INTERNATIONAL HELLENIC UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF INFORMATION AND ELECTRONIC ENGINEERING

DIPLOMA THESIS

Design and Development of an Educational Platform for Enhancing Soft Skills

Student:
Konstantinos Smaragdas
2019153

Supervisor:
Prof. Periklis Chatzimisios

Thessaloniki, February 2026

Τίτλος Δ.Ε. Σχεδιασμός και ανάπτυξη
εκπαιδευτικής πλατφόρμας για την ενίσχυση των ήπιων δεξιοτήτων
Κωδικός Δ.Ε. 25173
Ονοματεπώνυμο φοιτητή: Κωνσταντίνος Σμαραγδάς
Ονοματεπώνυμο εισηγητή: Καθηγητής Δρ. Περικλής Χατζημίσιος
Ημερομηνία ανάληψης: Δ.Ε. 18-03-2025
Ημερομηνία περάτωσης: Δ.Ε. 23-01-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Σε όσους πίστεψαν σε μένα όταν εγώ αμφέβαλλα.»

Abstract

This thesis presents the design and development of an educational platform aimed at strengthening and assessing key soft skills in engineering university students and early-career professionals, focusing on Communication, Teamwork, Leadership, and Problem Solving. Motivated by the practical difficulty of assessing behavioral and socio-cognitive skills in a consistent and scalable way, the study reviews recent work in AI in Education (AIED) and AI-supported assessment, with emphasis on hybrid Human–AI evaluation approaches, conversational agents, and critical issues such as ethics, bias or fairness, and transparency in automated scoring.

Building on this foundation, this thesis proposes and documents the SoftSkills AI system, a modular architecture consisting of: (a) a React-based frontend (SoftSkills User Quiz) that supports anonymous participation via unique tokens and manages PRE/POST assessment flow, (b) a backend (SoftSkills Bot) implemented with FastAPI and a PostgreSQL database that evaluates responses—especially open-ended answers—through a structured pipeline combining an LLM-based evaluation engine, an enriched heuristic scoring layer (GLMP), and fuzzy score calibration, and (c) a Rater UI that enables independent human rating for validation and comparison with AI-generated scores. The educational component includes structured learning modules organized into three proficiency levels (Beginner, Intermediate, Advanced) to support differentiated learning aligned with the participant's initial performance.

Methodologically, the work follows an experimental, longitudinal pre–post design. Participants complete 16 questions (multiple-choice and open-ended/scenario-based) across the four skill categories in both phases. The system generates a numerical score (0–10) and feedback per response, stores detailed scoring features, and supports analysis of (i) agreement between AI and human ratings (e.g., via inter-rater reliability measures such as ICC), (ii) pre–post-performance change using statistical testing, and (iii) participant perceptions of AI-based assessment and feedback. The overall design emphasizes anonymity and ethical handling of data in accordance with established principles and GDPR-oriented practices.

Σχεδιασμός και ανάπτυξη εκπαιδευτικής πλατφόρμας για την ενίσχυση των ήπιων δεξιοτήτων

Φοιτητής: Κωνσταντίνος Σμαραγδάς

Επιβλέπων: Καθηγητής Περικλής Χατζημίσιος

Περίληψη

Η παρούσα διπλωματική εργασία παρουσιάζει τον σχεδιασμό και την ανάπτυξη μιας εκπαιδευτικής πλατφόρμας που στοχεύει στην ενίσχυση και την αξιολόγηση βασικών ήπιων δεξιοτήτων (soft skills) σε φοιτητές/τριες μηχανικής και σε επαγγελματίες στα πρώτα στάδια της καριέρας τους, με έμφαση στις δεξιότητες Επικοινωνίας, Ομαδικής Εργασίας, Ηγεσίας και Επίλυσης Προβλημάτων. Με αφετηρία την πρακτική δυσκολία αξιολόγησης συμπεριφορικών και κοινωνικών/γνωστικών δεξιοτήτων, η μελέτη εξετάζει πρόσφατη βιβλιογραφία στον χώρο της Τεχνητής Νοημοσύνης στην Εκπαίδευση (AI in Education – AIED) και της αξιολόγησης με υποστήριξη ΤΝ, δίνοντας έμφαση σε υβριδικές προσεγγίσεις αξιολόγησης Ανθρώπου–ΤΝ, σε εφαρμογές ΤΝ που αλληλεπιδρούν με τον χρήστη μέσω διαλόγου, καθώς και σε κρίσιμα ζητήματα όπως η ηθική, η μεροληψία/δικαιοσύνη και η διαφάνεια στην αυτοματοποιημένη βαθμολόγηση.

Με βάση το παραπάνω θεωρητικό πλαίσιο, η εργασία προτείνει και τεκμηριώνει το σύστημα SoftSkills AI, μια αρθρωτή αρχιτεκτονική που αποτελείται από: (α) ένα frontend βασισμένο σε React (SoftSkills User Quiz), το οποίο υποστηρίζει ανώνυμη συμμετοχή μέσω μοναδικών token και διαχειρίζεται τη ροή αξιολόγησης PRE/POST, (β) ένα backend (SoftSkills Bot) υλοποιημένο με FastAPI και βάση δεδομένων PostgreSQL, το οποίο αξιολογεί τις απαντήσεις—ιδίως τις ανοικτού τύπου—μέσω ενός δομημένου pipeline που συνδυάζει μια μηχανή αξιολόγησης βασισμένη σε LLM, ένα ενισχυμένο ευρετικό επίπεδο βαθμολόγησης (GLMP) και fuzzy βαθμονόμηση σκορ, και (γ) ένα Rater UI που επιτρέπει ανεξάρτητη ανθρώπινη αξιολόγηση για επικύρωση και σύγκριση με τα σκορ που παράγει η ΤΝ. Το εκπαιδευτικό κομμάτι περιλαμβάνει δομημένες μαθησιακές ενότητες οργανωμένες σε τρία επίπεδα δεξιοτήτων (Beginner, Intermediate, Advanced), ώστε να υποστηρίζεται διαφοροποιημένη μάθηση σε συνάρτηση με την αρχική επίδοση του/της συμμετέχοντα/ουσας.

Μεθοδολογικά, η εργασία ακολουθεί έναν πειραματικό, διαχρονικό σχεδιασμό pre–post. Οι συμμετέχοντες/ουσες απαντούν σε 16 ερωτήσεις (πολλαπλής επιλογής και ανοικτού τύπου/σεναριακές) και στις δύο φάσεις, κατανεμημένες στις τέσσερις κατηγορίες δεξιοτήτων. Το σύστημα παράγει αριθμητικό σκορ (0–10) και ανατροφοδότηση ανά απάντηση, αποθηκεύει αναλυτικά χαρακτηριστικά βαθμολόγησης και υποστηρίζει την ανάλυση: (i) της συμφωνίας μεταξύ αξιολόγησης ΤΝ και ανθρώπου (π.χ. μέσω δεικτών αξιοπιστίας μεταξύ αξιολογητών όπως ICC), (ii) της μεταβολής επίδοσης pre–post μέσω στατιστικών δοκιμών και (iii) των αντιλήψεων των συμμετεχόντων/ουσών σχετικά με την αξιολόγηση και την ανατροφοδότηση που παρέχει η ΤΝ. Ο συνολικός σχεδιασμός δίνει έμφαση στην ανωνυμία και στην ηθική διαχείριση των δεδομένων, σύμφωνα με καθιερωμένες αρχές και πρακτικές προσανατολισμένες στον GDPR.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον επιβλέποντα καθηγητή μου κ. Περικλή Χατζημίσιο, για την καθοδήγηση, την επιστημονική υποστήριξη και τις πολύτιμες παρατηρήσεις του καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Η συμβολή του υπήρξε καθοριστική τόσο στη διαμόρφωση της ερευνητικής κατεύθυνσης όσο και στην ολοκλήρωση της εργασίας με συνέπεια και μεθοδικότητα.

Επιπλέον, ευχαριστώ θερμά τους διδάσκοντες του Τμήματος για τις γνώσεις και τα εφόδια που μου παρείχαν κατά τη διάρκεια των σπουδών μου, τα οποία συνέβαλαν ουσιαστικά στην εκπόνηση της παρούσας εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω στους συμμετέχοντες που έλαβαν μέρος στη διαδικασία αξιολόγησης, αφιερώνοντας χρόνο και προσφέροντας πολύτιμα δεδομένα, χωρίς τα οποία δεν θα ήταν δυνατή η πειραματική τεκμηρίωση των αποτελεσμάτων.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για τη συνεχή ηθική υποστήριξη, την κατανόηση και την ενθάρρυνση που μου προσέφεραν καθ' όλη τη διάρκεια των σπουδών μου και ειδικότερα κατά την περίοδο εκπόνησης της διπλωματικής εργασίας.

Table of Contents

List of Figures

List of Tables

Acronyms

2SAF — Soft skills assessment framework

IT — Information Technology

PjBL — Project-Based Learning

COM-B — Capability–Opportunity–Motivation

CDIO — Conceive–Design–Implement–Operate

AI — Artificial Intelligence

HR — Human Resources

BIP — Business-focused Inventory of Personality

CPSA — Computing Professional Skills Assessment

AIED — Artificial Intelligence in Education

IJAIED — International Journal of Artificial Intelligence in Education

IAIED — International AI in Education Society

EdTech — Educational Technology

ITS — Intelligent Tutoring Systems

AES — Automated Essay Scoring

AEE — Automated Essay Evaluation

QWK — Quadratic Weighted Kappa

LLM / LLMs — Large Language Model(s)

XAI-ED — Explainable AI in Education

SHAP — SHapley Additive exPlanations

LIME — Local Interpretable Model-agnostic Explanations

HCI — Human-Computer Interaction

ETS — Educational Testing Service

LMS — Learning Management System(s)

CV / CVs — Curriculum Vitae

ICC — Intraclass Correlation Coefficient

AI3 — Artificial Intelligence – Academic Integrity – Assessment Innovation

HASRL — Hybrid Human–AI Shared Regulation in Learning

NLP — Natural Language Processing

ADDIE — Analyze, Design, Develop, Implement, Evaluate

ASPIRE — Authoring Software Platform for Intelligent Resources in Education

L2 — Second Language

EFL — English as a Foreign Language

GLMP — Granular Linguistic Model of a Phenomenon

MAE — Mean Absolute Error

RQ — Research Question(s)

UI — User Interface

PRISMA — Preferred Reporting Items for Systematic Reviews and Meta-Analyses

ABET — Accreditation Board for Engineering and Technology

ACM — Association for Computing Machinery

CC2020 — Computing Curricula 2020

PMI — Project Management Institute

OE — Open-Ended

MC — Multiple-Choice

LSE — Leadership Self-Efficacy

ALAS — Active Listening Attitude Scale

ICCI — Interpersonal Communication Competence Inventory

OECD — Organisation for Economic Co-operation and Development

STEM — Science, Technology, Engineering and Mathematics

IEEE — Institute of Electrical and Electronics Engineers

CBL — Competency-Based Learning

PDF — Portable Document Format

CSV — Comma-Separated Values

API — Application Programming Interface

HTTP — HyperText Transfer Protocol

ORM — Object-Relational Mapping

DOM — Document Object Model

IIFE — Immediately Invoked Function Expression

URL — Uniform Resource Locator

SD — Standard Deviation

GDPR — General Data Protection Regulation

Chapter 1: Introduction

This thesis focuses on the design and development of an educational platform that aims to improve and assess basic soft skills in engineering students of the Department of Information and Electronic Engineering and professional engineers in the early stages of their careers. The motivation behind this work stems from a widely recognised challenge: while soft skills are considered crucial for academic progress and professional rehabilitation, their development and especially their assessment are fragmented and often difficult to scale in real educational conditions. The rapid entry of Artificial Intelligence (AI) technologies, specifically in education, creates a new field of possibilities but also requirements. On the one hand, AI can support more systematic processes of monitoring progress and providing feedback. On the other hand, applying AI to skills that are complex, multidimensional and highly context-dependent requires careful methodological documentation, human supervision and clear definition of the evaluation criteria.

1.1. Motivation and problem context

Soft skills, unlike technical skills, are characterized by the fact that they are not directly captured through standardized tests or closed-ended questions. They are skills that are expressed mainly through behavior, facial expressions or body movements, communication choices and through the way a person analyzes, organizes and supports their thoughts and opinions. In consequence, measuring soft skills requires evidence that derives from more complex forms of activity such as scenarios, open-ended responses, collaborative processes which do not have the same objectivity as a score or a knowledge test.

