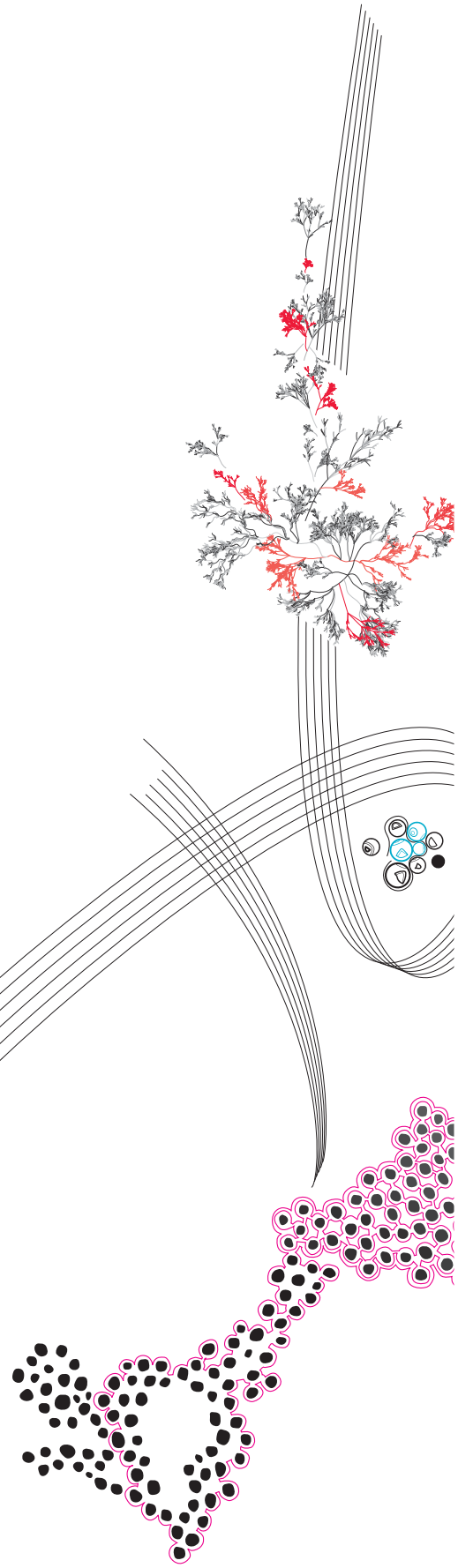

MSc Computer Science
Thesis

Talent Assistant: An AI System for Recruitment Support

Andrea Nappi

Supervisor: Professor Briassouli Alexia

October, 2026

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract **4**

1 Introduction **5**
1.1 Context: Hiring at scale and interview preparation in large firms and in PwC Italy . . . 5
1.2 Problem statement: Inefficient and inconsistent interview preparation . . . 7
1.3 Research questions . . . 7
1.4 Objective and contributions . . . 8
1.5 Scope, assumptions, and limitations of setting . . . 9
1.6 Thesis structure . . . 10

2 Background and related work **12**
2.1 AI-assisted hiring and interview support . . . 12
2.2 AI and LLM risks in job interviewing contexts . . . 13
2.3 Human oversight . . . 15
2.4 Expert evaluation . . . 16
2.5 Quality dimensions for interview question sets . . . 18

3 Requirements specification **20**
3.1 Purpose and scope of Talent Assistant . . . 20
3.2 Functional requirements . . . 20
3.3 Output quality requirements . . . 22
3.4 Non-functional requirements . . . 23

4 Evaluation design and methodology **25**
4.1 Study design overview . . . 25
4.2 Measures . . . 26
4.3 Analysis plan . . . 27

5 System design: Talent Assistant **30**
5.1 System overview and architecture . . . 30
5.2 UX flow and human-in-the-loop . . . 31
5.3 Intended usage of Talent Assistant . . . 35
5.4 Implementation details and prompting strategy . . . 36
5.5 Data governance and privacy . . . 39

6 Evaluation results and discussion **42**
6.1 Survey results . . . 42
6.1.1 Participants . . . 42
6.1.2 Perceived usefulness results . . . 43

6.1.3 Perceived quality results . . . 44
6.1.4 Usage and integration results . . . 45
6.2 Field usage metrics . . . 46
6.3 Observed issues and failure instances . . . 48
6.4 Threats to validity . . . 49

7 Conclusion **51**
7.1 Answering RQs . . . 51
7.2 Contributions and implications . . . 54
7.3 Future work . . . 55

Bibliography **56**

A Research interviews **61**
A.1 Interview with manager 1 . . . 61
A.2 Interview with manager 2 . . . 64
A.3 Interview with manager 3 . . . 67

B Evaluation rubric and participant instructions **70**
B.1 Survey for Interviewers: Evaluation of Talent Assistant . . . 70

C Survey results **75**

D Usage of generative AI in the thesis **80**

Abstract

Technical interview preparation in large professional services organizations is often time-consuming and inconsistent, as it depends on individual interviewer experience, informal question repositories, and manual tailoring to each candidate and role. This thesis addresses this problem through the design, implementation, and evaluation of Talent Assistant, an internally deployed, human-in-the-loop, agentic AI system that aims to support technical interview preparation at PwC Italy.

Talent Assistant requires a candidate's curriculum vitae, a job description, and (optionally) a seniority level to produce a structured package for interview support, containing a candidate profile and sets of experience-focused, behavioral, motivational, technical, and practical questions. The system is deliberately limited to preparation activities: it does not screen, rank, score, or recommend candidates, and it does not evaluate candidate answers. Interviewers remain responsible for reviewing, selecting, and adapting the generated material.

The system was implemented as a web application using a fixed multi-stage agentic pipeline and evaluated through a survey of its ten operational users, together with descriptive production data. Participants assessed the system relative to their usual preparation practice using a seven-point scale. Results provide evidence of positive perceived usefulness, output quality, usability, retained interviewer control, and consistency. The efficiency index received a mean score of 6.35 out of 7, the output-quality score was 6.20, and perceived interviewer control received the highest score, at 6.90. Production logs recorded 95 unique candidate–role packages and 133 generation runs between October 2025 and July 2026.

The findings suggest that bounded AI support can reduce perceived preparation effort and provide a structured, candidate–role-specific interview backbone while preserving human responsibility. However, the results are limited to a small user group within one organizational setting and rely mainly on perceived outcomes. They do not establish improvements in actual preparation time, factual correctness, fairness, candidate experience, or hiring outcomes.

Chapter 1

Introduction

1.1 Context: Hiring at scale and interview preparation in large firms and in PwC Italy

In large organizations, the hiring process is often a standardized, multi-stage pipeline that must be in place for many teams, locations, and roles at the same time, rather than an occasional activity [6]. High-volume recruitment is typically characterized by a large number of applicants per role, multiple open positions at the same time, and pressure on tight timelines, as hiring teams are expected to move candidates from application to offer quickly enough to remain competitive [2, 1]. This kind of pipeline often includes sequential steps such as application intake, CV screening via an Applicant Tracking System (ATS), one or more interviews led by HR, and one or more technical or domain-specific interviews, with each stage intended to filter candidates, preserving a manageable workload for hiring staff [8] (Appendices A.1, A.2, A.3). Technical interviews are found near the end of this pipeline in a high-cost, high-stakes position: they require valuable experts' time, but they are absolutely key to assessing a candidate's fit, communicating the nature of the role, and giving candidates an opportunity to show their skills to the employer [5].

This general structure, in the case of large firms, often involves multiple people, including HR professionals, technical interviewers, and final decision-makers [6] (Appendices A.1, A.2). Hiring staff often have to deal with hundreds of applications per open job position and need to schedule multiple interview rounds for shortlists that still involve many conversations, especially when organizations are growing or running seasonal hiring drives [2, 1]. For professional services companies, such as consulting firms, new roles are often opened when new projects start, often suddenly, and require a fast hiring response to avoid overloading employees or having to slow down the new project, all this with technical interviewers being busy working on billable projects; and managers involved in recruiting at PwC Italy describe similar patterns (Appendices A.1, A.2).

In this pipeline, scheduling and preparation for interviews can be complex, potentially becoming a main bottleneck, as it includes coordinating calendars for candidates and multiple interviewers, handling rescheduling, and running multi-round processes, potentially extending the hiring cycle [58]. Similar environments in terms of workload, such as medical education programs, report that program directors work evenings and weekends during interview periods, and that interview days disrupt normal service and teaching, showing how interview activity can interfere directly with regular work [4]. Managers interviewed at PwC Italy also mentioned this dynamic at firm level, with interviews being additional activities, preparation being often squeezed into gaps in the schedule, and time spent to prepare themselves for interviews often coming from after regular working hours (Appen-

dices A.1, A.2).

Regarding interview preparation, in large firms it generally focuses on planning a set of relevant questions that covers both technical and behavioral dimensions at an appropriate level of complexity, based on the candidate's profile and on the role they applied for [8]. In the technology domain at PwC Italy, a common practice for hiring is for staff to review the candidate's CV, note relevant experiences and projects, and then adapt a collection of standard questions or exercises to the candidate's background and the requirements expressed in the job description (JD), sometimes maintaining informal repositories or files of questions as a loose guideline for the conversation (Appendices A.1, A.2). This preparatory work usually results in interviews that include an introduction about the team, an open discussion of the candidate's study or career path and motivations, and more structured behavioral and technical questions.

About differences in interviews for roles at different seniority levels, managers at PwC Italy emphasize a problem-solving approach and basic conceptual understanding for junior interviews, focusing more on depth of expertise and coordination skills for senior ones (Appendices A.2, A.3), which is a consistent pattern with descriptions of different expectations across stages and roles in tech hiring pipelines in general [8].

Regardless of the candidate's seniority, the quality of individual interview questions is key to having effective interviews. In large firms, interview questions should be expected to be clearly job-related, behaviorally specific, and should have some observable indicators of the target competency [3, 9]. Generally, "good" interview questions are those that focus on past behavior or realistic job situations, ask all candidates about the same underlying competencies, and tend to be specific to the role and level, without going over sensitive or legally problematic topics (which is a requirement that often aligns with organizational policies) [9]. PwC Italy managers also mentioned that, based on their values and understanding, good technical questions are those grounded in the candidate's CV and in realistic project scenarios, which allow them to understand both how candidates think and what they have actually done (Appendices A.2, A.3). By contrast, weak or less useful questions are those that remain very generic, not very related to the job, or heavily reliant on improvisation, such as unstructured discussion of the candidate's background or classic tricky puzzles [15]. Indeed, unstructured interviews and unexpected trick questions have substantially lower predictive validity than structured behavioral or situational questions, because different candidates are asked different topics, and because, although they might feel engaging, they provide little systematic information about how a candidate would perform in the actual role [36].

Lastly, despite the advantages of structure, a common practice in many large organizations is still to leave interviewers with almost complete freedom in how they prepare and conduct their interviews [5], and even where global recruiting processes are standardized, the content and sequencing of interview questions often evolves organically, with different interviewers favoring different topics, and questions being adapted informally depending on local team culture, current projects, and personal preferences [23]. PwC Italy aligns with this framework (Appendices A.2, A.3), with interviewers reporting high-level corporate expectations (for example, avoiding sensitive personal topics and respecting confidentiality about clients and project details), but stating that the technical portion of the interview is mostly left to their judgment in terms of question choice, ordering, and depth.

Because the interview preparation problem is repetitive, information-heavy, and constrained by identifiable inputs such as the CV and the JD, it might be possible for an AI system to structure this information and turn it into a helpful interview backbone. By nature, such technology could have the potential to reduce the manual effort of starting

from scratch and to standardize the structure of preparation across interviewers, automating the transformation from raw candidate information and job specifics into a structured, reviewable, and adaptable interview package.

1.2 Problem statement: Inefficient and inconsistent interview preparation

As mentioned, technical interviews sit near the end of the hiring pipeline, require scarce expert time, and have a key impact on the final hiring decisions, yet the preparatory work preceding them is manual and dependent on individual interviewers' habits rather than on a shared official process. This configuration increases preparation effort and may reduce interview comparability, and at the same time, it does not ensure consistent question quality and topic coverage, making comparability across candidates for the same role more complex.

In the current practice at PwC Italy, every 30 minutes of interview preparation has to compete directly with billable client project work or with intended recovery time, creating time pressure problems when hiring volumes increase. This has implications beyond interviewer workload; indeed, from a recruitment perspective, this also slows down the pipeline, with extended time-to-hire leading to higher candidate dropout, more lost offers to faster competitors, and delayed project staffing [10, 56].

About the actual interview preparation, managers describe starting from the candidate's CV and JD, then consulting personal notes or informal files of questions, and adapting them on the fly depending on their own expertise and on how the conversation unfolds (Appendices A.1, A.2, A.3). This flexibility allows tailoring to the candidate, but it also means that some interviews rely heavily on generic or broad questions, while others include more specific behavioral or technical ones, with no guarantee that all questions are clearly related to the actual position. Variability in question quality risks wasting scarce interview time on "weak" questions and making it more difficult to compare candidates consistently.

Another potential issue is inconsistent topic coverage and structure across interviews for the same role and seniority level. Managers at PwC stated that the exact areas covered during an interview, the depth of questioning, and the order of topics depend heavily on who conducts the interview and on personal preferences (Appendices A.2, A.3), but the literature indicates that using a consistent structure with predefined competency domains and comparable question types is important for having a fair comparison downstream, because it ensures that all candidates are assessed on the same underlying constructs. In practice, this lack of structural standardization risks weakening both perceived fairness for candidates, who may receive very different interviews for the same job, and the organization's aim to hire the most suitable candidate for a given role.

1.3 Research questions

This thesis aims to investigate whether an AI-based system can assist technical interviewers in preparing structured, candidate–role specific interview materials while preserving their professional judgment and responsibility. Following on the problems identified in Section 1.2, these are the derived goals that should guide the design and evaluation of the system:

- Such a system should reduce the time and effort required to prepare technical interviews without lowering the quality of the resulting interview support materials. It

should shorten the path from a candidate's CV and a JD to a usable interview preparation package, allowing interviewers to spend more of their limited time reviewing and refining questions rather than starting from a blank page.

- Such a system should provide a more consistent structure for interview preparation. It should not impose identical questions on every candidate; instead, it should provide similarly structured packages with comparable levels of depth for candidates applying to similar roles at comparable seniority levels. At the same time, it should retain sufficient flexibility to account for differences in candidates' backgrounds and experience.
- Generated materials should be relevant to both the candidate profile and the target role, calibrated to the candidate's seniority level, sufficiently broad in their coverage of relevant competency areas, and non-redundant.
- The system should preserve human responsibility and support responsible deployment in a hiring-related setting. It should support interview preparation rather than screen, rank, score, recommend, or select candidates, and it should not evaluate candidate answers. Interviewers should remain responsible for reviewing, adapting, selecting, or disregarding generated questions, as well as for conducting the interview and making all hiring-related judgments.

Based on these design objectives, the following research questions were formulated:

1. RQ1: How can an agentic AI system be designed so that it meets hiring domain requirements?
   (a) RQ1.1: How can an agentic orchestration structure be designed to produce effective, structured interview support packages?
   (b) RQ1.2: Which inputs should be integrated into the pipeline, and how should they condition generation?
   (c) RQ1.3: What risks are tied to these requirements, and what do they imply for safe deployment of such a system?
2. RQ2: How is such a system perceived by early users to mitigate the key issues investigated compared with current practice?
   (a) RQ2.1: Does it improve the preparation time–quality ratio?
   (b) RQ2.2: Does it improve key quality dimensions for questions?
   (c) RQ2.3: Does it improve final output consistency across interviews for comparable candidates–roles?

1.4 Objective and contributions

The objective of this thesis is to present the design, implementation, and evaluation of Talent Assistant, an AI-based system for preparing technical interviews in the PwC Italy recruitment context. Talent Assistant supports technical interviewers by transforming information already used during preparation into a structured and reviewable interview-support package. This package includes a candidate profile and sets of experience-focused, behavioral, motivational, technical, and practical interview questions.

The system is designed as decision support rather than decision automation: interviewers remain responsible for reviewing the generated material and making all recruitment-related judgments.

This thesis makes four main contributions. First, it presents the design and implementation of Talent Assistant as a deployed system artifact. Rather than proposing only a conceptual design or a limited proof of concept, the thesis documents an operational web application developed for internal use at PwC Italy. It describes the system's architecture, interaction flow, generation pipeline, prompting strategy, persistence mechanisms, and logging instrumentation. Although the source code and selected operational details cannot be disclosed because of company-policy constraints, the thesis provides sufficient technical detail to explain the system's design rationale and functioning.

Second, the thesis derives a requirements-driven evaluation rubric for AI-generated interview support materials. The rubric is informed by the organizational context, interviews with PwC Italy managers, and relevant literature on structured interviews, AI-supported recruitment, and AI-assisted question generation. It translates the general notion of a useful interview question set into concrete dimensions, including relevance, specificity, calibration, coverage, diversity, and reliability.

Third, the thesis provides evidence from the system's deployment in its intended organizational setting. The evaluation combines structured feedback from technical interviewers who used Talent Assistant in real recruitment processes with descriptive usage data from the deployed application. This evidence assesses perceived usefulness, output quality, integration into interviewers' workflows, retained human control, and perceived consistency across comparable interview-preparation cases. The findings are interpreted as context-specific and formative; they do not establish causal effects on preparation time, hiring outcomes, fairness, or candidate experience.

Fourth, the thesis analyses the limitations, risks, and observed failure modes associated with generative AI support for interview preparation. By relating system observations to relevant literature on bias, hallucinations, automation bias, traceability, privacy, and human oversight, the thesis discusses implications for the responsible deployment and continued refinement of similar systems in hiring-related settings.

Overall, this thesis contributes both a practical system artifact and an evaluation of its use. It demonstrates how generative AI may support the preparation of technical interviews while preserving the professional judgment and decision-making responsibility of the human interviewer.

1.5 Scope, assumptions, and limitations of setting

Talent Assistant has been developed and evaluated in the organizational context of PwC Italy, and in particular, within technical and data-oriented teams in the organization. This introduced several limitations that constrained the way the system had to be designed and the way it could have been evaluated.

The system is intended only to support interviewers who prepare and conduct technical interviews near the end of the PwC Italy hiring pipeline, with other activities that appear elsewhere in the process being explicitly out of the scope of this thesis. Talent Assistant does not perform CV screening, candidate matching, or automatic shortlisting, and it will not provide algorithmic scoring or ranking of candidates for selection decisions.

Talent Assistant operates as a human-in-the-loop (HITL) supporting tool that generates a structured interview backbone consisting of behavioral and technical questions, which interviewers will be free to review, adapt, or ignore, so that the system does not

replace the interviewer's judgment about which topics to prioritize during the actual conversation. The system does not evaluate or grade candidates' answers during or after the interview; the focus is on supporting human interviewers in preparing a higher-quality, more consistent question set under time constraints, not on automating hiring decisions.

Regarding the content of the provided interview support package, the main content inputs affecting generation are the candidate's CV, the JD, and the seniority level, although the target language also influences the generated output. Given the development context, the system will only be able to operate under the assumption that both CVs and JDs are available in digital form, with the quality of the result depending also on the JD's content reliability and on the candidate's CV depth and precision.

The tool only focuses on interview preparation for candidates up to "manager" seniority, because these are the only positions sharing (to some extent) a standard interview structure, with a hybrid focus on soft and hard skills, although the balance changes depending on the seniority (with manager-level roles representing a boundary case). Higher seniority roles (such as senior managers, directors, or partners) require different evaluations, follow different processes, and are indeed typically handled directly by firm partners or similar (Appendix A.2).

Since the system is deployed and used internally at PwC Italy, where interviews as well as evaluation tasks are carried out almost always in Italian by Italians, the user interface is designed primarily for Italian-language use. The evaluation of Talent Assistant relies on a small purposive sample comprising all current operational users, who are expert interviewers from PwC Italy, who already participated in technical interviews for the relevant roles, which limits the generalizability of results beyond this firm, country, and role family. Data protection and confidentiality are handled directly at the corporate level by an authorized division, and they prevented the exploration of other deployment environments or governance models within the scope of this work.

1.6 Thesis structure

This thesis is organized into seven chapters, progressing from the organizational problem and research objectives to the design, evaluation, and discussion of Talent Assistant.

Chapter 2 reviews the background and related work relevant to AI-supported hiring and interview preparation. It discusses the use of AI in recruitment, the risks associated with LLM-based systems in hiring contexts, the role of human oversight, approaches for expert evaluation, and the quality dimensions that characterize useful interview question sets.

Chapter 3 translates the identified problem and background into requirements for Talent Assistant. It specifies the intended purpose and scope of the system, its functional and non-functional requirements, and the output quality dimensions that guide the generation of materials for interview support.

Chapter 4 presents the evaluation design and methodology. It describes how the system is assessed through feedback from technical interviewers and through field-usage data, outlining the measures and analysis approach used to address the research questions.

Chapter 5 describes the design and implementation of Talent Assistant. It presents the overall architecture, the user interaction flow, the HITL configuration, and the main technical choices behind the system's generation pipeline.

Chapter 6 reports and discusses the evaluation results. It combines survey responses, usage metrics, and observed issues to examine the perceived usefulness, quality, consistency, and practical integration of Talent Assistant in the interview-preparation workflow.

Finally, Chapter 7 concludes the thesis by answering the research questions, discussing the study's contributions and implications, and identifying directions for future work.

The appendices provide supporting material, including the research interviews, evaluation instrument, and detailed survey results.

Chapter 2

Background and related work

2.1 AI-assisted hiring and interview support

AI is becoming increasingly used in recruitment pipelines, from job-resume matching and applicant tracking to conversational agents and automated interview analytics [26, 46]. AI systems intended for recruitment or selection, such as systems that analyze applications or evaluate candidates, are listed as high-risk uses under the EU AI Act [22]. This makes the distinction between decision automation and bounded preparation support particularly important.

