceived transparency.

At the same time, the effectiveness of explanations may depend heavily on contextual and interaction-related factors, including explanation format, task complexity, user expertise, and the goals of the interaction [27, 29]. In evaluative settings, explanatory feedback may addition-

ally influence how users interpret assessments of their own performance and how they incorporate AI-generated evaluations into their self-assessment processes.

2.3.2 Limitations of Explanatory Feedback

However, growing evidence suggests that explanations do not always improve users' judgment, decision quality, or interaction outcomes [31]. Prior work shows that explanations can sometimes encourage over-reliance on automated systems, particularly when users perceive explanations as persuasive or authoritative rather than critically informative [9]. Similarly, explanations may increase users' perceived understanding of AI outputs without necessarily improving their ability to identify errors or independently evaluate output quality [27]. Recent work further suggests that explanations can influence users' perceived understanding, confidence, and trust in AI systems even when objective performance does not improve [5, 25].

In some cases, explanations may create an illusion of understanding, leading users to believe that they understand or evaluate AI-generated outputs more effectively than they actually do [6, 27]. Such effects may become particularly important in evaluative interaction settings, where users must simultaneously interpret AI-generated assessments and evaluate their performance. Under these conditions, persuasive or simplified explanations may unintentionally reduce critical reflection and encourage passive acceptance of automated evaluations [8, 9].

Together, these findings point to a potential disconnect between perceived understanding and accurate self-evaluation in explainable AI systems. Rather than uniformly improving user judgment, explanations may shape how users interpret, evaluate, and reflect on their performance in complex and context-dependent ways, particularly in evaluative human-AI interaction settings.

2.4 Calibration, Trust, and Cognitive Factors in Human-AI Interaction

2.4.1 User Calibration in Evaluative AI Systems

Calibration refers to the alignment between users' subjective self-assessments and objective performance assessments [1, 2]. In human-AI interaction, calibration plays an important role because it reflects users' ability to accurately evaluate their own performance relative to external assessment. Prior work shows that individuals frequently exhibit miscalibration, often in the form of overconfidence or underconfidence, particularly in cognitively demanding, uncertain, or feedback-driven tasks [3,32,33]. Such effects may become especially important in AI-supported environments, where automated feedback can shape how users interpret, evaluate, and reflect on their own performance.

In evaluative interaction settings, users must not only receive AI-generated feedback, but also

interpret and integrate it into their own judgments. As a result, calibration may be influenced not only by the accuracy of AI-generated evaluations, but also by how users perceive and incorporate those evaluations during self-assessment and decision-making processes. In interview preparation contexts, this challenge becomes particularly relevant because users are simultaneously required to interpret system feedback and assess their own interview abilities.

Together, these findings suggest that calibration represents a central challenge in evaluative human-AI interaction, particularly in contexts where users rely on AI-generated feedback to assess their own abilities and performance.

2.4.2 Trust and Reliance in Human-AI Interaction

A key factor shaping calibration in AI-supported environments is the distinction between trust and reliance. Trust reflects users' general attitudes toward an AI system, whereas reliance captures the extent to which users depend on AI outputs in practice [34]. Prior work further distinguishes trust from appropriate reliance, showing that trust in AI systems does not necessarily lead to effective decision-making or calibrated interaction behaviors [14, 34].

Similarly, users may rely on AI-generated outputs even when those outputs are incomplete, biased, or incorrectly interpreted [7, 8]. In some cases, overreliance on automated outputs may contribute to automation bias, reduced critical reflection, and diminished independent evaluation of system feedback [9]. Prior research additionally suggests that AI-generated feedback may influence not only task performance, but also users' confidence, perceived understanding, and judgments about their own abilities [5, 25].

These findings suggest that effective human-AI interaction depends not only on whether users trust AI systems, but also on whether users rely on automated feedback appropriately and critically during evaluative decision-making processes.

2.4.3 Stress and Cognitive Load in AI-assisted Tasks

Stress and cognitive load are important factors shaping judgment, decision-making, and self-evaluation processes. Prior research suggests that stressful conditions and increased cognitive pressure can reduce reflective reasoning and encourage greater reliance on heuristic processing, potentially leading to biased judgments and reduced calibration accuracy [10–12]. Cognitive load theory further suggests that individuals operating under high cognitive demands may have reduced capacity to critically process information and evaluate complex feedback [11, 33].

In AI-supported interactions, stressful conditions may additionally increase dependence on automated feedback while reducing users' ability to critically evaluate AI-generated outputs [9]. Under conditions of uncertainty, pressure, or limited cognitive resources, users may rely more

heavily on automated recommendations and less on independent reasoning or reflective self-assessment [7, 8]. These effects may become particularly important in evaluative interaction settings such as interview preparation, where users must simultaneously respond to performance pressure, interpret AI-generated evaluations, and assess their own abilities.

Taken together, prior work suggests that stressful and cognitively demanding conditions may substantially influence how users interpret and rely on AI-generated feedback, potentially shaping calibration and self-assessment accuracy in evaluative AI settings.

2.5 Research Gaps and Motivation

Despite growing interest in AI-supported interview preparation, explainable AI, and human-AI interaction, important gaps remain in understanding how users evaluate their own performance in AI-supported evaluative settings. Prior work has extensively examined explanations, trust, and automated feedback, yet these factors are often studied in isolation [23,26,27]. Most existing research has focused either on explanation design, trust in AI systems, or performance outcomes independently, rather than examining how these factors jointly influence users' self-assessment and calibration processes. Comparatively less attention has been given to how explanatory feedback influences users' calibration between subjective and objective performance assessments.

In addition, although stressful conditions and cognitive pressure are known to affect judgment and reflective reasoning [10, 12], their role in shaping interactions with AI-generated evaluative feedback remains relatively underexplored. Existing research has also largely overlooked how explanations and stress jointly influence user calibration in realistic feedback-driven contexts such as interview preparation. This limitation becomes particularly important in evaluative settings, where users must simultaneously interpret AI-generated assessments, reflect on their own performance, and form judgments about their abilities. Furthermore, relatively limited attention has been given to how adaptive conversational behavior and evaluative interaction design may shape users' interpretation of AI-generated feedback and self-assessment processes in controlled interview environments [23, 30].

Motivated by these gaps, the present work examines calibration as a central outcome of evaluative human-AI interaction in AI-supported interview preparation. Specifically, we investigate how explanatory feedback and stressful interaction conditions jointly influence users' ability to accurately assess their own performance. By examining calibration through the combined effects of explanations, evaluative stress, and AI-generated feedback, this work aims to contribute to a better understanding of how evaluative AI systems shape users' perceptions of their own abilities and performance.

2.6 Chapter Summary

This chapter reviewed prior research on AI-supported evaluative systems, explainable AI, and human-AI interaction. The discussion examined how AI-generated feedback, explanatory mechanisms, trust, reliance, and cognitive factors such as stress may shape users' interpretation of automated evaluations, self-assessment processes, and calibration accuracy. The chapter also identified important research gaps regarding calibration in AI-supported interview preparation, particularly under stressful evaluative conditions, thereby motivating the focus and research direction of the present study.

Chapter 3

Tools, Technologies and Architecture

3.1 Introduction

This chapter presents the tools, technologies, and system architecture of the proposed AI-supported interview preparation system. The technologies used for the development of the platform are first introduced, followed by an overview of the system architecture and the interaction between the frontend, backend, database, and AI components.

3.2 Technologies and Tools Used

For the design and implementation of the system, the following technologies and tools were used:

1. **Python:** Python was used as the primary programming language for backend development and AI integration due to its flexibility, readability, and extensive ecosystem of libraries.
2. **Visual Studio Code:** Visual Studio Code was used as the main development environment for code editing, debugging, and project management.
3. **Django:** The backend of the system was implemented using the Django web framework. Django was selected because it supports rapid development, modular design, and efficient database integration.
4. **Django REST Framework:** Django REST Framework (DRF) was used to support API-based communication between the frontend and backend components.
5. **PostgreSQL:** PostgreSQL was used as the relational database management system for storing user information, interview sessions, responses, evaluation results, and questionnaire data.

6. **Docker:** Docker was used to containerize the application and ensure consistent deployment across different environments.
7. **HTML, CSS, JavaScript, and Bootstrap 5:** These technologies were used to develop the frontend interface and support responsive user interaction.
8. **OpenAI API:** The OpenAI API was integrated into the system to generate context-sensitive interview questions, support controlled conversational interaction, and evaluate user responses using a large language model.
9. **faster-whisper:** The faster-whisper framework was used to perform speech-to-text transcription of users' voice responses.
10. **gTTS:** The gTTS library was used for text-to-speech generation, allowing interview questions to be delivered in audio format.
11. **JSON Web Tokens (JWT):** JWT authentication was used to support secure user authentication and session management.

3.3 System Architecture

The system follows a 3-tier client-server architecture consisting of the presentation tier, application tier, and data tier. This architecture was selected to support modularity, maintainability, and scalability. As illustrated in Figure 2, the frontend communicates with the Django backend through API requests, while AI and voice-processing components support transcription, evaluation, and audio generation functionalities.

Importantly, interview progression and interaction constraints were not controlled directly by the language model. Instead, the backend deterministically managed interview stages (introduction, motivation, experience, and closing stage), follow-up eligibility, condition-specific interaction behavior, and session progression rules. The language model operated within these externally enforced constraints, generating only localized conversational content for each interaction step. This separation between orchestration logic and language generation reduced uncontrolled variability across participants while preserving adaptive and context-sensitive interaction behavior.

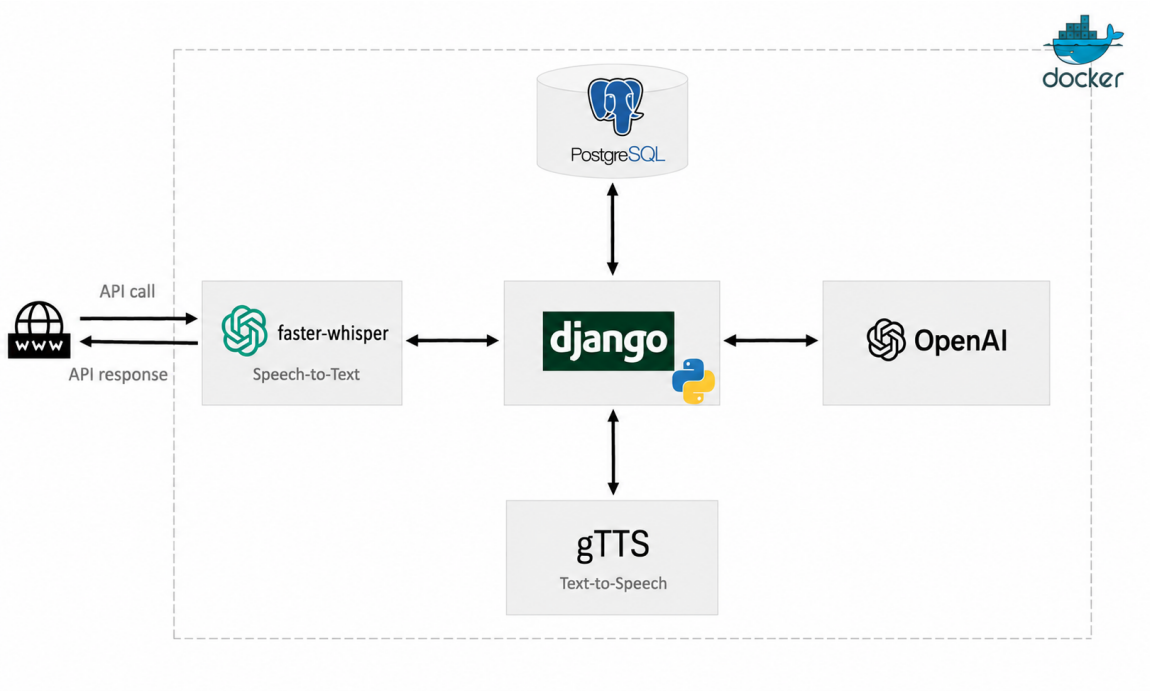

Figure 2: Overview of the proposed system architecture and interaction between the main system components.