A key obstacle is that the field of soft skills often shows conceptual diversity, different research or educational frameworks use different definitions, classifications and criteria sometimes even when referring to the same skill. This might lead to a lack of common standards meaning that the same skill category is assessed with different indicators by different evaluators or courses. In this way, inconsistent criteria and often incomparable results arise between courses, curricula or research approaches.

At the level of methodology, the above translates into three main problems:

1. Difficulty of operationalization:

To assess a soft skill, it must be transformed from an abstract concept into observable indications and descriptions of performance levels. For example, communication as a skill can include clarity, structure, coherence, documentation, adaptation to an audience and elements of collaboration such as coordination, information management, resolution of misunderstandings. Lack of clear criteria and quality levels lead to the risk of an unclear assessment or depends too much on the interpretation of the evaluator.

2. Limited documentation of progress over time:

Developing soft skills usually happens gradually and requires repeated efforts for practice and feedback. If the assessment is sporadic or without a stable framework, it is difficult to document measurable improvement and capture changes over time. Thus, an objective picture of the progress that can support targeted intervention is missing from the educational process.

3. Untimely feedback and difficulty in scaling:

The most efficient forms of assessment of soft skills (open-ended responses, scenarios, group work) are time-consuming and require significant human resources. This creates a practical limit to the scale of implementation because the more qualitative and contextual the assessment, the more difficult is to provide feedback on time to a large number of learners.

1.2. Why AI-supported and Human-AI hybrid approaches matter

The application of AI in educational assessment has emerged as an important direction mainly because it offers capabilities that have traditionally been difficult to achieve. For example, scaling of the assessment, uniformity of criteria, faster feedback and the ability to process a large volume of open-ended responses. In particular, when the assessment concerns written speech or argumentation, techniques based on modern language models can provide structured assessments, identify patterns and synthesize feedback in a consistent format.

However, soft skills assessment is not simply analysis of text. It requires connection to educational criteria, understanding of the context and possibly interpretation of the quality of arguments or behavioral choices. For this reason, the present work adopted a hybrid Human-AI approach in which:

- The AI generates the first result such as suggested score and short justification or feedback based on rubric,
- The human rater retains final control and
- The process maintains elements of control, traceability and comparison or calibration.

One extra practical advantage is that the hybrid approach supports the logic of "training - assessment - feedback" as a single natural cycle. It is not limited to the score as the final result but can offer the learner guided improvement. It can show what the learner did right, what is missing and how specific elements of the skill can be strengthened. This is particularly important for soft skills where progress usually results from repeated practice and targeted feedback.

Finally, the use of AI requires responsible design including data protection, avoidance of bias and a clear definition of the role of AI as a supporting tool, not as the final judge. Therefore, the value of AI-supported solutions lies both in automation and in their ability to enable a higher quality and more consistent assessment at scale, through human supervision and methodological documentation.

1.3. Aim and Objectives

Based on the challenges described in the previous sections, the aim of this thesis is to develop a solution that makes soft skills assessment more applicable in practice, without sacrificing its pedagogical value. Specifically, the goal is to create a modular SoftSkills AI platform which combines an assessment environment, educational targeting and a validation process through human scoring. The platform was designed to support a logic of longitudinal assessment (PRE/POST) which allows the monitoring of performance on the spot but also the observation of changes after exposure to educational material or activities.

The scope of this work lies on four skills which emerged as highly important in the literature review: communication, collaboration or teamwork, leadership and problem solving. The system was designed in such a way that the data collected is suitable for both educational use (user feedback) and research analysis (comparisons and correlations).

1.4. Structure of the thesis

This thesis is organized into chapters starting from the identification of the problem and the conceptual foundation of soft skills, continuing with the theoretical background of AI in education and assessment and ending with the description of the proposed platform and the experimental validation process. This structure connects the bibliography with the design and methodology to the final conclusions that are supported by both theory and data.

In Chapter 2, the theoretical framework of soft skills, the definitions and basic classifications as well as their importance in the field of engineering and computer science are presented. In addition, the dominant approaches to education and assessment are examined and the main difficulties related to the standardization, validity and reliability of assessment are highlighted. The chapter concludes with the gaps and limitations that necessitate the search for technologically supported solutions.

Chapter 3 presents the background of AI in Education (AIED) and the use of AI in assessment. The role of large language models (LLMs) in the processing of open-ended responses and their connection to rubric-based assessment and the concepts of explainability, human-in-the-loop and hybrid assessment are also analyzed. Finally, this chapter discusses the critical ethics and parameters (privacy, bias, fairness) which directly affect the applicability of AI in educational environments.

The design of the SoftSkills AI systems and its architecture (frontend, backend, data storage, scoring pipeline and human assessment environment) are described in Chapter 4. The research design, the data collection process and the metric used to draw conclusions are also presented. In addition, this chapter describes the validation framework through comparison of AI and human scoring and the procedures followed to document reliability.

In the final chapter of the thesis (Chapter 5), the results of the experimental validation are presented and discussed based on the collected assessment data. The chapter documents the full downstream analysis workflow, including descriptive statistics and paired PRE–POST comparisons at both the skill level and the overall composite level. In parallel, it reports reliability and agreement analyses that support the defensibility of the scoring process, such as inter-rater reliability between human evaluators and item-level alignment between LLM and human scoring using error metrics and correlation-based measures. The chapter concludes with an overall synthesis of findings, future research directions and improvements and a final consolidation on how the empirical results connect back to the thesis objectives and the proposed platform.

Chapter 2: Soft skills in engineering

2.1 Definitions and conceptual approaches to soft skills

The term soft skills is widely used in literature and is described as an “attractive but vague” term for which there is limited agreement on the meaning, scope, measurement tools and standardized education/training that accompanies it [1]. This polysemy is also documented in international policy and organizational contexts where soft skills sometimes appear as non-job specific skills and other times as intangible personal qualities. In practice the term ends up encompassing everything from personal characteristics and habits to attitudes and socio-emotional skills [2]. It has been pointed out that even the “soft versus hard” distinction is not always consistent, as what is considered “soft” or “hard” can vary depending on the profession and demands of the role, affecting comparability between studies and assessment frameworks [1], [3].

The conceptual ambiguity is also accompanied by criticism of the terminology itself. It has been argued that “hard-soft” distinction functions rhetorically, creating the impression that human-centered skills such as communication, collaboration and inclusion are secondary despite being critical professional competencies [4]. To avoid this devaluation and to better attribute their role in professional practice, alternative terms, such as professional skills or nontechnical skills are proposed [4].

A widely conceptual approach defines soft skills as intrapersonal and interpersonal (socio-emotional) skills meaning non-technical, domain-independent abilities that support behavior and collaboration in professional environments. This perspective emphasizes that there is no universally agreed set or official list of soft skills. For this reason, definitions often emerge from empirical research and are adapted to industries and professions [1], [5]. The distinction between skills that can be developed through training and practice and more stable traits is critical. For example, the 2SAF framework acknowledges that some approaches include traits within the broader concept of soft skills because they are more suitable for educational intervention and systematic development [5]. The need for clarity around this concept is reinforced by theoretical discussions that distinguish skills from related constructs such as dispositions, attitudes, beliefs and values as confusion between categories leads to unclear definitions and weak measurement tools [1].

Some approaches organize soft skills into broader categories in order to make the field more conceptually manageable. For example, a distinction has been proposed that separates personal qualities from interpersonal skills [3], while other frameworks describe “families” of soft skills such as interpersonal skills, thinking skills and personal self-management skills. Recurring core elements such as communication, self-regulation, leadership, emotional intelligence and problem solving frequently emerge across these categories [1].

In educational contexts, soft skills are often described as non-cognitive and non-technical qualities or behaviors that influence interaction, engagement and the learning experience. In a study focusing on teacher soft skills, soft skills are described as a set of “qualities, styles, competencies, habits and attributes” that are shaped through everyday interaction. A student-centered conceptual framework organized into four dimensions (approachability, rapport building, positive learning environment, communication) has been proposed as a response to the lack of tools that focus on non-disciplinary and interpersonal aspects of teaching [6]. This perspective shifts the emphasis away from soft skills as static characteristics towards soft skills as functional behaviors that gain meaning within specific relational, institutional and learning contexts.

Approaches like those linking soft skills to employability adopt a more ecosystemic conceptualization. Authors in [7] designed a conceptual competency framework which describes soft skills as a concept that extends beyond observable behaviors, to include attributes, attitudes, values, socio-emotional skills and forms of non-cognitive intelligence [7]. Organizational and qualitative studies emphasize that soft skills function complementary to hard skills and are linked to transversal competencies that influence collaboration, team effectiveness, organizational performance and outcomes [8].

In fields where more systematic mapping has been done, soft skills are more clearly organized into structures and categories. For example, in design education, soft skills are defined as transversal interpersonal, social and emotional competencies. A multi-axial classification between gateway and higher-order skills has been proposed as a form of supporting didactic design and development of evaluation indicators [9].

In summary, literature converges that soft skills are not a fixed list of characteristics but a category of skills that varies, depending on the context (education, job, industry) and theoretical perspective. For this reason, it is crucial to explicitly state (a) which conceptualization is adopted, (b) which sub-dimensions are included or excluded (e.g. traits vs trainable skills) and (c) how the conceptualization leads into observable indicators and measurement or assessment approaches [1], [2], [4], [5].

2.2 Theoretical frameworks and classification models of soft skills

Theoretical clarification and classification are necessary for the design of organized skill development programs, the selection of appropriate teaching practices and the performance of reliable assessments. This is even more important when soft skills are included in broader categories such as 21st-century skills, where different organizations use similar terms but mean different content. In such cases, there is the risk of the same terms circulating widely but with different meanings, creating ambiguity in educational design and complicating comparison between contexts. In cases where high risk examinations fail to assess such skills, educational practices tend to deprioritize them reinforcing the need for valid and practical ways of measuring them [10].

Given the cross-disciplinary use of the term, taxonomic approaches attempt to organize soft skills into functional categories. In organizational and managerial settings, taxonomies are proposed that separate skills into self-oriented (intrapersonal) and other-oriented (interpersonal), or between personal and social skills reflecting their role in shaping collaboration, team effectiveness and ultimately the performance of an organization. At a more institutional level, international and European initiatives have structured "generic skills" into categories such as instrumental, interpersonal and systemic competencies (as in the Tuning framework) providing a functional basis for mapping skills within curricula [8].

A common step in the theoretical classification of soft skills is their multi-axial organization because they are considered multidimensional constructs. For example, a scheme has been proposed that organizes them as (a) primarily collective or individual and (b) cognitive or meta-cognitive versus interpersonal or social. A hierarchical dimension is introduced by distinguishing between gateway skills and higher-order skills. The value of this approach is that it produces a conceptual map that captures relationships and interdependencies among skills, supporting both the design of learning experiences and the development of assessment indicators [9].

More recent theoretical efforts also attempt to bridge the hard-soft divide through unified models of skills. The Generic Skills Component Approach argues that all skills (whether hard or soft) can be analyzed into five basic components: knowledge, active cognition, conation, affection and sensory-motor abilities. The basic idea is that skills have a common composition and that differences between

hard and soft are not fundamental but relate to how they are observed and measured in specific contexts. This concept facilitates the transition from theoretical dimensions to concrete assessment criteria and indicators [11].

Beyond taxonomic classification, the literature also highlights theoretical frameworks that explain how soft skills are developed and how they transfer from educational or training environments to real practice. In organizational contexts, it has been pointed out that soft skills training does not necessarily lead to permanent behavioral change at work (soft skills transfer problem). To address this issue, the COMPASS framework has been proposed which combines the classic training transfer model (Baldwin & Ford) with the COM-B (Capability-Opportunity-Motivation) behavioral change model. This shifts the focus on the conditions under which newly acquired behaviors are implemented and consolidated in practice.

In educational settings, frameworks that link soft skills development to specific pedagogical methods play an important role. For example, a Project-Based Learning (PjBL) framework has been proposed for the integration of soft skills in technical schools and programs. Skill development is associated with structured dimensions of PjBL such as planning and preparation, implementation, learner commitment, assessment practices with emphasis on the learner-centred and constructivist logic of the method [12]. With this logic, the cultivation of soft skills is linked to authentic tasks, collaborative problem-solving and assessment within the flow of the project.

In higher education and employability contexts, practical classification is often achieved through graduate attribute frameworks which function as operational models for curriculum mapping. These frameworks include competences such as communication, creative problem solving, adaptability, autonomy or initiative, planning or organization and teamwork. Their value is that they make skills visible for curriculum mapping, but it remains crucial that they are linked to clear assessment mechanisms. Otherwise, graduate skills risk remaining aspirational statements rather than documented learning outcomes [8], [10].