Within the recruitment pipeline, AI is most present in finding candidates, screening them, and handling interactions during the initial interview stages. Machine learning models and natural language processing are widely used to analyze CVs and match them to job descriptions (JDs), extracting skills, experience, and education to produce ranked candidate lists or "fit scores" for recruiters [26, 46]. Applicant Tracking Systems (ATS) increasingly integrate these models as embedded services, allowing large organizations to process hundreds or thousands of applications per open role with very little manual effort [28]. Conversational agents and chatbots have also become common in pre-screening and candidate communication, where systems use chatbot interfaces to handle short assessments, or to ask standardized screening questions [47, 19].

Other than for early screening, AI is used to support or partially automate psychometric assessment and candidate evaluation. Asynchronous video interview platforms, for example, can record candidate answers to standardized questions and then apply machine learning models to extract both verbal and non-verbal features, which are then fed into predictive models of "hireability" or job performance [19, 25]. While such systems can increase consistency and scalability, studies have raised multiple concerns about validity, bias, and explainability when automated scores directly influence shortlisting decisions, opening an issue of "trustworthy AI" in recruitment and on governance mechanisms that preserve human oversight [24, 40].

Within this broad ecosystem, interview-specific AI is still under ongoing development. Many deployed tools focus either on candidate-side coaching or on downstream analytics after interviews have been conducted, rather than on the preparatory work that interviewers perform before meeting a candidate [61, 32]. Some systems, such as "EZInterviewer" [32], provide mock interviews that automatically generate questions, simulate recruiter behavior, and give feedback so that candidates can rehearse in a realistic but low-stakes environment. Others, such as "InterPilot" [61], are tools meant for recruiters that treat the interview as a human-led interaction, and interestingly aim to reduce interviewer workload during the conversation itself. InterPilot offers intelligent note-taking, adaptive question

suggestions, and skill–evidence mapping in real time, with evaluations suggesting that it reduces documentation burden, but also that it can increase tensions around attention management and trust when the AI proposes highly specific technical questions that interviewers cannot easily verify on the spot. At the planning stage, interview-assistant systems have been proposed that automatically select an optimal set of technical questions from a question bank [16], modeled as an integer linear programming problem to maximize coverage of desired knowledge areas under time constraints. These systems are explicitly framed as aids for constructing a structured interview plan for a given candidate and role, with the human interviewer remaining responsible for how questions are asked and interpreted. LLMs have been used for simulating interview dialogues or multi-party interactions, for example, by orchestrating multiple LLM agents to play both candidate and recruiter roles to generate realistic interview transcripts or to perform two-sided evaluation of candidate–job fit [17]. These systems demonstrate the feasibility of automating realistic interview-style conversations, but currently they are mainly used for research, training, or candidate practice rather than as decision-support tools embedded into real organizational workflows.

The question generation component is a core capability behind many of these systems meant for interviews, and there are a lot of examples, both inside and outside recruitment, of AI-assisted question generation and its evaluation. In education, for example, AI systems are used to generate assessment questions aiming towards specific learning outcomes. They often adopt quality dimensions such as relevance to the source material, tailored complexity level, clarity, and diversity, and they can rely on expert rubrics and mixed-methods evaluations to assess output quality [62, 13]. Similar dimensions are those required in recruitment contexts, where generated interview questions would need to be specific to the candidate and role, calibrated to the candidate's experience, possibly non-redundant, and free from obvious factual errors or inappropriate topics. However, compared to the education environment, fewer recruitment-focused studies provide a detailed, rubric-based analysis of question quality in terms that can be mapped directly to interviewer practice.

Across this landscape, another important design choice is whether AI aims to automate assessments and decisions or only to support human recruiters. Automated scoring and ranking systems for CVs, or video interviews, leave almost no human control: their outputs are often used directly as decision criteria, and documented issues around bias and validity have led to strong calls for transparency, accountability, and meaningful human oversight in line with the EU AI Act and broader AI-governance principles [22, 24, 40]. By contrast, decision-support systems are designed to preserve human agency by presenting AI outputs as inputs the human can build from, rather than having them as final verdicts. InterPilot, optimization-based interview-planning tools, and LLM-based systems for question generation that output candidate-specific question sets all fall, to some extent, into this category: they can structure or enrich interview preparation, but recruiters are expected to review, adapt, and select questions before and during the interview [61, 16, 17]. Talent Assistant belongs to the decision-support category, as it aims to ease the preparation stage of structured interviews, with interviewers being in charge of which questions to ask, how to adapt them, and how to interpret candidate answers.

2.2 AI and LLM risks in job interviewing contexts

As mentioned, the hiring landscape is classified as a high-risk environment according to the EU AI Act. In this context, LLMs introduce additional and partly distinct risks compared to traditional algorithmic scoring, and they should be understood when reasoning about

the safe deployment of AI tools.

A first major concern is about bias and fairness. Studies on automated CV screening and LLM-based matching have shown that models can reproduce or even amplify existing societal biases, for example, disadvantaging candidates with unconventional career paths or whose profiles hint at disabilities, when ranking CVs for shortlisting [24, 28]. Other studies highlight how biased training data, unclear feature representations, and the lack of explicit debiasing mechanisms can result in unfair treatment or impact across certain groups, and state that although fairness metrics and mitigation techniques are necessary, they might not be sufficient safeguards [11, 26, 40]. In an interview setting, these problems can show both in the criteria based on which candidates are invited to interview and in what they are asked during it: if AI-driven pipelines systematically emphasize certain types of experience or "cultural fit", they may steer interviewers towards criteria that can indirectly disadvantage specific groups, even when individual questions appear neutral [26, 40].

Another risk associated with generative AI is hallucination and limited content validity in LLM-generated outputs. By design, LLMs generate fluent text based on statistical patterns rather than grounded knowledge, and they can therefore produce outputs that contain factual inaccuracies, misinterpret documents, or rely on details that do not actually exist. In interview question generation, such hallucinations can manifest in the form of questions about skills or experiences that the candidate did not mention in the CV or that are not even required by the job position, general incorrect statements, or improper evaluation of the candidate's answers. Although these errors may not directly decide who is hired, they can distort what is being assessed during the interview, waste time on irrelevant topics, or create an impression of sloppiness and unfairness toward candidates if they notice that questions are based on false premises [25, 19].

The next issue presented is that of traceability. Traceability refers to the challenge of understanding the mechanics that led to a given AI output, so that the system's behavior can be audited and explained afterwards [29]. But once prompts and model parameters are fixed, it is difficult to fully predict or constrain the exact questions an LLM will generate, and small changes in inputs can lead to substantially different outputs. From a governance perspective, this raises questions about how to ensure that generated interview content remains within organizational policies (for example, avoiding sensitive topics such as family status or health), and how to log and reconstruct what the system proposed for an interview if concerns arise later [22]. In black-box video interview analytics, similar traceability problems have been documented: recruiters may receive numerical "hireability" scores without a clear explanation of which behavioral features actually drove them, making accountability more complex when decisions are challenged [25, 19]. For LLM-based question generation, traceability implies recording prompts, model versions, and generated artifacts in a way that supports later auditing while respecting privacy constraints [22, 30].

Confidentiality and data protection risks are also sensitive topics, especially when working with highly sensitive data, such as in the recruiting environment. CVs and interviews often include personal and private data, and feeding any of these through external services, including LLM providers, can raise concerns about data residency, retention, and possibly secondary use for model training [30, 40]. Studies focused on policy in AI in hiring contexts state that high-risk systems must comply not only with general data protection law but also with sector-specific expectations around confidentiality and professional secrecy, especially in domains such as consulting, where client information can be included in candidate projects [22, 26].

Risks around transparency, explainability, and poor candidate experience all cut across

the previously mentioned categories. When candidates are not clearly informed that AI is being used in generating questions or structuring interviews, they might perceive the process as not being transparent, especially if they encounter repeated or strangely phrased questions that do not seem aligned with their profile or with the position they applied for [47, 19]. Regulators and academics have argued that high-risk AI systems in hiring should at least provide high-level explanations of the role AI plays in the decision process, and that there should be mechanisms for candidates to contest outcomes or request human review [22, 40]. In the context of interview support, this could raise questions about whether and how to disclose AI involvement in question preparation, how to maintain consistency across interviewers using and not using such tools, and how to ensure that generated content respects organizational values and legal constraints.

2.3 Human oversight

A human-in-the-loop (HITL) system is one where AI does not act autonomously from input to decision, but rather it produces outputs (in the form of predictions, rankings, explanations, and so on) that human operators can review before deciding what to do [26]. Typical tasks for the human in these setups include checking whether the AI output is plausible, correcting obvious errors, adapting generic outputs to the specifics of a case, and deciding whether to accept, partially use, or completely discard the system's suggestions [26]. The EU AI Act states the importance of this control being effective, and not just a formality: human reviewers should have the competence, time, and interface support needed to understand the role of AI in the workflow and to meaningfully intervene when something appears wrong [22].

More specifically, literature distinguishes among different configuration modalities of human oversight in algorithmic decision-making [20]. In "human **on** the loop" configurations, the system operates mainly automatically, and humans monitor aggregate behavior or a stream of cases, intervening only when alarms are raised or when there are obvious anomalies. In "human-**in**-the-loop" configurations, humans are systematically involved in each decision, for example, by reviewing every proposed output, editing it, or explicitly approving it before it is used. A third configuration is that of humans being "in command", where people are responsible for deciding whether and how an AI system may be used at all in a given context, for example, by setting policies and thresholds, but then not necessarily reviewing individual outputs. In particular, among the recruitment and interviewing tools that generate question sets or provide real-time suggestions during interviews, the typical configuration is the HITL model, with the systems producing artifacts for each candidate, and interviewers being expected to review, adapt, or reject these artifacts before and during the interview [16, 61].

HITL practices have been used in several areas of machine learning and decision support, for example, in interactive recommending systems, medical diagnosis support, credit risk assessment, and content moderation, where human experts validate or correct model outputs before they have external effects [26]. In these domains, the human review step is well established: humans are assumed to bring domain knowledge, contextual information, and judgment that models are not expected to have, and to be able to catch errors or norm violations that can be missed by statistical systems [26, 20]. Job interviewing is a field that shares many of these traits: decisions are very important and context-dependent, where there can be AI systems to improve structure and efficiency, with humans remaining responsible for assessing candidates and ensuring that the process aligns with legal, ethical, and organizational norms [40].

By design, a HITL system can help mitigate some risks, including some of those outlined in Section 2.2. Human review may help identify hallucinations and factual errors, with their effectiveness depending on reviewer expertise, time, and interface design, and possibly notice when AI-generated content systematically privileges certain backgrounds or traits, potentially mitigating some fairness issues even though the underlying model is not fully unbiased [60, 26, 40]. Other than that, HITL designs have a much clearer accountability: if AI outputs are advisory rather than binding, responsibility for final decisions remains on the human decision-makers, which aligns with both legal requirements and organizational norms in high-stakes domains [22, 20]. However, HITL systems also come with many limitations, for example, when people are placed in a position to check AI recommendations, they can exhibit automation bias, meaning they are more likely to accept algorithmic suggestions than to challenge them, even when those suggestions are inaccurate or conflict with other available information, with this being also documented in recent experiments where biased LLM hiring recommendations systematically shape human decisions and reduce their autonomy [40, 59]. In some experiments, introducing a HITL increased the acceptance rate of algorithmic recommendations but actually decreased overall decision accuracy, with participants intervening more on relatively accurate recommendations and less on correcting the least accurate ones [50]. Similar analyses come from situations where decision support systems provide risk scores or rankings: users have the tendency to see the model as more objective or informed than themselves, and consequently align with it, especially under time pressure or heavy workload [26, 40]. Another risk is the one of gradual deskilling: if interviewers always rely on AI-generated content over the long term, they may lose their sharpness at structuring interviews themselves and become more likely to internalize the model's framing of what "good" interviews look like [26, 40].

Overall, whether an HITL design is effective or not depends strongly on contextual factors. Oversight is more likely to be meaningful when humans have domain expertise, understand the strengths and limitations of the system, have access to relevant contextual information (which in the hiring interview support environment can be full CVs, for example), and are given enough time and suitable interfaces to properly review AI outputs [26, 20]. Instead, when review is rushed, when users are not properly aware of how the system works, or when interfaces make it much easier to accept AI suggestions than to question them, HITL configurations can lose their strengths, with humans remaining "in the loop" in theory but, in practice, rarely overriding the system [20, 50].

2.4 Expert evaluation

Evaluating an AI interview-support system, such as Talent Assistant, is different from evaluating a fully automated scoring model. The system is meant to support experienced interviewers inside an existing hiring process, not to replace them, so there are several plausible evaluation strategies that focus on expert judgment, interaction quality, and real-world use rather than only on predictive accuracy [34]. Before selecting an evaluation approach appropriate to the system and its deployment setting, it is useful to explore the main strategies used for comparable systems, including their strengths, goals, and weaknesses.

A first option is to run controlled experiments with candidates, where outcomes are compared between conditions with and without AI support [32]. In mock-interview systems for candidates, this has been done by randomly assigning students or job seekers to practice with an AI coach or a baseline condition, then comparing later interview scores,

self-reported preparedness, or confidence between groups. In principle, a similar design could be applied to a tool for interviewers: some interviews would be prepared using Talent Assistant and others using current practice only, and candidate ratings or hiring decisions would be compared across conditions. The main strength of this approach is that, under suitable conditions, it can provide relatively strong causal evidence about the effect of AI support on measurable outcomes at the candidate level [27]. However, it requires access to large numbers of comparable candidates, adequate control or randomization of preparation practices, and a willingness to expose real candidates to different preparation conditions over time (all of which is especially complex to implement in corporate hiring).

A second kind of approach could be to ask for expert evaluations [27]. This design invites people who have actually used the tool to reflect on its usefulness and implications, and to provide structured survey and qualitative feedback [49]. In the context of Talent Assistant, this would mean that expert interviewers exposed to the tool would be asked to score its outputs on relevant dimensions and on its practical usefulness, alongside any other relevant comments. The main strength of an expert evaluation is that judgments are based on real usage and workflow constraints [49], which in Talent Assistant's case would mean that experts would rate the tool based on situations where they actually had to integrate it into their preparation under time pressure, with real candidates and real organizational expectations. At the same time, this approach comes with limitations: it usually involves small samples of people from similar (or the same) settings; evaluations are retrospective and subjective, and they can be influenced by a few salient positive or negative cases [7, 12]. On their own, these studies say more about perceived quality and usefulness than about long-term impact on hiring outcomes, so they should be intended as formative evidence rather than as definitive performance claims [48].

Another evaluation design system is to run task-based user studies, in which professionals use the system on realistic but still simulated tasks. Some domain-relevant examples, such as InterPilot [61], an AI-assisted system for HR professionals during interviews, were evaluated by asking interviewers to conduct simulated interviews using the tool, while researchers observed how they interacted with the interface and collected standard usability, workload, and perceived-support measures. For Talent Assistant, such a study could involve asking managers to prepare interview plans for fictional or anonymized candidates using Talent Assistant, logging which features they use, how they edit or ignore generated questions, and how long they take, followed by standard usability scales and open-ended questions about trust and workload. Generally, task-based user studies are strong at revealing interaction issues and trust dynamics, such as parts of the workflow where the tool does not fit, elements that cause confusion, or points where users hesitate to follow AI suggestions [53]. Their main limitations are that they usually involve a small number of participants, are based on simulated or reconstructed cases rather than live interviews, and are very labor-intensive to run and analyze [49].

One more evaluation strategy used in other domains would be to rely on log-based observational and field studies of deployed systems, sometimes combined with user feedback. In clinical and care settings [42, 31], for example, decision-support systems have been evaluated by instrumenting them in real use, collecting statistics on how often and in which situations they are used, which suggestions are accepted or overridden, and how long tasks take, and then combining these data with interviews or surveys of professionals about perceived usefulness and impact on their work. For interview-support tools, a similar approach might track how many interview packages are generated with Talent Assistant, for which roles and seniority levels, and more, depending on available logs. This log-based observational approach is strong because it captures how the system behaves under real or-

ganizational constraints and competing priorities, but it cannot answer detailed questions about output quality or fairness [31].

Because each of these strategies focuses on different aspects of system behavior, evaluations of tools sometimes combine elements from several of them in mixed-method designs [48, 34]. The appropriate choice often depends on resource availability, on context, and on the result the study aims to achieve.

Regarding the research sample size, larger samples are required when the goal is to estimate effect sizes with precision and to draw statistically powerful, generalizable conclusions about a broader population, because sample size directly influences both the margin of error and the probability of detecting true effects [51]. When talking about a system's evaluation, smaller samples, especially if based on expert users and rich contextual data, can also be suitable for formative aims, to get an initial understanding of its strengths and weaknesses, and understand interaction quality, but they cannot support precise hypothesis testing or strong claims about external validity [55, 57].

2.5 Quality dimensions for interview question sets

Because one aim of this thesis is to evaluate the quality of generated question sets, this section identifies the key dimensions used for that purpose. These dimensions are based on the problems described, the expectations and practice of managers at PwC Italy as captured in the research interviews in Appendix A, and prior work on structured interviews [33] and AI-assisted question generation [41]. The study focuses on the following main dimensions:

Relevance

Relevance captures the extent to which questions are clearly job-related and grounded in the specific role and candidate profile, instead of being generic or only loosely connected to the position. This dimension reflects findings from the structured interview literature, which states that questions should be derived from a job analysis and target competencies that predict performance, as well as PwC Italy's managers' emphasis (Appendices A.1, A.3) on questions that connect directly to the JD and to what candidates have actually done. PwC managers (Appendices A.1, A.3) confirmed also that the problem is relevant, describing a tendency to reuse a narrow set of standard questions across candidates.

Specificity

Specificity is about how concrete and precise questions are, and whether they aim to gather precise information rather than high-level answers. Structured interview guidelines and the question-generation study also identify this concept as an important element to check quality, recommending questions that refer to identifiable situations and avoiding vague prompts, which aligns with how managers describe "good" questions as clear and related to real projects (Appendices A.2, A.3).

Calibration

Calibration refers to the match between the difficulty and depth of questions and the candidate's seniority level and target role. Both the interviews with managers (Appendices A.1, A.3) and the work on structured interviews highlight that there are different expectations

for different seniority levels, and so miscalibrated questions are either trivial or unfairly advanced.

Coverage

Coverage aims to check whether the interview question set covers the core relevant macro-areas that should be assessed for a given role, such as the candidate's past experience, their personal values and motivation, and their skills. This wraps around the current practice described by managers (Appendices A.1, A.2, and A.3), and aims to replicate their operational flow as closely as possible, since Talent Assistant is intended as support.

Diversity

Diversity is about the extent to which questions cover complementary aspects of the candidate and the job position, rather than asking for similar information in slightly different forms. Question-generation research treats diversity, alongside relevance and specificity, as an independent quality dimension, showing that redundant questions waste limited interaction time and reduce usefulness.

Reliability

Reliability refers to the correctness and plausibility of the information in the questions and model answers—especially technical content—and to their interview readiness, meaning the amount of checking and revision required before use. This is informative because more reliable generations can save the user time and effort, while very unreliable questions risk being useless or even counterproductive.

Taken together, these dimensions articulate what it means for a generated interview question set to be good in this organizational setting: they should be aligned with the role and candidate (relevance, specificity, calibration), provide a structured yet flexible backbone for the interview (coverage, diversity), and possibly alleviate the time and effort pressures documented in current practice (reliability).

Chapter 3

Requirements specification

3.1 Purpose and scope of Talent Assistant

Talent Assistant will be designed and developed as a HITL interview preparation system to support technical interviewers at PwC Italy in speeding up the structuring of the interview backbone, with its scope being restricted to the preparation of late-stage technical interviews for roles up to "manager" level, without automating candidate ranking or hiring decisions. The idea is that manual interview preparation currently requires scarce expert time and leads to variability in question quality and topic coverage across interviewers and candidates. Talent Assistant is intended to shorten the path from a candidate's CV and the relevant JD to a usable interview-preparation package. It should provide a structured starting point while preserving the interviewer's autonomy to review, adapt, or disregard the generated material. The system should also support the quality, privacy, and governance constraints associated with interview preparation in a hiring-related setting.

Operationally, the system will have to start from a candidate CV (in digital form), a JD associated with a specific role, and, optionally, a seniority level, and from these inputs, produce two main artifacts: a summary of the candidate profile, providing an overview of the candidate's background, experience, skills, and alignment with the target role; and a set of interview questions, including personal (behavioral and experience) ones and technical (theoretical and practical) ones, relevant for the candidate and the role, and calibrated to the candidate's seniority. Any other auxiliary features able to improve the user experience should be considered worth adding.

3.2 Functional requirements

Building on top of the identified domain needs, this section sets down Talent Assistant's functional requirements in detail, grouping them into three areas: input handling and validation, interview artifact generation, and processing and usage.

Input handling and validation

1. *CV ingestion and text extraction:* The system shall ingest candidate CVs in PDF format and robustly extract their content, handling multi-page documents and common layout patterns.

2. *Job description management:* The system shall allow users to enter and update JDs, as well as access them in the future. JDs are expected to include a text description section and a key requirements section.

3. *Seniority handling:* The system shall support four seniority levels (intern, associate, senior associate, and manager) that are relevant for the targeted part of the PwC Italy hiring pipeline. For each candidate, Talent Assistant should be able to infer an appropriate seniority level from the CV and JD, but also to accept a user-provided override. The resulting seniority level must be considered to set question difficulty and focus consistently.

4. *Language configuration:* The system shall accept a target language parameter for the generated content, and all generated artifacts must be produced in the specified language.