Each tier is responsible for different functionalities within the platform:

1. **Presentation tier:** This layer represents the frontend interface through which users interact with the system. Users can upload their CV and job description, create interview sessions, answer interview questions, and view evaluation results and feedback. The interface follows a chat-based conversational interaction design supporting sequential interview progression, adaptive questioning behavior, and voice-based interaction.
2. **Application tier:** This layer contains the main backend orchestration logic implemented in Django. Beyond managing user authentication and API communication, the backend controls interview progression, session state management, stage transitions, and condition-specific interaction behavior. Interview questions and follow-up prompts are generated through a large language model accessed via the OpenAI API, while the backend externally enforces interaction constraints, including follow-up eligibility, stage persistence, and experimental condition rules. User responses are additionally processed through the speech-to-text pipeline and evaluated using the AI-based assessment framework before the generated feedback and evaluation results are returned to the frontend interface.
3. **Data tier:** This layer is responsible for persistent data storage. PostgreSQL was used to store users, uploaded CVs, job descriptions, interview sessions, question-answer pairs, evaluation scores, generated feedback, questionnaire responses, and additional interaction metadata such as audio-based emotion analysis outputs.

The frontend communicates with the backend through API requests, while the backend coordinates interactions between the AI services and the database. When a user submits a voice response, the audio is first transcribed into text using the speech-to-text component. The transcription is then evaluated through the AI pipeline, and the generated evaluation results and feedback are returned to the frontend interface.

The separation of the system into independent tiers improves maintainability and allows future extensions of the platform without affecting the overall architecture. Docker containerization additionally supports portability and consistent deployment across environments.

3.4 Chapter Summary

This chapter presented the tools, technologies, and overall architecture of the proposed AI-supported interview preparation system. The discussion described the main technologies used for backend development, frontend interaction, database management, AI integration, and speech processing functionalities. In addition, the chapter presented the system architecture and explained how the frontend, backend, database, and AI components interact to support interview orchestration, automated evaluation, explanatory feedback generation, and controlled conversational interaction across experimental conditions.

Chapter 4

Design and Implementation

4.1 Introduction

This chapter presents the design and implementation of the proposed AI-supported interview simulation platform. The system enables users to participate in structured mock interview sessions based on their uploaded CV and a selected job description while interacting with an AI interviewer through a conversational interface.

The platform was designed to support the experimental conditions of this study by enabling controlled manipulation of explanatory feedback and stress-related interaction behavior. More specifically, the system supports controlled conversational interaction, backend-guided interview orchestration, AI-based evaluation, and the collection of both objective and subjective assessment data across experimental conditions. These components were integrated to examine how explanatory feedback and evaluative stress influence users' calibration, trust, perceived usefulness, reliance on AI-generated feedback, and overall interaction behavior.

The following sections describe the system architecture, database structure, interaction workflow, and implementation decisions underlying the development of the platform. Particular em-

phasis is placed on the orchestration of interview stages, evaluation pipelines, speech-processing components, and the integration of AI-generated feedback within the experimental workflow. The complete prompt specifications, parameterization strategy, and condition-specific configurations are provided in Appendix I.2.

4.2 Database Design

The system is supported by a structured relational database designed to manage user information, interview sessions, interaction history, and evaluation data. The database architecture enables efficient storage and retrieval of both static user information and dynamic data generated during interview simulations. An overview of the database architecture is illustrated in Figure 3.

User entity, which represents each participant. Users can upload their CVs and corresponding job descriptions, which are stored as separate entities and later used to generate personalized interview questions. The CV and JobDescription entities store uploaded documents together with their extracted textual content. This extracted information is subsequently used as contextual input for the large language model (LLM) during interview question generation.

At the core of the interaction workflow is the Session entity, which represents a complete interview instance associated with a user, CV, and job description. In addition, sessions store configuration parameters related to the experimental conditions, including the interview mode (e.g., normal or pressure condition) and whether explanatory feedback is enabled. Aggregate evaluation metrics, such as clarity, structure, depth, confidence, relevance, and overall score, are maintained at the session level to support analysis of interview performance across conditions.

Within each session, the interaction is modeled as a sequence of Question-Answer (QA) pairs. Each QA entity stores the generated interview question, the user's response, and detailed evaluation results. These evaluation results include numerical assessment scores and structured feedback components, such as strengths, areas for improvement, missing points, highlighted excerpts, and example improved answers. This structure allows the system to maintain detailed records of the interaction process while supporting consistent AI-based evaluation across all interview sessions.

To further support analysis of interaction behavior during interview sessions, the system incorporated an AudioEmotion entity storing lightweight emotion-related metadata derived from users' voice responses, including inferred emotional cues, intensity estimates, confidence scores, and timestamps. These signals were not used for automated evaluation or scoring, but instead served as auxiliary contextual information for limited conversational adaptation during interview question generation and interaction flow. More specifically, affective cues could influence localized interviewer tone and probing behavior while preserving the predefined interview structure and

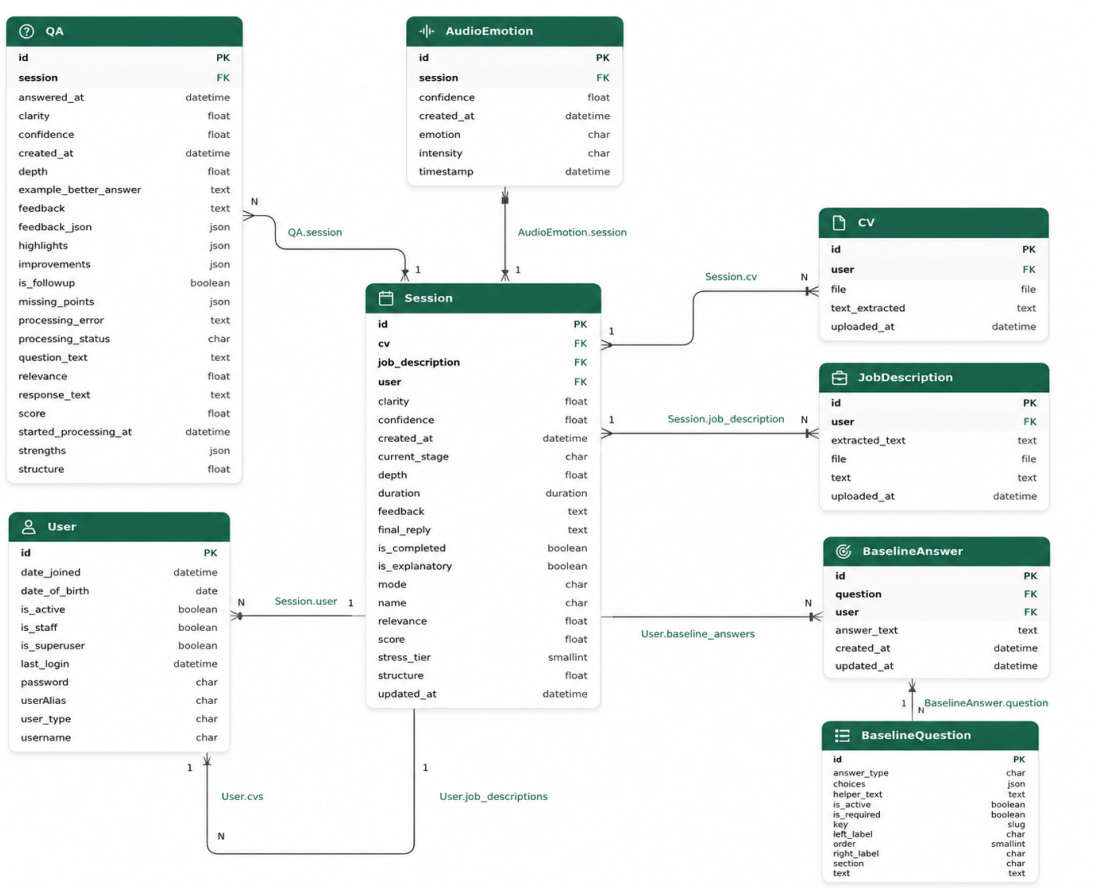

Figure 3: Overview of the database architecture and relationships between the main system entities.

experimental constraints. The collected metadata may additionally support future multimodal analysis of stress and interaction behavior.

In addition to interview interaction data, the system also includes BaselineQuestion and BaselineAnswer entities. These entities are used to store users' responses to post-session questionnaires and support the collection of subjective self-assessment data. The baseline questions capture aspects such as perceived performance, confidence, stress, trust, and overall user experience, while the corresponding answers provide insight into users' self-evaluation after completing an interview session.

Overall, the database design supports both the functional requirements of the platform and the experimental requirements of the study, enabling detailed tracking of user interactions, evaluation outcomes, and calibration-related measures.

4.3 Interaction Flow

This section describes the interaction flow of the system, outlining the main steps users follow during their interaction with the platform. The workflow includes account access, document upload, session configuration, interview interaction, and post-session evaluation. The interaction process was designed to follow a structured and consistent workflow, supporting a realistic interview experience while maintaining the experimental conditions required for the study.

4.3.1 User Registration and Login

Users begin their interaction with the platform by either creating a new account or logging into an existing one. The registration process requires users to provide basic information, including a username, user alias, date of birth, and password. Password creation follows predefined security requirements, such as minimum length constraints and the inclusion of alphanumeric characters.

As illustrated in Figure 4, the registration interface was designed to be simple and user-friendly, guiding users through the required fields while clearly presenting validation constraints. This design supports accessibility and minimizes friction during the onboarding process. In addition, the final two characters of the user alias were used internally by the system to assign participants to one of the four experimental conditions of the study. This mechanism supported controlled distribution of participants across conditions while preserving a consistent registration workflow from the user's perspective.

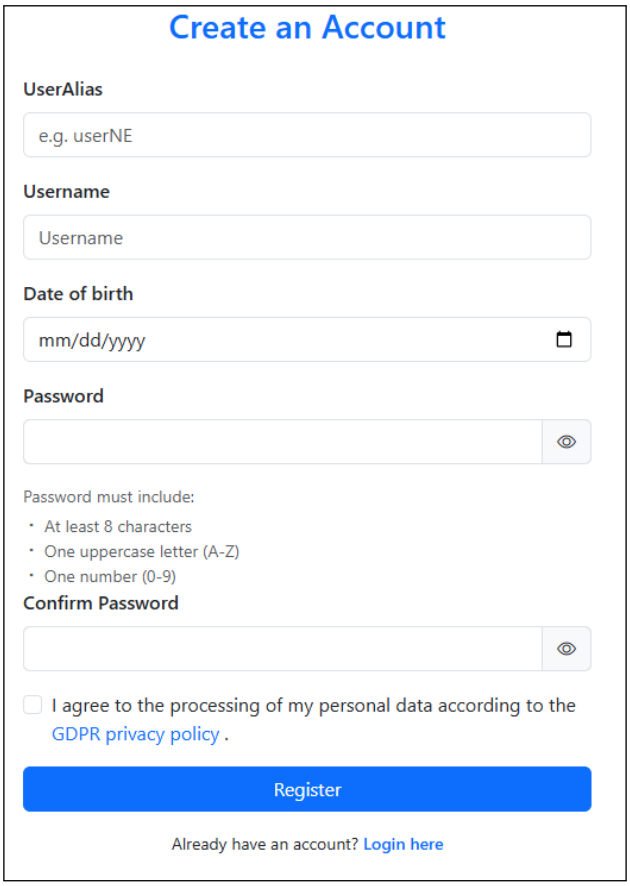

Figure 4: User registration interface.