In summary, theoretical frameworks and classification models converge on the basic idea that soft skills are multidimensional, interconnected and context dependent. This requires (a) taxonomies that allow for operational organization and mapping of relationships such as multi-axial and hierarchical schemes, (b) development theories that translate “generic” skills into pedagogically actionable mechanisms such as PjBL and (c) models of transfer and behavioral change that explain when and why training leads to actual application in real contexts such as COM-B and COMPASS. Overall, the value of classification frameworks lies in whether they support the design of learning outcomes and development processes so that they function as practical tools and not as descriptive constructs [8], [11], [12].

2.3 Soft Skills in engineering and IT professions

In engineering and computer science professions, literature highlights that professional effectiveness depends on technical knowledge, skills that support collaboration and the application of knowledge in real-world projects. In the modern project environment, work is done in teams with complex chains of collaborations and constant contact with clients or users. Skills such as communication, collaboration, adaptability, leadership and problem-solving act as enablers of technical knowledge in practice [13], [14].

In the field of engineering, evidence from professionals shows that technical competence alone is not sufficient for long-term success. Engineers are required to operate in interdisciplinary and multicultural environments, coordinate with different stakeholders and contribute to solutions with organizational and

social implications [13]. It has been argued that the increase in project complexity and organizational integration is shifting professional requirements towards more people-centered competencies. As a result, employability increasingly depends on interpersonal skills and emotional intelligence alongside technical knowledge. Among these, communication appears to be particularly critical while teamwork and time management are also identified as important for the successful completion of engineering projects [14].

From the perspective of engineering students, empirical data show that there is a market pressure for graduates who are trained in practice with a demand for skills such as communication and stress management. Pedagogical approaches based on collaborative activities and small group work are proposed to enhance social skills including collaboration, open communication and empathy, as well as elements of self-assessment related to professional functioning [15]. This is linked to the observation that soft skills are often expected to develop “on the job” and not in a systematic way within the curriculum contributing to the well-known skills gap during transition from education to professional practice [13], [14].

In the field of software development, the demand for soft skills is also documented through the analyses of job postings. In an analysis of 500 job postings for IT roles (system analyst, designer, programmer, tester), employers explicitly distinguish between hard and soft skills and define soft skills as traits that guide professional behavior and complement technical competencies. Different phases of the software development lifecycle are associated with distinct soft skills profiles. The same study organizes soft skills into nine core skills and shows that some appear in very high demand such as communication, while others, including interpersonal skills and innovation, are underrepresented despite being considered important [16].

In software engineering education, it has been suggested that soft skills can be functionally integrated into technical courses through project-based learning. A study reports an intervention where students worked with real clients on real projects in teams of 4-5 people covering the full project lifecycle. Assessment was done through formal technical reviews and activities targeted skills such as teamwork, leadership, supervision, client communication, conflict management, decision-making, report writing and presentation of results. This approach considered soft skills as core elements of collaboration, accountability and delivery of project results and not as separate learning objectives [17].

Similarly, in engineering and computer science education frameworks linked to CDIO, it has been argued that preparing for team-based professional environments requires systematic focus on soft skill learning outcomes. It is proposed that “professionalism” courses can be organized around domains such as personal effectiveness, personal development, social competence and the professional role of the engineer with an emphasis on reflective writing and structured forms of group learning and discussion [18].

Finally, in an Industry 4.0 environment, the importance of soft skills is further increased as employers continuously value skills that support teamwork, communication and co-creation of innovations with customers and external stakeholders. Furthermore, data-driven operations and cyber-physical systems increase the need for collaboration across complex networks and for skills such as sharing knowledge, cognitive flexibility, project management, negotiation and the ability to translate technical information for non-technical audiences. Consequently, soft skills function as key adaptative capabilities in new organizational and technological ecosystems [19], [20].

In summary, in engineering and information technology professions, soft skills appear to be crucial for the application and transfer of technical knowledge to real-world projects highlighted by labor market analyses and educational interventions that simulate authentic work conditions [16], [17], [18]. At the same time, the tendency to develop soft skills in practice after recruitment reinforces the need for their more systematic integration into the curriculum supported by authentic learning experiences and clearer mechanisms for documenting progress [14], [15].

2.4 Relevance of soft skills in contemporary engineering education

The importance of soft skills in modern engineering education stems from the fact that the professional competence of engineers does not depend only on technical expertise, but also on abilities that allow its effective application in complex socio-technical environments. Recent literature describes soft skills as an element that can differentiate graduates with similar technical qualifications especially in projects that require collaboration, clear communication, ethical judgment and management of relationships with multiple stakeholders [13]. A large-scale systematic literature review identifies six key clusters of skills: problem solving and critical thinking, communication, teamwork, ethical perspective, emotional intelligence and creative thinking. These indicate that "market readiness" is a complex professional functioning capability and not just a technical issue [21].

This emphasis is reinforced by international discussion on quality assurance of engineering programs which have explicitly incorporated non-technical skills as intended learning outcomes. For example, the Washington Accord framework mentions that graduate attributes include elements associated with traditional soft skills such as individual autonomy, effective teamwork, oral and written communication, project management and preparation for lifelong learning [22]. It has also pointed out that despite the general recognition of their importance, the integration of soft skills into engineering curricula often remains slow or fragmented, maintaining a gap between technical training and the real requirements of professional practice [13].

The need for systematic integration of soft skills becomes even more pronounced in the context of Industry 4.0 and the Fourth Industrial Revolution, where rapid technological change increases the demands for collaboration, adaptability and creative problem solving in data-driven environments of high complexity. Relevant interventions argue that professional preparation requires the deliberate cultivation of both cognitive (hard) and socio-emotional (soft) skills through active learning methodologies [23]. Following that, the adoption of a project-based "spine" is proposed as a functional way of integrating soft skills into the curriculum because projects create authentic conditions for collaboration, communication, role negotiation and accountability already from the first years of studies without being disconnected from technical learning objectives [22].

The development of soft skills is also linked to the humanistic and ethical dimension of engineering practice. A common direction is the integration of elements of the humanities such as literature into engineering communication courses to enhance awareness and application of leadership teamwork, critical thinking and ethical judgment in technical decision-making [13]. Furthermore, it has been argued that preparation for professional engineering roles includes skills related to documentation, self-regulation and adaptability further highlighting the need for targeted educational interventions focusing on soft skills alongside technical skills [24].

Overall, the sources converge that modern engineering education needs to treat soft skills as a core component of professional competence. This requires clear learning outcomes and the use of active

learning experiences that enable evidence-based feedback and systematic monitoring of student progress [13], [21], [22], [23].

2.5 Challenges in measuring and assessing soft skills

The measurement and assessment of soft skills have long been one of the most challenging aspects of modern engineering and computing education. A difficulty lies in the fact that these skills cannot be assessed in binary terms of right or wrong like with many technical skills. To reduce ambiguity, assessment requires clear descriptions of performance levels and procedures that ensure common understanding between evaluators, for example through the use of analytical rubrics and score normalization practices. Without these, the comparability of results decreases and discrepancies between evaluators increase [25]. At a broader level, European mappings of soft skills teaching show that heterogeneity in terminology and taxonomic frameworks leads to inconsistency of criteria and tools, making meaningful comparisons between courses, programs and institutions difficult [26].

Another challenge concerns the readiness of teachers and more broadly the "assessment culture" of departments. Soft skills assessment requires evidence of different types than traditional written exams such as observation of interaction, evaluation of group deliverables or reflective assignments. In technical disciplines, teaching staff may not always have sufficient experience or training to assess transferable professional skills [25]. This is also seen in practice with students who have strong technical performance but may have difficulty demonstrating their competence in contexts that involve discussion, role negotiation and teamwork [27].

At the methodological level, many assessment tools are either fragmentary or indirect. A common choice is questionnaires such as pre- and post-measurements before and after Project-Based Learning, which facilitate data collection and allow for basic comparisons. However, these instruments mainly capture perceptions and self-assessment rather than actual performance in authentic situations. Similarly, tools that target "skill gaps" based on HR requirements can be practically useful but the conclusions depend strongly on how the tool translates a skill into specific questions [28].

The same logic is seen in large-scale assessment tools. For example, instruments such as the Business-focused Inventory of Personality (BIP) have been used to profile student groups and make comparisons between groups. While their value is in their standardization and comparability, the interpretation of the results remains closely tied to the theoretical model of the tool and to the transparency of the mapping between subscales and soft skills they are assumed to represent [29].

Another practical problem is that even when soft skills are stated as learning outcomes, traditional tests including quizzes and exams, often do not adequately capture them. As a result, it is suggested to integrate assessment into long-term group projects with explicit soft skills requirements and assessment into the flow of the project. This approach produces more authentic evidence of skill development, but increases the demands of planning, monitoring and grading, requiring realistic time allocations and more structured assessment protocols to be sustainable [30].

Finally, a critical issue is the documentation of the quality of the assessment. In rubric-based procedures, there is a need for interrater reliability and documentation that the criteria actually correspond to the skill we want to measure addressing issues of construct and content validity). This is highlighted in frameworks such as the Computing Professional Skills Assessment (CPSA) where the emphasis is on rater training, shared interpretation of criteria and systematic assessment protocols [25]. In online or

hybrid learning environments where direct observation of collaboration is more difficult, assessment practices often shift towards more learning-oriented approaches emphasizing authenticity, meaningful feedback, self-assessment and reflection [27].

Overall, the main difficulties concern (i) conceptual ambiguity and heterogeneity of frameworks, (ii) the need for clear and commonly applicable assessment criteria, (iii) the limitations of indirect measurement tools such as questionnaires and (iv) the practical feasibility of more authentic but demanding forms of assessment.

2.6 Summary and transition to AI-based approaches

Sections 1.1-1.5 have shown that soft skills are a multidimensional and context-dependent category of skills which is defined differently depending on the scope and domain including education, employability and professional practice. At the same time, analysis has highlighted that the term “soft skills” itself is not neutral: the hard-soft distinction can devalue skills that are critical for professional competence. This makes it necessary for each study to clearly state which conceptualization is adopted and what the term exactly includes [2], [4].

In the fields of engineering and IT, it has been documented that soft skills function as a necessary complementary pillar to technical knowledge, because they allow its effective application in complex projects, interdisciplinary team environments and conditions of rapid organizational and technological change. However, it has also been shown that their systematic and comparable assessment of these skills remains both methodologically and organizationally demanding, especially when assessment is based on indirect evidence or on practices that are difficult to scale in a consistent manner [25].

Based on the above, the question arises of how more continuous, evidence-based and scalable forms of soft skills assessment can be supported without compromising pedagogical quality. The next chapter addresses this question by focusing on AI-based approaches where the use of interaction data, linguistic and conversational signals and where applicable, multimodal cues can contribute to more systematic documentation of performance and learning progress. At the same time, this transition introduces increased requirements for accountability, placing issues of validity, transparency, bias, privacy and human supervision at the center of any responsible application of AI in soft skills assessment.

Chapter 3: Artificial Intelligence in education and learning assessment

3.1 Artificial Intelligence in Education (AIED)

Artificial Intelligence in Education (AIED) is a multidisciplinary field of research & development, which investigates the potential of intelligent computational systems to support, improve or even transform learning, teaching and assessment processes in educational environments. In recent years, the growing interest in AIED is closely linked to the general evolution of AI towards systems that utilize large amounts of data and learn from it. Techniques such as machine learning, deep learning and reinforcement learning play a key role in this shift. As a result, both the design of educational technologies and educational decision-making practices are increasingly shaped by data-driven approaches [31].

At the application level, AI has already been incorporated into a wide range of educational technologies and services including digital or online platforms, pedagogical agents, chatbots and in some cases robotic solutions. These applications support both administrative tasks (e.g., grading assistance or assessment organization) and instructional practices such as personalized content delivery and adaptive support for learners. The key differentiation compared to older educational technologies is that many of these systems can adjust their behavior based on interaction data, namely information derived from how the user works and responds within the learning environment [32].

As an area of research, AIED has been established as a distinct and coherent academic field since at least the 1980s. Significant steps along this journey include the founding of the International Journal of Artificial Intelligence in Education (IJAIED) in 1989 and the founding of the International AI in Education Society (IAIED), in 1993. The contemporary phase of AIED has emerged through a dynamic interplay between academic research, the commercial educational technology (EdTech) sector, the growing influence of large technology companies and the increasing adoption of data-driven policies and practices in educational governance and management [31].

The development of the field is often characterized by the coexistence of two complementary trajectories. On one hand, an “evolutionary” approach focuses on the gradual improvement of technologies and practices that can be smoothly integrated into existing educational structures and routines. On the other hand, a more “radical” direction where attempts are made to fundamentally change how learning and support for students are organized within technologically rich environments and learning communities [33].