Interview artifact generation

1. *Candidate profile overview:* Given a validated CV, Talent Assistant shall generate a structured candidate profile that includes basic personal details (if available), a summarized academic and professional history, relevant projects and experiences, a skill summary, and a high-level mapping between the candidate's profile and the JD's relevant traits. This panel is intended to give interviewers a compact but informative overview they can use to orient themselves quickly before selecting questions.

2. *Experience-focused questions:* The system shall provide a set of questions about the candidate's past experiences, based on their CV, and designed to elicit insights appropriate to their seniority. For junior candidates, questions should emphasize academic and internship experiences; for more senior candidates, they should target professional achievements, impact, and increasing responsibility.

3. *Behavioral and soft skills questions:* The system shall provide behavioral questions that probe soft skills and ways of working (e.g., communication, collaboration, dealing with uncertainty). These questions should be aligned with the JD's expectations and the candidate's background.

4. *Motivational questions:* The system shall provide motivational questions exploring the candidate's reasons for applying, longer-term aspirations, and fit with the role and organization, again conditioned on the JD and seniority level.

5. *Technical theoretical questions:* The system shall provide theoretical questions targeting domain-specific concepts drawn from the JD and possibly matched by the candidate's profile. Question difficulty and depth must scale with seniority: from basic conceptual understanding at junior levels to deeper, cross-cutting knowledge and judgment at senior levels.

6. *Practical exercises and scenarios:* The system shall provide short practical domain-relevant exercises. Exercises should be reasonably addressable within a few minutes, keeping in mind the interview time is limited. Exercises must be domain-specific (e.g., coding exercises for software engineers, financial modeling for analysts, and comparable role-specific exercises) and tailored to the candidate's seniority.

Processing and usage

1. *Interview configuration and triggering:* The user interface shall allow interviewers to easily configure a generation run by uploading a CV, selecting a JD, specifying (or accepting a suggestion for) seniority, and choosing the target language.

2. *Persistent storage of artifacts:* Relevant input and output information shall be stored in a database to support retrieval, traceability, and later analysis.

3. *Access control by space:* All read and write operations shall be restricted according to the user's authorized spaces. Users shall only be able to access candidates, JDs, and interview artifacts associated with spaces to which they belong.

4. *Agentic orchestration:* The agentic orchestration structure shall rely on specialized agents for each area, and allow parallel execution where possible to minimize latency.

5. *Model-agnostic LLM integration:* It shall be possible to change the underlying model configuration (e.g., switching to a different LLM) without altering the contract with the frontend or the overall pipeline. This allows easy integration of newer, better-performing models as they are released, without altering the stable functioning of the application.

3.3 Output quality requirements

After having described what the system shall produce via functional requirements, output quality requirements are built on top of the quality dimensions defined previously in Section 2.5, translating them into concrete expectations for Talent Assistant's outputs, linking back to RQ2.2.

Relevance

Questions must be clearly job-related and grounded in both the JD and the candidate's CV, rather than generic questions that could be asked in any interview. For example, experience questions should refer to specific projects or responsibilities mentioned in the CV; technical questions should be drawn from the technologies, domains, and tasks specified in the JD; and motivational questions should connect to the role's actual content and growth opportunities.

Specificity

Behavioral questions should reference realistic situations, and technical questions should be precise enough that an answer can be evaluated as more or less complete, rather than eliciting only high-level opinions.

Calibration

The aggregate difficulty and depth must be aligned with the candidate's seniority level. For interns and associates, questions should primarily test foundational understanding and problem-solving approaches; for senior associates and managers, they should probe deeper expertise, ability to prioritize and make trade-offs, and (for managers) elements of technical leadership and coordination. Miscalibrated questions that are consistently trivial or unfairly advanced can be considered quality defects.

Coverage

The generated interview package must cover the core macro-areas that interviewers typically consider essential for a given role: past experience, technical skills, soft skills and

ways of working, and motivation and values. This coverage requirement is not intended to impose a rigid script but rather to ensure that the backbone proposed by the system gives interviewers a structured starting point that mirrors the flows they described in the research interviews.

Diversity

Within each subset of questions, and across the entire interview preparation package, questions should target complementary aspects of the candidate and role, avoiding asking for essentially the same information. Redundancy can be acceptable only when it serves a clear purpose (e.g., probing the same area from a different angle).

Reliability

Questions and model answers must be free from obvious factual errors, contradictions, and hallucinations, especially in technical content. Questions must not presuppose skills, experiences, or constraints that are not supported by the CV or JD, and model answers must reflect plausible, correct solutions at an appropriate level of detail.

In general, generated questions and model answers should be close to "interview-ready", meaning that an experienced interviewer can adopt them with only minor edits for wording or emphasis. If outputs require heavy adjustments systematically, Talent Assistant's effectiveness may decrease or even disappear.

3.4 Non-functional requirements

Other than functional behavior and output quality, Talent Assistant must also satisfy other non-functional requirements given the constraints and expectations of its deployment context within PwC Italy, which will ultimately affect whether the tool is adopted in practice or not.

Usability

The application has to fit smoothly into interviewers' existing preparation routines: the interface should be simple, consistent with familiar web applications, and make core actions such as uploading the CV, selecting a JD, triggering generation, and reviewing the resulting artifacts discoverable with minimal navigation and without requiring much training. Also, end-to-end generation time for a full interview support package (candidate profile and questions) has to be sufficiently responsive for use during routine interview preparation, including when preparation takes place shortly before an interview, or when it is squeezed between other activities.

Availability

Reliability and availability are important because recruitment processes can involve tight timelines and clustered interview periods. The system is expected to behave stably during these peaks, avoiding failures such as lost sessions or incomplete generations.

Governance

Handling CVs, JDs, and interview artifacts that contain personal and potentially sensitive information imposes strict requirements around privacy, confidentiality, and data protection. Access to Talent Assistant must therefore be limited to authenticated users, governed by space-based access control (as mentioned), and aligned with corporate guidelines on data residency and retention, with LLM calls routed through private and governed endpoints rather than public services. To support both governance and the analysis of issues and failure modes, the system also needs to provide traceability through logs showing which inputs and outputs are associated with each use.

Other non-functional requirements

Other non-functional aspects are those of maintainability, scalability, and cost, which can influence adoption independently of model performance. The architecture and codebase should be structured so that prompts, agents, and storage schemas can be updated or extended with limited impact on the rest of the application. This should allow the system to accommodate changes in organizational requirements and in the underlying LLM infrastructure. At the infrastructure level, the system should support increased usage without unacceptable latency while keeping computational and storage costs manageable. Finally, compatibility with the firm's existing cloud environment, authentication mechanisms, and deployment, including the ability to operate separate test and production instances, are also important, so that Talent Assistant can become a supported internal tool rather than remaining just a prototype.

Chapter 4

Evaluation design and methodology

4.1 Study design overview

The aim of the evaluation conducted in this study will be to assess Talent Assistant as an interview-support tool within the PwC Italy hiring process, with a particular focus on the aspects captured by RQ2, such as perceived usefulness, output quality, and alignment with existing interview practices. The study is intended to show not only whether Talent Assistant is considered helpful by interviewers, but also which concrete obstacles and enablers appear when the tool is integrated into day-to-day technical interviews.

The evaluation uses a hybrid design that combines elements of expert evaluation with elements of log-based field studies. Expert evaluations rely on domain experts who systematically inspect a system and rate it along predefined criteria, instead of conducting large-sample user experiments. Log-based studies, on the other hand, draw on naturally occurring usage traces to describe how a system is used in practice and to uncover patterns that can complement self-reported data. In this project, this hybrid design is a pragmatic response to the availability of a small yet highly relevant group of technical interviewers, the limited time and availability of business stakeholders, and the internal organizational setting of the deployment, which together make large-scale quantitative testing less feasible. The study instead aims to obtain practice-grounded evidence.

Within this design, the expert evaluation forms the main component of the study. Ten technical interviewers who have actually used Talent Assistant in live interview processes at PwC Italy will be invited to rate the tool using a structured rubric that directly builds on the evaluation dimensions introduced in Section 3.3 and, through them, on the research questions. The number of available technical interviewers is limited by the company's policy guidelines, which create frictions between teams to share resources without proper formalized resource expenditure and usage tracking, and because of this, the accessible population comprises all ten interviewers who have been granted an operational account at the moment of the study. The rubric also includes a small set of additional questions that aim to gather information on how interviewers experience Talent Assistant. This expert perspective is expected to yield detailed and context-aware judgments about the tool's strengths and weaknesses that would be difficult to obtain from generic usability tests alone.

The second component consists of a moderate log-based analysis of basic usage statistics from the internally deployed system, used as field-usage evidence to complement the expert ratings, illustrating how Talent Assistant is actually taken up in the PwC Italy hiring pipeline under the given organizational constraints.

4.2 Measures

As mentioned, the evaluation combines two main sources of evidence: a structured survey filled in by technical interviewers who have actually used Talent Assistant in real PwC Italy processes (Appendix B), and log-based field usage metrics from the deployed system.

In Sections B, C, and D of the survey, all closed-ended questions use the same seven-point Likert scale, from 1 ("strongly disagree") to 7 ("strongly agree"). Interviewers are asked to treat the middle point 4 ("neither agree nor disagree") as roughly equal to how they would rate their current way of preparing and conducting interviews without Talent Assistant, so scores above 4 generally indicate an improvement over their baseline practice, and scores below 4 indicate shortcomings. This anchoring allows the participants to interpret rubric scores directly in terms of perceived added value or degradation compared to existing practice.

Section A collects basic background information about each interviewer, such as their position at PwC Italy, service line and team, main technical domains, approximate years of experience with technical interviews, and the number of interviews they handle per year. It also asks which roles they interview for (from intern to manager), how many candidates they have interviewed using Talent Assistant, in which phases of the process they have used the tool (before, during, after the interview), and how familiar they feel with Talent Assistant overall. These background questions help contextualize the participants and how close they are to the part of the pipeline that Talent Assistant affects, while practically supporting exploratory checks that help interpret the main results and threats to validity.

Section B focuses on perceived usefulness and efficiency, and it is the main way in which the survey measures RQ2.1. It includes five statements about how Talent Assistant affects preparation time, perceived effort, and the overall value of the tool for interviewers. Items B1 and B2 capture perceived gains in efficiency: B1 asks whether Talent Assistant helps interviewers prepare interviews more quickly than they otherwise could, and B2 asks whether it helps them prepare more efficiently, with less effort on their side. These two items connect directly to the problem described in Chapter 1, where interview preparation is described as time-consuming, often pushed into evenings, and competing with billable work. Items B3, B4, and B5 move from pure time and effort to perceived added value coming from the tool. B3 asks whether Talent Assistant helps generate relevant technical questions or prompts that interviewers might not have thought of on their own, reflecting the idea that the tool can broaden the question set and reduce reliance on the same generic prompts. B4 asks whether Talent Assistant improves the quality of technical interviews overall, and B5 asks whether the tool adds value compared to interviewers' previous way of preparing and running interviews without it.

Section C is the main rubric for output quality and directly reflects the quality dimensions defined earlier in Section 3.3: relevance, specificity, calibration, coverage, diversity, and reliability. Items C1 and C2 focus on relevance, checking whether the tool incorporates the information extracted from both the CV and the JD appropriately to produce the questions. Item C3 captures specificity, checking whether questions are concrete enough that candidates can give detailed answers, in line with structured interview guidelines that recommend asking about identifiable situations rather than vague opinions. Items C4 and C5 are intended to measure calibration. C4 asks whether the questions are appropriate for the candidate's seniority level (not too simple and not too complex), and C5 asks whether they focus on the right aspects for that seniority, such as problem-solving approaches for juniors and deep expertise or technical leadership for more senior figures. Items C6 and C7 capture, respectively, coverage and diversity. C6 asks if, taken together, the questions for a

given candidate cover all the core macro-areas that should be assessed for that role, and C7 asks whether the questions target complementary aspects without overlapping too much. Items C8 and C9 focus on reliability and overall interview readiness. C8 asks whether the questions and prompts are technically correct and free from major conceptual errors, which is the survey-level way of checking the reliability requirement that outputs should not contain obvious hallucinations or contradictions, especially on technical content; C9 asks whether the generated questions can be considered "interview-ready with minor or no adjustments". Finally, item C10 asks interviewers how satisfied they are with the quality of the content that Talent Assistant produces, aiming to sum up the overall satisfaction level across all categories. Taken together, scores on Section C form the primary quantitative answer to RQ2.2 on output quality, again using 4 as an approximate reference benchmark. In addition, because different interviewers may rate outputs for similar roles or comparable candidates, we can later look at how much these scores vary across raters to say something about perceived consistency as well.

Section D items also use the same 1–7 scale and focus on other relevant areas that have not been covered yet, including ease of use, control, and consistency. D1 asks whether learning to operate Talent Assistant was easy, D2 asks whether the tool fits reasonably well into the interviewer's existing preparation workflow, and D3 asks whether the interviewer still feels fully in control and responsible for which questions and evaluation criteria they use in the interview. These items connect to non-functional requirements and to the HITL configuration, where it is important to ensure people are included in the pipeline, and in this case, are the ones responsible for deciding what to ask and how to interpret answers. Items D4 and D5, instead, are more directly linked to RQ2.3 on consistency. D4 asks whether, when Talent Assistant is used for comparable candidates in the same role, it tends to suggest similar types of questions across interviews, and D5 asks whether differences in the interviews across comparable candidates mostly reflect deliberate changes that interviewers make rather than random system output variation.

Closing out the survey, section E contains three open questions where interviewers can describe, in their own words, any clear failure modes or issues they have seen when using Talent Assistant (item E1), situations where they prefer not to use it (item E2), and any other relevant aspects of their experience (item E3). Answers to these questions can be integrated into the discussion of the results and the conclusion.

Finally, the evaluation uses field usage metrics taken from the deployed Talent Assistant instances. These metrics include high-level counts such as the number of interview packages generated, the number of candidates for whom Talent Assistant has been used, the distribution of roles and seniority levels covered, and the usage frequency across the observed time period.

By combining survey data and field usage metrics, the study can link interviewers' perceptions about efficiency (RQ2.1), question quality (RQ2.2), and consistency (RQ2.3) to actual patterns of deployment, and use this combined evidence to reason about obstacles, enablers, and future work in the later chapters.

4.3 Analysis plan

Given the small population of ten expert interviewers, the focus is on summarizing patterns and linking them back to the research questions, rather than on formal hypothesis testing or generalization beyond PwC Italy. The intent is to explicitly relate observed patterns to the quality dimensions and requirements defined earlier in the thesis, wherever possible.

Background information from Section A will be analyzed using simple descriptive statis-

tics and tabular summaries, such as counts, percentages, and ranges for the available variables, aiming to provide a comprehensive overview of the interviewer population, rather than to exhibit outcome measures in themselves. When interpreting results for other sections, Section A variables will not be used to explore patterns, for example, by inspecting whether highly experienced interviewers tend to give generally higher or lower ratings than those with fewer years of experience, or whether heavy users of the tool tend to rate it differently from occasional users, as survey results will be considered in aggregate form only.

For Section B, each item will be summarized with standard descriptive statistics: mean, median, standard deviation, minimum, and maximum. Because the Likert scale is explicitly anchored such that 4 ("neither agree nor disagree") corresponds approximately to the interviewers' own current way of preparing and conducting interviews without Talent Assistant, the simplest and most interpretable analysis is to compare the observed means and medians to this reference point. In practice, items with clear averages meaningfully above 4 will be interpreted as areas where interviewers tend to perceive Talent Assistant as an improvement over their usual practice; items around 4 will indicate that the tool is seen as roughly equivalent to their baseline; and items below 4 will signal perceived shortcomings. To provide a more compact view, two composite indices will also be calculated as averages across items: an "efficiency" index based on B1 and B2 (time and effort), and a broader "perceived added value" index based on B3, B4, and B5. Taken together, these measures provide a perception-based assessment relevant to RQ2.1 by examining whether interviewers experienced gains in preparation speed and effort alongside perceived added value and quality.

For Section C, the analysis will again begin with the same item-level descriptive statistics as for Section B, for each of the ten items. Since Section C is structured to mirror the output quality requirements defined in Section 3.3, items will be grouped accordingly to provide per-dimension summaries: relevance (C1 and C2), specificity (C3), calibration (C4 and C5), coverage (C6), diversity (C7), and reliability (C8–C9), complemented by the overall satisfaction item (C10). Within each group, averages will be used to obtain dimension-level scores, which will again be interpreted relative to the reference value 4. It will also be reported, for each dimension, the proportion of interviewers whose individual scores are above, around, or below 4, other than the overall means, making it easier to see whether strong averages are driven by broad agreement or by a few very positive ratings, which matters given the small population.

Responses to Section D will be analyzed in two complementary ways. First, as with previous sections, each item will be summarized descriptively, with means and dispersion measures interpreted relative to the reference value 4. Items D1 (ease of learning), D2 (tool integration), and D3 (feeling in control) will be used primarily to assess whether the non-functional requirements around usability and the HITL configuration are met in the eyes of interviewers. High scores on these items would suggest that Talent Assistant can realistically be integrated into existing routines with little disruption. Items D4 and D5 (consistency) are instead directly linked to RQ2.3, which covers the consistency of Talent Assistant's outputs across interviews for comparable candidates and roles. Here the analysis will again focus on where the items' combined mean falls relative to 4 and on the spread of ratings: if most interviewers agree that comparable uses of the tool lead to similar types of questions (D4), and that observed differences across interviews mostly reflect deliberate human adjustments rather than unpredictable system behavior (D5), this will be interpreted as evidence that Talent Assistant behaves in a reasonably consistent and controllable way. Conversely, if ratings on these items are closer to neutrality or show

strong disagreement, this will indicate that interviewers perceive the system's behavior as more variable or difficult to anticipate.

Open-ended answers from Section E might provide additional concrete, case-based evidence for the RQs, especially for RQ1.3. Also, they might help interpret the quantitative findings, potentially serving as a textual explanation of the reasons behind some of the scores: for instance, if one quality dimension in Section C receives lower average scores, Section E answers may clarify whether this reflects specific, repeated problems (such as hallucinated technical details) or more general concerns (such as lack of trust in AI-assisted question generation). Given the small number of participants, qualitative evidence will be treated as contextual input to the discussion, rather than as exhaustive coverage of all possible issues.

Field usage metrics will be analyzed descriptively, reporting basic counts and distributions. These metrics cannot, on their own, answer questions about output quality or fairness, nor can they establish causal impact on hiring outcomes. Their main role is to contextualize the survey results and provide field evidence related to adoption and integration into practice.

Chapter 5

System design: Talent Assistant

5.1 System overview and architecture

Talent Assistant is deployed as a web-based internal application within the PwC Italy environment. At a high level, it follows a client–server architecture (Fig. 1) in which a frontend allows interviewers to configure and trigger interview preparation, while a backend service orchestrates the interaction with LLMs, handles data, and applies the logic of the interview support workflow.

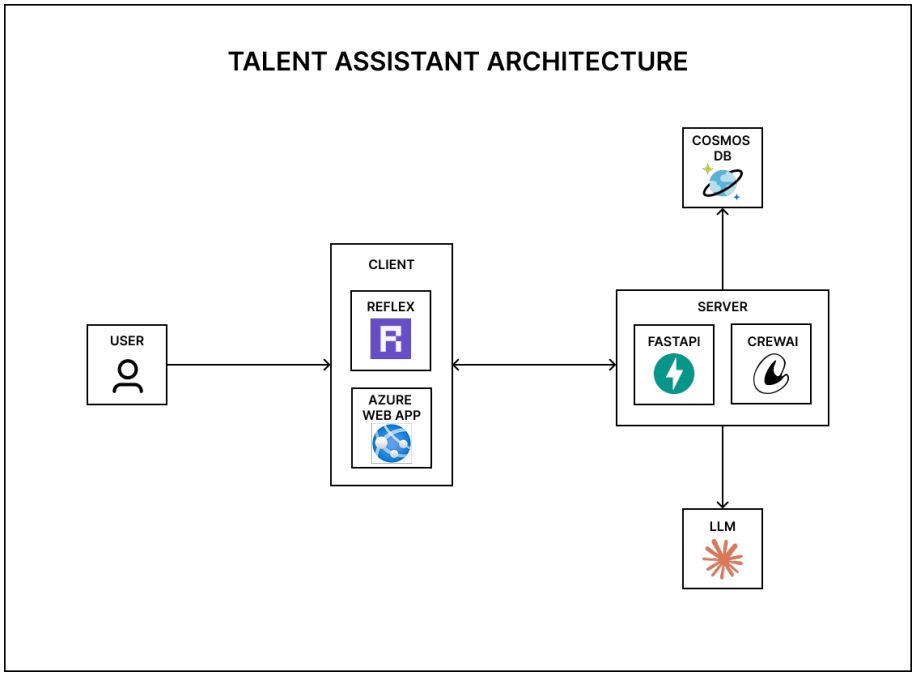

FIGURE 1: Talent Assistant Architecture

The frontend, which is the user interface through which interviewers access Talent Assistant, is implemented in Reflex [45], a framework that allows working in Python [43] while compiling to React [35]. An earlier version of Talent Assistant was built using Streamlit [54], which was good enough for rapid prototyping and initial pilots, but as the tool evolved into a more structured application, intended for longer-term use, the frontend was migrated to Reflex, which is a more adequate choice for web apps requiring clear navigation and state management rather than notebook-style dashboards. The frontend communicates with the

backend through HTTP APIs, sending calls needed to trigger analysis and retrieve content, and receiving structured (JSON) responses. Talent Assistant's visual design is in line with the PwC corporate environment, with a color palette relying on black and orange, and a professional aesthetic suiting the system's operational area.