After successful registration, users can log into the system using their credentials. The login interface, shown in Figure 5, follows a minimal and straightforward design, allowing users to quickly access the platform and begin interacting with the system.

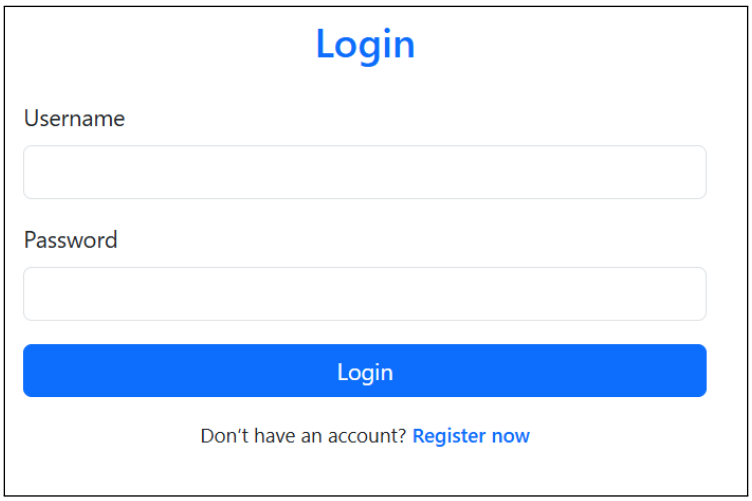

Figure 5: User login interface.

4.3.2 Dashboard and Session Management

After logging into the system, users are directed to the main dashboard, which serves as the central hub of the platform. Through this interface, users can manage interview sessions, access uploaded documents, review previous sessions, and navigate to additional functionalities.

As shown in Figure 6, the dashboard presents a simple and organized interface that highlights the option to create a new interview session. If no sessions have been created yet, the platform provides a prompt encouraging users to begin their first session. This design helps users quickly understand the main functionality of the system while maintaining a clear and intuitive navigation structure.

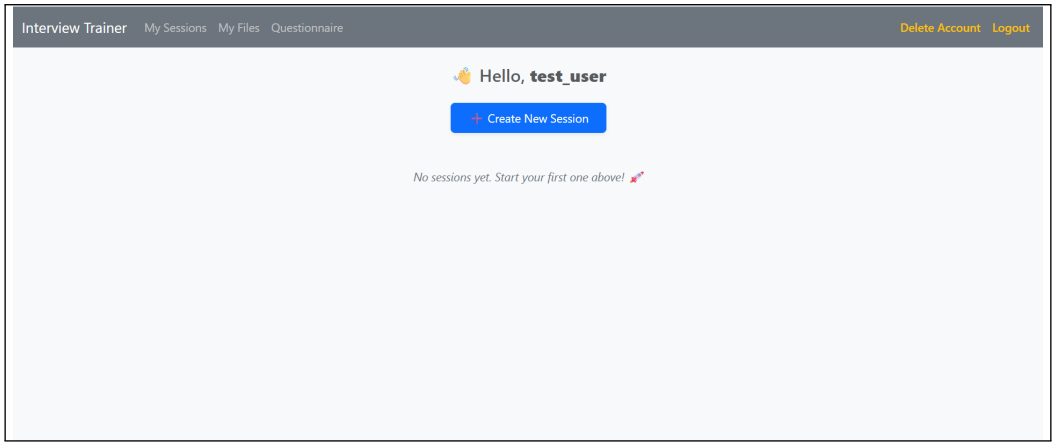

Figure 6: Main dashboard interface after user login.

From the dashboard, users can initiate a new session, review previously completed sessions, access uploaded CVs and job descriptions, or navigate to the post-session questionnaire. Centralizing these functionalities within a single interface improves usability and supports a consistent interaction flow throughout the platform.

4.3.3 Document Upload Management

Before creating a new interview session, users must upload the documents required for interview personalization. The platform provides a dedicated upload management page where users can upload and manage their CVs and job descriptions. These documents form the basis for generating personalized interview questions during the session creation process.

As shown in Figure 7, the interface is divided into two main sections: one for CV uploads and one for job description uploads. Both sections support PDF files and include upload controls, file validation information, and options for deleting previously uploaded documents. This design allows users to clearly distinguish between personal profile information and target job information while maintaining a simple and organized interaction flow.

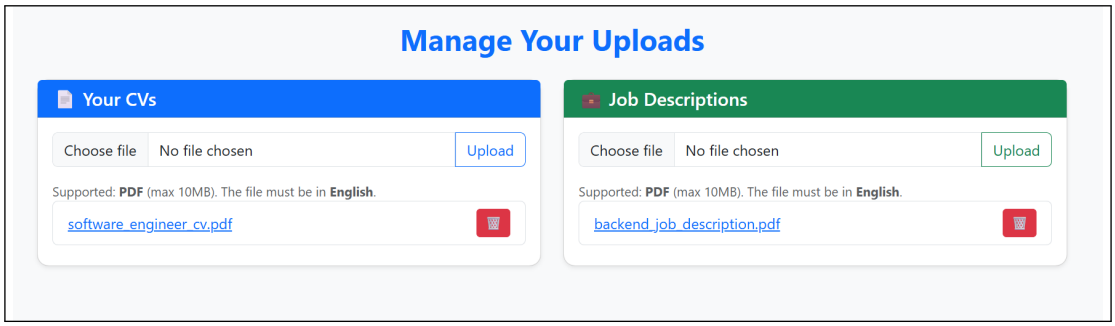

Figure 7: Document upload mechanism for CVs and job descriptions.

The uploaded CV and job description are stored in the database and later made available during session creation. Separating document upload from session creation allows users to reuse previously uploaded files across multiple interview sessions, improving flexibility and reducing repetitive input during interaction with the platform.

4.3.4 Session Creation

After uploading the required documents, users can initiate a new interview session through the dashboard by selecting the session creation option. This process allows users to configure the interview before beginning the interaction with the AI interviewer.

As illustrated in Figure 8, users are required to provide a session name and select a previously uploaded CV and job description. These selections determine the context of the interview and allow the system to generate personalized interview questions tailored to the user's profile and target position.

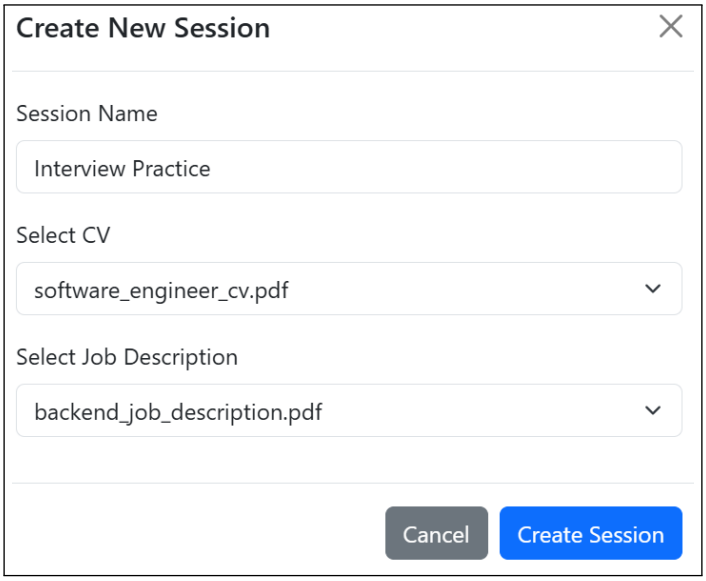

Figure 8: Session creation interface.

In addition to document selection, each session is associated with specific interaction conditions that influence the interview experience. These include the presence or absence of explanatory feedback as well as the interaction mode related to stress-inducing interviewer behavior. Although these parameters are configured internally by the system, they directly affect the interaction flow and feedback presentation experienced by users during the interview session.

Overall, the session creation process establishes the foundation for the interview interaction by linking user-provided information with the experimental configuration of the system. This ensures that each session remains both personalized and aligned with the study design.

4.3.5 Session Overview Interface

After creating a session, users are redirected to the session overview interface, which provides a summary of the current interview state before continuing the interaction. The interface presents general session information, including interview duration, overall evaluation score, session status, and completion progress.

As illustrated in Figure 9, the interface additionally provides expandable sections containing detailed evaluation information, generated feedback, and linked session documents. This design allows users to monitor their interview progress and review session-related information through a centralized interface.

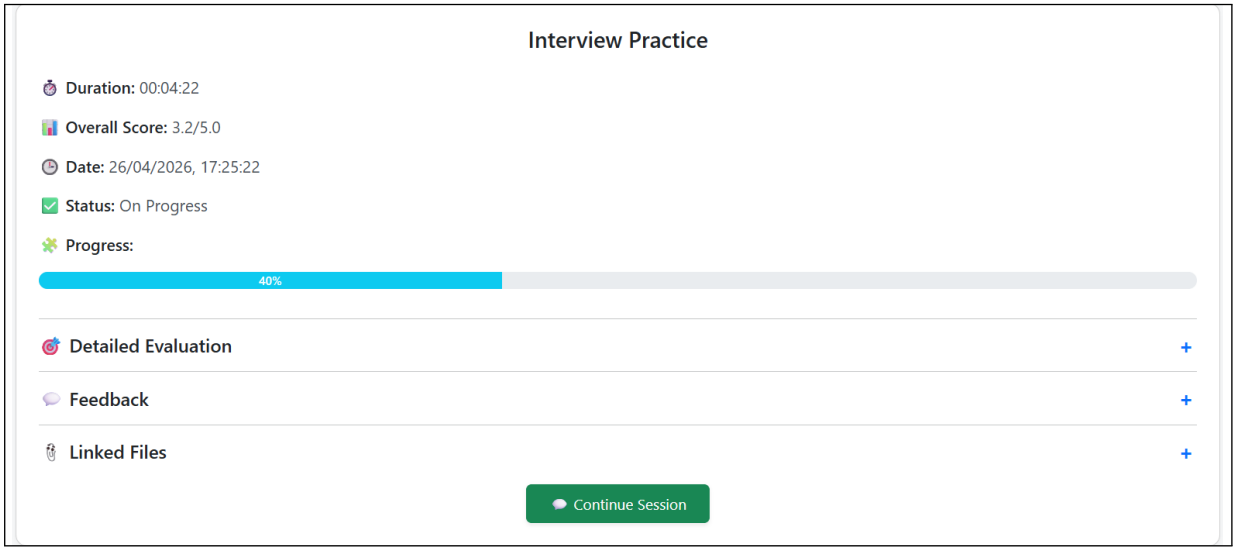

Figure 9: Session overview interface displaying interview progress, evaluation information, and session-related details.

The progress bar visually indicates interview completion status, while the overall score provides users with a cumulative assessment of their performance during the session. Through this interface, users can continue an active interview session and access additional evaluation information

generated throughout the interaction process.

As illustrated in Figure 10, the session overview interface additionally provides aggregated evaluation summaries and overall interview feedback. The platform presents cumulative assessment scores across predefined evaluation dimensions, including clarity, structure, depth, confidence, and relevance.