At the core of AIED lies the long-standing goal of achieving personalization at scale, inspired by the well-established effectiveness of one-to-one tutoring. However, recent research points out that today’s educational challenges, such as the development of 21st century skills, engagement with authentic and complex problems, collaborative learning, increased learner agency and diverse learning contexts, demand a more precise description of what AIED can realistically offer. This need for greater “specificity” is essential for clarifying the possibilities, conditions of application and limitations so that the adoption of AIED goes beyond mere administrative efficiency and leads to meaningful improvements in learning outcomes [34].

One of the most well-established application areas within AIED is that of Intelligent Tutoring Systems (ITS) which provide adaptive guidance and immediate feedback. Many ITS are described using a "double-loop" architecture, in which an outer loop organizes and selects activities or problems and an inner loop monitors the learner's problem-solving process step by step [35]. The inner loop evaluates user actions and delivers appropriate hints or feedback as needed. These systems often rely on models of skills or knowledge components, along with techniques such as knowledge tracing and model tracing. Moreover, the widespread use of ITS generates large volumes of integration data which can be leveraged to refine underlying models, increase system adaptability and optimize feedback [35].

In higher education, AIED does not refer to a single, unified technology but to a diverse set of applications serving different purposes. Systematic mapping studies categorize the use of AI in higher education along four main dimensions: (i) profiling and prediction, (ii) assessment and evaluation, (iii) adaptive systems and personalization and (iv) intelligent tutoring systems. This classification shows that the field ranges from personalized learning support to institution-level functions related to organization and management. As a result, the adoption of AIED in higher education is closely linked to a set of pedagogical, organizational and ethical considerations that must be carefully addressed [36].

The combination of learning and assessment within digital environments is a key element of the theoretical foundation of AIED. In such settings, assessment can be embedded directly within the learning process rather than occurring as a periodic, external event, such as a traditional test. One prominent approach in this context is stealth assessment, which is a methodology that continuously collects process data in real time, often within games or simulations [37]. By capturing observable indicators ("clues") linked to underlying, latent abilities (unobservables), stealth assessment supports ongoing formative evaluation and enables the adaptive adjustment of the learning environment.

Finally, the integration of AI in education cannot be evaluated only in terms of technical effectiveness. Research emphasizes that AIED systems should function as a supportive infrastructure for human-centered pedagogical practices with their effectiveness largely determined by how well they are integrated into authentic educational contexts. At the same time, social and ethical challenges such as privacy, surveillance, learner profiling, inequality and bias are highlighted especially when AI is used for automated assessment or other highly sensitive applications [31], [38]. From a critical standpoint, these considerations emphasize the need for transparency, accountability and clear pedagogical purpose as the socio-technical implications of AIED constitute a fundamental and structural dimension of the field.

3.2 AI in educational assessment

The use of Artificial Intelligence in educational assessment mainly refers to the extraction of reliable indicators of learning from data generated during educational activities. In this sense, AI extends beyond the notion of "mechanical grading" and can instead support the continuous monitoring of learning progress, the early detection of learning difficulties and the linking of assessment to targeted pedagogical interventions within digital learning environments [39].

Assessment applications, within the tradition of AIED research, are clearly closely intertwined with ITS and more generally to methods that model the performance and knowledge development of learners longitudinally. One good example of such is the knowledge tracing approaches, continuously updating an estimate of a learner's "knowledge state" about interaction data such as correct or incorrect answers and solution strategies. This estimate is subsequently used to select subsequent learning activities or to

adapt the type and level of support provided. In this framework, the role of assessment does not act as a discrete or independent phase but as integral to the learning cycle itself [35].

One of the important components of the use of artificial intelligence in assessment is open-ended response including report and essay assessment that has become well known as Automated Essay Scoring (AES) or Automated Essay Evaluation (AEE) [40]. Traditional systems such as e-rater are described as models that identify linguistic and textual features and learn to generate scores that approximate those assigned by human raters [41]. However, achieving a high level of agreement between human and algorithmic scoring is not enough on its own. Equally important is transparent documentation of what the system actually measures and the extent to which it avoids reliance on superficial or construct-irrelevant features that do not reflect the true quality of the written response [40].

Based on AES literature reports, model performance is often assessed using agreement metrics, such as Quadratic Weighted Kappa (QWK) which accounts not only for agreement but also for the magnitude of score discrepancies. The ability of a model to generalize beyond a specific topic (cross-topic scoring) is of particular interest as this is critical for real-world applications. Similar approaches emphasize practices that protect the integrity of the assessment, including score normalization procedures and checks for potential information leakage from topic-specific keywords that could artificially inflate performance. Limitations regarding transfer across different text genres, possible biases and computational cost are frequently acknowledged. Consequently, most works propose the use of hybrid strategies where AI is meant to aid, together with human judgments, rather than replace [42].

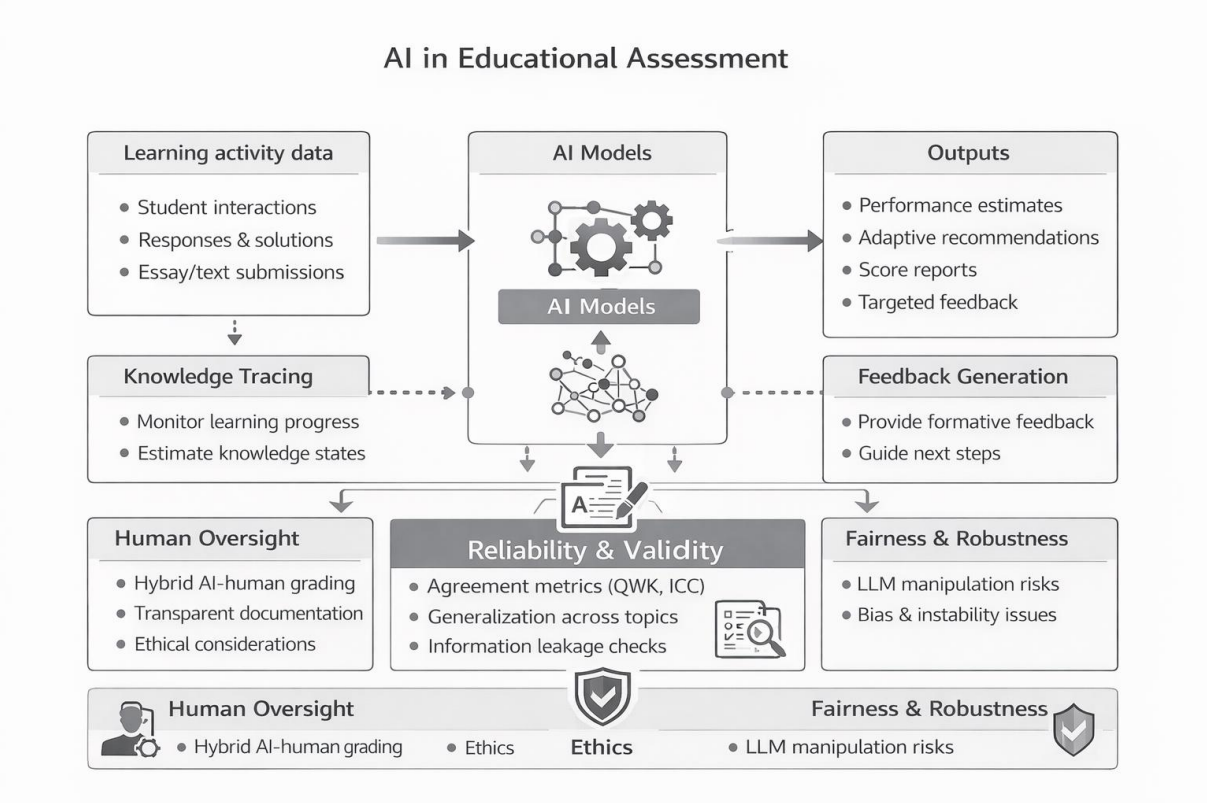

Figure 3.1 – AI in educational assessment: an overview of data sources, model components, outputs, and quality/ethics safeguards.

Because educational assessment requires reliability and reproducibility, discussions of AI-based scoring are increasingly grounded in established psychometric methods. The Intraclass Correlation Coefficient (ICC) can be considered a significant factor where consistency or agreement between raters or methods of assessment such as human and algorithmic scoring, are being compared [43]. For ICC results to be meaningful, it is imperative to detail which type of coefficient might have been used with a specified definition based on agreement or consistency, in addition to confidence intervals. However, the generation of a score is only part of the overall validity argument. The key question is whether the resulting scores can be interpreted appropriately and used responsibly within a given educational context [40].

In addition to the function of grading, feedback generation is a crucial one, particularly for formative assessment settings. Research on effective feedback shows that it is most impactful when it is clear, aligned with learning goals and directs learners towards next steps, rather than offering vague or judgmental comments. These principles provide important design criteria for AI-based assessment systems that aim to deliver timely and personalized support [44].

The proliferation of Large Language Models (LLMs) is further transforming the landscape of automated assessment, as these systems handle more complex language tasks. However, they also exhibit forms of instability and errors that become problematic when used in evaluative roles [45]. Studies on model robustness indicate that in open-input environments such as free-text or code submissions, small changes or deliberate “manipulations” of the input can affect the output. These phenomena cause significant concerns with respect to the fairness of the assessments in the context of automated scoring [39]. Consequently, the use of AI in educational assessment cannot be treated as a neutral technical choice but it requires careful documentation, ongoing human oversight and clearly defined conditions of use, especially when assessment outcomes carry important consequences [46].

3.3 Large Language Models in educational assessment

LLMs introduce a new class of tools for educational assessment because they can process natural language at scale. This capability enables different types of applications including support for grading, generating feedback, summarizing arguments, organizing responses and evaluating open-ended questions based on explicit criteria such as rubrics. The literature describes LLMs as technologies with strong transformative potential and emphasizes that their educational value depends on how they are pedagogically integrated, the presence of human supervision and the development of appropriate AI literacies. These factors are necessary for limiting errors, biases and misuse [47].

At the level of scalable grading, LLMs differ from earlier automated assessment approaches because they can engage with conceptual and explanatory responses and not just superficial textual features. This enables semi-automated assessment workflows in which the model proposes a grade and/or qualitative feedback that are subsequently reviewed by human raters (see also 3.6 for the hybrid framework). Research shows that the performance of LLMs in grading depends to a large extent on the way the prompts are formulated and on the use of clear, detailed rubrics. On the other hand, hybrid utilization with human confirmation emerges as a more realistic and safe practical direction [48], [49].

One of the main questions to be answered is whether LLMs can function as consistent and valid evaluators. To answer this question, studies focus on the stability of model outputs under variations in prompt wording, when the same assessment is repeated or when the evaluation is performed at different points in time. Moreover, reliability and agreement measures such as the ICC are used to assess whether a model produces stable ratings under specific conditions and prompts. Comparative studies in AES

further show that performance varies across different models. In some cases, high intra-rater consistency may be observed while agreement measures may vary over time which is crucial for operational use where longitudinal stability is required [50].

In written assessment contexts such as essays, reports, open-ended responses, the application of LLMs has practical limitations. Reliability decreases when assessment criteria are unclear, when decisions are high-stakes (e.g. certification) or when evaluation involves complex constructs such as the quality and consistency of argumentation [51]. While the use of explicit rubrics and analytical justification can improve the structure of the process and facilitate control, these measures alone do not guarantee that assessments are valid or fair across different learner populations.

Beyond traditional essay scoring, recent work also explores the evaluation of long, multidimensional answers known as Automated Long Answer Grading (ALAG). More analytical and criterion-based approaches are proposed including "rubric entailment" where assessment focuses on whether a response explicitly addresses the elements specified in a rubric. Nevertheless, long answers remain a more demanding task even for powerful models because they often require coherence, sufficient evidence and deep understanding of the content [49].