The backend is built with FastAPI [44] and uses CrewAI [14], Talent Assistant's core framework, to structure the generative AI logic. FastAPI is a web framework for building APIs in Python, designed around standard HTTP and JSON, which uses Python type hints to define and validate request and response models. CrewAI is an open-source agentic AI framework designed to build and coordinate teams of specialized AI agents to solve complex tasks. It is structured around two complementary elements, crews and flows: crews group agents and tasks into autonomous "teams" that collaborate toward a shared objective, while flows provide an event-driven orchestration layer that manages state, triggers, and execution paths. Within a crew, each agent is powered by an LLM and defined by a role, a goal, and a backstory, and is assigned one or more tasks with clear inputs and expected outputs. Flows connect crews to traditional code through events, listeners, and routers, enabling hybrid pipelines where deterministic logic and agentic reasoning are combined in a single structured workflow.

Both backend and frontend are packaged as Docker images [18] and deployed as Microsoft Azure [38] App Service applications, which is a fully managed cloud platform that allows building, deploying, and scaling web applications without having to manage underlying servers or infrastructure. Azure is the standard cloud platform used by the PwC Italy team in which Talent Assistant has been developed, since the corporation has an ongoing strategic partnership with Microsoft [37], and this allows the services to be deployed and operate within the firm's existing infrastructure and governance framework (more on this in Section 5.5). Currently, PwC Italy operates two deployed instances of Talent Assistant, a test one, used to test the proper functioning of new incoming features and to work on bug fixing, and a production one, always ready to be used at any given moment.

All persistent data is stored in an instance of Azure Cosmos DB [39], a fully managed NoSQL database service that provides flexible JSON document storage with automatic indexing and horizontal scalability. This choice follows naturally from using Azure because Cosmos DB is natively integrated with the platform.

The LLM backbone of the system is currently Claude 4.5 Haiku, accessed via PwC Italy's private API system "Shared Service", PwC's internally approved gateway for invoking AI models when processing confidential material. Haiku is a relatively fast and cost-efficient model that still provides competitive performance for the kind of structured, text-heavy outputs required in interview preparation. The architecture has been built to be model-agnostic, meaning that the backend can easily be reconfigured to call different LLMs if organizational constraints or performance requirements change (especially given the fast pace at which new LLMs are being released as of 2026), without altering the overall flow or user experience.

Overall, this architecture is designed to support the requirements identified earlier in the document. It centralizes the interview preparation logic in a single backend service, stores interview artifacts (and more) in an easily accessible database, and enables a consistent and usable interaction flow by providing a frontend built in Reflex.

5.2 UX flow and human-in-the-loop

Talent Assistant is deployed only as a web-based internal application within the PwC Italy environment. The interface is currently available only in Italian, and user accounts

(username and password) are created and managed by the development team upon being contacted directly. Once users access the link and authenticate, they land on the home page (Fig. 2), which presents a sidebar navigation menu with three main sections: "Interviews", "Candidates", and "Job Descriptions", and lists the logged-in username, accessible spaces, and a logout button in its lower section. Spaces represent operational areas that group a specific set of candidates and JDs; multiple interviewers can share a space, and any information uploaded by one user in a space is visible to all users who belong to that space. Individual users can be associated with multiple spaces, typically reflecting their department or practice within the firm.



FIGURE 2: Talent Assistant – Homepage

The default view is the "Interviews" tab, where users can initiate and re-access interview preparation content. To generate a new interview support package, the user has to upload a candidate's CV (in PDF format), then select one of the spaces they have access to, choose a specific JD available within that space, and specify the language the interview support package should be generated in. By default, during the execution of the backend logic, Talent Assistant infers an appropriate seniority level for the candidate (e.g., intern, associate, senior associate, and manager, which are all the roles that require at least some degree of technical knowledge) based on the CV and the target role. However, the user can explicitly override this choice before triggering the whole generation pipeline, for example, when HR constraints make certain seniority levels undesirable (a plausible scenario is when the candidate's pay expectations or past experience make it unrealistic to consider them for very junior positions), or when the team is seeking a specific-seniority figure to fit a client's offer. The resulting seniority level, whether inferred or forced, will influence the calibration of the interview support package. After parameters are set, the user triggers generation and typically waits around three to four minutes for Talent Assistant to display the generated content.

When the content becomes available, the "Interviews" view displays three main panels. The first one, "Candidate Profile", includes basic personal information (where available) such as name, email, links to GitHub and LinkedIn profiles, and location; a brief summary of the candidate, and the suggested seniority level (regardless of whether it had been forced or not). The first panel also includes structured background information, such as academic history, previous work experience, and relevant projects carried out; a summary of skills,

including technical and soft skills; spoken languages and certifications; and an overview of the job-position key traits (tags) that the candidate matches well. This panel is intended to give interviewers a compact but informative picture of the candidate, so they can quickly recall the profile even if they do not recall the CV precisely.

The second panel, "Interview Questions" (Fig. 3), contains the core question sets that will likely be used as a backbone for the interview. It is divided into a "personal" and a "technical" section, each with multiple questions and mock answers. Personal questions include a set of questions focused on past experiences, one targeting soft skills, and one looking into personal motivation. Technical questions include a set of domain-specific theoretical questions, and a set of domain-specific exercises (which can be code-related tasks, scenario-based cases, or other role-dependent exercises). Each set includes several alternative questions that follow the firm's general interview structure, so that interviewers can choose which ones to use while having room for choice and a general overview.

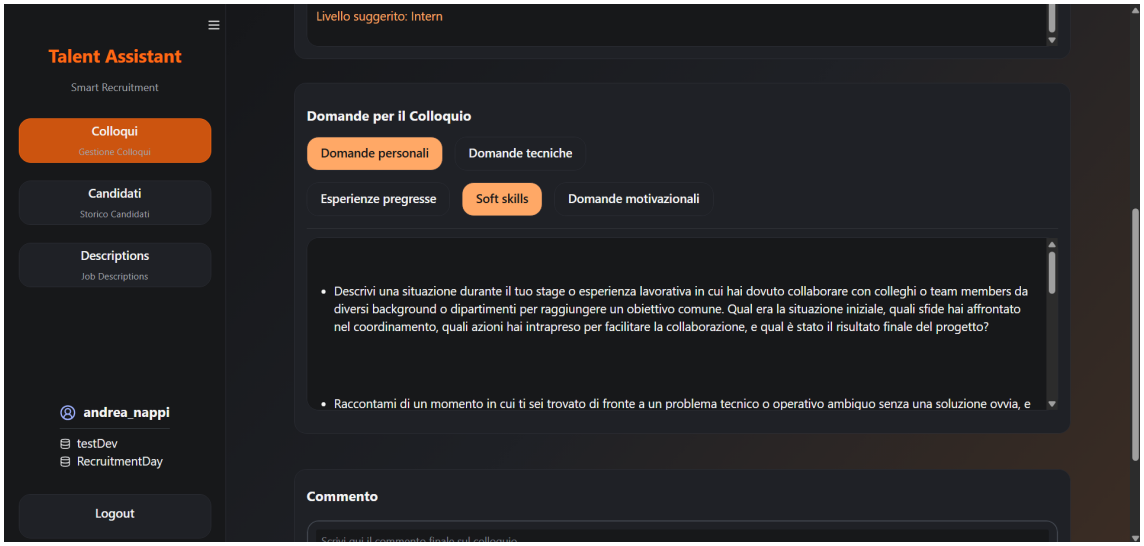

FIGURE 3: Talent Assistant – Interviews Tab

The third panel serves as a notebook that users can return to in order to write comments on the interview questions or notes during or after the conversation with the candidate; these notes are stored together with the interview artifacts and can be accessed later, possibly during an evaluation phase, for reference.

The intended workflow is for interviewers to use the candidate profile panel to orient themselves on the candidate's strengths, weaknesses, and traits, and then select some questions from the personal and technical sections that are a good fit for the candidate's profile, the time constraints of the interview calls, and the topics the interviewer wishes to cover. In practice, users are not expected to ask every single question, and indeed, the multiple-choice structure of each subset is meant to offer flexibility: interviewers can skip questions they personally consider less relevant or inappropriate, and use alternative ones instead. If users are not satisfied with the proposed sets as a whole (for example, if they feel that the technical section does not align well with the candidate's experience), they can trigger a new generation, which re-runs the same pipeline on the same inputs and yields an alternative set of questions. Major differences between runs are not guaranteed, as discrepancies are only to be attributed to the non-deterministic nature of the underlying LLM. Importantly, the human interviewer remains fully responsible for the interview. Although Talent Assistant's content might bias the user's perspective, it does not enforce

the use of any part of the interview support package. There is no explicit "accept" or "reject" operation at the level of individual questions in the interface, and interviewers retain control over which questions to ask, how to adapt them, how to follow up on candidate answers, and when to depart from the suggested script entirely. If a question is perceived as inappropriate or irrelevant, they can simply skip it. At present, there is no built-in mechanism for flagging problematic questions directly in the user interface; feedback on issues is typically communicated informally to the development team.

The other two tabs provide complementary views of the underlying data. The "Candidates" tab lists all candidates for whom an interview has been generated in any of the spaces the user can access (Fig. 4). For each entry, the list shows summary information such as the candidate's name, the spaces they belong to, the assigned seniority level, and an overview of associated tags, and it supports filtering based on tags. Selecting an entry from this list redirects back to the "Interviews" tab, populated with the previously generated interview and where applicable, with associated comments, allowing users to revisit past interview preparations and supporting later analysis of usage at the level of individual cases and sessions. The "Job Descriptions" tab enables users to create and maintain JDs for their spaces. In the "Job Descriptions" tab, users can create a new JD (Fig. 5) by providing a name, a description of the role, the main requirements, and specifying the space in which it should be stored. Talent Assistant then pairs the JD together with a one-time-generated set of tailored tags, which are reused for all candidates associated with that role, so that interview preparations share a consistent structure and vocabulary. Most of the time, the users do not have to come up with the specifics for all fields, as a standard JD already exists for most roles, and it is the one publicly available in open roles on likely many recruitment platforms. The same tab also supports updating or deleting existing JDs, enabling ongoing maintenance as role definitions evolve.

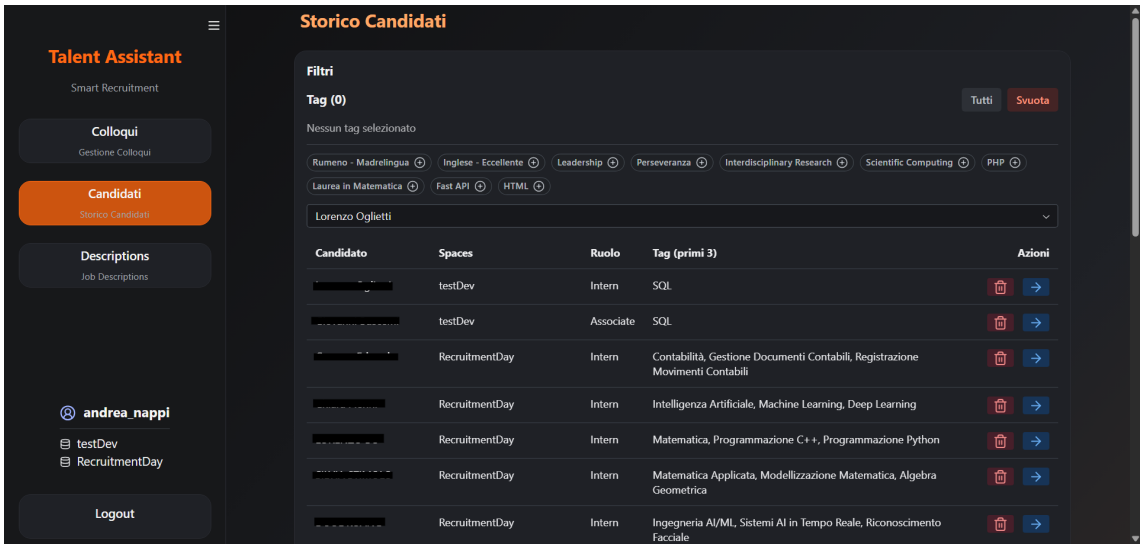

FIGURE 4: Talent Assistant – Candidates Tab

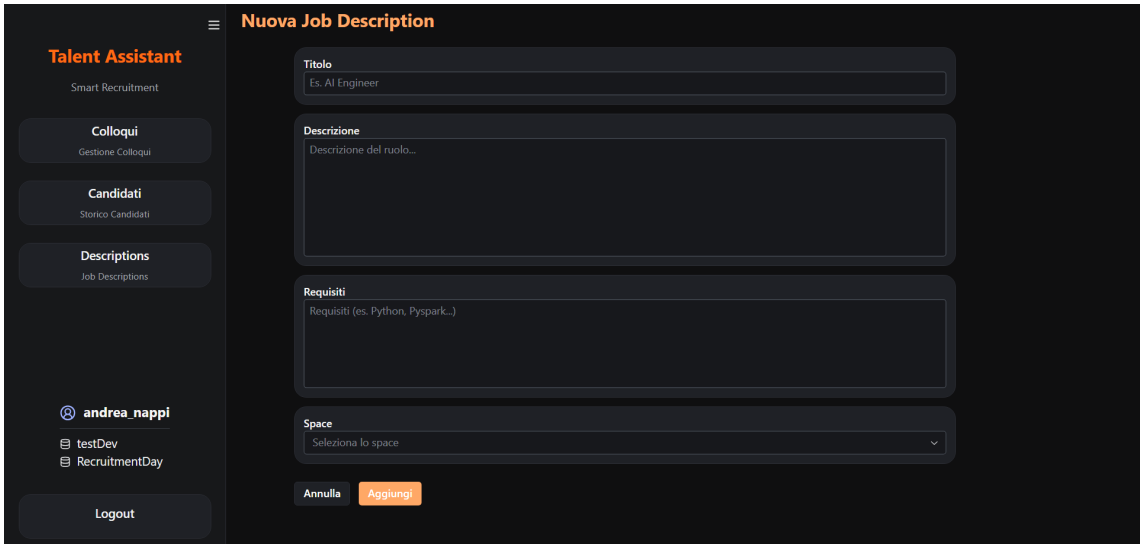

FIGURE 5: Talent Assistant – Job Descriptions Tab

5.3 Intended usage of Talent Assistant

Talent Assistant is meant to be used only after the organization has already received the candidate's CV and identified the candidate as someone for whom a technical interview should be prepared. As shown in Figure 6, the tool is designed to be positioned primarily before the technical interview itself (interviewers may still consult the generated material during the interview and use the notebook during or after it), and should not play a role during sourcing, initial screening, shortlisting, or final selection stages; and it should not be interpreted as an alternative applicant tracking system, a CV-screening tool, or an automated candidate-evaluation mechanism. All of these activities taking place before and after the Talent Assistant-supported preparation remain led by humans, with the focus on preparation representing a deliberate boundary, because the system is instead intended to help the interviewer prepare more effectively for a conversation.

The intended users are technical interviewers, which is the group already involved in assessing a candidate for a defined role. They should choose to use the system once a technical interview is planned, and so they need to prepare for it. Its usage is particularly intended for situations in which the interviewer needs to familiarize themselves with a candidate profile, identify relevant topics to explore, and construct an interview plan within limited preparation time. The generated interview-support package aims to provide the interviewer with a structured draft, rather than with a script that should be followed to the letter; and the interviewer is definitely not expected to ask every generated question: the multiple-question structure is designed to give users a broad pool of possible directions from which they can select the material most relevant to the candidate, the target role, the available interview time, or even the way the interview is progressing. Before using the material in an interview, the interviewer should still review the generated package critically. In particular, the interviewer should check whether the candidate profile accurately reflects the original CV without introducing unsupported or potentially biased interpretations, whether questions are actually grounded in the candidate's background and the selected JD, whether the difficulty of the technical questions is appropriate for the intended seniority level, and whether the proposed technical content is correct and relevant. It is also on the user to remove, adapt, or ignore questions that are vague, repetitive, inappropriate

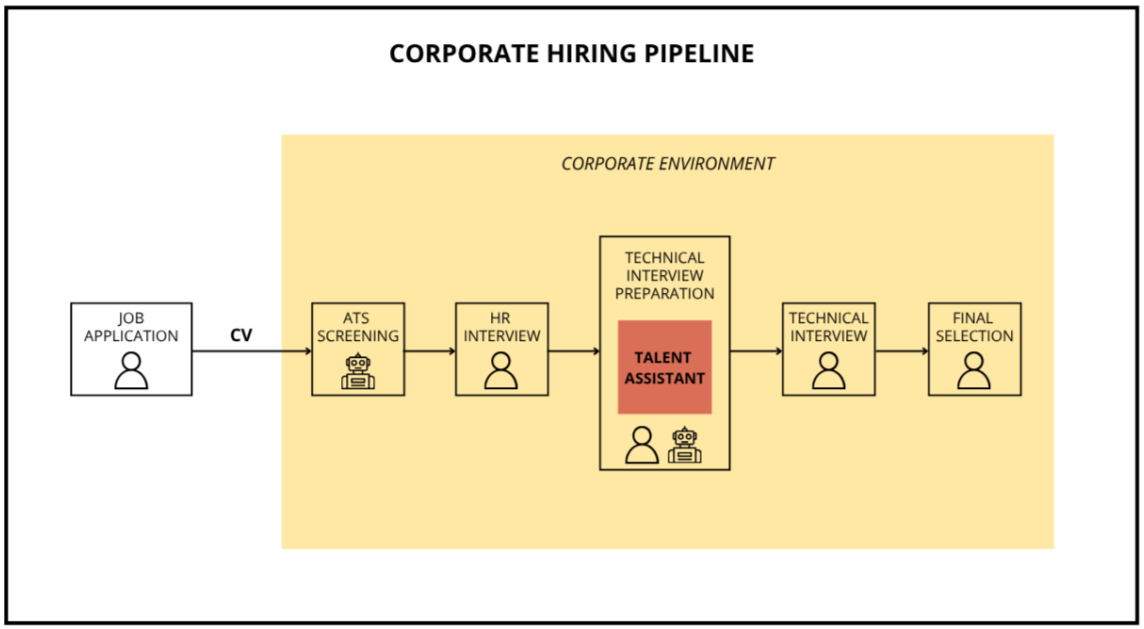

FIGURE 6: Proposed corporate hiring pipeline including Talent Assistant

for the interview context, based on unsupported assumptions, or otherwise unsuitable for other reasons. During the interview, the interviewer is encouraged to ask follow-up questions building on the proposed ones, and is free to depart from the generated content entirely in response to the actual conversation with the candidate. Ultimately, the goal is not to standardize every interview mechanically, but to make a sufficiently broad yet relevant interview plan more readily available. During or even after the interview, the notes can be used to record comments or observations worth remembering for future human consultation.

Talent Assistant is a HITL system. Although it generates suggestions, no responsibility shall be delegated to Talent Assistant, with the interviewer remaining responsible for deciding how the generated material is used and for any judgments they contribute to the established recruitment process. Talent Assistant should not be used as a substitute for technical expertise, and it is most appropriate when users have sufficient knowledge of the technical domain and of the target role to recognize incorrect, incomplete, overly generic, or inappropriate output.

5.4 Implementation details and prompting strategy

Talent Assistant's backend is implemented as a Python 3.11 application built on the FastAPI framework. The project is organized into a main application module and a nested package that includes all generative-AI logic, database access, session management, and utility functions. Among the top-level files, there is one that defines the FastAPI application and all HTTP endpoints; one that contains the Pydantic request and response models used for any input validation; and one that provides helper functions for credential verification and candidate profile storage. The inner package is divided into a directory containing two CrewAI crew definitions and their YAML configuration (containing prompts); a utils directory housing the logic for PDF processing, text extraction, CV validation, and tag management; and several standalone modules for configuration loading, Cosmos DB con-

nection management, optimized database operations with retry logic, and session lifecycle management.

The FastAPI application exposes ten endpoints covering authentication, JD management, CV upload and processing, interview generation, and retrieval of previously generated interviews and notes. Authentication is implemented through a two-layer scheme: a static X-Token header is required for the login endpoint, which verifies user credentials against those stored in Cosmos DB and returns a JSON Web Token (JWT), then all endpoints coming after require a valid JWT in the Authorization header. The JWT payload encodes all information needed to uniquely identify a user, together with an expiration timestamp for session management. All interactions with the database are filtered by the user's spaces, so that users can access only candidates and JDs associated with their own spaces. From the API perspective, the main content generation flow is split into three steps. First, a "POST" request accepts a base64-encoded CV PDF, creates a processing session identified by a UUID, stores the file in a dedicated session directory, and returns a session identifier and an initial progress status tracker. Second, a "GET" request triggers the asynchronous generation pipeline as a FastAPI background task, returning HTTP 202 with the session identifier so that the frontend can poll without blocking. Third, two "GET" requests combined allow the frontend to poll for completion and retrieve the final interview preparation content package once processing has completed. Additional endpoints support listing and filtering candidates by space and tags, listing tags, retrieving previous interview runs, posting interviewer comments, and creating, updating, and listing JDs, backing up all user-facing actions described in Section 5.2.

Persistent data is stored in Azure Cosmos DB in three logical containers corresponding to users, candidates, and JDs. The "users" container stores user credentials, including username, bcrypt-hashed password, and assigned spaces. The "candidates" container stores candidate profiles, with each document being identified by a deterministic hash computed from the candidate's first name, last name, email address, and associated JD identifier. Each candidate record contains structured fields for personal information, assigned spaces, selected skill tags, interviewer comments, and an "interview data" object that includes the output content (as JSON) of every agent in the pipeline. The third container, the "job descriptions" one, stores JDs with title, description, requirements, visibility flag, generated skill tags, and creation/update timestamps. The application uses a single shared Cosmos DB client instance and performs write operations with retry logic and exponential backoff to safely recover from temporary database failures or race conditions.