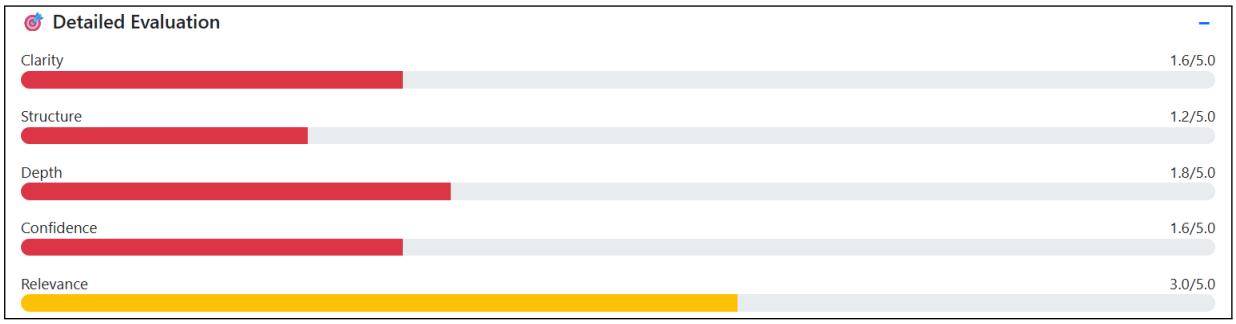

Figure 10: Aggregated session-level evaluation scores across predefined assessment dimensions.

In addition to numerical assessment summaries, the system generates structured session-level feedback describing strengths, weaknesses, and areas for improvement throughout the interview session. As shown in Figure 11, this feedback summarizes interaction patterns and provides personalized recommendations intended to support reflection and interview preparation.

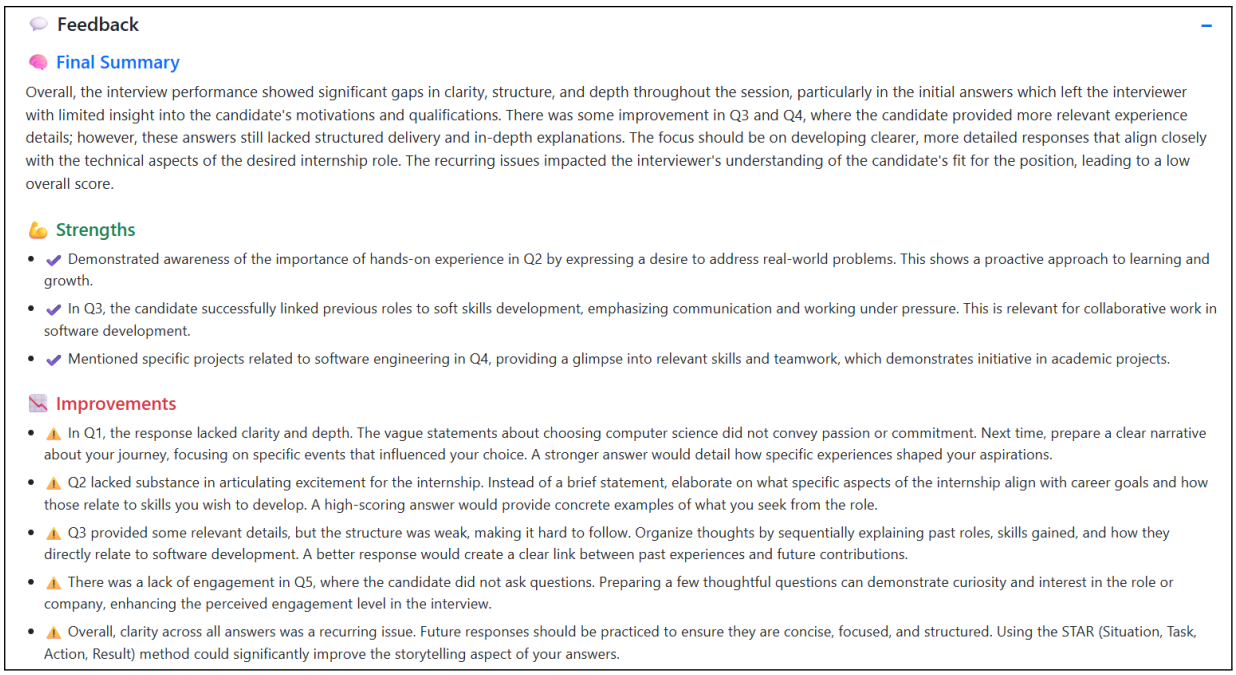

Figure 11: Session-level feedback summarizing interview strengths, weaknesses, and improvement suggestions.

These session-level summaries allow users to review their overall interview performance beyond individual question evaluations, supporting broader reflection on communication quality, answer structure, and preparedness throughout the interview interaction.

4.3.6 Interview Interaction Interface

After starting or continuing a session, users are redirected to the main interview interaction interface, where they communicate with the AI interviewer through a conversational chat-based environment. The interface was designed to simulate a realistic interview experience while supporting a voice-based interaction.

As illustrated in Figure 12, each interview question is presented sequentially together with an audio playback option, allowing users to hear the generated question through the text-to-speech component. Users can respond verbally through the integrated voice recording functionality, while their responses are displayed within the conversational interface after transcription. Although interview questions were dynamically generated, interview progression remained backend-controlled throughout the interaction. The system externally enforced interview stages, follow-up eligibility, and condition-specific interaction constraints, while the language model generated localized conversational content within these predefined boundaries.

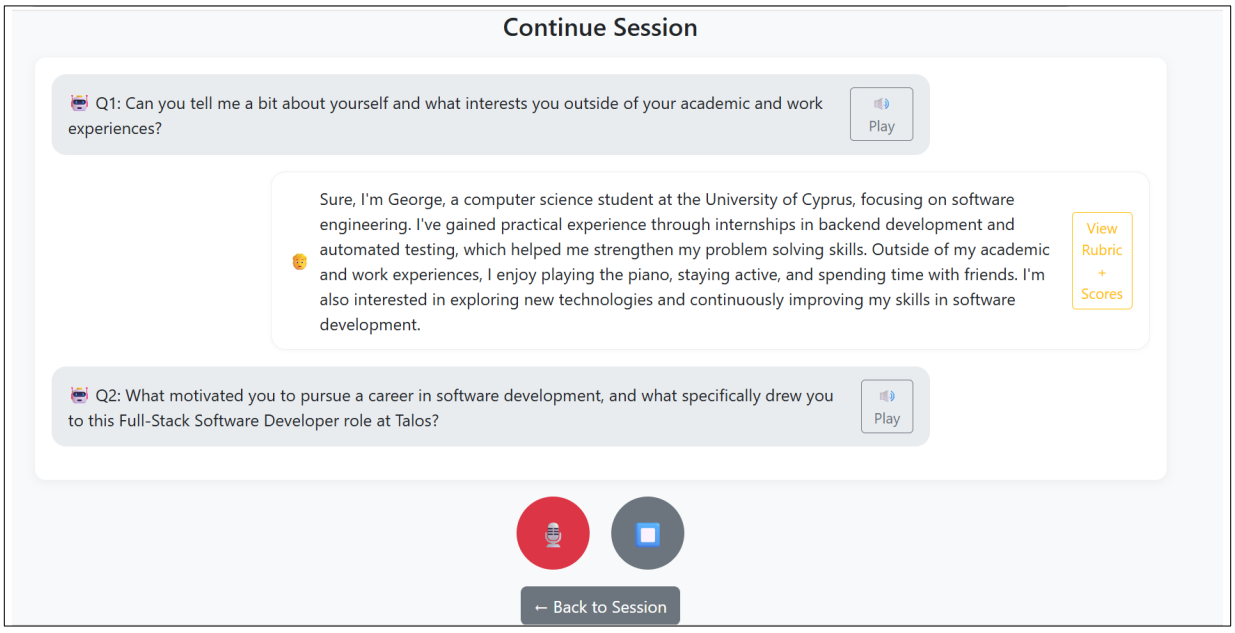

Figure 12: Main interview interaction interface during an active interview session.

The interface additionally provides access to automated evaluation results through a dedicated evaluation button associated with each response. This interaction flow allows users to review feedback and scores while maintaining a structured conversational interview experience.

4.3.7 Evaluation Interface

After each response submission, the system generates automated evaluation results based on predefined assessment dimensions. These results are presented through a dedicated evaluation interface, illustrated in Figure 13.

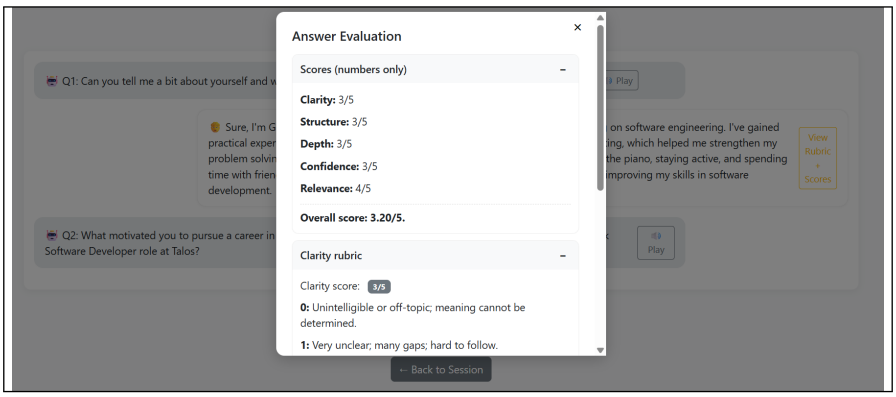

Figure 13: Evaluation interface presenting rubric-based scores and automated assessment results.

The evaluation interface displays assessment scores across multiple dimensions, including clarity, structure, depth, confidence, and relevance, together with an overall performance score. Depending on the experimental condition, the interface may additionally present explanatory evaluative feedback grounded in specific excerpts of the user's response.

This design enables users to review both quantitative assessment results and qualitative feedback while maintaining consistency across interview sessions and experimental conditions.

4.3.8 Experimental Condition Differences

The platform was designed to support different experimental conditions through controlled variations in explanatory feedback and interviewer behavior. Specifically, the study manipulated two main factors: the presence of explanatory feedback and the level of interview stress. Figure 14 presents examples from the interview interaction interface, while Figure 15 shows the corresponding differences in evaluation feedback.

In the explanatory feedback condition, the system provided response-grounded evaluative support linked to specific excerpts of the user's answer. As shown in Figure 14A, relevant parts of the response were highlighted to indicate the evidence used by the system during evaluation. This design was informed by prior work on explainable AI and instance-level explanations in human-AI interaction [26, 27, 29].

The platform additionally supported stress-inducing interview conditions through controlled variations in interviewer behavior. In the Stress OFF condition, participants interacted with a

neutral interview process consisting of predefined interview questions without additional probing or evaluative pressure. In contrast, the Stress ON condition introduced a more demanding interaction style involving stricter interviewer behavior and probing follow-up questions for incomplete or vague responses. As illustrated in Figure 14B, the system generated additional clarification questions intended to increase evaluative pressure and cognitive load during the interview interaction [11, 33].

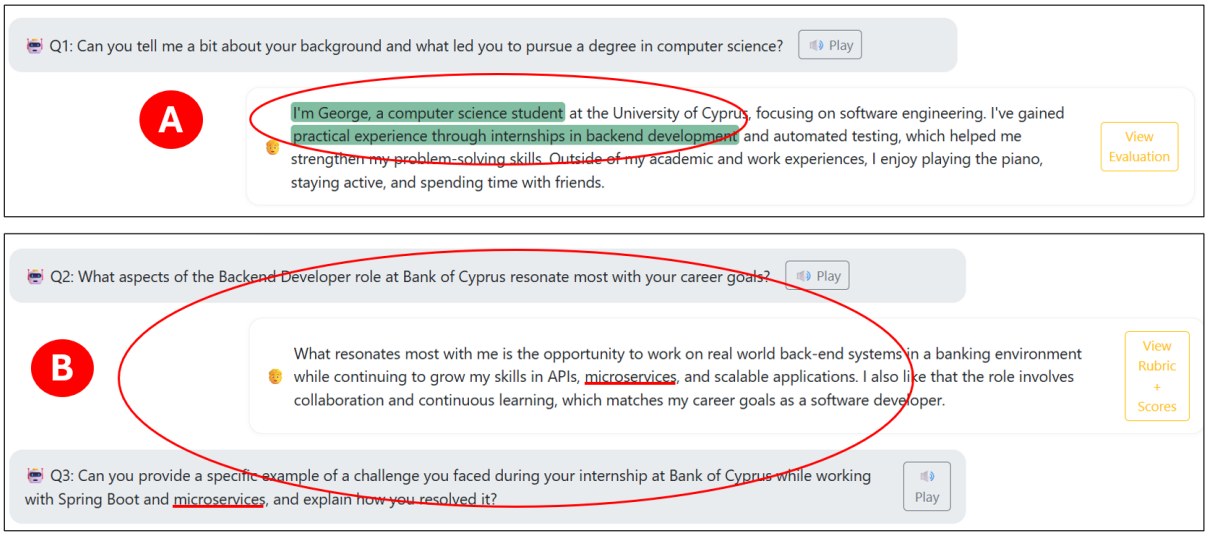

Figure 14: Representative interaction differences across conditions: Explanation ON highlighted response-grounded evidence excerpts (A), while Stress ON introduced probing follow-up questions (B).