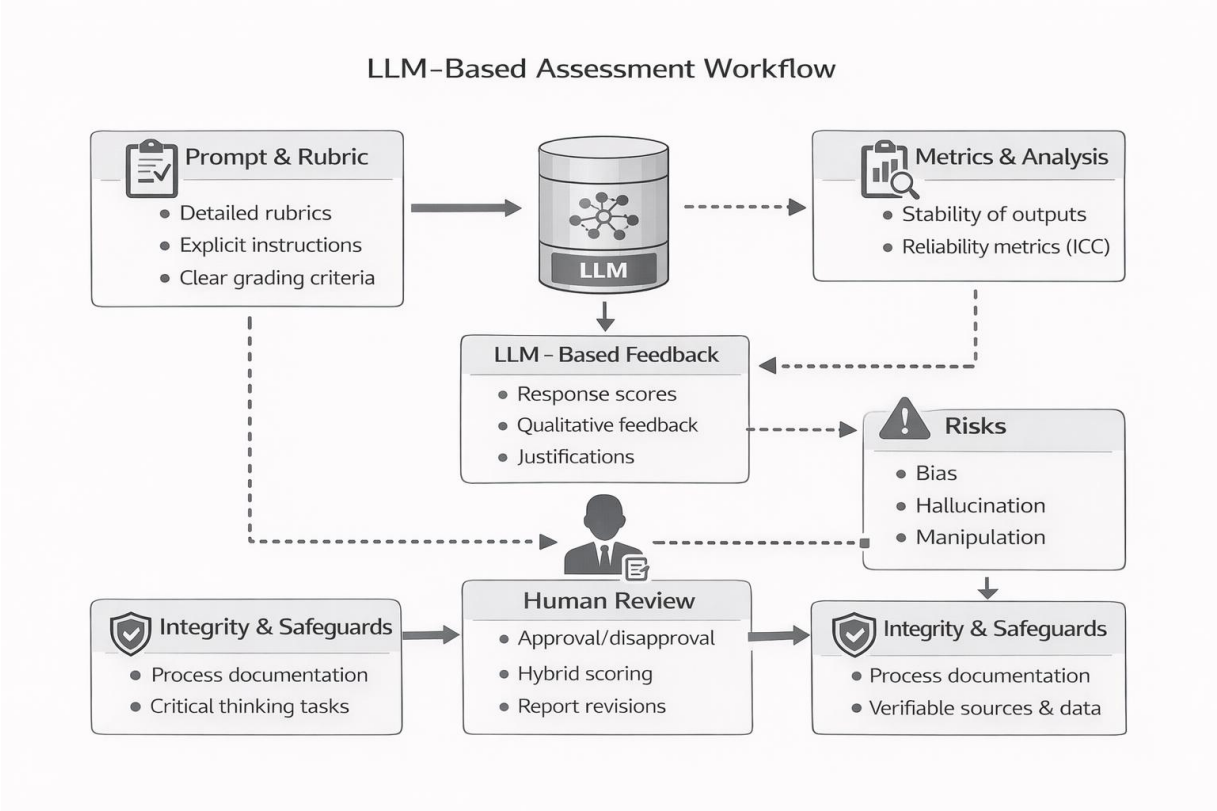

Figure 3.2 – LLM-based assessment workflow illustrating rubric-guided prompting, model-generated scoring/feedback, human review, evaluation metrics (stability/ICC), and integrity safeguards.

Another interesting aspect is the use of LLMs for assessment analytics through text representations. By transforming texts into numerical representations, LLM can support predictive models for educational and psychosocial outcomes, often compared to established indicators such as teacher evaluations. This direction suggests that LLMs function both as tools for scoring and infrastructures for feature extraction given that transparency, fairness and responsible use are ensured [52].

In fields where evaluation is strictly based on the correctness of results and reasoning such as physics or mathematics, important limitations are observed. LLMs can produce explanations that seem convincing but are incorrect and often struggle with numerical calculations, symbolic manipulation and reproducibility. Research work on introductory course material have recorded cases in which a model could marginally "pass" an evaluation process but with errors that have direct consequences if used as an evaluator or a solution provider [53].

One of the most pressing issues regarding LLM-based assessment is academic integrity. Since it becomes more difficult to distinguish between human- and machine-generated text, there is a need for developing detection tools despite their known limitations or redesign of assessment methods. In practice, this involves assessments that are not based solely on the final text but ask for evidence of the learning process such as drafts, intermediate steps, reflections, oral justifications or the use of verifiable sources and data. Assessments that require critical thinking, contextual reasoning and personal engagement are considered more robust under these conditions [47], [54].

Overall, recent studies agree that the use of LLMs in assessment ranges from supportive functions such as pre-scoring or formative feedback to more automated workflows. In all cases, responsibility for final decisions should remain with human evaluators, especially when they are of high risk. Before large-scale adoption, issues related to data quality, clarity of criteria, stability of results and fairness across learner groups require systematic validation and governance mechanisms. Finally, the need to document evaluative decisions, especially when they come from "black-box" models, directly motivates the discussion of explainability, human understanding and accountability which is further developed in 3.4 [48].

3.4 Explainable AI in Education (XAI-ED)

AI systems are increasingly used for evaluative and predictive decisions in education so there is a growing need for their outputs to be understandable and verifiable by humans. In this context, Explainable AI in Education (XAI-ED) concerns methods, interfaces and design practices that make the decisions and reasoning of educational AI systems transparent and interpretable. XAI-ED is highly important because the widespread use of complex models may function as "black boxes" while their outputs can have direct consequences for learners, for example through scores, risk predictions or adaptive interventions. When the rationale behind such decisions is unclear, trust, user control and responsible use become more difficult [55], [56].

From a design perspective, the goal of XAI-ED is to produce explanations that have pedagogical meaning and are useful within real educational contexts. This means that explanations should be aligned with learning objectives, be comprehensible for the intended audience and be adapted to different user roles, including learners, teachers, administrators and system designers [55]. The importance of human supervision and shared responsibility is also highlighted, meaning that there should be a clear distinction between what the system recommends and what the human approves or acts upon. There are also practical challenges in ensuring compliance with regulatory frameworks (such as data protection and AI governance regulations without explanations becoming either overly technical or so abstract that they fail to support informed decision-making [57].

In terms of technical approaches, post-hoc explanation methods are widely used in XAI-ED because they can be applied to powerful but non-transparent models. Techniques such as SHAP and LIME generate explanations after a model has been trained and can provide either global insights into which factors influence predictions or local explanations relevant to a specific case [58]. However, the field is

heterogeneous in terms of definitions, objectives and evaluation criteria for explanations. Open challenges remain not only in technical implementation, but also in interface design and Human-Computer Interaction (HCI) [59].

A critical requirement for explainability is that an explanation must be both understandable and reliable. Relevant literature discusses concepts such as fidelity which refers to the extent to which an explanation reflects the actual decision-making process of the model and stability, which concerns whether explanations remain consistent under small changes in data or conditions. It has been highlighted that post-hoc explanation methods may show variability in the feature importance and may not always faithfully capture the structure learned by the model. These limitations become problematic in high-stakes scenarios where incorrect or misleading explanations can undermine accountability. Alternative approaches such as neuro-symbolic methods and rule extraction techniques are discussed as potential ways to combine strong predictive performance with more structure and interpretable explanations under certain conditions [60].

Finally, the application of XAI in automated assessment has pedagogical value as it can transform a score from a final outcome into usable information for learning. A well-designed explanation can show the elements that contributed to an assessment result and pinpoint areas for improvement, enhancing both the legitimacy of the evaluative judgement and its formative usefulness for learners. In this sense, explainability functions as a prerequisite for maintaining meaningful human supervision and accountability in AI-supported educational practices. The need for explainability becomes even more pronounced when AI-based assessment targets complex and indirect constructs such as soft skills where indicators derive from patterns of behavior and interaction rather than clearly defined right or wrong outcomes. In such cases, the ethical and legal implications are particularly sensitive, and these challenges are further discussed in 3.8 [55].

3.5 Conceptualization and measurement of soft skills in AI-based systems

The measurement of soft skills through AI-based systems presupposes a conceptualization of what constitutes a “skill” and which observable indicators can serve as valid evidence of its presence. In contrast to “hard” skills which are often captured by direct and well-defined performance indicators, soft skills are associated with non-technical aspects of human functioning such as social interaction, communication, collaboration and self-regulation. In the literature, these skills are found under different names including people skills, non-technical skills, social skills and emotional intelligence. In higher education and professional environments, they are often described as non-cognitive skills or attitudes that shape how individuals learn, collaborate and interact. Commonly cited dimensions include emotional awareness, positivity, effective interaction, people management, conflict management, strategic thinking and the ability to learn quickly. Emotional intelligence is presented as a foundational construct, since dimensions like self-awareness and self-regulation are linked to higher-level skills such as leadership and problem-solving [61].

To reduce the ambiguity associated with broad or overlapping terms and to enable the operationalization of soft skills into measurable dimensions, skill taxonomies play a crucial role. For example, the ETS Skills Taxonomy 2025 defines a skill as a developed characteristic that supports human performance along a continuum of competence within a specific domain. It also highlights that many existing taxonomies remain insufficiently precise or poorly connected to assessment practices, which limits their practical applicability. The framework proposes eight broad skill categories (Interpersonal, Intrapersonal, Essential Cognitive, Advanced Cognitive, Digital, Societal, Learning, Remote Work) and

accompanies them with definitions and assessment considerations. This is particularly useful for AI-based systems because it acts as a bridge between abstract concepts such as teamwork, empathy or conflict management, and observable data sources including text, dialogue, video or behavioral traces captured during targeted activities [62].

One major challenge is that in real educational and professional settings, skills are often described in unstructured formats, such as professional standards, job descriptions or high-level learning objectives. The absence of this structure complicates standardization and coherent assessment, specifically when organizations are attempting to align educational programs with labor market needs. At this point, AI-driven approaches have been proposed that use LLMs for the transformation of unstructured skill descriptions into structured competency frameworks and learning outcomes. Such approaches can support integration with learning management systems, such as Canvas or Moodle and facilitate clearer alignment between requirements, instructional activities and assessment practices. As far as soft skills are concerned, this allows for a transition from vague, generic labels to more precise and assessable competency statements that can be linked to scenarios, rubrics or evidence-producing tasks [63].

The measurement of soft skills in AI-based systems is often indirect relying on performance data, interaction traces or multimodal signals rather than right or wrong responses. In the educational field, intelligent learning technologies are characterized by adaptability, automation and analytics, supporting features such as personalization, progress monitoring and real-time adaptation [64]. This means that these systems continuously produce “traces” of interaction which can be used as indicators of skills related to collaboration, persistence, problem-solving strategies or communication. However, these skills are inferred rather than directly observed, the same issues of reliability and validity that arise in AI-supported text assessment and grading (see Section 3.3) become equally critical [49].

In professional and organizational settings, AI is widely used in personnel selection and evaluation processes, including the analysis of CVs, psychometric tests, chat- or video-based interviews and performance in structured tasks or simulations. However, even when there is evidence that AI can capture certain aspects of interpersonal and behavioral skills, the social acceptance of these assessments plays a decisive role. Research shows that many individuals perceive AI as less capable than humans in assessing interpersonal skills and that these perceptions influence whether managers trust AI-generated results or how candidates choose to present their qualifications in AI-mediated selection processes. Therefore, measuring soft skills is not only a technical challenge but also a socio-technical one. The usefulness of an indicator depends both on its methodological robustness and on whether it is understood and accepted in the context in which it is applied [65].

Finally, rigorous psychometric documentation of reliability and validity is necessary for AI-based measurements of soft skills to be considered trustworthy. The ICC is commonly used in cases of interrater, test-retest and intrarater reliability because it captures not only correlation but also agreement between measurements. To ensure a scientifically robust application, it is necessary to specify the model used, the definition applied (agreement vs consistency) and the corresponding 95% confidence intervals [66]. The importance of such practices becomes evident when AI systems are used as evaluators or as support mechanisms within assessment flows (see Section 3.3) when both initial validation and ongoing monitoring of stability under real conditions are required [49].

3.6 Hybrid human–AI models in educational assessment

Hybrid human-AI models in educational assessment describe approaches in which AI does not function as an autonomous grader but as part of a collaborative system that combines algorithmic capabilities

with human pedagogical judgment. In such models, AI contributes strengths such as pattern recognition, large-scale data processing and generation of preliminary scores, insights or feedback while humans remain responsible for interpretation, contextual understanding and accountability. The principle of augmentation is central where technology aims to improve the ability of teachers and learners to understand learning processes and make informed decisions without removing human judgment at critical points in the assessment process [67]. Literature connecting AI in education with learning analytics highlights that assessment cannot be limited to mere "tool use" but needs to be seen as a dynamic relationship between data, algorithmic processing and pedagogical goals so that automation supports the understanding of learning and not simply the mechanical production of results [68].