The generation pipeline is orchestrated through the "ResumeFlow" CrewAI Flow class, which defines a fixed sequence of processing steps connected by CrewAI listener and router decorators. Looking from a final outcome perspective, each major component of the interview support package is produced by a dedicated agent, and agents are grouped into two crews: a support crew that infers seniority and selects tags, and an analysis crew that generates the interview content questions, potentially leveraging the information coming out of the first crew (hence the separation). The pipeline goes through multiple stages. First, the uploaded CV, stored as a PDF in the session directory, is converted to PNG images, with multi-page documents being merged into a single vertical image, and this is because having the PDF converted to an image yielded better results in the upcoming phase. Then, the CV image is processed by the dedicated LLM, Claude Haiku 4.5, which looks at the image and returns the full textual content of the CV; the extracted text is saved and is passed to a further LLM call that determines whether the document is plausibly a CV or not; this acts as a gateway, and if the document fails validation, the pipeline terminates early and returns an error to the caller, saving time and resources. Otherwise, if the document has

been classified as a CV, the first CrewAI crew, "RoleCrew", which contains two agents and two corresponding tasks executed sequentially, starts running. Here, the first agent, the "Entry-Level Role Classifier" agent, analyses the CV text and assigns the candidate to one of four seniority levels (intern, associate, senior associate, or manager) based on years of experience, relevance of previous roles, educational background, certifications, and notable achievements. Then, the "Tag Creator" agent receives the list of job-specific skill tags (that had been extracted from the JD at upload time) and selects those that genuinely represent the candidate's profile, based on their CV content. Although this second piece of information, unlike the predicted seniority, is not used by the second crew, it has still been included in the first running crew to support potential future extensions. Once the "RoleCrew" is done processing, the main crew, "CrewAiInterview", starts running, configured with five agents and five tasks acting mostly in parallel to save time. The "Experience Question Generator" agent receives the CV text and seniority level and produces five open-ended questions about the candidate's past experiences, tailoring them to the career stage. The "Soft Skills Interviewer" receives the same inputs and produces five behavioral questions framed using the STAR method (Situation, Task, Action, Result), again tailored to the seniority tier. The "Technical Theory Interviewer" receives the CV text, JD, and seniority level and produces five domain-specific technical knowledge questions with model answers whose complexity scales with seniority. The "Technical Exercise Generator" uses the same three inputs to produce three practical exercises, each designed to be completable in about five minutes: coding exercises for technical roles (e.g., Python or SQL) and domain-specific scenarios (e.g., contract clause analysis, financial metrics interpretation, or business problem structuring) for non-technical roles, each with a problem description, input example, constraints, and a fully commented possible solution. Finally, the "Motivational Interviewer" receives the CV text, JD, and seniority level and produces three in-depth motivational questions that probe intrinsic motivation, career aspirations, and alignment with the target role. After all agents are done, the system runs some checks and verifies that all expected output files exist in the session's output directory and contain valid JSON. In a post-processing section, personal information and selected tags are extracted, the deterministic candidate identifier is computed, and the previously mentioned profile document is assembled, containing all agent outputs.

Prompting plays a key role in controlling agent behavior. Generally, the agent's role provides a short professional title (for example, "Universal Soft Skills Interview Question Generator"), the goal describes its objective in detail, including the dimensions to evaluate, and the backstory frames the agent as an experienced, fair, and unbiased interviewer or domain expert. Prompts for all agents follow a common pattern that includes: a high-level instruction describing the task; the relevant inputs (CV text, JD where applicable, seniority level, and target language); and explicit constraints derived from the PwC interview structure and expectations. The JD text is injected into the prompts for the technical theory, technical exercise, and motivational interview agents so that they can derive domain-specific topics and realistic scenarios from the JD, which allows the same pipeline to support a broad range of roles.

Seniority is handled explicitly in the prompting strategy. The seniority level is injected into each task description, and every task prompt contains conditional guidance that differentiates between the four seniority levels: experience questions focus more on academic history and internships for entry-level candidates, professional achievements and project impact for mid-level candidates, advanced expertise and cross-functional collaboration for senior candidates, and leadership and strategic decision-making for manager-level profiles. A similar structure is present in the soft skills, technical theory, technical exercise, and

motivational tasks, so that the same agent logic can flexibly adapt to different career stages. Language is controlled by a template variable, which defaults to "Italian"; each task description ends with an instruction to produce all output in the specified language, enabling generation in any model-supported language while being used predominantly in Italian and English in practice.

The prompts are designed to be domain-agnostic while still leveraging domain cues from the CV and JD. They explicitly enumerate multiple business domains, such as technology, finance, legal, marketing, operations, HR, consulting, healthcare, and more; and instruct each agent to infer the relevant domain from the inputs and tailor questions accordingly, ensuring that the same pipeline can serve diverse practice areas without specialized variants. A global task prompt, appended to every task description, provides explicit JSON formatting rules to avoid having issues with format checks while exchanging information, with endpoints or frontend elements expecting precise, valid formats.

For all LLM calls, Claude 4.5 Haiku is accessed via PwC Italy's internal shared service endpoint. No explicit values are set for temperature, maximum tokens, top-p, or similar generation parameters; instead, the system uses the defaults provided by the CrewAI framework and the shared service. All agents share a single LLM client instance, keeping configuration simple, also given the low number of calls per generation, maintaining a thin abstraction layer that would allow the underlying model to be swapped in case of changes of organizational or performance requirements.

Regarding safeguards, as described, the implementation prioritizes input validation, access control, and prompt design rather than hard programmatic content filters. All API inputs are validated using Pydantic models, constraining the length of free-text fields and enforcing expected formats. After the initial technical and access-related controls, at the prompt level, agent backstories and task descriptions consistently emphasize fairness, inclusivity, and job relevance. There are no explicit content filters that detect and block questions about sensitive attributes such as age, gender, family status, or ethnicity, and there is no automated bias or redundancy detection integrated into the pipeline; potentially problematic questions are expected to be spotted by the interviewer during review.

5.5 Data governance and privacy

Talent Assistant processes CVs, JDs, and generated interview material, which contains personal and, in part, potentially sensitive data handled within the hiring context. Section 3.4 identified governance requirements related to authenticated access, space-based separation, data residency, the use of governed model endpoints, and traceability of generated artifacts. This section describes the main data flows and application-level controls implemented in Talent Assistant. In particular, it outlines how candidate material enters and is stored by the system, the organizational privacy context in which the tool operates, how data is routed to the LLM through PwC Italy's internal shared service, and how access is restricted across organizational spaces. Broader legal, contractual, retention, and provider-side governance arrangements remain managed at corporate level and are not independently assessed in this thesis.

Data inventory and flows

The first point to mention about data inventory is that the application does not retrieve candidate data from any corporate system: every CV enters the system because an interviewer uploads it manually, which keeps the tool disconnected from the applicant tracking

system used for recruitment at PwC Italy and limits the volume of personal data held. Second, all persistent storage is concentrated in a single Azure Cosmos DB account [39] deployed in the West Europe (Ireland) region, alongside the two Azure App Service instances [38] that host the application, so that the application layer and the database remain within the European Union. The only point at which candidate data leaves this perimeter is the model call itself, which is discussed separately below.

Data protection framing

Recruitment data at PwC Italy is processed by "Servizi Aziendali PricewaterhouseCoopers S.r.l." as data controller, jointly with the Italian entities of the PwC network under agreements based on Article 26 of the GDPR [21, 52]. The purposes declared in the candidate privacy notice include the management of the application and the assessment of the CV against the job offer, for current and future open positions [52]. Talent Assistant supports interview preparation within that broader recruitment activity (note that this thesis does not provide an independent legal assessment of purpose compatibility, lawful basis, or the adequacy of the organization's governance arrangements): it supports an activity, the preparation of a technical interview, that is part of the selection process for which the CV was collected. The system introduces no new category of data subject and no new collection channel, since interviewers upload material that the firm already holds.

Data transmitted to the LLM

All model calls are routed through PwC Italy's internal shared service rather than through a public model provider API. The shared service follows the established enterprise gateway pattern: the calling application authenticates with a composite credential that encodes the target model, an application key, and a tenant identifier, while the gateway holds the provider credentials, determines which models a given tenant may reach, and attributes consumption to the requesting application. Three governance properties follow. Talent Assistant never stores or transmits a provider API key, so the compromise of the application would not expose credentials usable outside the corporate perimeter. The set of reachable models is controlled centrally, which is what makes the model-agnostic design described in Section 5.1 a governed choice rather than an arbitrary one. And the use of the shared service places the system inside the firm's existing AI governance framework, since this is the channel sanctioned internally for processing all confidential material, even highly sensitive client data. Still, there is the concern of where the data is processed. The application and the database remain in Ireland, but the model runs on infrastructure operated by the model provider. This is not in conflict with the candidate privacy notice, which already covers processing on servers located inside and outside the European Union under the safeguards required for such transfers [52], but it does mean that the system cannot be confidently described as one in which candidate data never leaves the European environment.

Access control and confidentiality within the firm

Access to Talent Assistant is limited to accounts created by the development team, which maintains an administrative interface through which accounts and space assignments are managed. The application is reachable over the public internet, but no endpoint returns data without authorization: the login endpoint requires a static token and verifies credentials against bcrypt hashes stored in Cosmos DB, and every subsequent endpoint requires a

valid JWT whose payload identifies the user and carries an expiration timestamp. Authorization is enforced server-side on every query, since all database interactions are filtered by the spaces associated with the requesting user, so a user cannot retrieve a candidate or a JD belonging to a space they do not belong to, irrespective of the request they construct. Confidentiality between organizational units is therefore a property of the data access layer rather than of the interface.

Chapter 6

Evaluation results and discussion

6.1 Survey results

This section reports the evaluation of Talent Assistant in its intended organizational setting, coming from the interpretation of the survey submitted to technical interviewers. Given the small expert population, the results are interpreted as formative and descriptive evidence. They should not be read as estimates of Talent Assistant's effect on interview outcomes or as findings that can be generalized beyond the studied setting.

As mentioned in Chapter 4, for all closed-ended survey items, participants used the seven-point scale specified in Appendix B. The midpoint, 4, was defined as approximately equivalent to the participant's usual practice without Talent Assistant. Scores above 4 are interpreted as perceived improvement over the existing practice, scores close to 4 as approximate equivalence, and scores below 4 as perceived shortcomings. Item-level results are presented using the mean (M), median (Mdn), population standard deviation (SD), and observed range. Because of the small population size, no inferential tests are performed.

6.1.1 Participants

All ten users who had been granted an operational Talent Assistant account (meaning they are to some extent responsible for preparing and conducting interviews at PwC Italy) at the time of the study were invited to participate. This group is relevant because its members work directly in the part of the interview pipeline affected by Talent Assistant. They therefore have relevant practical experience and are the users whose standards the system must meet. All the participants completed the survey after providing consent, meaning the response rate was 100% ($N = 10$), and although this does not make the population representative of technical interviewers outside the setting, it means that the survey captures the total population of current users.

The participant group consisted of five senior associates, three managers, one senior manager, and one director. All participants belonged to Data & AI-related teams, including "AI Engineering", "Agentic AI", "Palantir CoE", and broader Data & AI units, and their primary interview domains included mainly data engineering and artificial intelligence, but also AI strategy and data governance (Appendix C).

As shown in Table 1, the interviewers had different interview experience and workload. One participant had less than one year of technical-interview experience, five had two to three years, two had four to six years, one had seven to ten years, and one had more than ten years. Nine participants had interviewed interns and nine had interviewed associates; four had interviewed senior associates and one had interviewed managers. This indicates

how the subjects had experience mainly with junior and mid-level roles that fall within Talent Assistant's intended scope, based on the roles handled by early users.

TABLE 1: Characteristics of survey participants ($N = 10$).

Characteristic	**Category**	n
Position	Senior Associate	5
	Manager	3
	Senior Manager	1
	Director	1
Technical-interview experience	Less than 1 year	1
	2–3 years	5
	4–6 years	2
	7–10 years	1
	More than 10 years	1
Technical interviews per year	Fewer than 10	1
	10–24	5
	25–49	1
	50–99	1
	100 or more	2
Candidates interviewed using Talent Assistant	Fewer than 10	2
	10–19	5
	30 or more	3
Self-reported tool familiarity	5 – Very familiar	8
	4 – Moderately to very familiar	2

About their exposure to the system, five respondents had used Talent Assistant for 10-19 candidates, three had used it for 30 or more, and two had used it for less than 10 interviews. Nine respondents mainly used the system before the interview, seven used it during the live interview, and three reported use after the interview (Appendix C).

Overall, these metrics demonstrate that the upcoming evaluation is based on practical, although moderate, use rather than first impressions only.

6.1.2 Perceived usefulness results

Table 2 reports the results for perceived usefulness. All five items had means above the reference value of 4. The efficiency index, calculated as the unweighted average of B1 and B2 items, was $M = 6.35$ ($Mdn = 6.50$, $SD = 0.55$). The broader perceived-added-value index, calculated from B3–B5, was $M = 6.17$ ($Mdn = 6.33$, $SD = 0.54$). All ten participants scored both indices above 4.

The results suggest that participants perceived Talent Assistant as reducing the effort required to prepare an interview. The result is particularly relevant to the problem identified in Chapter 1, where preparation was described as an activity competing with project work and often conducted under time pressure. However, this finding concerns perceived efficiency; the present evaluation did not measure preparation time directly and cannot establish the number of minutes saved per interview.

The strongest usefulness result was B5 ($M = 6.50$), indicating broad agreement that the tool adds value relative to previous preparation and interview practices. B1 and B4

TABLE 2: Perceived usefulness and efficiency results ($N = 10$; scale 1–7).

Item	M	Mdn	SD	Min	Max
B1: Enables faster interview preparation	6.40	6.50	0.66	5	7
B2: Enables preparation with less effort	6.30	6.50	0.78	5	7
B3: Provides relevant technical prompts not otherwise considered	5.80	6.00	1.16	3	7
B4: Improves the overall quality of technical interviews	6.20	6.00	0.40	6	7
B5: Adds value over previous practice	6.50	6.50	0.50	6	7

had means of $M = 6.40$ and $M = 6.20$, respectively, with all respondents selecting either 6 or 7 for B4. These findings indicate that participants did not view the system only as a faster way to reproduce their usual process; they also perceived it as contributing positively to the technical interview.

The B3 item had the largest dispersion of this section ($SD = 1.16$) and it is the only one which included a rating below the baseline. Thus, although most participants believed that Talent Assistant provides relevant questions or prompts they might not have considered on their own, this benefit seems to be more dependent on interviewer experience, role, or use case than the general value of the system. Taken together, the results provide positive formative evidence for RQ2.1.

6.1.3 Perceived quality results

Table 3 presents the results for the output quality rubric. The mean for every item was above the benchmark of 4. The unweighted aggregate across C1–C9 was $M = 6.20$ ($Mdn = 6.05$, $SD = 0.46$), and every participant's aggregate score was above 4. The results therefore indicate consistently positive perceived quality across the dimensions extracted from the requirements in Section 3.3.

For relevance, the combined C1–C2 score was also $M = 6.30$. This indicates that participants generally perceived the generated questions as grounded both in the information stated by the candidate in their CV and in the requirements of the target JD. This result is consistent with the intended CV–JD conditioning of the pipeline and directly addresses the relevance requirement.

Specificity was also rated positively (C3: $M = 6.30$), meaning participants generally considered the questions sufficiently clear and precise. Calibration received a combined C4–C5 score of $M = 6.25$, suggesting that the seniority-based calibration in the generation process was generally perceived as adequate both in complexity and in the topics being assessed. These results support the claim that the system provides an interview backbone tailored to the candidate and target role, although they do not independently validate the actual quality of each generated question.

Coverage and diversity were both positive (C6: $M = 6.20$; C7: $M = 5.90$). One participant rated coverage at the midpoint, and C6 showed relatively greater variation ($SD = 0.87$). This suggests that the structured output manages to cover the relevant macro-areas, with limited overlap between questions. Although C7 had the lowest mean in the section, its average score was still fairly high (M = 5.90).

The perceived-reliability index, calculated from C8 and C9, was $M = 6.15$. Participants rated technical correctness highly (C8: $M = 6.30$), and they generally viewed the material as interview-ready (C9: $M = 6.00$). The latter item included one midpoint rating,

TABLE 3: Perceived output quality results ($N = 10$; scale 1–7).

Item	M	Mdn	SD	Min	Max
C1: Questions focus on CV-declared skills and experience	6.30	6.50	0.78	5	7
C2: Questions align with the job position and requirements	6.30	6.00	0.64	5	7
C3: Questions are clear and elicit precise, specific answers	6.30	6.00	0.64	5	7
C4: Questions are appropriate to candidate seniority	6.30	6.50	0.78	5	7
C5: Questions target appropriate seniority-specific aspects	6.20	6.00	0.60	5	7
C6: Set covers the relevant assessment macro-areas	6.20	6.00	0.87	4	7
C7: Set covers complementary rather than overlapping aspects	5.90	6.00	0.70	5	7
C8: Questions and prompts are technically correct	6.30	6.00	0.64	5	7
C9: Questions are interview-ready with minor or no adjustments	6.00	6.00	0.89	4	7
C10: Overall satisfaction with output quality	6.40	6.00	0.49	6	7

making it important not to claim that outputs always require no adaptation. Rather, the result supports the intended HITL role of Talent Assistant as a support element where interviewers remain responsible for reviewing, selecting, and adapting. Finally, all participants rated overall satisfaction with output quality at either 6 or 7 (C10: $M = 6.40$). Overall, these findings provide positive expert-based evidence for RQ2.2.

6.1.4 Usage and integration results

Table 4 presents the results for usability, workflow integration, retained control, and consistency. The index composed of items D1, D2, and D3 had the highest scores in the survey ($M = 6.73$, $Mdn = 6.67$, $SD = 0.25$), with every participant scoring every item above 4 (either 6 or 7).

TABLE 4: Usage, integration, human control, and consistency results ($N = 10$; scale 1–7).

Item	M	Mdn	SD	Min	Max
D1: Learning to operate the system was easy	6.80	7.00	0.40	6	7
D2: The tool fits the existing preparation workflow	6.50	6.50	0.50	6	7
D3: The interviewer remains fully in control and responsible	6.90	7.00	0.30	6	7
D4: Comparable cases receive similar types of questions	6.20	6.00	0.60	5	7
D5: Differences mainly reflect deliberate adjustment rather than random variation	6.00	6.00	0.77	4	7

Ease of learning received a mean of 6.80, and perceived workflow fit received a mean

of 6.50. These results suggest that the web application and generated interview artifacts can be integrated into established preparation routines without high learning barriers, at least for experienced, technically oriented users such as the ones in the group studied here, while little can be concluded about less technically familiar users or other organizational contexts.

The strongest single result in the entire survey was retained interviewer control (D3: $M = 6.90$). This is an important empirical indication that the system's design is experienced as decision support rather than as an automated decision-maker. It supports the key requirement that users both retain responsibility and feel responsible for deciding which questions to ask, how to interpret answers, and how to evaluate candidates. Nevertheless, perceived control does not eliminate the potential for automation bias; this risk remains relevant because users may still be influenced by the system's framing even when they retain the formal ability to disregard it.

The consistency index, based on D4 and D5, had $M = 6.10$ ($Mdn = 6.00$, $SD = 0.62$). Respondents agreed that comparable candidates in the same role tend to receive similar question types (D4: $M = 6.20$). They also generally agreed that observed differences mainly reflect deliberate interviewer adjustments rather than random model variation (D5: $M = 6.00$), although this item included one neutral response. The findings suggest that users perceived a repeatable structure across comparable cases while retaining the ability to adapt questions. It is still important to state that they do not constitute a direct content-level comparison of question sets or an objective measurement of inter-interview consistency.

6.2 Field usage metrics

This section reports descriptive usage data extracted from the production Talent Assistant database for the period from 1 October 2025 to 31 July 2026. These data complement the survey results by documenting metrics on system usage in the organizational setting. They describe adoption and the distribution of generated interview-support packages; they do not measure question quality, fairness, preparation time saved, or causal effects on interview or hiring outcomes.

During the observation period, Talent Assistant produced interview-support packages for 95 unique candidate–JD pairs. In total, 133 generation runs were processed. The difference of 38 runs reflects that a candidate–JD pair could be generated more than once, for example when an interviewer re-ran the generation process or when different interviewers initiated a generation for the same pair. On average, this corresponds to 1.40 processed generation runs per unique pair. This figure should be interpreted only as evidence of re-use or repeated generation, because the available logs do not state the precise reason for each additional run and therefore cannot identify whether it resulted from dissatisfaction with a prior output, a deliberate search for alternatives, or just multi-interviewer use.

Table 6 shows the distribution of the 95 unique candidate–JD packages across seniority levels. Usage was concentrated among intern and associate candidates, which together accounted for 73 of the 95 packages (76.8%). This shows that the system supports mostly technical interviews for early and mid career positions. Senior Associate packages represented 14.7% of all packages, while manager-level packages represented 8.4%. The smaller manager-level share should be considered when interpreting the survey results and when considering any extension of the system to more senior roles. This distribution is overall coherent with the typical role pyramid structure that often characterizes large consulting

TABLE 5: Field usage of Talent Assistant, 1 October 2025 - 31 July 2026.

Metric	Value
Observation period	1 October 2025 - 31 July 2026
Unique candidate/JD packages generated	95
Total processed generation runs	133
Additional runs beyond the first generation per pair	38
Mean generation runs per candidate–JD pair	1.40

firms.

TABLE 6: Unique candidate–JD packages by seniority level ($N = 95$).