The feedback interface also differed depending on whether explanatory feedback was enabled. In the Explanation ON condition, the evaluation interface presented detailed rubric-based feedback describing strengths, weaknesses, and areas for improvement across multiple evaluation dimensions. As shown in Figure 15C, this feedback was connected to the user's response and provided qualitative explanations in addition to scores.

In contrast, the Explanation OFF condition provided numerical assessment scores together with static rubric descriptions, without personalized response-grounded explanations. As illustrated in Figure 15D, users received quantitative evaluation results while explanatory evaluative content was omitted. This condition was designed to preserve a consistent evaluation structure while limiting interpretability-related support during feedback presentation.

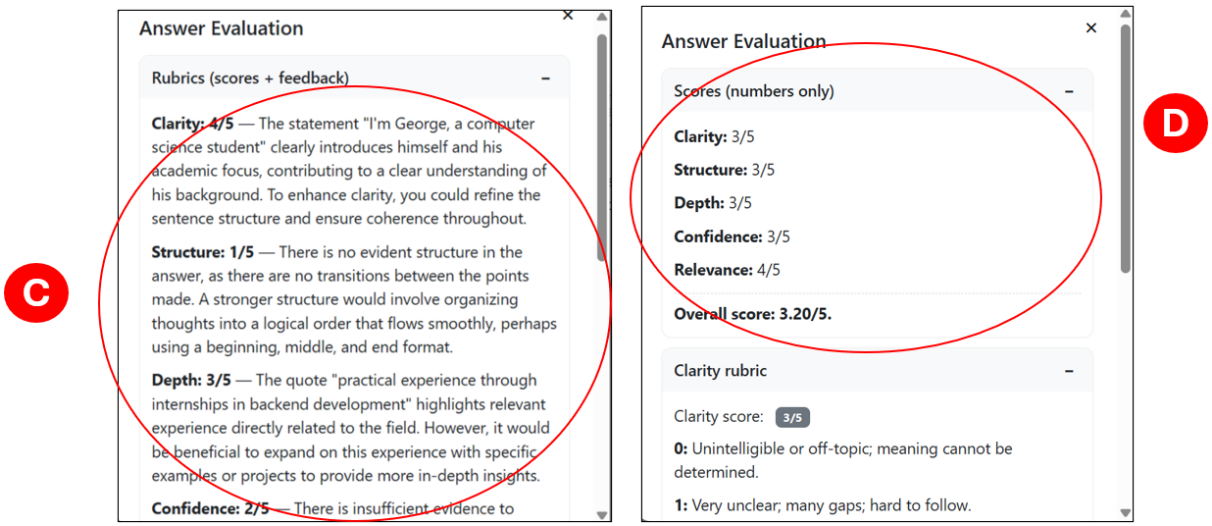

Figure 15: Comparison of evaluation feedback across explanation conditions: Explanation ON provided response-grounded feedback (C), whereas Explanation OFF provided numerical scores and static rubric references (D).

The experimental conditions were implemented through parameterized prompt configurations while preserving a consistent interview structure across participants. For reproducibility purposes, the complete interview, evaluation, and feedback prompts, together with the parameterization strategy used across experimental conditions, are provided in Appendix I.2.

Overall, these controlled differences enabled a systematic investigation of how explanatory evaluative feedback and interview stress influence users' interpretation of automated assessment results and their subjective self-assessment during AI-supported interview preparation.

4.3.9 Post-session Questionnaire

After completing the interview session, users are redirected to the post-session questionnaire. This questionnaire was designed to collect subjective self-assessment data and user perceptions after interacting with the AI-supported interview system.

As shown in Figure 16, the questionnaire interface presents items grouped into different measurement categories using a Likert-scale format ranging from strong disagreement to strong agreement. The questionnaire incorporated multiple validated measurement scales related to perceived competence, effort and importance, pressure and tension, perceived usefulness, trust in AI, reliance on AI feedback, and overall user experience [35–37].

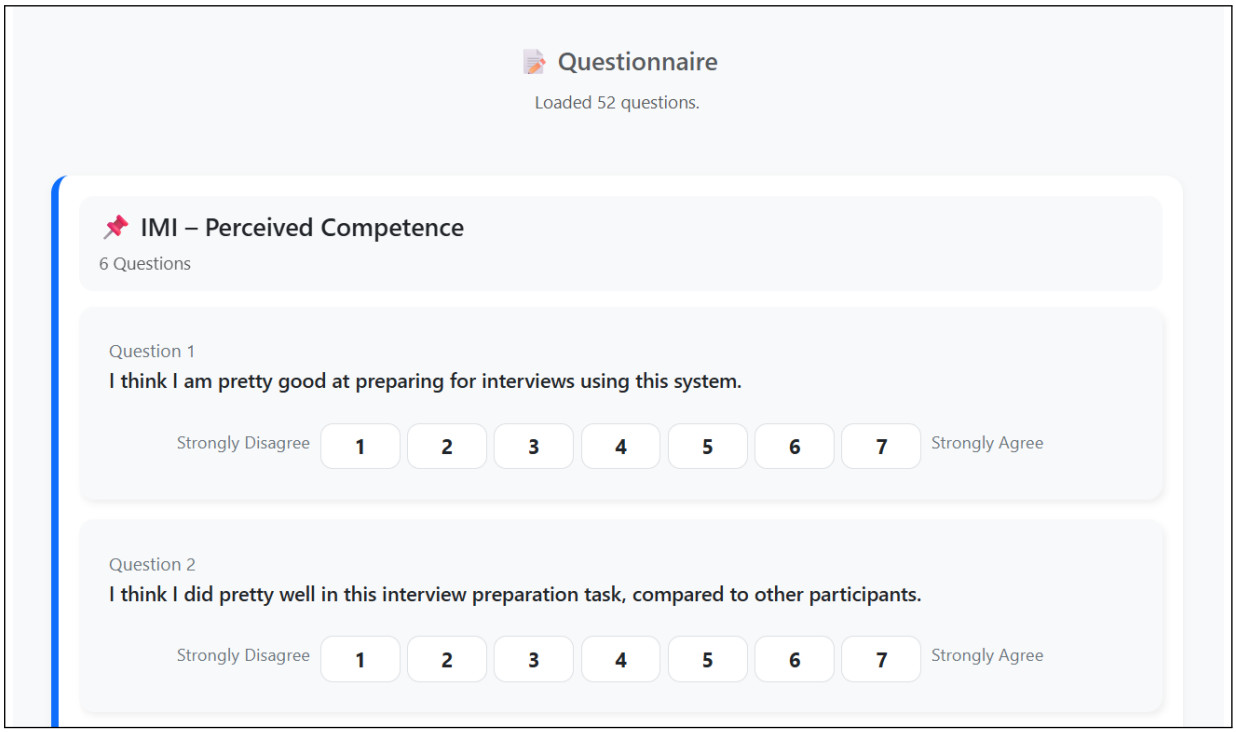

Figure 16: Post-session questionnaire interface used to collect subjective self-assessment and user perception measures.

More specifically, the questionnaire included adapted subscales from the Intrinsic Motivation Inventory (IMI) to assess perceived competence, effort, pressure, and perceived value of the interview preparation process [35]. In addition, the system collected measures related to trust in AI-generated feedback [14, 36], reliance on automated evaluation [9, 34], and overall user experience through the UEQ-S questionnaire [37]. Open-ended questions were also included to capture qualitative reflections regarding the perceived reliability and usefulness of the AI-generated feedback.

The questionnaire responses were stored in the database and later used for experimental analysis. These data enabled comparison between users' subjective self-assessment and objective assessment, supporting the analysis of calibration across the different experimental conditions [1, 2].

4.4 Chapter Summary

This chapter presented the design and implementation of the proposed AI-supported interview preparation platform. The discussion described the database architecture, interaction workflow, interview orchestration process, evaluation pipeline, and speech-processing integration supporting the experimental study. In addition, the chapter presented the main user interfaces and explained how the system implemented controlled variations in explanatory feedback and evalua-

tive stress across experimental conditions. Overall, the implemented platform enabled structured conversational interview interaction, automated evaluation, and the collection of both objective and subjective assessment data required for the investigation of user calibration in AI-supported interview preparation.

Chapter 5

Evaluation

5.1 Introduction

This chapter presents the evaluation of the AI-supported interview preparation system developed in this thesis. The evaluation focused on how users assess their own interview performance after interacting with AI-generated feedback, and how this process is shaped by explanatory feedback and evaluative stress.

The study examined user calibration, defined as the alignment between participants' subjective self-assessments and objective performance assessments [1, 2]. In this context, calibration captures whether users accurately judge their interview performance after receiving feedback from the system. This is particularly important in AI-supported evaluative settings, where users may trust, rely on, or question automated feedback when forming judgments about their abilities.

To investigate these effects, a controlled user study was conducted using a 2×2 between-subjects design. The study manipulated two factors: explanatory feedback and stress. Explanatory feedback was either enabled or disabled, while the interview interaction was either neutral or stress-inducing. The main goal was to examine whether explanations improve calibration, whether

stress disrupts accurate self-assessment, and how user perceptions such as trust, perceived usefulness, and reliance relate to calibration outcomes.

5.2 Study

5.2.1 Study Design

The evaluation was conducted as a between-subjects experiment using a 2×2 factorial design. The two independent variables were explanatory feedback and stress. Explanatory feedback had two levels: Explanation ON and Explanation OFF. Stress also had two levels: Stress ON and Stress OFF. This design resulted in four experimental conditions:

- **Explanation OFF + Stress OFF**
- **Explanation OFF + Stress ON**
- **Explanation ON + Stress OFF**
- **Explanation ON + Stress ON**

Across all conditions, participants interacted with the same web-based interview preparation system. The interface structure, interview workflow, evaluation pipeline, and core task remained consistent across groups. This ensured that differences in participant responses and calibration outcomes could be attributed primarily to the experimental manipulations rather than to changes in the interface or procedure. Figure 17 provides an overview of the study procedure, illustrating the main stages followed by participants.