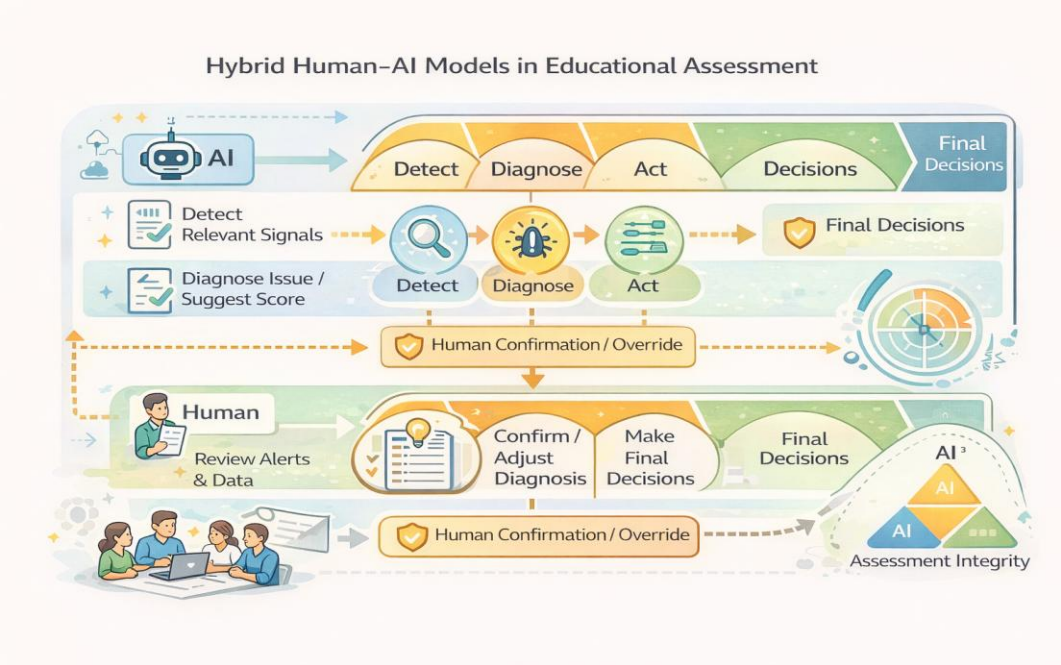

Figure 3.3 – Hybrid Assessment Architecture: AI Signal Detection and Scoring Support with Human Final Decisions

For a hybrid model to function effectively in practice, a clear separation of the role of the AI system and the human actor is needed. Authors in [67] propose the detect-diagnose-act scheme along with the concept of different levels of automation so that it is clear how responsibilities are distributed. In this scheme, AI may detect relevant signals or summarize data, support the diagnosis and interpretation of a learning situation or contribute to actions that follow from the assessment. These actions may remain entirely human-led or be partially automated, for example through adaptive system responses. This is particularly important in assessment contexts where grading and evaluating often involve high risk decisions. In such cases, full automation increases the possibility of errors, bias or unfair outcomes. For this reason, literature favors conditional automation in which automated processes are activated under clearly defined conditions and include explicit points for human confirmation, correction or override, preventing humans from being placed "out of the loop" [67].

The value of the hybrid model becomes even more evident nowadays where generative AI directly influences how assessments are designed and implemented. The AI3 model (AI–academic integrity–assessment innovation) frames assessment decision-making as a balance between technological

possibilities, the protection of academic integrity and meaningful pedagogical innovation. It shows that assessment cannot be treated as a one-dimensional problem. If the emphasis is only on technology without integrity or without significant pedagogical renewal, practical and ethical issues arise. If technology is ignored, assessment practices risk becoming unrealistic or ineffective in contemporary educational environments [69]. Within this framework, hybridity functions as a pragmatic bridge with AI reducing the workload, surface relevant patterns or provide initial drafts and indicators while the final interpretation, justification and accountability remain at the human level especially when decisions have significant consequences for the learner.

Finally, in hybrid models, human-AI collaboration should not be assumed as inherently beneficial but as a relationship that needs to be designed, monitored and evaluated. Beyond the question of "who decides" (human-in-the-loop), the literature addresses "how" humans and AI collaborate. Approaches have been proposed to describe or measure the degree of synergy between human and algorithmic contributions, as well as frameworks that model co-regulation processes in learning and assessment, such as the activation and evolution of regulatory "triggers" in human-AI supported regulation of learning (HASRL) [70], [71]. In this sense, hybrid human-AI models provide a practical foundation for more controllable and transparent assessment processes where algorithmic support is combined with human judgement and responsibility [72].

3.7 Conversational agents and educational chatbots

Conversational agents or pedagogical conversational agents and more specifically educational chatbots are digital systems that interact with the user through natural language, written or spoken. Their purpose is to support automated dialogue that simulates a human interlocutor and they are commonly integrated into online learning platforms, digital assistants or messaging applications [73]. In the educational context, chatbots are mainly associated with providing immediate support and feedback, offering a practical solution to situations where personalized teacher guidance is limited such as the well-documented "teach-bandwidth" problem. In recent years, research has focused on chatbots as a very common type of dialogue systems. Systematic reviews describe chatbots as tools that can support personalized learning and various educational functions, providing immediate responses and guidance through conversation, leveraging the familiarity of learners with everyday digital communication tools [74].

In higher education, educational chatbots are used for various purposes and often adopt different pedagogical roles or "identities". A review of 66 studies reports that chatbots have been implemented in roles such as assistant, tutor, student and virtual patient. Overall findings suggest mainly positive outcomes, although there is little evidence on whether these benefits are long-term. Reported benefits concern fast access to information and support without the need for time-consuming searches. Significant challenges are identified, including technical limitations related to NLP and knowledge base coverage as well as methodological such as small or homogeneous samples and difficulties in distinguishing between short-term novelty effects and sustained engagement [75]. Similar conclusions are drawn in another systematic review of 53 studies which highlights the need for further research on technical improvements, ethical principles and usability testing to support effective and sustainable adoption of chatbots in education [74].

Evaluation of chatbots is considered a critical issue since evaluation metrics are often based on subjective data with student questionnaires using Likert-type scales as the dominant tool. More objective indicators such as learning outcomes or knowledge gains are used less frequently. There is no

internationally accepted framework for assessing the quality of chatbots or user satisfaction resulting in different methods being applied that make direct comparisons between studies difficult [75]. Moreover, inconsistent terminology and classification of chatbot types further complicate the synthesis and interpretation of findings. According to these, practical guidelines for chatbot integration are proposed including user-centered design approaches, the use of instructional models such as ADDIE or ASPIRE, improvements in NLP capabilities and data security when embedding chatbots into institutional support infrastructures [75].

The use of pedagogical chatbots as conversational agents in second-language (L2) learning contexts is of particular interest, as conversation is not just a support mechanism but a central component of the learning process. A study involving English as a Foreign Language (EFL) learners found that a text-based chatbot was generally considered understandable and useful for practicing language skills and improving written expression while also helping to reduce writing anxiety. The study also recorded limitations in the interactive and teaching capacity of the chatbot including difficulties in sustaining long or complex conversations, reduced effectiveness when learners produced erroneous input, occasional irrelevant responses and the presence of a potential novelty effect where initial enthusiasm declines over time without appropriate pedagogical design [76]. Therefore, the use of more guided activities and a variety of tasks are suggested in order to sustain learner engagement. It is also highlighted that learners often pay more attention to linguistic accuracy in order to be understood by the chatbot which can be leveraged pedagogically [76].

Finally, reviews agree that the evaluation of educational chatbots should not be limited to measuring user satisfaction but should examine whether the system serves effectively its implementation goals and whether it can support more advanced functions such as mentoring and adaptive interaction. Key directions for future research on chatbots in education include aligning chatbot evaluation with their implementation goals, exploring possibilities for student mentoring and exploiting/exploring adaptive capabilities [73]. These priorities are in agreement with other findings which highlight the fact that lack of shared quality criteria and dominance of self-reported measures make it difficult to compare systems in an evidence-based manner and draw robust conclusions about what works, for whom and under which conditions [75].

3.8 Ethical considerations, bias and fairness in AI-based soft skills assessment

While the assessment of soft skills using AI systems is particularly based on multimodal data, such as speech and text, intonation, or non-verbal cues, this use shifts the focus from what learners know to how they behave and interact. This will have direct ethical implications, given that it entails the collection and analysis of highly sensitive personal data-from linguistic patterns to facial expression and body language. As a result, the application of such systems requires robust frameworks for transparency and informed consent, so that individuals know what data is collected, for what purpose, for how long and with what potential consequences [77]. In educational contexts, these risks become associated with those of surveillance practices and with limiting the autonomy of learners, especially if systems extract behavioral metadata without clarity and a shared understanding of their use [38].

At the same time, AI systems cannot be considered neutral because both training data and design choices affect their outputs. This may pose a risk that historical inequalities and social biases are sedimented and reproduced by algorithmic decisions. In soft skill assessment, this is an even more critical issue since interpretation of human behavior depends on the context-dependent and differs across linguistic, cultural or social settings [38]. These variations increase the likelihood of biased inferences and unfair

discrimination against specific groups. To reduce this risk, literature emphasizes the importance of representative datasets, systematic bias audits and continuous monitoring of the model after its implementation to detect and correct problematic model behavior [77].

Within this framework, fairness cannot remain an abstract principle but must be defined using concrete and measurable terms. A useful distinction is made between individual-, group- and multi-level fairness. In educational assessment, unfairness can occur either when individuals with a similar profile are treated differently or when entire groups, (e.g. defined by language, gender or socio-economic background) consistently receive less favorable outcomes. Group fairness aims to prevent systematic disadvantage and can be evaluated using specific indicators such as statistical parity or equality of opportunity. However, relying on group-level metrics alone is not sufficient as disparities affecting subgroups or individuals may be obscured by aggregate measures, necessitating complementary analyses across multiple levels [78].

Implementing fair assessment practices in real-world settings presents additional challenges which directly affect soft skills. Issues such as partially observed or censored outcomes as well as trade-offs between fairness and predictive utility often require context-sensitive decisions that depend on the intended use and the level of risk associated with assessment outcomes. Furthermore, detecting and addressing biases requires appropriate assessment resources including data, tools and protocols. Fairness depends both on the algorithmic design and the conditions of access and use [78]. For example, the digital divide can exacerbate educational inequalities when access to AI infrastructure and tools is uneven. For this, policy-oriented instruments have been proposed for assessment contexts such as structured AI assessment scales that clarify permissible uses, documentation requirements and expectations for critical engagement. These approaches highlight that fairness also concerns the maintenance of equal conditions of participation, particularly when advanced AI systems are available only through paid or restricted access [79].

Finally, ethical responsibility in AI-based soft skills assessment extends beyond technical bias mitigation to encompass governance, accountability and principles of design and use [38]. For soft skills assessments, this means clear definition of the assessment purposes, limiting data collection to what is strictly necessary, documented procedures for human oversight (see Section 3.6) and providing mechanisms for contesting or reviewing decisions [80]. Moreover, since soft skills indicators are derived from indirect and inferential evidence, their interpretation needs to be accompanied by explanations that are appropriate to the educational context (see Section 3.4) [77]. Through such measures, assessment outcomes will function as transparent, contextualized judgments with clearly articulated limits and conditions of use.

3.9 Emerging trends and future directions in AI-supported soft skills assessment

Recent developments show that AI-supported soft skills assessment is gradually moving away from unidimensional approaches such as isolated text analysis or final outcome scoring, towards richer and more integrated systems that combine multiple data sources and place greater emphasis on explainability. This shift aims to capture human behavior and interaction more realistically, reducing the ambiguity that characterizes soft skills. A representative example is the development of a multimodal and explainable frameworks that combine video, audio and text analysis with fuzzy logic and linguistic perception models (GLMP) allowing soft skills to be decomposed into higher-level subcomponents while explicitly addressing the uncertainty associated with behavioral and emotional signals. By combining multiple modalities, such systems can improve the robustness of soft skill assessments and

make the process more transparent as they provide more interpretable justifications for evaluation outcomes [81]. In terms of future direction, progress in this area is increasingly defined not only by whether a model predicts accurately but also by how well systems document their estimates, communicate uncertainty and allow human users to understand and control assessment decisions.

Another strong trend concerns the scaling and standardization of AI-supported assessment, particularly in contexts such as large-scale examinations and transnational educational programs. An OECD study on AI-based scoring of constructed-response items shows how deep learning models pre-trained on large multilingual corpora can leverage historical assessment data to produce scores that closely align with human ratings, both in terms of score distributions and core psychometric properties. This work emphasizes the need to take into account the uncertainty in automated scoring ensuring that the assessment outputs are not treated as definitive decisions without appropriate confidence information. Presenting both a score and an associated level of certainty allows cases with high uncertainty to be flagged for human review, supporting hybrid assessment workflows and reducing the risk of inappropriate automation [82].