Seniority level	Packages	Share
Intern	47	49.5%
Associate	26	27.4%
Senior Associate	14	14.7%
Manager	8	8.4%
Total	95	100.0%

Table 7 displays the monthly distribution of the 95 unique candidate–JD packages from October 2025 to July 2026 (Table 5). Use peaked in November 2025, with 29 packages (30.5%), followed by October 2025 with 13 packages (13.7%). Ten packages were generated in each of January and February 2026. Usage then decreased over the final months of the observation period, reaching three packages in July 2026. This pattern indicates that deployment was not limited to a single day or an isolated demonstration; however, without data on job openings, candidate flow, or interview demand by month, it cannot be used to infer a trend in adoption, user satisfaction, or organizational hiring volume.

TABLE 7: Monthly distribution of unique candidate–JD packages ($N = 95$).

Month	Packages	Share
October 2025	13	13.7%
November 2025	29	30.5%
December 2025	7	7.4%
January 2026	10	10.5%
February 2026	10	10.5%
March 2026	7	7.4%
April 2026	5	5.3%
May 2026	6	6.3%
June 2026	5	5.3%
July 2026	3	3.2%
Total	95	100.0%

Taken together, the field data establish that Talent Assistant was used on a significant number of distinct candidate–JD pairs, across all four supported seniority categories and throughout a period of 10 months. They provide contextual evidence of operational use,

complementing the survey's positive evidence about perceived usefulness, quality, workflow fit, control, and consistency. The logs do not, however, provide evidence about the quality of questions, output correctness, bias, precise preparation time saved, or downstream recruitment outcomes. These limitations are addressed further in Section 6.4.

6.3 Observed issues and failure instances

The survey's open-ended responses (Appendix C) provide contextual evidence about the failure modes observed by users. Most respondents reported no clear technical, ethical, or alignment-related issue during their use of Talent Assistant. One respondent noted that the system could sometimes generate questions that were similar or overly standard across different interviews. Individual respondents also raised boundaries around use beyond strictly technical roles or for manager-level candidates, and one participant emphasized the need to continue paying attention to bias and imposed limitations on AI tools.

These observations should not be treated as evidence that Talent Assistant is free of errors, bias, or other risks. First, the survey relies on user observation and recall, and a user may not detect an inaccurate statement, potential stereotypes, or problematic omissions that can be somewhat hidden in apparently fine question sets. Second, the number of respondents and observed cases is limited. Third, the system is used in a high-risk employment context, where even rare failures matter. Anyway, the largely positive responses instead indicate that there seems to be no common failure mechanism.

The observed issue of standard or similar questions is consistent with the system's design trade-off. A repeatable pipeline and shared JD conditioning are intended to promote common structure and coverage across comparable interviews, but excessive similarity can reduce perceived tailoring and may lower the information value of a limited interview period. This finding should therefore be read alongside the positive consistency result in Section 6.1.4: consistency is beneficial when it reflects common competency coverage, but it becomes undesirable when it turns into avoidable redundancy or generic questioning.

Many other risks, as identified in Chapter 2, remain relevant even though they were not frequently reported in the survey. These include hallucinated or unsupported questions, biased emphasis on candidate characteristics, inappropriate references to sensitive attributes, confidentiality concerns arising from CV content, and automation bias by interviewers. The current implementation attempts to reduce these risks through controlled access, space-based data separation, the use of internal endpoints, structured prompts (which cannot be altered via endpoints), and interviewer review. However, the implementation does not include an explicit automated filter for sensitive topics, an automated bias audit, or a systematic question-level flagging mechanism. Human review is therefore a necessary safeguard, although it is not a guarantee of fairness or correctness.

The results point to several practical refinements. First, the generation pipeline could include post-generation checks for repeated or semantically overlapping questions, unsupported claims about the candidate, and sensitive or legally problematic topics. Second, the interface could allow users to flag problematic questions and record a structured reason, converting informal feedback into useful evidence for later refinement. Third, manager-level and non-technical use cases should be evaluated separately before broadening the system's scope, because even though the system is designed to support these cases, the current survey contains limited evidence for these contexts. Finally, periodic review by domain experts should test generated question sets against CV–JD evidence and organizational interview policy, rather than relying only on user satisfaction.

6.4 Threats to validity

Internal validity

The study does not use a control group, random assignment, or direct before-after measurement of preparation time and interview outcomes. The positive ratings therefore cannot establish the extent of the improvements reported for Talent Assistant. Participants may have compared the tool with different personal baselines, may have used it for different roles and candidate profiles, and may have encountered different versions or outputs of the system. Exposure also varied substantially, from fewer than ten to 30 or more candidates. More experienced or more frequent users may assess the tool differently from occasional users, but the sample is too small for reliable subgroup analysis.

Learning and novelty effects are also possible. Participants who were already familiar with the application may have found it easier to use and more valuable than newly introduced users would (although every user stated that learning to use the application was very easy). Conversely, respondents could have become more critical as they encountered repeated generations. The evaluation did not track how ratings developed over time. Finally, because the system operates in a live organizational setting, changes in hiring volume, role requirements, interviewer workload, and JD quality may all have influenced the experience of using the tool.

Construct validity

The survey measures participants' perceptions of usefulness, quality, consistency, and control. It does not directly measure actual preparation time, factual correctness of every generated question, predictive validity of interviews, fairness across demographic groups, or candidate experience. For example, a high score on C8 indicates that users perceived the questions as technically correct; it is not an independent expert audit of factual correctness. Similarly, the D4–D5 results measure perceived consistency, not a formal similarity analysis of question sets generated for matched candidate/role cases.

The use of means and standard deviations for Likert responses is appropriate here as a transparent descriptive convention, but these values should not be interpreted as interval-scale measurements with high precision. The explicit midpoint anchor improves comparability to usual practice, while still relying on each participant's subjective understanding of that baseline.

Self-reporting introduces possible recall bias, and because participants work in the same organization in which the system was developed and deployed, they may have been inclined to provide favorable ratings, and although making the survey anonymous and the inclusion of open-ended questions were intended to make criticism possible, they cannot eliminate such bias.

External validity

The sample contains the entire current operational user population, and the 100% response rate removes non-response bias within that defined population. However, this population is small and situated within PwC Italy's Data & AI-related technical teams. Participants were technically literate, accustomed to technical interviewing, and familiar with Talent Assistant. The results may therefore not generalize to other PwC service lines, non-technical recruiting contexts, other countries, organizations with different hiring practices, or users with lower technological familiarity.

The study also concerns preparation for late-stage technical interviews for roles from intern to manager. It does not evaluate CV screening, candidate ranking, automated hiring decisions, candidate-facing AI, or interview outcomes after employment.

The one respondent who indicated possible hesitation for manager-level use reinforces the need for caution when extending the conclusions beyond the best-represented junior and mid-level contexts.

Chapter 7

Conclusion

This thesis addressed the problem of inefficient and inconsistent preparation for late-stage technical interviews in the PwC Italy context. To address it, Talent Assistant was designed and deployed as a bounded, HITL agentic system that transforms a candidate's CV, a JD, and, optionally, a seniority level into a structured interview-support package. The system supports preparation only: interviewers remain responsible for selecting questions, conducting interviews, and making recruitment-related judgments.

A survey of the ten operational users, complemented by descriptive production logs, provided formative evidence of positive perceived usefulness, output quality, workflow integration, and retained interviewer control. These findings should not be interpreted as evidence of objectively reduced preparation time, factual correctness, fairness, or improved hiring outcomes.

7.1 Answering RQs

RQ1: How can an agentic AI system be designed so that it meets hiring domain requirements?

The first research question is about the design of an agentic AI system capable of supporting technical interview preparation while meeting practical, quality, and governance requirements.

RQ1.1: How can an agentic orchestration structure be designed to produce effective, structured interview support packages?

The implemented system demonstrates that a structured orchestration approach can be used to produce a consistent interview-support package without requiring the system to make hiring decisions. Talent Assistant uses a fixed, multi-stage pipeline that combines deterministic software logic with specialized LLM-based agents. The pipeline first extracts and validates CV content, infers or receives a seniority level, and then coordinates dedicated agents for experience, behavioral, motivational, technical-theory, and practical-exercise question generation. This separation of responsibilities makes the generation process more traceable and maintainable than a single unconstrained prompt. It also enables the system to return a repeatable output structure across candidates while also preserving candidate-specific adaptation through the CV and JD inputs.

The orchestration is therefore agentic in the sense that multiple specialized agents are responsible for different sub-tasks, but it is not autonomous in the sense of independently deciding how candidates should be assessed or selected. The workflow is deliberately

bounded: the system follows a predefined sequence, generates structured outputs, and leaves the substantive decision of what to ask and how to interpret candidate answers to the interviewer. This design is appropriate for the high-stakes hiring setting, where fully automated candidate assessment would introduce considerably greater governance, fairness, and accountability concerns.

RQ1.2: Which inputs should be integrated into the pipeline, and how should they condition generation?

The design shows that three inputs are particularly important for conditioning generation: the candidate CV, the JD, and the seniority level. The CV is the source of information for the candidate's academic background, professional experience, projects, skills, and career stage. The JD presents the role-specific requirements, technologies, tasks, and domain context for which the candidate should be assessed. Finally, the seniority level determines the expected complexity and focus of the generated questions. For example, questions for interns and associates should primarily check for foundational understanding and learning ability, while questions for senior associates and managers should increasingly examine depth of expertise, trade-off reasoning, coordination, and leadership aspects.

The evaluation results provide supportive evidence that these inputs were successfully reflected in the generated content. Participants rated both CV-grounding and JD alignment highly, with both relevance items receiving a mean score of $M = 6.30$ on the seven point scale. Similarly, the two seniority calibration items received a combined mean of $M = 6.25$. These results do not independently verify that every question was correctly grounded or calibrated, but they indicate that experienced users generally perceived the input-conditioned generation strategy as appropriate in their real work for interview preparation.

RQ1.3: What risks are tied to these requirements, and what do they imply for safe deployment of such a system?

The design and evaluation indicate that AI-based interview preparation remains associated with relevant risks that cannot be eliminated only through a HITL configuration. Relevant risks include hallucinated or unsupported claims about the candidate, technically incorrect questions or model answers, repetition and over-standardization, biased emphasis on certain candidate characteristics, inappropriate references to sensitive attributes, confidentiality concerns related to CV data, and automation bias among interviewers. The system addresses some of these risks, other than by design, being a HITL system, through controlled internal deployment, authenticated access, space-based data separation, CV validation, structured prompting, and the preservation of interviewer control. The current implementation does not include automated filters for sensitive topics, automated bias auditing, semantic redundancy detection, or a built-in reporting mechanism for specific elements.

The findings therefore suggest that safe deployment should be understood as a complex ongoing social and technical process rather than as a property guaranteed by the system architecture alone. Human review is necessary because the interviewer can reject, adapt, or ignore generated material, but it is not sufficient to ensure correctness or fairness. Future versions might work towards supplementing human review with explicit post-generation checks, structured feedback mechanisms, periodic expert audits, and clearer evidence trails linking generated questions to the CV and JD content from which they were derived.

RQ2: How is such a system perceived by early users to mitigate the key issues investigated compared with current practice?

The second research question examined whether Talent Assistant, according to its early users, mitigates the key problems of manual interview preparation: high preparation effort, variable output quality, and inconsistent interview structure.

RQ2.1: Does it improve the preparation time–quality ratio?

The findings provide positive formative evidence that users perceived lower preparation effort and faster preparation, which addresses part of RQ2.1. The efficiency index, based on the items concerning faster preparation and lower effort, received a mean score of $M = 6.35$, compared with the reference value of 4 representing participants' usual practice without the system. Participants also rated the system's overall added value relative to previous practice highly, with the broader index for perceived added value receiving $M = 6.17$. The strongest individual usefulness item was the perception that Talent Assistant adds value over previous preparation practice ($M = 6.50$).

These findings indicate that users did not view Talent Assistant as a different interface for producing the same interview materials only; instead, they generally perceived it as reducing the effort required to prepare an interview while contributing positively to the overall process. However, this conclusion concerns perceived efficiency rather than directly measured preparation time. The study did not record baseline and post-adoption preparation durations, and so it cannot quantify the amount of time saved per interview or establish a causal effect on time-to-hire.

RQ2.2: Does it improve key quality dimensions for questions?

Users evaluated the quality of the generated interview materials positively across all assessed dimensions. The aggregate output quality score across items C1–C9 was $M = 6.20$, and every participant's aggregate score was above the reference value of 4. Relevance was rated strongly, with a combined score of $M = 6.30$ for CV-grounding and JD alignment. Specificity also received $M = 6.30$, suggesting that users generally considered the questions sufficiently clear and concrete to elicit informative answers. The results for coverage ($M = 6.20$), diversity ($M = 5.90$), technical correctness ($M = 6.30$), and interview readiness ($M = 6.00$) were also positive.

The results suggest that the system succeeds in producing a structured and usable starting point for interview preparation. In particular, the high ratings for relevance, calibration, and reliability are coherent with the intended design of the generation pipeline, which combines candidate information, job requirements, and seniority-specific instructions. At the same time, the lower mean for diversity relative to the other quality dimensions, together with the open ended observation that some questions can be similar or overly standard, identifies repetition as a potential area for further refinement. A system that is sufficiently structured to improve consistency should not become so standardized that it reduces meaningful tailoring to individual candidates.

RQ2.3: Does it improve final output consistency across interviews for comparable candidates–roles?

The findings provide positive evidence (although based on perception) that Talent Assistant is perceived to improve consistency across comparable interview preparations. The

consistency index, based on items D4 and D5, received a mean score of $M = 6.10$. Participants generally agreed that comparable candidates applying for the same role receive similar types of questions ($M = 6.20$), and that differences across interviews mainly reflect deliberate interviewer adjustments rather than unpredictable model variation ($M = 6.00$). This suggests that the fixed output structure and shared JD conditioning provide a common interview backbone while still leaving room for professional judgment and adaptation to the candidate.

Consistency in this context should not be interpreted as asking every candidate identical questions. Rather, it is about providing comparable coverage of relevant competency areas and similar levels of expected complexity for candidates applying to similar roles. Talent Assistant appears to support this objective by producing common categories of questions and a repeatable workflow. Nevertheless, this study did not conduct a direct content comparison of question sets across matched candidate–role cases, so it cannot establish an objective degree of semantic similarity, fairness, or comparability across interviews.

7.2 Contributions and implications

This thesis makes four main contributions.

First, it contributes Talent Assistant as a deployed system artifact for AI-supported technical interview preparation. Unlike a conceptual proposal or isolated prototype, the system was implemented as an internal web application and integrated into the organizational environment in which it was evaluated. Its architecture combines an interface in Reflex, a FastAPI backend, CrewAI-based orchestration, Azure deployment, Cosmos DB persistence, and access to an internally governed LLM endpoint. The implementation shows how LLMs can be embedded in an operational workflow while maintaining a clear separation between automated preparation support and human hiring responsibility.

Second, the thesis contributes a requirements-driven evaluation rubric for AI-generated interview-support materials. The rubric includes six relevant dimensions of question quality: relevance, specificity, calibration, coverage, diversity, and reliability. By connecting these dimensions to both structured-interview literature and the local needs expressed by PwC Italy interviewers, the rubric offers a reusable base for evaluating similar systems in settings of recruitment support. It is particularly useful because it evaluates the practical quality of generated materials instead of only relying on generic usability or model-performance metrics.

Third, the thesis contributes formative empirical evidence from early real-world usage. All ten operational users of Talent Assistant completed the survey, resulting in a 100% response rate within the defined user population. Production data recorded 95 unique candidate–JD packages and 133 total generation runs between October 2025 and July 2026. Use occurred across all four supported seniority levels, although it was concentrated among intern and associate candidates. Together, the survey and usage data demonstrate that the system was used beyond a single demonstration and that its early users generally perceived it as useful, usable, and aligned with their workflow.

Fourth, the thesis contributes practical insights into the responsible deployment of agentic AI in recruitment-related settings. The central implication is that AI can provide value before an interview without directly automating the assessment or selection of candidates. This narrower form of support can be relevant especially in high-stakes domains: it reduces the burden of preparing interviews while preserving the interviewer's authority over the actual assessment process. Also, the study shows that a HITL design must be supported by more than formal human control. Users need sufficient time, domain knowledge,

awareness of system limitations, and possibly other mechanisms for control.

For organizations considering similar systems, the findings suggest that deployment should begin with a bounded use case, clear restrictions on automated decision-making, controlled access to sensitive data, and a structured workflow that users can understand and review. Evaluation can combine user perceptions with technical audits and, where feasible, objective process measures. High user satisfaction is valuable evidence of adoption potential, but it should not be treated as evidence that a system is unbiased, factually correct in every case, legally compliant, or predictive of improved hiring outcomes.

7.3 Future work

Starting from the findings and limitations of this thesis, several directions for future work emerge.

First, future evaluations should measure outcomes more directly. A study could compare preparation duration before and after adoption, record the extent to which generated questions are retained or edited, and analyze how use changes over time. A carefully designed comparison between preparation with Talent Assistant and current practice could provide stronger evidence about efficiency, output quality, and interview consistency. Such research would need to be designed carefully because hiring decisions are influenced by many factors beyond interview preparation.

Future work could also focus on strengthening question quality assurance. The system could include automated post-generation checks for unsupported claims, factual inconsistencies, sensitive or legally problematic topics, and semantically overlapping questions. A traceability layer could show which CV passages and JD requirements motivated each generated question. This would make outputs easier to verify, help interviewers challenge incorrect assumptions, and support later auditing of problematic cases.

The current feedback process should be formalized within the interface. Interviewers could be enabled to flag a question as irrelevant, repetitive, technically incorrect, overly difficult, insufficiently tailored, or potentially sensitive. Structured feedback would create an evidence base for prompt refinement and model monitoring, while also making the human-review step more active and auditable. Such functionality could reduce the dependence on informal feedback to the development team.

Further research should test the transferability of the system and its evaluation rubric. The present evidence concerns technically literate users in Data & AI related teams within PwC Italy and primarily concerns intern and associate positions. Additional studies should evaluate the tool with different service lines, non-technical roles, other national contexts, and users with lower familiarity with AI systems, especially given that the system is designed to handle these cases already. Manager-level roles should be examined separately because they were less represented in both the field data and the survey evidence, and because their assessment may require more emphasis on strategic judgment, leadership, and organizational fit.

Finally, future work should examine the longer-term effects of routine AI support on interviewer practice. Although Talent Assistant can reduce preparation effort and improve access to a structured interview backbone, continuous use may also shape what interviewers consider relevant evidence, which questions they ask, and how actively they develop their own interviewing capabilities. Research should therefore investigate automation bias, over-reliance, deskilling, and the conditions under which AI support strengthens rather than weakens professional judgment.

Bibliography

[1] Aerotek. The 4 biggest high-volume recruiting challenges and how to overcome them. https://www.aerotek.com/en/insights/4-biggest-high-volume-recruiting-challenges, 2022. Aerotek insights blog.

[2] Jeff Aplin. How to manage high-volume hiring (without losing your mind). https://www.aplin.com/blog/how-to-manage-high-volume-hiring/, 2019. Aplin Recruitment blog.

[3] Ilana Bergelson, Chad Tracy, and Elizabeth Takacs. Best practices for reducing bias in the interview process. *Current Urology Reports*, 23(11):319–325, 2022. doi:10.1007/s11934-022-01116-7.

[4] Jeffrey S Berger and Anne Cioletti. Viewpoint from 2 graduate medical education deans application overload in the residency match process. *Journal of Graduate Medical Education*, 8(3):317–321, 2016. doi:10.4300/JGME-D-16-00239.1.

[5] Vikram R Bhargava and Pooria Assadi. Hiring, algorithms, and choice: Why interviews still matter. *Business Ethics Quarterly*, 34(2):201–230, 2024. doi:10.1017/beq.2022.41.

[6] Maggie Blehar. What is high volume recruiting and what are the challenges? https://www.phenom.com/blog/high-volume-recruiting-challenges, April 2025. Phenom blog.

[7] Magdalena Borys and Maciej Laskowski. Expert vs novice evaluators—comparison of heuristic evaluation assessment. In *International Conference on Enterprise Information Systems*, volume 2, pages 144–149. SCITEPRESS, 2014. doi:10.5220/0004970901440149.

[8] Chris Brown and Swanand Vaishampayan. Towards evidence-based tech hiring pipelines. *arXiv preprint arXiv:2504.06387*, 2025. arXiv:2504.06387, doi:10.48550/arXiv.2504.06387.

[9] Canadian Centre on Substance Use and Addiction. Behavioural competencies for Canada's substance use workforce v.2: Guide to behavioural competency-based interviewing, 2021. URL: https://www.ccsa.ca/sites/default/files/2021-05/CCSA-WFDB-Behavioural-Competency-Based-Interviewing-Guide-2021-en_0.pdf.

[10] Carelever. The cost of delay: How lengthy hiring processes increase candidate dropout rates. https://www.carelever.com/the-cost-of-delay/, March 2025.

[11] Emilio J. Castilla. AI is reinventing hiring—with the same old biases. Here's how to avoid that trap. *MIT Sloan Ideas Made to Matter*, December

2025. URL: https://mitsloan.mit.edu/ideas-made-to-matter/ai-reinventing-hiring-same-old-biases-heres-how-to-avoid-trap.

[12] Hwayoung Cho, Gail Keenan, Olatunde O Madandola, Fabiana Cristina Dos Santos, Tamara GR Macieira, Ragnhildur I Bjarnadottir, Karen JB Priola, and Karen Dunn Lopez. Assessing the usability of a clinical decision support system: heuristic evaluation. *JMIR Human Factors*, 9(2):e31758, 2022. doi:10.2196/31758.