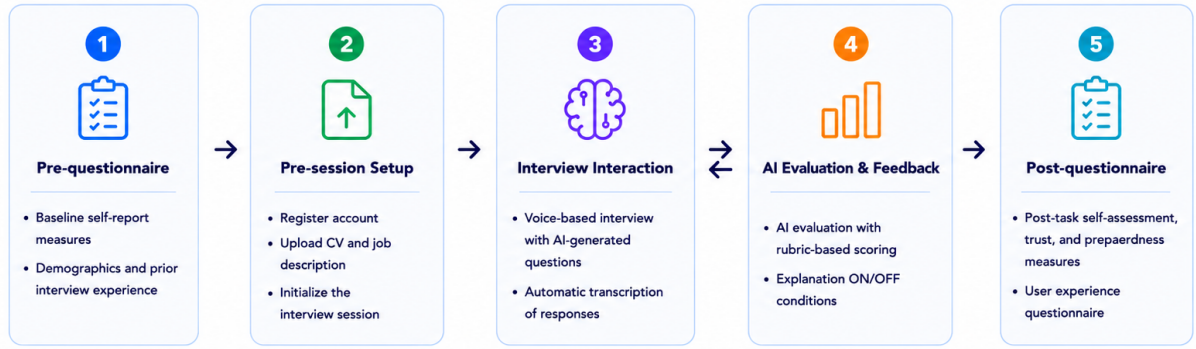

Figure 17: Overview of the experimental workflow

5.2.2 Experimental Tasks

At the beginning of the session, participants uploaded a CV and a target job description in PDF format. To ensure compatibility and consistency across sessions, uploaded documents were required to be written in English and smaller than 10MB. The system automatically processed the

uploaded documents using AI-based extraction and structured summarization to identify relevant information regarding the participant's background and the target role, including skills, experience, education, qualifications, responsibilities, and required technologies. These summaries were then used to contextualize the interview and generate questions tailored to the participant's experience and target role.

Each participant answered a core set of five interview questions covering typical interview stages, including introduction, motivation, experience, and closing. The interview was conducted by the AI-supported conversational system, which acted as the interviewer while the backend controlled the overall interview flow and stage progression to preserve comparable interview experiences across participants and conditions. In the Stress OFF condition, the interviewer used a neutral tone and presented only the fixed set of five questions. In the Stress ON condition, the interviewer adopted a stricter and more fast-paced tone, and the system could generate up to three additional follow-up questions when a previous response was evaluated as incomplete, vague, or insufficiently specific. Follow-up eligibility was determined by the backend based on the previous response score and remained within the same interview stage. These follow-up questions were intended to increase evaluative pressure and cognitive load during the interview interaction [11, 33].

The explanation manipulation changed the type of feedback shown to participants after each answer. In the Explanation OFF condition, participants received numerical scores and static rubric information. In the Explanation ON condition, participants received response-grounded explanations, including highlighted excerpts, strengths, weaknesses, missing points, and example improved answers. This design was informed by prior work on explainable AI and instance-level explanations in human-AI interaction [26, 27].

As discussed in Chapter 4, the experimental conditions differed in explanatory feedback and interviewer behavior while maintaining a consistent interview structure across conditions. This design helped ensure that observed differences were primarily associated with the experimental manipulations rather than variations in the interaction process.

5.2.3 Recruitment and Procedure

Participants were recruited through university communities and personal networks. Participation was voluntary, and informed consent was obtained before the study. The study was conducted remotely, and participants used their own devices. Standardized written instructions were provided to ensure consistency across sessions.

After signing up through a Google Form, participants received an email containing the study instructions, consent form, and a unique participant code. They were then randomly assigned to one of the four experimental conditions. Before beginning the interview, participants completed

a baseline questionnaire that collected demographic information, prior interview experience, and initial perceptions of interview performance.

Participants then created an account on the platform, uploaded their CV and target job description, and started the interview session. After each response, the system evaluated the answer and presented feedback according to the assigned explanation condition. At the end of the interview, participants completed a post-session questionnaire assessing their subjective performance, trust in AI-generated feedback, perceived usefulness, reliance on feedback, user experience, perceived competence, effort, and pressure.

Sessions in the Stress OFF condition typically lasted approximately 25-30 minutes. Sessions in the Stress ON condition generally lasted approximately 35-40 minutes due to the additional follow-up questions.

5.2.4 Participants

A total of $N = 87$ participants took part in the study. Participants were between 19 and 28 years old ($M = 21.62$, $SD = 1.59$). The sample included 58 male participants and 29 female participants. Most participants were undergraduate students, while a smaller number were postgraduate or graduate-level students.

All participants were required to be proficient in English, as the interview interaction and questionnaires were conducted in English. Self-reported English proficiency ranged from basic to advanced levels. Most participants had limited work experience and had attended one or two interviews during the previous year. Participants also reported moderate to high familiarity with AI tools such as ChatGPT, Gemini, and Claude. In addition, 49 participants indicated that they were actively seeking employment.

5.2.5 Outcome Measures

The study collected both subjective and objective measures. The primary dependent variable was calibration, operationalized as Delta Assessment. Delta Assessment was calculated as the difference between objective assessment and subjective assessment:

$$\text{Delta Assessment} = \text{Objective Assessment} - \text{Subjective Assessment}$$

Positive values indicate that participants underestimated their performance, while negative values indicate overestimation. Values closer to zero indicate better calibration.

Objective assessment was based on evaluations provided by three independent expert raters using a standardized preparedness rubric. More specifically, experts evaluated each interview

session across dimensions related to perceived competence, effort and importance, pressure and tension, and perceived value/usefulness. Scores were provided on a 7-point Likert scale and aggregated by computing the mean across raters. Inter-rater reliability was assessed using the intraclass correlation coefficient before aggregating the expert scores [38].

Subjective assessment was collected after the interview session through post-session questionnaires measuring participants' perceived competence, effort and importance, pressure and tension, and perceived value/usefulness of the interview preparation process. Additional measures examined trust in AI-generated feedback, reliance on AI feedback, explanation usefulness, and overall user experience. IMI-based measures were used for perceived competence, effort, pressure/tension, and value/usefulness [35], while UEQ-S was used to assess overall user experience [37]. Measures related to trust and reliance were informed by prior work on human-AI trust and reliance [14, 34, 36]. The complete pre-session and post-session questionnaires are provided in Appendix I.1.

Table 1 summarizes the main dependent variables and questionnaire-based measures collected during the study. These measures were used to evaluate participants' calibration, subjective perceptions, and interaction behaviors throughout the interview process.

Table 1: Overview of dependent variables and measures used in the evaluation.

Variable	**Description**
Delta Assessment	Difference between objective and subjective assessment. Positive values indicate underestimation, while negative values indicate overestimation.
Objective Assessment	Expert-rated interview performance using a standardized evaluation form.
Subjective Assessment	Self-reported performance assessment after the interview session.
Perceived Competence	IMI-based self-reported competence.
Effort / Importance	IMI-based effort and importance measure.
Pressure / Tension	IMI-based stress measure used as a manipulation check.
Perceived Usefulness	Participants' perceived usefulness of the system.
Explanation Satisfaction	Perceived quality of explanations in the Explanation ON condition.
Trust in AI	Trust in AI-generated feedback.
User Experience	Usability and experience measured using UEQ-S.

5.2.6 Data Analysis

The quantitative analysis examined how explanatory feedback and stress influenced user calibration in AI-supported interview preparation. The main dependent variable was Delta Assessment. To evaluate the effects of the experimental manipulations, a two-way between-subjects ANOVA was conducted with Explanation and Stress as independent variables.

Before conducting the ANOVA, assumptions of normality and homogeneity of variance were assessed using Shapiro-Wilk and Levene tests. No significant violations were observed, supporting the use of parametric tests. Statistical significance was evaluated at $\alpha = .05$. Effect sizes were reported using eta-squared (η^2) [39, 40].

Secondary analyses examined relationships between calibration and user perception measures. Pearson correlations were used to assess associations between Delta Assessment, trust in AI, perceived usefulness, and reliance on AI feedback.

5.3 Results

5.3.1 Main Effects of Stress and Explanation on Calibration

This analysis addresses RQ1, which investigated how explanatory feedback and evaluative stress influence users' calibration between subjective and objective interview performance assessments. To examine these effects, a two-way between-subjects ANOVA was conducted with Stress and Explanation as independent variables. Table 2 presents descriptive statistics for Delta Assessment across the four experimental conditions. Positive values indicate underestimation relative to objective performance assessment.

Table 2: Descriptive statistics for Delta Assessment across the four experimental conditions

Condition	Mean	SD	95% CI
Explanation OFF + No Stress	0.49	1.28	[-0.03, 1.02]
Explanation OFF + Stress	2.16	1.21	[1.61, 2.70]
Explanation ON + No Stress	0.80	0.94	[0.40, 1.19]
Explanation ON + Stress	1.50	1.20	[1.00, 2.00]

The analysis revealed a significant main effect of Stress, $F(1, 82) = 21.68$, $p < .001$, $\eta^2 = .21$. Participants in stress conditions showed higher positive Delta Assessment values compared to participants in no-stress conditions. This indicates that stress increased underestimation relative to objective performance assessment and disrupted accurate self-assessment.

In contrast, there was no significant main effect of Explanation, $F(1, 82) = 0.37$, $p = .546$, $\eta^2 = .004$. This suggests that explanatory feedback alone did not substantially improve alignment between subjective and objective assessment. Overall, these findings indicate that stress had a strong effect on calibration, while explanations by themselves were insufficient to reduce calibration discrepancies.

5.3.2 Moderating Effect of Explanation on Stress

This analysis addresses RQ2, which examined whether explanatory feedback moderates the effect of stress on users' calibration. A two-way ANOVA revealed a marginally significant interaction between Explanation and Stress, suggesting that the influence of stress on calibration may vary depending on the presence of explanatory feedback. Although the interaction did not reach conventional levels of statistical significance, it approached significance, $F(1, 82) = 3.63$, $p = .060$, $\eta^2 = .04$.

As illustrated in Figure 18, stress appeared to have a stronger negative effect on calibration when explanatory feedback was absent. In the Explanation OFF condition, participants showed substantially higher Delta Assessment values under stress. In contrast, the difference between stress and no-stress conditions was smaller when explanatory feedback was enabled.

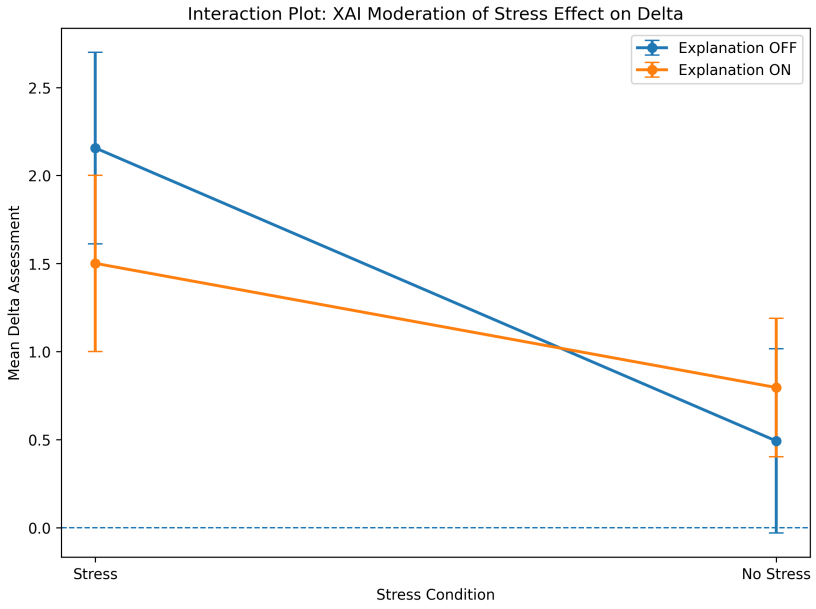

Figure 18: Interaction effect of stress and explanatory feedback on Delta Assessment.

Although exploratory, this pattern suggests that explanatory feedback may partially reduce the calibration-disrupting effects of stress. Rather than uniformly improving calibration, explanations may provide interpretive support under cognitively demanding interaction conditions.

5.3.3 User Perceptions and Calibration

This analysis addresses RQ3, which examined the relationship between users' perceptions of the system, their reliance on AI feedback, and calibration outcomes. Pearson correlation analyses were conducted to assess the associations between Delta Assessment, perceived usefulness, reliance on AI feedback, and trust in AI.