In addition to technical maturity, the future of AI-supported soft skills assessment will depend to a large extent on social acceptance and perceived legitimacy. Soft skills are often regarded as deeply human attributes and research on lay beliefs shows that even when there is evidence that AI assessment can predict interpersonal skills in specific contexts, individuals tend to believe that AI is less capable than humans at assessing interpersonal skills. These beliefs influence practical outcomes such as the willingness of managers to assign tasks that require interpersonal skills to people selected by AI and the strategy of candidates who may under-emphasize these skills when the process is AI-assessed. Therefore, a critical future direction is to improve algorithms and design more transparent communication processes about what AI-based assessments measure, how results should be interpreted and what the limitations are. Evidence suggests that clear information about advanced AI assessment technologies can reduce negative perceptions and increase acceptance in specific contexts [65].

Emerging trends also include a shift towards more experiential and dynamic assessment environments often supported by gamification and AI-enhanced e-learning platforms. In these settings, soft skills are assessed not only based on the final outputs but also through participation patterns, decision-making processes and behavioral indicators within realistic learning scenarios. A cross-study synthesis reports that AI-driven personalization can enhance engagement and academic performance while gamification is associated with the development of soft skills and with more interactive forms of assessment compared to traditional methods. It is also highlighted that data privacy and technological accessibility are persistent challenges and long-term adoption will depend on whether systems effectively address these issues and on whether research examines their broader implications for learner adaptability in rapidly changing educational and work environments [83]. This underscores the need for assessment systems that balance dynamic behavioral data collection with interpretability, governance mechanisms and clearly defined limits of use.

Finally, technological innovation in AI-supported soft skills assessment must be considered in relation to broader structural and infrastructural conditions. The [84] report describes that AI innovation remains largely concentrated in high-income countries while trends in adopting “small AI” are emerging in low- and middle-income countries. Furthermore, the report identifies key AI foundations including connectivity, access to computational resources, data and language availability and skills development as critical policy and infrastructure dimensions. For AI-supported soft skills assessment, this translates into a very practical condition in which the more multimodal and computationally intensive systems

become, the more they will be affected by inequalities in access to technology, data and expertise. Therefore, future progress should be evaluated not only in terms of algorithmic sophistication, but also with respect to sustainability, scalability and applicability across diverse educational systems and socio-economic contexts [84].

Chapter 4: Architecture and methodology of soft skill assessment system

4.1 Introduction to methodological approach

This chapter presents the methodology applied for the design, development and assessment of the proposed system for the determination and improvement of the level of students in certain soft skills. In this section, the aim is to describe step-by-step the procedure followed in this thesis to design, implement and configure the proposed system ensuring replicability and reproducibility.

The methodological approach combined the creation of a user-friendly tool and an integrated, reproducible framework for assessing soft skills by leveraging artificial intelligence. The goal was to develop an innovative system which offers objective assessment and personalized learning/feedback through a robust experimental procedure for investigating its validity and reliability.

In this context, a digital learning and assessment environment was created which allows participants to answer questions related to their competencies, receive feedback and recommendations for further practice, as well as personalized material for their improvement. Moreover, a system for automatic (LLM-based) and human assessment was also integrated into the digital environment in order to compare the results and monitor the progress of the participants.

The following chapter includes a detailed description of the research/study design, data collection, system architecture, tools and indicators used for the analysis and interpretation of the results

4.2 Research design

The current research design followed a mixed-method approach which combined quantitative and qualitative data for achieving a better understanding of the functionality and utility of the proposed tool. The ultimate goal was to perform a deep investigation of the validity, reliability and educational effectiveness of the soft skills assessment system.

4.2.1 Study type

The present study has an experimental and longitudinal character following the logic of pre- and post-test which is increasingly observed in recent research as a largely reliable technique. Every participant was assessed twice, once for the first time using the platform (pre-test) and once after (post-test). This structure allowed us to determine any changes in the performance and, thus, assess the potential effect of the use of this tool as an educational intervention.

The quantitative part of the study was based on numerical scoring assigned by using both AI-based ruling and human raters. These values were later used to statistically compare the two types of assessment in terms of agreement and consistency. To do so, indicators such as the ICC and Mean Absolute Error (MAE) were calculated in the data analysis stage.

The qualitative part of the study focused on the analysis of open-ended questions provided to the participants during the test and feedback automatically given by the system based on their answers. This analysis aimed at examining the structure, clarity and content of the verbal responses, as well as the way the AI model interprets and comments on the performance of the participants. In this way, we examined

if and how accurate automatic assessment was in attributing the exact meaning of the participants' answers.

4.2.2 Objectives and research questions

The main objectives of this study are the following:

1. To assess the robustness of automatic grading as performed by the AI model. Examine the agreement between the AI- and human-generated rating.
2. To estimate the improvement of the soft skills of the participants through the PRE-POST scheme and determine if the proposed system leads to actual progress of the users.
3. To investigate how the participants perceive the use of the platform and the feedback given by the system, as well as their degree of satisfaction regarding the assessment of their skills using artificial intelligence.

Based on these objectives, the following research questions (RQ) were formed:

RQ1: To what extent the AI-generated grades agree with the scores assigned by the human raters?

RQ2: Is statistically significant improvement of the participants' performance observed between the pre- and post-test?

RQ3: How do the participants perceive the assessment and the provided feedback?

To guide the statistical analysis and the interpretation of the results, the following research hypotheses (H) were formed:

H1: The grades of the automatic assessment show strong correlation with the human-generated grades ($ICC > 0.75$).

H2: The participants show statistically significant improvement between pre- and post-test ($p < 0.05$).

H3: The participants receive positively the assessment performed by AI and they consider personalized feedback useful.

4.2.3 Participants

The sample consists of university students and early career professionals of the Department of Information and Electronic Engineering who volunteered to participate in the study. The studied population was chosen because it includes individuals who are familiar with technology and potentially need to develop their soft skills in order to meet the demands of the modern job market.

Each participant received a unique anonymous token through Google Forms which was used for login and matching the data of the pre- and post-tests while keeping a secret identity. The use of tokens secures complete anonymity and allows monitoring the progress of each participant without violating privacy.

During the experimental process, each student answered 16 questions per phase (8 multiple choice and 8 open-ended) for each of the four studied soft skills: Communication, Teamwork, Leadership and Problem Solving. Answers were submitted through the digital environment 'SoftSkills User Quiz' and were graded both by AI and later by instructors via Rater UI.

All participants were informed for the aim of the study, their rights and terms for data usage. Their participation was completely voluntary without any academic or professional benefit or consequence. Data were securely collected and stored, used exclusively for research purposes and analyzed anonymously.

4.2.4 Summary

Overall, the research design of this study combines experimental data, mixed methodology and longitudinal monitoring to examine in a robust manner the reliability and the educational value of the soft skill assessment system. The combination of AI- and human-generated rating, as well as the quantitative and qualitative analysis, offer a complete overview of the system performance and its effect on the participants.

4.3 Experimental procedure

The experimental procedure was designed to examine in practice the functionality, accuracy and effectiveness of the proposed system in soft skill assessment. It consists of two distinct phases – the initial assessment (PRE) and the final assessment (POST) – with an intermediate feedback and practice period (coaching phase). During this period, participants had the opportunity to study the results of the first phase, reflect on the received feedback, study the learning material and implement the proposed practices in their daily activity for improving their soft skills.

4.3.1 Soft skill selection and question generation

A systematic literature review was performed and the following soft skills were selected: Communication, Teamwork, Leadership and Problem-Solving. According to PRISMA guidelines these four soft skills were the most frequently cited and were considered the most critical among engineers and IT professionals [85]. Meta-analysis studies in the field of employability and project team performance show that these soft skills rank high with high frequency in research projects that examine professional effectiveness in technical environments. In particular, Communication is recognized as a fundamental skill for the clear transmission of technical information and collaboration between interdisciplinary teams, a key criterion in the Accreditation Board for Engineering and Technology (ABET) Criteria.

Previous studies show that teamwork is significantly related to the performance and effectiveness in a team and directly linked to the success of software and engineering projects as highlighted in the Association for Computing Machinery (ACM) Curricula (CC2020). In the case of leadership, it is a critical skill for decision-making and the management of complex and dynamic work environments according to international and professional standards such as those of the Project Management Institute (PMI). Finally, Problem-Solving has been a central skill in all engineering certification organizations as it is strongly connected to the ability to address complex technical challenges.

For each skill area, a set of open-ended (OE) and scenario-based multiple-choice (MC) questions were generated and stored in a question bank (database) This bank consisted of 160 questions in total, i.e. 20 questions per skill category and per type (OE and MC). The questions were adapted from open-access, peer-reviewed studies that had developed and validated psychometry tools for assessing soft skills in educational and/or professional settings. Studies were chosen only if they had published full item sets and provided sufficient methodology to ensure accurate adaptation. Specifically, questions measuring leadership were based on the Leadership Self-Efficacy (LSE) scale [86] and the Leadership Learning

Agility scale [87] and teamwork on the Team-Q Survey [88] and the Teamwork Competency scale [89]. Problem-solving items were based on the Engineering-Modified Problem-Solving inventory [90] adapted from Heppner & Petersen (1982) [91]. Questions evaluating communication were generated based on Active Listening Attitude Scale (ALAS) [92] and the Interpersonal Communication Competence Inventory (ICCI) [93]. Questions were either retained as in the original studies when appropriate or reformulated to match the educational context of engineering students and professionals. The MC questions were reformulated in a way that aimed at assessing the knowledge of the students and the OE questions were scenario-based to assess their attitude and reasoning in their daily and professional life, as suggested by OECD (2023) [94] for complex skills evaluation. In any case, theoretical links with the original psychometric tools were preserved.

4.3.2 Development of educational material

4.3.2.1 Source material

As mentioned above, personalized educational material was developed for each soft skill and given to the users during the interim period to study and improve their skills. To create a strong theoretical basis for the personalized material, a systematic literature search was performed using a combination of key words and phrases such as “soft skills in engineering education”, “transferrable and transversal skills in STEM”, “developing soft skills in engineering”. Academic databases including Google Scholar, IEEE Xplore and ScienceDirect were used to find peer-reviewed articles and conference proceedings relevant to the development of soft skills in engineering and related fields. In addition, the material of the course “Technical Writing” taught in the 2nd semester in the Department of Information and Electronic Engineering was used in creating the theoretical basis of the learning material.

The literature review led to the development of one document for each skill with a similar structure. Each skill was introduced with the definition, key terminology and a short explanation of its components. This was followed by a section on the importance and relevance using evidence from literature on why the skill is necessary for engineering students and professionals. Descriptions of the different aspects or types of each soft skill, for example types of communication or leadership styles were also included in the educational material to highlight the complexity of each skill. In certain cases, factors that influence negatively the effective demonstration of the skill were described to raise awareness of potential challenges students might face when applying the skill in practice. Techniques for applying a specific soft skill and examples from engineering education contexts (e.g. group projects, hackathon) were also included to show how each soft skill is applied in real life. This stage of the methodology resulted in the development of structured documentation based on academic literature that served as reference for the differentiated content created in the next phases.

4.3.2.2 Soft skill learning modules

Following the creation of the theoretical material for each soft skill, the next step was to transform this content using a format that could support personalized learning. To achieve this, the source material was further adapted into three competence levels: beginner, intermediate and advanced. The level assigned to each student was based on their initial performance in the PRE test. The aim was to create targeted materials each with their own learning objectives and content complexity. These documents are provided in PDF format responding to the student’s level and allowing them to train and improve their soft skills during the 10-day study period between pre- and post-assessments.

Table 1. The level for each soft skill was defined based on core behaviors, observable actions, level of autonomy expected and level of responsibility within a team. The table shows certain characteristics students display at each level.

Beginner	**Intermediate**	**Advanced**
Limited experience	Basic reliability	Flexibility
Low confidence	Partial autonomy	Autonomy
Fragmented understanding	Ability to apply skills in familiar context	Ability to guide ambiguity
High need for structure	Readiness for guided practice	Strategic use of the soft skill

To ensure that each level met the needs of the students with different competence levels, several established educational frameworks were applied and are described below:

- Competency-based learning (CBL): Defines clear competencies-what the learner should be able to do or demonstrate. It emphasizes that students' progress at their own pace once they can show they have mastered a specific skill. It is a learning approach that focuses on what the students can actually do with what they have learned and not how long they have been studying [95]. In this thesis, CBL was used to report observable behaviors that show someone has mastered the skill at each level. For example, for the soft skill "Communication" a beginner might just be able to define what communication is while an intermediate student might be able to actively listen in a group meeting and an advanced student might be able to lead a difficult conversation (Table 1).
- Bloom's revised taxonomy: It classifies thinking according to six cognitive levels of complexity. The categories are hierarchically organized as follows: Remember, Understand, Apply, Analyze, Evaluate and Create [96], [97]. Here, Bloom's revised taxonomy was used to align the learning activities in each module with the appropriate level. For example, beginners were required to remember and understand theoretical definitions, intermediate students were asked to apply and analyze these concepts and advanced tasks involved evaluating and creating solutions. In this way, there was a gradual increase in challenging students across levels while matching their stage of development.
- Instructional scaffolding: Is the support given to a student to promote learning which is gradually removed as students develop autonomous learning skills [98]. In this thesis, the learning modules were designed using scaffolding offering high support for beginners with step-by-step instructions, clear examples and defined roles. Support was moderate for intermediates encouraging greater independence with collaborative tasks and minimal in advanced modules suggesting mentoring and team coordination activities (high level of autonomy).
- Differentiated Instruction: It uses different means to adjust the learning experience acknowledging various student backgrounds, readiness levels and learning needs [99]. In this thesis, the learning modules had different complexity, instructional tone, task design and depth of reflection even though the theoretical content was common across all levels.