[13] Cheng-Yu Chung, I-Han Hsiao, and Yi-Ling Lin. AI-assisted programming question generation: Constructing semantic networks of programming knowledge by local knowledge graph and abstract syntax tree. *Journal of Research on Technology in Education*, 55(1):94–110, 2023. doi:10.1080/15391523.2022.2123872.

[14] CrewAI Contributors. CrewAI. https://github.com/crewAIInc/crewAI, 2023.

[15] Jason Dana, Robyn Dawes, and Nathanial Peterson. Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making*, 8(5):512–520, 2013. doi:10.1017/S1930297500003612.

[16] Soham Datta, Prabir Mallick, Sangameshwar Patil, Indrajit Bhattacharya, and Girish Palshikar. Generating an optimal interview question plan using a knowledge graph and integer linear programming. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1996–2005, 2021. doi:10.18653/v1/2021.naacl-main.160.

[17] Joachim De Baer, A Seza Doğruöz, Thomas Demeester, and Chris Develder. Single- vs. dual-prompt dialogue generation with LLMs for job interviews in human resources. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 947–957, 2025.

[18] Docker, Inc. Docker. https://www.docker.com/, 2013.

[19] Danilo Dukanovic and Dario Krpan. Comparing chatbots to psychometric tests in hiring: reduced social desirability bias, but lower predictive validity. *Frontiers in Psychology*, 16:1564979, 2025. doi:10.3389/fpsyg.2025.1564979.

[20] Therese Enarsson, Lena Enqvist, and Markus Naarttijärvi. Approaching the human in the loop—legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, 31(1):123–153, 2022. doi:10.1080/13600834.2021.1958860.

[21] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Official Journal of the European Union, L 119, 1–88, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[22] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). OJ L, 2024/1689, 12.7.2024, 2024. ELI: http://data.europa.eu/eli/reg/2024/1689/oj. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[23] FoxMatrix. From chaos to consistency: Why interview quality is now a market differentiator, 2025. URL: https://www.linkedin.com/pulse/from-chaos-consistency-why-interview-quality-now-market-differentiator-kd2dc.

[24] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. Identifying and improving disability bias in GPT-based resume screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 687–700, 2024. doi:10.1145/3630106.3658933.

[25] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. HireNet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 573–581, 2019. doi:10.1609/aaai.v33i01.3301573.

[26] S. Jain and S. Saxena. Exploring the implications of artificial intelligence in the recruitment process: A systematic literature review. *International Journal of Advanced Research in Commerce, Management & Social Science*, 8(4(I)):182–190, 2025. doi:10.62823/ijarcmss/8.4(i).8207.

[27] Athanasis Karoulis, Stavros Demetriadis, and Andreas Pombortsis. Comparison of expert-based and empirical evaluation methodologies in the case of a CBL environment: the "Orestis" experience. *Computers & Education*, 47(2):172–185, 2006. doi:10.1016/j.compedu.2004.09.002.

[28] Nam Ho Koh, Joseph Plata, and Joyce Chai. BAD: BiAs detection for large language models in the context of candidate screening. *arXiv preprint arXiv:2305.10407*, 2023. arXiv:2305.10407, doi:10.48550/arXiv.2305.10407.

[29] Joshua A Kroll. Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 758–771, 2021. doi:10.1145/3442188.3445937.

[30] Linchi Kwok, Ning-Kuang Chuang, and Yung-Kuei Huang. Staying ahead: A critical assessment of the GenAI-assisted hiring process using the OECD AI principles. *Journal of Hospitality & Tourism Research*, page 10963480251391278, 2025. doi:10.1177/10963480251391278.

[31] Linda Lapp, Kieren Egan, Lisa McCann, Moira Mackenzie, Ann Wales, and Roma Maguire. Decision support tools in adult long-term care facilities: scoping review. *Journal of Medical Internet Research*, 24(9):e39681, 2022. doi:10.2196/39681.

[32] Mingzhe Li, Xiuying Chen, Weiheng Liao, Yang Song, Tao Zhang, Dongyan Zhao, and Rui Yan. EZInterviewer: To improve job interview performance with mock interview generator. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1102–1110, 2023. doi:10.1145/3539597.3570476.

[33] Steven Lu. Structured interviews: How to run them and why they work, March 2026. URL: https://www.pin.com/blog/structured-interviews-guide/.

[34] Sean Maynard, Frada Burstein, and David Arnott. A multi-faceted decision support system evaluation approach. *Journal of Decision Systems*, 10(3-4):395–428, 2001. doi:10.3166/jds.10.395-428.

[35] Meta Platforms, Inc. React. https://react.dev/, 2013.

[36] Metaintro. Why tricky job interview questions are useless, according to experts, 2026. URL: https://www.metaintro.com/blog/tricky-job-interview-questions-useless-experts.

[37] Microsoft Corporation. Microsoft Corporation. https://www.microsoft.com/, 1975. Technology company. Accessed: 2026-02-24.

[38] Microsoft Corporation. Microsoft Azure. https://azure.microsoft.com/, 2010.

[39] Microsoft Corporation. Azure Cosmos DB. https://azure.microsoft.com/en-us/services/cosmos-db/, 2017.

[40] Dena F Mujtaba and Nihar R Mahapatra. Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions. *arXiv preprint arXiv:2405.19699*, 2024. arXiv:2405.19699, doi:10.48550/arXiv.2405.19699.

[41] Nikahat Mulla and Prachi Gharpure. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32, 2023. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9886210/, doi:10.1007/s13748-023-00295-9.

[42] Henk Herman Nap, Nathalie E Stolwijk, Sima Ipakchian Askari, Dirk RM Lukkien, Bob M Hofstede, Nicole Morresi, Sara Casaccia, Giulio Amabili, Roberta Bevilacqua, Arianna Margaritini, et al. The evaluation of a decision support system integrating assistive technology for people with dementia at home. *Frontiers in Dementia*, 3:1400624, 2024. doi:10.3389/frdem.2024.1400624.

[43] Python Software Foundation. Python programming language. https://www.python.org/, 1991.

[44] Sebastián Ramírez. FastAPI, 2018. URL: https://github.com/fastapi/fastapi.

[45] Reflex. Reflex: Web apps in pure Python, 2022. URL: https://github.com/reflex-dev/reflex.

[46] Sergio Rojas-Galeano, Jorge Posada, and Esteban Ordoñez. A bibliometric perspective on AI research for job-résumé matching. *The Scientific World Journal*, 2022(1):8002363, 2022. doi:10.1155/2022/8002363.

[47] Aaradhana Rukadikar and Komal Khandelwal. Navigating change: a qualitative exploration of chatbot adoption in recruitment. *Cogent Business & Management*, 11(1):2345759, 2024. doi:10.1080/23311975.2024.2345759.

[48] Carlos Sáenz-Royo and Francisco Chiclana. The value of expert judgments in decision support systems. *Applied Soft Computing*, 171:112806, 2025. URL: https://www.sciencedirect.com/science/article/pii/S1568494625001176, doi:10.1016/j.asoc.2025.112806.

[49] Ramteja Sajja, Yusuf Sermet, Brian Fodale, and Ibrahim Demir. Evaluating AI-powered learning assistants in engineering higher education: Student engagement, ethical challenges, and policy implications. *arXiv preprint arXiv:2506.05699*, 2025. arXiv:2506.05699, doi:10.48550/arXiv.2506.05699.

[50] Daniela Sele and Marina Chugunova. Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE*, 19(2):e0298037, 2024. doi:10.1371/journal.pone.0298037.

[51] Ceyhan Ceran Serdar, Murat Cihan, Doğan Yücel, and Muhittin A Serdar. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, 31(1):27–53, 2021. doi:10.11613/BM.2021.010502.

[52] Servizi Aziendali PricewaterhouseCoopers S.r.l. ATS privacy statement: Privacy notice given pursuant to article 13, GDPR. PwC Italy careers portal, 2026. Accessed: 2026-09-16. URL: https://jobs-it.pwc.com/it/it/privacy-policy.

[53] Stanford University. CS329X lecture 8: Evaluate human-AI interaction. https://web.stanford.edu/class/cs329x/slides/Lecture8_evaluate_HAI.pdf, 2025.

[54] Streamlit Inc. Streamlit. https://streamlit.io/, 2019.

[55] KP Suresh and Sachin Chandrashekara. Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences*, 5(1):7–13, 2012. doi:10.4103/0974-1208.97779.

[56] TalentHub. Time to hire: What it is, why it matters, and how to reduce it. https://www.talenthub.eu/blog/time-to-hire-what-it-is-why-it-matters-and-how-to-reduce-it, May 2025.

[57] CR Wilson VanVoorhis and Betsy L Morgan. Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2):43–50, 2007. doi:10.20982/tqmp.03.2.p043.

[58] Drew Whitehurst. Why interview scheduling is the biggest bottleneck in hiring (and how to fix it), 2026. Interviewstream blog. URL: https://interviewstream.com/resources/blog/interview-scheduling-bottleneck-how-to-fix-it/.

[59] Kyra Wilson, Mattea Sim, Anna-Maria Gueorguieva, and Aylin Caliskan. No thoughts just AI: Biased LLM hiring recommendations alter human decision making and limit human autonomy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2692–2704, 2025. doi:10.1609/aies.v8i3.36749.

[60] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. If I hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. doi:10.1145/3313831.3376424.

[61] Zhengtao Xu, Zimo Xia, Zicheng Zhu, Nattapat Boonprakong, Yu-An Chen, Rabih Zbib, Casimiro Pio Carrino, and Yi-Chieh Lee. InterPilot: Exploring the design space of AI-assisted job interview support for HR professionals. *arXiv preprint arXiv:2602.20891*, 2026. arXiv:2602.20891, doi:10.48550/arXiv.2602.20891.

[62] Yusuf Zorlu. The role of artificial intelligence (AI)-assisted question generation on preservice science teachers' 21st-century teaching competencies. *Journal of Baltic Science Education*, 24(5):984–1005, 2025. doi:10.33225/jbse/25.24.984.

Appendix A

Research interviews

In this Appendix, three semi-structured interviews conducted by the author of this thesis, Andrea Nappi, with technical managers involved in hiring interviews at PwC Italy are presented. The aim was to develop a grounded understanding of the current hiring pipeline, including its functioning, structure, problems, and main characteristics. The interviews covered the following topics:

- Brief introduction.
- Overview of the hiring pipeline.
- Personal preferences in interview structure.
- Interview preparation.
- Perceived usefulness of a system such as the one discussed in the thesis.

The interviews were anonymized and lightly polished from the original transcripts to improve readability, while preserving their meaning and content entirely.

A.1 Interview with manager 1

Metadata

Interviewee:	manager 1
Role:	manager, Data & AI
Interview date:	31-03-2026
Interview language:	English
Duration:	00h 21m 05s

Transcript

Interviewer: Could you introduce yourself briefly?
manager 1: Sure, I am a manager in the Data & AI team. I have been a manager for a couple of years, and I have been involved in recruiting during this period, because before that, I was not allowed.

Interviewer: Thank you. Could you now explain what the stages of the hiring process are for the roles you usually deal with, starting from when the position is opened, or more generally when there is a need for a new resource, all the way to the final decision?

manager 1: Let's say that the pipeline in general starts with a request at the team level to have someone inside the team with a certain seniority, usually an Intern, an associate, a senior associate, and so on. Then it has to be validated at a higher level, at the partner level, because we are not the ones who decide on hiring. Once it is decided to open the position, HR opens the position both on LinkedIn and by directly searching for candidates, I think, through the various portals they have available. Usually, they first send us the CVs, or, if they know that maybe we need to cover a large number of positions, they directly schedule the interview with the profiles that seem aligned. The first step is an interview with HR, so each candidate has an individual interview with HR, and immediately after that, they have a technical interview with us managers. Then it depends on the role the person will have to cover, because usually HR, after its interview, classifies the candidate according to a role, so it tells you whether, depending on experience, they can be classified as an Intern, as an associate, or senior associate, and so on. Our interview is a 45-minute technical interview; after that, we report our feedback to HR. If it is positive, HR asks the partner for permission to send out the hiring proposal, and then it goes back to the candidate as an offer.

Interviewer: Thanks. Regarding the seniority part that is sometimes communicated by HR, does that always happen or only sometimes?
manager 1: A proposal is always communicated by HR. Obviously, we also always form an idea from the CV, because you can see the years of experience, and you also know what you need within the team, so often you see CVs of recent graduates or people with little experience. Then we confirm it or not through our technical interview. Sometimes HR says: "We would have placed them in a higher role", and instead, after the interview, we do not consider it appropriate to place them in such a high role. So a lower proposal is made, and then you see whether the candidate accepts it, or the opposite, although that is rarer.

Interviewer: Got it. Personally, for the part you deal with, how do you structure the interviews?
manager 1: They are pretty standard, actually. There is an initial introductory part about the team, so who we are and what we do. Then there is a second part where we let the candidate speak: we ask them who they are, what they have done, and why they applied for this position, and we see whether their interests are aligned with what we explained. After that, we move on to some questions that are first a bit more general about their background, and maybe we notice some points in their profile that refer either to projects like ours or to something that is a bit different from what we deal with. There, we try to understand whether they can actually be aligned with what we do or not, and then we move on to slightly more technical questions about what we deal with, based on the level that has been communicated.

Interviewer: Do you know if there are any PwC Italy standards or guidelines to follow during interviews, or if there are optional or mandatory parts, or any direction in general?
manager 1: Nothing in particular, really, I think. I know we are very free on the technical side, at least. Then, of course, there are the standard things, which are also a bit common sense, such as not bringing up sensitive topics.

Interviewer: Are there differences, maybe structural or even just technical, for interviews at different seniority levels?
manager 1: The differences I have seen are mostly technical. Structurally, in my opinion, things only change for roles above senior associate, so if you go into a role like manager, then it is not just the technical side that comes into play, but also the relationship and management side, and for roles above that, deep technical knowledge is almost never ever asked. But that is handled directly by the partners or directors.

Interviewer: The next question is: how do you generally prepare for interviews, if you prepare at all?
manager 1: Well, what I do is look at the candidate's CV, so I see whether they already have

work-related skills or whether they are a recent graduate or still studying, and I try to understand roughly what stage of experience they are at. I look at what their strengths are, what the main projects they have done are, and whether there is anything related to what we do directly. Or, based on the degree, maybe in general computer science, they may not yet have touched on the topics we work with, so I try to understand whether they already have experience in common with what we do. Based on that, I think of more technical questions, which a person who studied computer science may know, or otherwise slightly more specific questions about our field, like AI at this moment. If you want to frame it as preparation, you think about the questions. Maybe you look for something to structure them better, and sometimes a file of questions is created, with questions and answers already there, so that you have a guideline and you are not just going by feel. So yes, I would tell you the structure of a file with questions and answers, and if you find something more specific, depending on the CV, you put it in there.

Interviewer: Can you tell me about some of the difficulties you have encountered in this process, if there are any particular ones? Not necessarily technical difficulties, but also difficulties such as finding the time.
manager 1: Yes, many times interviews are something above all the other activities, they are not the main activity you do. So, unfortunately, sometimes you are a bit chasing things, so you find yourself with interviews and think: "I have to do an interview", and you sort of end up rushing to look at the CV, while you would need a bit more time. You are chasing to find the questions and go back over them, so yes, this is definitely an issue that unfortunately exists, because you are not dedicated only to that.

Interviewer: Regarding the questions in an interview, in your opinion, what is a good set of questions? What characteristics does it have, and which questions are perhaps not effective?
manager 1: Just like this, it is not easy to say, because obviously it has to be calibrated based on the person in front of you, based on their experience. You cannot ask them about something they have never seen, maybe a technology, just because you use it and say: "I will ask you this, and if you know it, good, if not, no". Then you have to be clear, not too cryptic, to the point of making them wonder what you mean. So, in my opinion, a good question is a clear question.

Interviewer: I would like to propose a scenario now. Imagine having access to an AI tool that generates ideal sets of questions for each candidate, ideal meaning based on their profile, as far as what can be inferred from the CV, and also based on the position they are being interviewed for. It would also be calibrated to seniority, so this tool would propose both behavioural and technical questions and could also provide an example of an adequate answer that is appropriate to the question. If you had the chance to use such a tool, do you think you would use it?
manager 1: Yes, I definitely think so. In fact, keep me updated, because I am interested.

Interviewer: If you said you would use it, do you think you would use it more as a starting point for ideas for some questions, or would you rely on it a lot, assuming the questions are valid in themselves?
manager 1: It depends. I imagine that having the questions and answers already prepared helps a lot because you can say: "Today I did not have time to prepare my usual questions and answers, I looked at the CV, I read it, the questions are relevant, I select them, and I already have the answer". So sometimes you could do that. Other times, instead, you want to understand certain topics more deeply, so you say: "OK, I will let myself be inspired".

Interviewer: Still on this tool, in your opinion, are there topics that should be avoided in the questions, things this tool should take into account? Within the limits of common sense, obviously.
manager 1: Yes, there are certainly out-of-place questions at a personal level that can maybe put the candidate in difficulty and, in fact, do not bring any added value to the interview.

Interviewer: All right, thanks. If you have nothing else to add, we are done.
manager 1: All right, thank you.

A.2 Interview with manager 2

Metadata

Interviewee:	manager 2
Role:	manager, Data & AI
Interview date:	31-03-2026
Interview language:	English
Duration:	00h 25m 46s

Transcript

Interviewer: Good morning, first of all, I'd like to ask you to introduce yourself, so who you are and what role you have in the company.
manager 2: OK, so I've been working at PwC for four years, and I'm currently a manager in the AI Engineering team, and I'm responsible for leading the project side in the field of generative AI and agentic AI within the team. At the same time, I also handle everything related to the team's budget and staffing.

Interviewer: Perfect, thank you. What are the stages of the hiring process for the roles you usually deal with, starting from when the position becomes available or open until the final decision?
manager 2: OK, basically every year at the beginning of the year, we have a budget of available positions, by role, for the people we can hire. What does this involve? It means that every year, we create a job description for each of the three roles I deal with, namely Intern, associate, and senior associate, which gets published, so that people can obviously apply. Once that is done, the first step that starts the selection process is that the candidate opens the job position, sends their CV, and applies. The CV is first screened by HR based on some minimum basic skills that must be met and that are included in the job description, and at this stage, there is no intervention, no real knowledge of the candidate, so there are no relations with the possible candidate on my part or on the part of the other managers on my team. HR screens and enters one or more CVs that they consider suitable and tag with our team inside an Excel file, so then I go there and I have the person with an initial seniority level that is taken from the CV and I can see the CV, so I analyze the CV, look at it together with the other managers, we understand whether the CV is compliant or not with what we are looking for, and we ask HR to contact the person to set up an interview. The interview usually lasts an hour and a half, where in the first half hour we managers do not take part, only HR does, and they ask motivational questions and questions about the economic level, meaning about the salary the person currently receives if they have previous work experience, about the salary expectation that the person has, and from that HR hypothesizes a possible entry role. OK, after that comes our interview, which is divided into three moments: a first moment in which we explain what we do, how our projects are organized, how our team is structured, and the training activities that take place. A second part in which the candidate asks questions about what has been said and talks about themselves, and a third part in which technical questions are asked, together with practical exercises. Once this step is finished, we give our feedback to HR. If the feedback is positive, HR makes a request to the partner to have the budget approved, and if the partner approves, an offer is made to the candidate. If the candidate accepts, the selection process ends with the hiring; otherwise, obviously not.

Interviewer: OK. You mentioned that seniority is sometimes influenced by the previous role the person held or is otherwise communicated by HR before hiring. How often does it actually happen

that a seniority level is imposed regardless?
manager 2: Never. I mean, HR obviously evaluates a candidate's entry seniority not so much on actual experience or actual knowledge, but on their previous salary. This information is fundamental for us because we calibrate the interview exactly on the seniority that HR indicates to us. But very often they have indicated, for example, people with only a few years of experience, very few years of experience, as senior associates, and for us, usually that is a no. Because to be a senior, you need at least two years of experience, even if you are technically good, you still need soft skills that take time to develop. If you have more than two years of experience, then we go into the technical side, because the senior role, for example, requires a certain technical leadership, and therefore, if a candidate is proposed to us as a senior associate, we have to assess their actual seniority. And sometimes they communicate it to us, and other times they do not.

Interviewer: OK, perfect. Regarding the part you handle within the hiring process, so the more technical interview section, are there any general PwC Italy standards or guidelines that need to be followed?
manager 2: Definitely, we are, let's say, required to explain the organizational structure of PwC in general, the organizational structure of our service line, which specifically is advisory. We also have guidelines like avoiding giving client names, avoiding giving project amounts, and avoiding giving too much information on data that we cannot share with third parties. On the technical part of the interview, we have total freedom.

Interviewer: OK, good. And are there big differences in interviews for candidates at different seniority levels?
manager 2: Yes, yes, there are differences, because for senior associate, more complex questions are asked, obviously aimed at assessing specific knowledge. Whereas for Interns and also associates who maybe have little experience, the tendency is less to assess specific knowledge of a topic and more the way they think and approach a problem, while for a senior associate you do not expect the way they approach the problem, but you expect full knowledge of that technology, because they have to handle possible technical problems, have to propose technical solutions, and so there is a completely different paradigm, obviously for Interns it is totally linked to the way they reason.

Interviewer: OK. How do you prepare for an interview, if you prepare at all?
manager 2: Yes, I prepare for every interview, and what I do is I have already studied the CV at the initial stage, when I ask HR to interview, to schedule the interview with the candidate, so there I had already formed a small idea of the person I have to interview. Inside an Excel file, I list all the people I have to speak with, together with my colleagues, and we have a notes field where we write down peculiarities or things we may want to focus on during the interview. This, combined with a set of questions that we have in a Word file and that we randomly select during the interview phase, and then adapt, let's say, based on how the discussion evolves. Very often, candidates may tell us something that is not in the CV, which sparks our curiosity, and then we go a bit more into detail about that.