Calibration was moderately negatively correlated with perceived usefulness ($r = -0.41$, $p < .01$) and reliance on AI feedback ($r = -0.39$, $p < .01$). As illustrated in Figure 19, participants who perceived the system as more useful and relied more on its feedback tended to show smaller discrepancies between subjective and objective assessment.

In contrast, trust in AI was not significantly associated with calibration ($r = -0.01$, $p > .05$). This finding suggests that general trust in the system did not necessarily correspond to more accurate self-assessment. Instead, calibration appeared to be more closely related to how participants used and relied on the feedback during reflection.

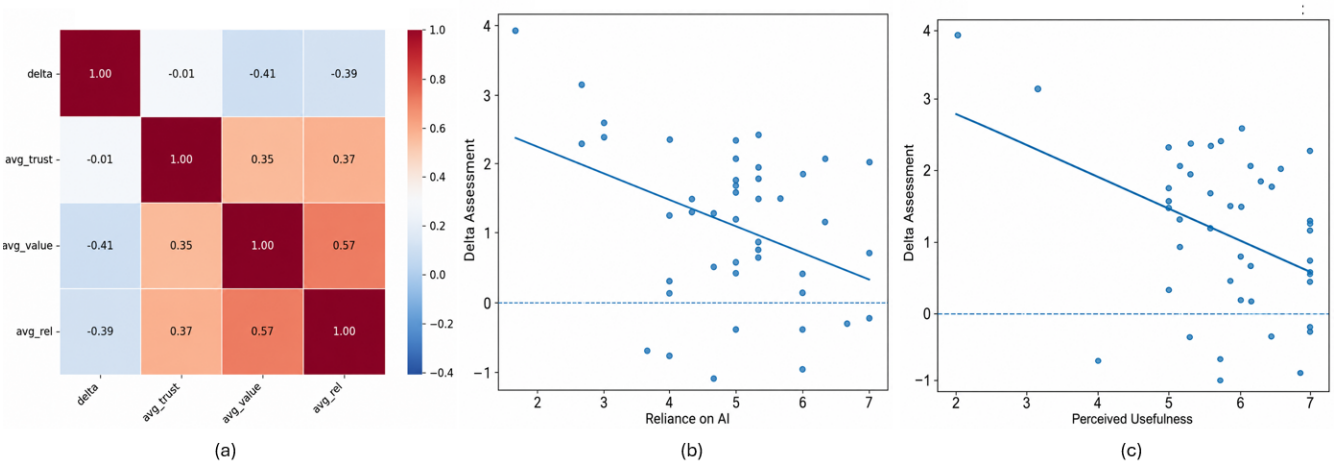

Figure 19: (a) Correlation matrix between calibration (Delta Assessment), trust in AI, perceived usefulness, and reliance on AI feedback. (b) Relationship between perceived usefulness and calibration. (c) Relationship between reliance on AI feedback and calibration.

5.4 Chapter Summary

This chapter presented the evaluation of the proposed AI-supported interview preparation system and examined how explanatory feedback and evaluative stress influenced users' calibration during interview preparation. The results showed that stress significantly disrupted accurate self-assessment, leading participants to underestimate their performance relative to expert evaluations. Although explanatory feedback did not produce a significant main effect on calibration, the observed interaction pattern suggested that explanations may partially attenuate the negative effects of stress under cognitively demanding conditions. In addition, the findings demonstrated

that calibration was more strongly associated with perceived usefulness and reliance on AI-generated feedback than with general trust in the system, highlighting the importance of how users interpret and utilize evaluative AI feedback during self-assessment.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis investigated how explanatory feedback and evaluative stress influence users' ability to accurately assess their performance in AI-supported interview preparation. Framing interview preparation as a feedback-driven human-AI interaction process, the study examined how users interpret and utilize AI-generated evaluative feedback across different interview stress conditions. More specifically, the work focused on calibration, defined as the alignment between subjective self-assessments and objective performance assessments [1, 2].

To support this investigation, a web-based interview preparation system was designed and implemented using Large Language Models (LLMs), automated speech processing, and conversational interaction techniques. The platform enabled users to participate in structured mock interviews, receive AI-generated feedback, and reflect on their own performance through post-session self-assessment.

The findings showed that stress significantly impaired users' ability to accurately evaluate their performance, systematically increasing underestimation relative to objective performance assessments. Although explanatory feedback did not directly improve calibration overall, the observed interaction patterns suggested that explanations may partially reduce the negative effects of stress under cognitively demanding conditions.

The study also highlighted important distinctions between trust, reliance, and calibration in human-AI interaction. While trust in AI-generated feedback was not significantly associated with calibration accuracy, perceived usefulness and reliance on AI feedback were more strongly related to alignment between subjective and objective assessment [14, 34]. These findings suggest that users may trust AI systems without necessarily forming accurate self-assessments.

Overall, this thesis contributes to Human-Computer Interaction (HCI) and AI-supported learning research by positioning calibration as an important dimension of interaction quality in evaluative AI systems. The findings further demonstrate that the effectiveness of AI-generated feedback depends not only on the information presented, but also on the cognitive conditions under which users interpret that feedback [11, 33].

6.2 Contributions

This thesis further contributes to Human-Computer Interaction (HCI) and AI-supported evaluative systems research by empirically examining calibration in AI-supported interview preparation settings [1, 2]. While prior work has often focused on usability, trust, or perceived system effectiveness, this study investigates how accurately users align their subjective self-assessments with objective performance evaluations during AI-supported interview preparation.

First, the thesis provides empirical evidence that evaluative stress significantly affects users' calibration. The findings showed that stressful interaction conditions increased underestimation relative to objective performance assessments, suggesting that cognitive and emotional factors influence how users interpret evaluative AI feedback and assess their performance [11, 12].

Second, the study contributes to Explainable AI research by showing that explanatory feedback does not necessarily improve calibration across contexts [27, 29]. Although explanations did not directly improve calibration, the observed interaction patterns suggest that explanatory feedback may function as support under cognitively demanding conditions.

Third, the thesis highlights an important distinction between trust, reliance, and calibration in human-AI interaction [9, 34]. While trust in AI-generated feedback was not significantly associated with improved calibration, perceived usefulness and reliance on AI feedback were more strongly related to lower calibration error. These findings suggest that trust alone may not necessarily support accurate self-assessment during evaluative interaction.

Finally, this work contributes a web-based AI-supported interview preparation platform integrating conversational interaction, automated speech processing, LLM-based evaluation, and response-grounded explanatory feedback. The system provides a practical foundation for future research on adaptive, stress-aware, and reflective AI-supported evaluative environments.

6.3 Limitations

Although this study provides useful insights into AI-supported interview preparation and calibration, several limitations should be acknowledged.

First, participants were recruited primarily through convenience sampling and mainly consisted of university students familiar with AI technologies. As a result, the findings may not fully generalize to broader populations or real professional interview settings.

Second, the evaluation was conducted in a simulated interview environment rather than a real hiring context. Although the stress manipulation introduced evaluative pressure, it may not fully capture the psychological demands of real-world interviews.

Third, the study focused primarily on calibration error (Delta Assessment), which reflects discrepancies between subjective and objective assessments. While this measure captures calibration differences, it does not fully represent the broader reflective and metacognitive processes involved in self-assessment [1, 2].

Finally, the findings are tied to the specific explanation design and interaction setup used in the study. Different explanation styles, prompting strategies, or interaction designs may influence calibration and self-assessment differently [27, 29].

6.4 Future Work

Future research could further explore adaptive and personalized forms of explanatory feedback in AI-supported interview preparation. For example, explanation content and presentation could dynamically adapt according to users' expertise, confidence, or stress levels.

Another important direction involves longitudinal studies examining how calibration evolves through repeated interaction with AI-generated feedback over time. Such studies could provide deeper insights into whether AI-supported reflection improves long-term self-assessment accuracy and interview preparedness.

Future work could also investigate multimodal and stress-aware interaction approaches by incorporating behavioral or physiological signals such as speech characteristics, facial expressions, or gaze behavior. These signals may help AI systems better estimate user stress and provide more personalized evaluative feedback.

Finally, future research should continue examining calibration as a central component of effective human-AI collaboration, particularly in high-stakes evaluative settings where users must accurately assess their own performance while interacting with AI-generated feedback.

6.5 Chapter Summary

This chapter summarized the main findings, contributions, limitations, and future research directions of the thesis. The findings highlighted how evaluative stress significantly influenced user calibration in AI-supported interview preparation, while explanatory feedback showed context-

dependent effects under cognitively demanding conditions. The chapter further discussed the thesis contributions to HCI and AI-supported evaluative systems research, particularly regarding calibration, explanatory feedback, and human-AI interaction. Finally, the limitations of the study and potential directions for future work were presented, including adaptive explanatory feedback, longitudinal evaluation, and multimodal stress-aware AI interaction.

BIBLIOGRAPHY

[1] J. Dunlosky and J. Metcalfe, *Metacognition*. Sage Publications, 2008.

[2] S. M. Fleming and R. J. Dolan, "The neural basis of metacognitive ability," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1594, pp. 1338–1349, 2012.

[3] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[4] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International journal of human-computer studies*, vol. 146, p. 102551, 2021.

[5] J. Kim, H. Maathuis, and D. Sent, "Human-centered evaluation of explainable ai applications: a systematic review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1456486, 2024.

[6] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2390–2395.

[7] S. Alon-Barkat and M. Busuioc, "Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice," *Journal of Public Administration Research and Theory*, vol. 33, no. 1, pp. 153–169, 2023.

[8] M. De-Arteaga, R. Fogliato, and A. Chouldechova, "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–12.

[9] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "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*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[10] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[11] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

[12] K. Starcke and M. Brand, "Decision making under stress: a selective review," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1228–1248, 2012.

[13] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM international joint conference on*

Pervasive and ubiquitous computing, 2013, pp. 697–706.

[14] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 624–635.

[15] M. A. Campion, D. K. Palmer, and J. E. Campion, "A review of structure in the selection interview," *Personnel psychology*, vol. 50, no. 3, pp. 655–702, 1997.

[16] J. Levashina, C. J. Hartwell, F. P. Morgeson, and M. A. Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Personnel psychology*, vol. 67, no. 1, pp. 241–293, 2014.

[17] M. A. Campion, J. E. Campion, and J. P. Hudson Jr, "Structured interviewing: A note on incremental validity and alternative question types." *Journal of Applied Psychology*, vol. 79, no. 6, p. 998, 1994.

[18] E. Panadero and A. Jonsson, "The use of scoring rubrics for formative assessment purposes revisited: A review," *Educational research review*, vol. 9, pp. 129–144, 2013.

[19] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational research review*, vol. 2, no. 2, pp. 130–144, 2007.

[20] C. C. Liem, M. Langer, A. Demetriou, A. M. Hiemstra, A. Sukma Wicaksana, M. P. Born, and C. J. König, "Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening," in *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 197–253.

[21] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[22] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "" help me help the ai": Understanding how explainability can support human-ai interaction," in *proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–17.

[23] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–13.

[24] Q. V. Liao and K. R. Varshney, "Human-centered explainable ai (xai): From algorithms to user experiences," *arXiv preprint arXiv:2110.10790*, 2021.

[25] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "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, 2020, pp. 295–305.

[26] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[27] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[29] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, pp. 2119–2128.

[30] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *Proceedings of the 23rd international conference on intelligent user interfaces*, 2018, pp. 211–223.

[31] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "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*, 2021, pp. 1–16.

[32] D. A. Moore and P. J. Healy, "The trouble with overconfidence." *Psychological review*, vol. 115, no. 2, p. 502, 2008.

[33] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[34] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[35] R. M. Ryan, "Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory." *Journal of personality and social psychology*, vol. 43, no. 3, p. 450, 1982.

[36] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[37] A. Hinderks, "Design and evaluation of a short version of the user experience questionnaire (ueq-s)," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2017.

[38] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp.

155–163, 2016.

[39] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

[40] D. C. Montgomery, *Design and analysis of experiments*. John wiley & sons, 2017.

APPENDIX I

Questionnaires and Prompt Design

I.1 Questionnaires

This appendix presents the questionnaires administered before and after the interview preparation session. The instruments capture participants' background, baseline perceptions, and post-session evaluations of the AI system.

I.1.1 Pre-Session Questionnaire

The pre-session questionnaire collected demographic information and baseline measures of interview preparedness prior to interaction with the system.

I.1.1.1 Demographics and Background

1. Gender
2. Age
3. Year of studies
4. Level of English proficiency
5. Have you previously participated in this study?
6. How would you rate your overall experience with AI-based systems?
7. How many job interviews have you attended in the last year?
8. How many years of work experience do you have?
9. Have you previously used any AI-based tools for interview preparation? If yes, please briefly describe the tool(s) and how you used them.
10. Are you currently actively seeking a job?

I.1.1.2 Baseline Interview Preparedness (IMI-adapted)

Participants responded using a 7-point Likert scale: 1 = Not at all true, 7 = Very true.

I.1.1.2.1 Perceived Competence

1. I think I am pretty good at preparing for job interviews.

2. I believe I generally perform well in interview preparation tasks compared to others.
3. I feel competent about participating in a real job interview.
4. I am satisfied with my current interview preparation skills.
5. I feel that I am skilled at preparing for interviews.
6. Preparing for job interviews is something I do not do very well. (R)

I.1.1.2.2 Effort / Importance

1. I usually put a lot of effort into preparing for job interviews.
2. I do not try very hard when preparing for job interviews. (R)
3. I try very hard when preparing for interviews.
4. It is important for me to perform well in job interviews.
5. I do not put much energy into interview preparation. (R)

I.1.1.2.3 Pressure / Tension

1. I usually do not feel nervous when preparing for job interviews. (R)
2. I tend to feel tense when preparing for job interviews.
3. I usually feel relaxed when preparing for job interviews. (R)
4. I feel anxious when thinking about job interviews.
5. I feel pressured when preparing for job interviews.

I.1.1.2.4 Value / Usefulness

1. Preparing for job interviews is valuable to me.
2. I believe interview preparation helps improve my interview performance.
3. I think interview preparation helps me understand how to improve my answers.
4. I am willing to spend time preparing for interviews because it is useful to me.
5. Interview preparation helps me identify strengths and weaknesses in my responses.
6. I believe preparing for interviews is beneficial to me.
7. I think interview preparation is an important activity.

I.1.1.3 Logistics

1. When do you expect to start the interview session?

2. Please provide your email address for study communication purposes.

I.1.2 Post-Session Questionnaire

The post-session questionnaire assessed participants' perceptions of performance, interaction experience, and attitudes toward the AI system.

Unless otherwise stated, items were rated on a 7-point Likert scale: 1 = Strongly disagree, 7 = Strongly agree.

I.1.2.1 Intrinsic Motivation Inventory (IMI)

I.1.2.1.1 Perceived Competence

1. I think I am pretty good at preparing for interviews using this system.
2. I think I did pretty well in this interview preparation task, compared to other participants.
3. After working on this interview preparation task for a while, I felt quite competent for a real interview.
4. I am satisfied with my performance during the interview preparation task.
5. I felt that I was pretty skilled at this interview preparation activity.
6. This was an interview preparation task that I could not do very well. (R)

I.1.2.1.2 Effort / Importance

1. I put a lot of effort into preparing for the interview using this system.
2. I did not try very hard to do well during the interview preparation task. (R)
3. I tried very hard during this interview preparation activity.
4. It was important to me to perform well in this interview preparation task.
5. I did not put much energy into preparing for the interview. (R)

I.1.2.1.3 Pressure / Tension

1. I did not feel nervous at all while preparing for the interview using this system. (R)
2. I felt very tense while preparing for the interview using the system.
3. I felt very relaxed while preparing for the interview using this system. (R)
4. I felt anxious while working on this interview preparation task.
5. I felt pressured while preparing for the interview using this system.

I.1.2.1.4 Value / Usefulness

1. I believe that preparing for interviews using this system could be valuable to me.
2. I think that using this system for interview preparation is useful for improving my interview performance.
3. I think this is an important activity because it helps me better understand how to improve my interview answers.
4. I would be willing to use this interview preparation system again because it has value to me.
5. I think preparing for interviews using this system could help me identify strengths and weaknesses in my interview responses.
6. I believe that using this system for interview preparation could be beneficial to me.
7. I think this is an important interview preparation activity.

I.1.2.2 Explanation Satisfaction

I.1.2.2.1 Perceived Usefulness of Explanations

1. The explanations helped me understand what I need to improve in my answers.
2. The explanations helped me prepare better for a real interview.
3. The explanations helped me learn something new.
4. I would like to use similar explanations in future interview preparation sessions.

I.1.2.3 Reliance on AI Feedback

1. I relied on the system's explanations to judge how well I performed.
2. Even when I had doubts, I trusted the system's evaluations.
3. I would follow the system's advice even if it differed from my own opinion.

I.1.2.4 Trust in AI System

1. The system is deceptive.
2. The system behaves in an underhanded manner.
3. I am suspicious of the system's intent, action, or outputs.
4. I am wary of the system.
5. The system's actions will have a harmful or injurious outcome.

6. I am confident in the system.
7. The system provides security.
8. The system has integrity.
9. The system is dependable.
10. The system is reliable.
11. I can trust the system.
12. I am familiar with the system.
13. I found the AI-generated interview feedback to be trustworthy.
14. The information and feedback provided by the AI system felt accurate and relevant for interview preparation.
15. I would rely on this AI system to support my interview preparation in the future.
16. I was skeptical about the AI-generated interview feedback.

I.1.2.4.1 Open-ended Questions

1. What factors made you trust or distrust the feedback generated by the AI during interview preparation?
2. Was there anything in the AI-generated feedback that seemed inaccurate, unclear, or misleading?

I.2 Prompt Design and Parameterization

I.2.1 Prompting Strategy

The system uses structured prompts to control the interview interaction, response evaluation, and feedback generation. All prompts follow a standardized high-level structure consisting of three core components: Role, Context, and Task, extended with additional constraints such as stage definitions, CV usage rules, interaction parameters, and output format requirements. This design ensures consistency across participants while supporting controlled adaptive conversational behavior across experimental conditions.

All experimental conditions are implemented through parameterization of the same base prompts rather than separate prompt designs.

I.2.2 Interview Prompt

The following prompt is used to generate and control the interview interaction:

```
You are an AI interviewer conducting a structured interview session.

ROLE:
Act as a professional interviewer.

CONTEXT:
The participant is answering interview questions in a simulated
interview setting.

TASK:
- Ask one question at a time
- If follow-ups are enabled, clarification follow-up questions may
be generated based on the quality of the participant's previous
response
- If follow-ups are disabled, proceed directly to the next scheduled
question
- Maintain the specified tone

TONE: {tone}
FOLLOW_UPS: {followups_enabled}
```

I.2.2.1 Stage-Based Interview Structure

The interview follows a predefined five-question structure. The stage of each question is determined by the backend based on the question index rather than by the language model itself. The language model operates within these predefined orchestration constraints and is responsible only for generating localized conversational behavior for each interaction step. The fixed sequence consists of introduction, motivation, experience-focused stages, and closing.

This structure ensures that all participants progress through the same overall interview flow. In the introduction stage, the system asks a high-level warm-up question. In the motivation stage, it asks about role or career motivation. In the experience-focused stages, it asks about concrete experience, projects, responsibilities, or skills extracted from the participant's CV. In the closing stage, the system asks a final wrap-up question.

The stage-based structure is applied uniformly across all experimental conditions and is not manipulated as part of the study.

I.2.2.2 Tone Manipulation

The tone parameter controls the interaction style. In the neutral condition, the interviewer uses calm, supportive, and non-judgmental phrasing. In the strict condition, the interviewer adopts more direct, concise, and evaluative phrasing, including clarification-oriented probing behavior

intended to increase evaluative pressure during the interaction.

I.2.2.3 Follow-Up Question Logic

Follow-up behavior is controlled by both the experimental condition and the quality of the participant's previous response. In conditions where follow-ups are disabled, the system proceeds directly to the next scheduled main question. In conditions where follow-ups are enabled, a clarification-oriented follow-up question may be generated within the same interview stage only when the average rubric score of the participant's previous response falls below 3 out of 5. Follow-up eligibility and stage persistence are externally enforced by the backend rather than determined autonomously by the language model. This design preserves comparable interview structure and interaction consistency across participants and conditions while enabling controlled adaptive probing behavior.

I.2.2.4 Affective Context Integration

To support more natural conversational interaction, the system additionally incorporated lightweight affective context derived from users' voice responses. These emotion-related cues were not used for automated evaluation or scoring, but instead served as auxiliary contextual information for localized interviewer tone adaptation and probing behavior during question generation. Importantly, affective context did not alter interview stage progression, evaluation criteria, or task difficulty.

I.2.3 Evaluation and Feedback Prompt

The following prompt is used to evaluate participant responses. The evaluation process remains identical across all experimental conditions and is not influenced by interview tone or interaction style.

Specifically, the same rubric dimensions (clarity, structure, depth, confidence, and relevance), scoring scale, and feedback generation procedure are applied uniformly across all conditions.

```
You are an AI interview evaluator.

ROLE:
Act as an expert evaluator assessing interview performance.

CONTEXT:
The participant has provided an answer to an interview question.
Evaluation is based on predefined rubric dimensions.

TASK:
- Assign scores for each rubric dimension from 0 to 5
```

```
- Evaluate clarity, structure, depth, confidence, and relevance
- Provide an overall score
- If explanation mode is enabled:
  - Explain strengths and weaknesses
  - Identify missing elements
  - Suggest improvements
  - Provide an example of a better answer

EXPLANATION_MODE: {explanation_mode}
```

Keeping the evaluation prompt constant ensures that observed differences are attributable to the interaction design rather than changes in evaluator behavior.

I.2.4 Feedback Highlight Extraction

When detailed feedback is enabled, the system first extracts concise evidence of strengths and weaknesses from the participant response. This step is used internally to support grounded feedback generation.

```
Extract key strengths and weaknesses from the response.

TASK:
- Identify strong points
- Identify issues or missing elements
- Return structured evidence (short quotes and reasons) for
strengths and issues across rubric dimensions
```

I.2.5 Condition Parameterization

All experimental conditions are implemented through parameter values applied to the same base prompts. Interview stage progression, follow-up eligibility, and termination behavior remained backend-controlled across all conditions to preserve consistent interview structure and experimental comparability. The manipulated parameters are:

- `tone`: controls the conversational style of the interviewer.
- `followups_enabled`: controls whether clarification-oriented follow-up questions may be generated.
- `explanation_mode`: controls whether detailed response-grounded feedback is generated.

The four experimental conditions are defined as follows:

The stress condition is operationalized through stricter interviewer tone, clarification-oriented

Table 3: Prompt parameters used across experimental conditions.

Condition	Explanation	Tone	Follow-ups
C1	OFF	neutral	false
C2	OFF	strict	true
C3	ON	neutral	false
C4	ON	strict	true

probing behavior, and controlled follow-up questioning. The evaluation process remains unchanged across all conditions.

I.2.6 Reproducibility Note

All prompts were applied consistently across participants. Condition-specific interaction behavior was controlled through predefined parameter values and backend orchestration constraints, while all other aspects of the prompt structure, evaluation process, and output formatting remained constant.

I.3 Source Code Repository

The implementation developed for this thesis is available at: https://github.com/RafaelChrysanthou/RafaelThesis.git

The repository contains the source code of the interview preparation platform.