Interviewer: Sure. Still, regarding preparation for interviews in general, how much time does it take you, and how much does having a more or less full schedule cause you problems?
manager 2: Yes, so generally what I do to prepare my interviews is one of the activities I do after regular working hours, so I check whether I have interviews to do the next day, and I spend about half an hour identifying the points I should focus on in the interview for the next day. So in terms of time, preparation is about half an hour per candidate, if I had to estimate it.

Interviewer: Perfect. And regarding the questions asked during an interview, what do you think are the attributes that let you decide whether a question is actually a good question or a bad one?
manager 2: OK, so let's say that the first rationale I take into consideration, also looking at the standard questions we have, is definitely how much one question compared to another can fit

the candidate's competence. Competence can be of two types: work-related competence or basic competence. Very often, we interview Interns, so there are questions you cannot ask them because you can only ask questions related to their academic path. Obviously, we do not all have the same academic path, so certain questions you cannot ask maybe a mathematical engineering student the same way you would ask an IT engineering student, because your basic competence is different. So you need to try to adapt a bit what you have in terms of written questions to what the candidate can actually know or not know, also adapting to possible academic changes in what is studied over time. So this is also an important parameter because very often we find ourselves asking questions and discovering that certain things are no longer studied, and so a good question in my opinion is one that can be identified like this. Having a set of questions makes it complex. In fact, usually the tendency is to ask the same ones and more general ones, which is not positive because then you do not go into detail, nor maybe into the specifics of some peculiarities of the person in front of you.

Interviewer: Thank you. I'd like to propose a scenario now: imagine having access to an AI-based tool that generates an ideal set of questions for each candidate, partly based on their profile, inferred from the CV, and partly based on the position they are being interviewed for. This tool would propose both behavioral questions and technical questions, and they would also be calibrated to seniority, and would also provide possible examples of suitable answers. If you had such a tool available, would you use it?
manager 2: Yes, absolutely. Maybe something like that can help both the candidate and the evaluator for two reasons: on the evaluator side, because it gives you the chance to ask targeted questions based on the candidate's actual experience and what you are looking for, and it also makes the process as general as possible even with respect to the evaluator themselves, so it is not necessary to always have the most technically expert person present during interviews, and you can also divide the various interviews fairly and guarantee attention, as well as make the recruitment process more efficient, with even less technical colleagues being enabled to ask advanced questions maybe. On the candidate side, you might have questions asked being totally relevant to them, and so the answers they can give have to come from their own background, from their experience, and they do not have to answer questions about something they do not know, and risk seeing their interview affected by that aspect.

Interviewer: Regarding the type of use, do you think you would use it as a starting point to structure an interview, or would you directly use the questions provided after evaluating them?
manager 2: Well, given the structure of AI applications today, I would still want to make the final decision myself on the question, meaning the first thing I would do is study the generated questions and understand from that set which ones are the most suitable. It could also be that I think those questions are not suitable, so I would use them as the main support, but I still want the final decision to remain mine.

Interviewer: Perfect. Regarding this tool, in your opinion, should there be topics that absolutely must be avoided, maybe during an interview, things that this tool should take into account?
manager 2: Yes, definitely. It should avoid topics concerning gender, political views, and, if they can somehow emerge from volunteer activities included in the CV, also religious matters, so based on nationality and everything. In short, everything that concerns the person themselves and is not related to the work sphere must be left out.

Interviewer: OK, perfect, then actually we're done, so if there is anything else you think would be useful to know or that could be interesting for this interview, let me know, otherwise we're done.
manager 2: I think I've said everything, yes.

A.3 Interview with manager 3

Metadata

Interviewee:	manager 3
Role:	manager, Data & AI
Interview date:	31-03-2026
Interview language:	English
Duration:	00h 19m 20s

Transcript

Interviewer: So, the first thing I'd like to ask is a brief introduction to who you are and what role you have within PwC Italy.
manager 3: So, I've been a manager in Data & AI for 4/5 years. I mainly work on projects related to data and artificial intelligence.

Interviewer: Perfect. Can you give me a description of the stages of the hiring process for the roles you usually deal with, starting from when the position is open, all the way to the final decision, whether positive or negative? How does the whole process work?
manager 3: Yes, so we signal to HR the need to bring in a new resource. HR then takes care of placing the ad on the various channels, such as LinkedIn, the PwC portal, and so on, collects a series of CVs, contacts us, and gives us the list and the CVs of the people who applied to the ad. After that, we on the technical recruiting side review them, so we manually take each CV, read it, and for those that in our view reflect the requirements we are looking for, we then ask HR to schedule the interview. HR then organizes these interviews, usually two of them: one on the softer-skill side, handled by HR, and then the technical one, which is done by the other manager colleagues and me. Once the interview is finished, if there is positive feedback from both sides, HR eventually handles the offer.

Interviewer: Okay. Regarding entry seniority, who decides it?
manager 3: Well, when we do technical interviews, we more or less realize the level a person has. And we give our opinion to HR, so we interview a candidate, and if they performed well, were good, and also had previous experience, then in that case we might say that, in our opinion, they can enter as a senior figure. Then, after the interview, there is a discussion with HR, HR gives us its feedback, there is a discussion, and we try to come out with that role together; that's it.

Interviewer: Personally, how do you structure the part of the interview you handle?
manager 3: Usually, we always start with a presentation of the group, the day-by-day activities we do, and the whole learning side we have, by learning I mean all the paths, including getting certifications, with the possibility of taking courses in both soft and hard skills. Then we give the floor to the candidate, the candidate introduces themselves, and after that we move to a slightly more technical phase where, based on the candidate's experience, we ask targeted questions. And recently, we have also introduced an even more technical part where there are exercises to be done in Python, pseudocode, or at least reasoning exercises, to see how the candidate behaves when dealing with certain types of problems.

Interviewer: OK, and as far as you know, are there any company standards or guidelines to follow during an interview? Or any parts that are optional or mandatory?
manager 3: There is nothing mandatory. There are some courses that they give you to enable you to conduct interviews, but there isn't a fixed sequence you have to follow. You can decide the structure a bit yourself.

Interviewer: OK, perfect. And are there major differences in interviews for different seniority levels?
manager 3: Yes, meaning that obviously if you are interviewing a guy who has just graduated from university, the level of technical questions will definitely be lower than for a higher seniority, where more specific skills are required. And also other types of skills, so yes.

Interviewer: OK. Regarding the stage before an interview, do you usually prepare for an interview? If yes, how?
manager 3: Well, generally, I just look at what they have done and the technologies they have worked on. And I go back over them just to be able to ask questions and make sure the candidate actually has the skills they put in the CV.

Interviewer: OK. Do you think that if you had a less intense calendar schedule, you could spend more time preparing or preparing more deeply, or do you think it is not too necessary?
manager 3: No, for the profiles we look for, I don't think it is necessary.

Interviewer: OK, perfect. Then the next thing I'd like to ask is whether you could tell me some characteristics that might help distinguish good interview questions from bad ones, and from not-so-good questions.
manager 3: In my opinion, it depends. You should not ask questions that are too basic, but rather questions centered on the person's CV, because it is useless to ask about things the candidate may not have seen.

Interviewer: OK, how important do you think it is to be, let's say, aligned in terms of complexity level across interviews for people of the same seniority? And how do interviews for different seniority levels change?
manager 3: Well, in our group, it changes because, as you go further, more things are required of you. Then for people of the same seniority we try to be as consistent as possible, then there may be someone who knows a little more and so you may be more inclined to go into more depth, but we try to keep things fairly standard also to have a comparison, in the sense that if we interview three senior profiles we want to have a method to judge all three of them, so if you do three completely different interviews afterward you would have no way of really comparing them with each other.

Interviewer: OK, now I'll propose a scenario. Imagine having access to an AI tool that generates ideal sets of questions for each candidate based on their profile and the position they are applying for. This tool could give you both behavioral questions and technical questions, calibrated to seniority as well, with examples of plausible, appropriate answers. Do you think you would use it?
manager 3: Yes, definitely, as a guide to make my job a bit easier, even if I think I would still want to add something of my own. Maybe I can start from those questions, and if I see that the candidate knows them well, and so on, then maybe I can decide to ask additional questions to go a bit deeper into the topic, or decide to move on to the next questions you made available to me. So I would definitely use it as a starting point, and then, based on how the interview develops, understand how to go forward. That's it.

Interviewer: OK. Regarding this tool that proposes questions, do you think there are topics that it should absolutely avoid, maybe during an interview, things this tool should take into account?
manager 3: No, the important thing is that it focuses on the technical aspect and on the things the person has done, rather than going into topics that have nothing to do with it.

Interviewer: OK, perfect, then we're done. I don't know if there is anything else you think would be useful to share about the hiring process, otherwise we're good.
manager 3: No, that's all good.

Appendix B

Evaluation rubric and participant instructions

This Chapter displays the full survey instrument that was given to technical interviewers at PwC Italy who used Talent Assistant.

B.1 Survey for Interviewers: Evaluation of Talent Assistant

Informed consent

You are invited to participate in a research survey about Talent Assistant, an AI system used to support technical interviews at PwC Italy. This study is conducted as part of a Master's thesis at the University of Twente and aims to understand how Talent Assistant is used in practice, how useful it is for interviewers, and which limitations or improvement opportunities may exist.

The survey is intended only for technical interviewers at PwC Italy who have used Talent Assistant in at least one real interview process. Completing the survey should take approximately 10 minutes.

Your responses will be used for academic research, with the results being analyzed and reported only in aggregate form, and no individual respondent will be identifiable.

The survey does not ask for special category personal data, and you are asked not to include unnecessary personal or highly sensitive information in any open-ended responses. Any personal data collected through this survey will be handled confidentially and processed only for the purposes described above. All raw survey data will be retained within PwC Italy and accessed only by authorized individuals involved in this research.

If you have questions about the study, you may contact the researcher and, if appropriate, the thesis supervisor. Their names and contact details are listed below this consent text.

Researcher: Andrea Nappi, a.nappi@student.utwente.nl
Supervisor: Alexia Briassouli, a.briassouli@utwente.nl

By selecting "I agree" below and proceeding to the survey, you confirm that:

- You have read and understood the information above.
- You are 18 years of age or older.

- You have experience using Talent Assistant in at least one real PwC Italy interview process.
- You agree to participate in this study.

Do you agree to participate in this survey under the conditions described above?

- Yes, I agree to participate.
- No, I do not agree to participate.

Section A: Background

1. **Your current position at PwC Italy** (select one)
   - Associate
   - Senior Associate
   - Manager
   - Senior Manager
   - Director
   - Partner
2. **Your Line of Service and Team section** (e.g., Data & AI, AI Engineering)
   - _____
3. **Your primary technical domain(s) for your interviews** (e.g., data engineering, artificial intelligence, private equity, legal audit)
   - _____
4. **Approximate total years of experience conducting technical interviews you have (at any employer)**
   - Less than 1 year
   - 2 – 3 years
   - 4 – 6 years
   - 7 – 10 years
   - More than 10 years
5. **Approximate number of technical interviews you conduct per year at PwC Italy**
   - Fewer than 10
   - 10 – 24
   - 25 – 49
   - 50 – 99
   - 100 or more
6. **Roles for which you have interviewed** (select all that apply)
   - Intern

- Associate
- Senior Associate
- Manager

7. **How many candidates have you interviewed so far using Talent Assistant?**
   - Fewer than 10 candidates
   - 10 – 19 candidates
   - 20 – 29 candidates
   - 30 candidates or more
8. **In which phases of the interview process have you mainly used Talent Assistant?** (select all that apply)
   - Before conducting the interview (e.g., helping prepare questions or topics)
   - During the live interview (e.g., relying on the provided information, questions, or answers)
   - After the interview (e.g., structuring feedback or evaluations)
   - Other (please specify): _____
9. **How familiar do you feel with Talent Assistant overall?**
   - 1 – Not at all familiar
   - 2
   - 3 – Moderately familiar
   - 4
   - 5 – Very familiar

Section B: Perceived usefulness for the interview process

Please indicate to what extent you agree or disagree with each of the following statements using the following 1 – 7 scale, where 4 (neither agree nor disagree) can be interpreted as the expected benchmark of your way of preparing and conducting interviews without Talent Assistant.

1 – Strongly disagree 2 – Disagree 3 – Slightly disagree 4 – Neither agree nor disagree 5 – Slightly agree 6 – Agree 7 – Strongly agree

1. Using Talent Assistant helps me prepare interviews more quickly than I otherwise could.
2. Using Talent Assistant helps me prepare interviews more efficiently, with less effort on my side.
3. Talent Assistant helps me generate relevant technical questions or prompts that I might not have thought of on my own.
4. Overall, Talent Assistant improves the quality of my technical interviews.
5. Talent Assistant adds value compared to my previous way of preparing and running interviews without the tool.

Section C: Perceived quality of Talent Assistant's outputs

Please use the 1 – 7 scale provided in section B to answer the following questions.

1. The questions generated by Talent Assistant focus directly on the candidate's declared (in the resume) relevant skills and experiences for the role.
2. The questions generated by Talent Assistant are aligned with the job position and requirements we are hiring for.
3. The questions generated by Talent Assistant are clear and relevant, and expected to be met with precise and specific answers.
4. The questions generated by Talent Assistant are appropriate for the seniority level of the interviewed candidate, and are neither too simple nor too complex.
5. The questions generated by Talent Assistant focus on the right aspects for the candidates, in line with their seniority (e.g., on solution approach for junior figures and on deep expertise for senior figures).
6. Considering all generated questions by Talent Assistant for a given candidate, the provided set covers all the core relevant macro-areas that should be assessed for the role.
7. Considering all generated questions by Talent Assistant for a given candidate, they cover complementary aspects of the candidate's experience and job position, and don't overlap.
8. The questions and prompts generated by Talent Assistant are technically correct (contain no major conceptual errors).
9. The questions generated by Talent Assistant can be considered interview-ready with minor to no adjustments.
10. Overall, I am satisfied with the quality of the content that Talent Assistant produces.

Section D: Usage and integration into workflow

Please use the 1 – 7 scale provided in section B to answer the following questions.

1. Learning to operate Talent Assistant was easy for me.
2. Talent Assistant fits reasonably well into my existing preparation workflow.
3. When using Talent Assistant, I still feel fully in control and responsible for which questions and evaluation system I use in the interview.
4. When I use Talent Assistant for comparable candidates in the same role, it suggests similar types of questions across interviews.
5. Differences in Talent Assistant's suggestions across comparable candidates mostly reflect deliberate adjustments I make, rather than random variation in the tool's outputs.

Section E: Open-ended questions

Please answer the following questions in your own words. Short, concrete answers are fully sufficient.

1. Have you observed any clear failure modes or issues when using Talent Assistant? For example: technical issues, misleading suggestions, bias of any kind, poor alignment with PwC's ethical or technical guidelines, etc.

   _____

   _____

2. Are there specific types of interviews or candidates for which you prefer not to use Talent Assistant? Why?

   _____

   _____

3. Is there anything else about your experience with Talent Assistant that you would like to share?

   _____

   _____

End of survey. Thank you for your time and input.

Appendix C

Survey results

The survey was administered online using Microsoft Forms between 5 and 18 August 2026. It was distributed to all ten operational Talent Assistant users who met the inclusion criteria. Participants completed the survey after providing informed consent. Responses were collected confidentially. The survey instrument was presented in English, and was only accessible via the users' official PwC account.

TABLE 8: Aggregated survey responses for the evaluation of Talent Assistant

Question	**Option / response**	**N**
Do you agree to participate in this survey under the conditions described above?	Yes, I agree to participate.	10
Section A: Background		
Your current position at PwC Italy	Senior Associate	5
	Manager	3
	Senior Manager	1
	Director	1
Your Line of Service and Team section	Data & AI	3
	Data & AI, AI Engineering	2
	AI Engineering	3
	Data & AI, Palantir CoE	1
	Advisory | Data & AI - Agentic AI	1
Your primary technical domain(s) for your interviews	Data engineering	4
	Artificial intelligence	7
	Data governance	1
	AI Strategy	1
Approximate total years of experience conducting technical interviews you have (at any employer)	0–1 years	1
	2–3 years	5
	4–6 years	2
	7–10 years	1
	More than 10 years	1

Continued on next page

Question	Option / response	N
Approximate number of technical interviews you conduct per year at PwC Italy	Fewer than 10	1
	10–24	5
	25–49	1
	50–99	1
	100 or more	2
Roles for which you have interviewed	Intern	9
	Associate	9
	Senior Associate	4
	Manager	1
How many candidates have you interviewed so far using Talent Assistant?	Fewer than 10 candidates	2
	10–19 candidates	5
	30 candidates or more	3
In which phases of the interview process have you mainly used Talent Assistant?	Before conducting the interview	9
	During the live interview	7
	After the interview	3
How familiar do you feel with Talent Assistant overall?	4	2
	5 – Very familiar	8
Section B: Perceived usefulness for the interview process		
Using Talent Assistant helps me prepare interviews more quickly than I otherwise could.	5	1
	6	4
	7	5
Using Talent Assistant helps me prepare interviews more efficiently, with less effort on my side.	5	2
	6	3
	7	5
Talent Assistant helps me generate relevant technical questions or prompts that I might not have thought of on my own.	3	1
	5	2
	6	4
	7	3
Overall, Talent Assistant improves the quality of my technical interviews.	6	8
	7	2
Talent Assistant adds value compared to my previous way of preparing and running interviews without the tool.	6	5
	7	5

Continued on next page

Question	Option / response	N
Section C: Perceived quality of Talent Assistant's outputs		
The questions generated by Talent Assistant focus directly on the candidate's declared (in the resume) relevant skills and experiences for the role.	5	2
	6	3
	7	5
The questions generated by Talent Assistant are aligned with the job position and requirements we are hiring for.	5	1
	6	5
	7	4
The questions generated by Talent Assistant are clear and relevant, and expected to be met with precise and specific answers.	5	1
	6	5
	7	4
The questions generated by Talent Assistant are appropriate for the seniority level of the interviewed candidate, and are not too simple nor too complex.	5	2
	6	3
	7	5
The questions generated by Talent Assistant focus on the right aspects for the candidates, in line with their seniority	5	1
	6	6
	7	3
Considering all generated questions by Talent Assistant for a given candidate, the provided set covers all the core relevant macro-areas that should be assessed for the role.	4	1
	6	5
	7	4
Considering all generated questions by Talent Assistant for a given candidate, they cover complementary aspects of the candidate's experience and job position, and don't overlap.	5	3
	6	5
	7	2
The questions and prompts generated by Talent Assistant are technically correct	5	1
	6	5
	7	4
The questions generated by Talent Assistant can be considered interview-ready with minor to no adjustments	4	1
	5	1
	6	5
	7	3

Continued on next page

Question	Option / response	N
Overall, I am satisfied with the quality of the content that Talent Assistant produces.	6	6
	7	4
Section D: Usage and integration into workflow		
Learning to operate Talent Assistant was easy for me.	6	2
	7	8
Talent Assistant fits reasonably well into my existing preparation workflow.	6	5
	7	5
When using Talent Assistant, I still feel fully in control and responsible for which questions and evaluation system I use in the interview.	6	1
	7	9
When I use Talent Assistant for comparable candidates in the same role, it suggests similar types of questions across interviews.	5	1
	6	6
	7	3
Differences in Talent Assistant's suggestions across comparable candidates mostly reflect deliberate adjustments I make, rather than random variation in the tool's outputs.	4	1
	6	7
	7	2
Section E: Open-ended questions		
Have you observed any clear failure modes or issues when using Talent Assistant? For example: technical issues, misleading suggestions, bias of any kind, poor alignment with PwC's ethical or technical guidelines, etc.	No	5
	I never observed any kind or bias or issues	1
	I've never encountered a technical issue that prevented me from using the app	1
	Sometimes similar or too standard questions were generated across different interviews	1
	In the last release I didn't find any kind of issue	1
	All good	1
Are there specific types of interviews or candidates for which you prefer not to use Talent Assistant? Why?	No	4

Continued on next page

Question	Option / response	N
	I will use and feel comfortable to use Talent Assistant with every type of candidate as it adjusts the questions according to the roles and job position	1
	No, I use it for all candidates	1
	Perhaps in roles that aren't strictly technical	1
	No, it can help in the different domain of AI Engineering	1
	I usually use it for everyone	1
	Maybe if I have to interview a manager	1
Is there anything else about your experience with Talent Assistant that you would like to share?	No	6
	Talent Assistant is very useful, especially for interview preparation. It gives me an overview of the candidate by highlighting the key points of their CV and suggesting a list of questions related to their experience and the job description. It significantly reduces the time and effort required to prepare for the interview.	1
	Pay more attention to bias and tool limitations	1
	Great tool	1

Appendix D

Usage of generative AI in the thesis

During the preparation of this thesis, Anthropic Claude models, including Claude Opus versions 4.6, 4.7, and 4.8, have been used as writing support tools. The tools were used to support the restructuring of sections, the development and refinement of outlines, and improvements to the clarity and English of selected passages. The prompting strategy always included a brief prompt explaining the context of the document, a statement of the polishing goal, and one or more previous sections for reference to ensure output consistency. The same models were also used to support the formulation and refinement of survey items.

Generative AI was not used to collect, fabricate, modify, or analyze the empirical data reported in this thesis, including interview material, survey responses, system-usage logs, or statistical results. It was not used as a substitute for the author's judgment or for independent verification of sources. All AI-assisted content was critically reviewed, checked, and finally edited by the author. The author takes full responsibility for the research design, system implementation, data collection, analysis, interpretation, references, and final content of this thesis.