4.3.2.3 Generation of training documents

For each soft skill (Leadership, Communication, Teamwork, Problem-Solving), three educational documents were created including a unified theoretical section and a distinct learning module corresponding to the proficiency level (Beginner, Intermediate and Advanced) of the student determined by the PRE test.

The learning material for all soft skills and competence levels had a similar structure to ensure that all students receive a consistent learning experience. At the beginning of each learning module the aim and a short explanation of the skill were included to help students understand their proficiency level. This was followed by key behaviors to develop and basic activities to perform in order to improve each skill. In addition, recommendations on which project roles would fit best based on the proficiency level and weekly reflection points were also included to help students monitor their progress. At the end of each document, indicators were reported for the students to assess progression to the next competence level.

4.3.3 Preparation and anonymous participation

For secure and anonymous participation, the SoftSkills Tokens Utility subsystem was used to create two versions (PRE and POST) of unique participation links for every user. Each link contained an encrypted token which allowed the matching of the data of the same participant between the two phases without the use of personal information.

The tokens were distributed using a Google form through which users filled in their email address, selected one of the available link-tokens and registered it. The link was automatically removed from the list after selection to ensure that each token would correspond exclusively to one participant.

After submitting the form, the system automatically forwarded the unique token to the user's email address using a Google Form extension. In this way, absolute anonymity was ensured, double participation was avoided and easy monitoring of the progress in PRE and POST phases was performed without storing personal information in the database.

4.3.3.1 Participant recruitment and guidelines

Participants were recruited via email sent to all members of the department which included information about the purpose of the study, the participation process and the use terms of their data. The message included a brief description of the research objectives, the duration of the tests and the link to the registration form (Google Form). It was emphasized that participation was completely voluntary, anonymous and without any academic consequence (positive or negative). Furthermore, it was clarified that the data would be used exclusively for research purposes and in an anonymous form. Communication through the official channel of the department contributed to enhancing the transparency and trust of students in the process, as well as ensuring the ethical validity of this study.

4.3.4 PRE phase – Initial assessment

During the first phase (PRE), participants followed the personal PRE link and entered the 'SoftSkills User Quiz' platform. Before starting the test, users could choose which soft skill category they would like to start with (Communication, Teamwork, Leadership or Problem Solving) via a selection field on the left of the home screen (Figure 4.1).

Regardless of which category was chosen first, each user completed a total of 16 questions evenly distributed across all four skills. The test flow was adaptive:

1. Selection of starting category. The participant selects the category to start with and clicks "Start".
2. First "block" of 4 questions in the selected category. The system randomly loads 4 questions belonging to the selected category (2 MC and 2 OE). Answers to MC questions are immediately stored while OE questions are sent to the backend for processing.
3. Level assessment. After scoring the first four questions, the system calculates the average of the scores for each category and classifies the participant in one of three levels: low, mid and high. The level is displayed in the user interface as an indication (LEVEL) and is used to give the user a first impression of how the quiz scoring is structured (Figure 4.1).
4. Continue with the remaining categories. After the 'competence' level for the first category is assessed, the system automatically loads the rest of the "blocks" of questions for the other three soft skill categories. The order of the categories and the appearance of the questions are determined by the quiz engine. For the user, the experience remains a single test with a clear indication of the current category and progress at the web interface (e.g. "Category: Leadership").
5. Automatic response evaluation. Each answer to an open-ended question is sent to the backend, where the SoftSkills Bot processes it based on the combination of LLM + GLMP + fuzzy logic. For each question, the following is produced:
    I. Numerical score on a scale of 0-10,
    II. Verbal summary/comment
    III. Internal features which are later used for the

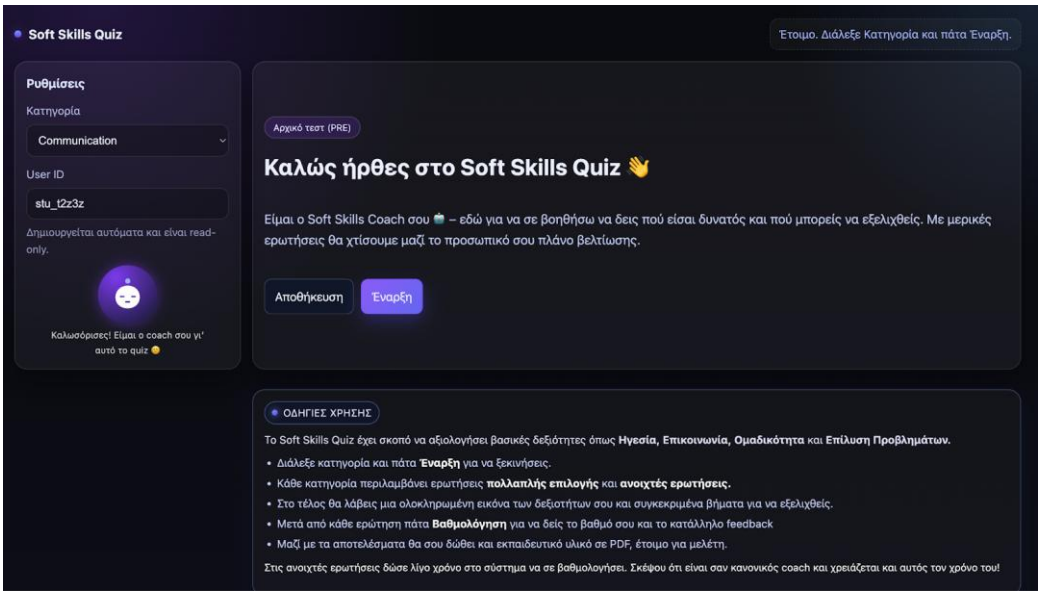

Figure 4.1 – Soft Skills Bot user interface

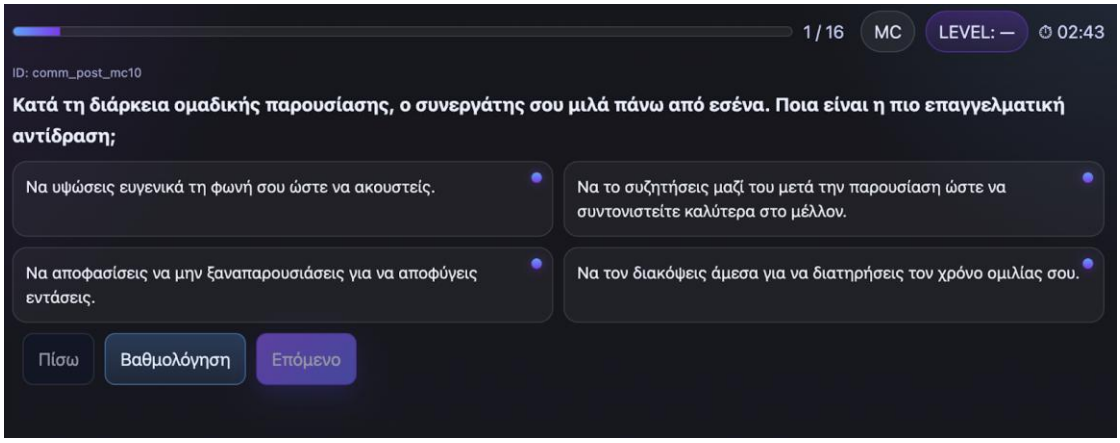

Figure 4.2 – Multiple-Choice Question Screen within the Adaptive Quiz Flow

All results are stored in the PostgreSQL database and linked to the anonymous token of the participant.

- After completing the 16 questions of the PRE test, the participant is taken to a Summary Page which displays the level of proficiency and an overview of the user's performance:
- Competence level (LEVEL): The overall level (low/mid/high) as derived from the question scores.
- Average total score: The average of all quiz scores.
- Averages per category: The average of all questions for each of the four soft skills in a list format so that the participant can see in which areas they performed best and in which they might lag.
- Weakest category: The system highlights in which category performance was the weakest and uses it as a central targeting point in the improvement plan.
- 2-week personalized coaching plan: A personalized plan which includes a brief description of the goal, steps and actions to apply in their daily life and practical suggestions to improve their skills.
- Supporting material (PDFs & course pack): Files in PDF format which include learning material linked to the participant's level and soft skill category. Recommended pages from the educational material of the Technical Writing course (2nd semester) are given (with reference to course pack) so that the student knows exactly where to focus their study.
- Ability to export results: Ability to download the results in CSV format to store locally or use them for later analysis. Ability to download the complete plan for any future reference.

The PRE phase serves as a baseline for quantitative comparison with the POST phase, as well as for the personal self-awareness of the participant providing an initial overview of strengths and weaknesses along with personalized suggestions for improvement. After completing the PRE phase, the OE responses were extracted from the database and subjected to an additional assessment by human raters through the Rater UI subsystem. This phase was carried out after the end of the PRE test and aimed at verifying the accuracy of the automatic scoring system.

4.3.5 Interim period (‘study period’)

After the completion of the initial phase (PRE), an interim period of several days followed during which the participants had the opportunity to reflect on the test results and study the personalized development plan provided by the system.

This period had a dual purpose, on the one hand to allow students to internalize the findings of the assessment and on the other hand to function as a phase for implementing improvement practices before the second assessment (POST).

The coaching plan provided in the PRE phase served as a personalized practice guide and included:

- Behavioral suggestions that enhance collaboration.
- Active listening practices.
- Effective leadership strategies.
- Tips to manage conflicts and empathy.

Participants were encouraged to implement the suggested actions in their daily academic and personal lives without any external supervision or guidance. This ensured that the POST phase would capture real and spontaneous change in behavior and self-perception of skills and not the result of temporary influence or guidance.

4.3.6 POST phase – Re-evaluation

The second phase of the process (POST) was also carried out via a personal token link distributed after the interim period. The link was different from that of the PRE phase, but the distribution happened the same way and corresponded to the same anonymous participant identifier.

Upon entering the POST link, the system recognized both the token and the phase and loaded a new set of questions from the question bank. In the present implementation, there was no technical possibility of strictly separating the question into “PRE only” and “POST only” maintaining one pool of randomly selected questions. This means that in certain cases participants may have encountered same questions in the two phases which is a methodological limitation but still allows for the comparison of scores at the skill level.

The flow of POST phase was similar to that of the PRE phase:

1. The participant receives and enters the POST link with their personal token.
2. The system loads 16 questions (8 MC and 8 OE), 4 per soft skill.
3. The MC answers are entered directly into the backend while the OE answers are evaluated by the SoftSkills Bot automatically via LLM + GLMP + fuzzy logic.
4. The results (scores per question and category) are stored in the PostgreSQL database linked to the same token but with a “post” phase tag.

After completing the POST test, the participant is also taken to a summary page which presents:

- Competence level (LEVEL)
- Average total score
- Average score per soft skill category
- Weakest category

The POST phase summary page does not include supporting material because it focuses mainly on assessing progress after the practice period without new intervention. Similarly, downloading the results in CSV format and the complete plan or summary is possible.

PRE-POST comparison and statistical analysis

Comparison of the results between PRE and POST phase is not performed automatically by the system. The platform undertakes the collection and storage of all necessary data (scores per question, category and phase) with consistent use of the token. Downstream analysis of the effectiveness of the system was performed in the framework of the thesis to determine:

- PRE-POST differences per skill
- Improvement indicators at the participant level
- Reliability indicators (e.g. ICC, MAE) for the comparison of automatic and human assessment

In this way, the POST phase provides the second measurement point needed to estimate the extent the use of the platform and the interim period are associated with a substantial improvement in soft skills. Similarly, the OE responses of the POST phase were extracted and subjected to human evaluation via the Rater UI subsystem to assess the consistency and accuracy of the automatic scoring of the system.

4.3.7 Human rating (Rater UI)

To verify and gain a deeper understanding of the automated rating process, all OE answers from both PRE and POST phases were subjected to human rating through the Rater UI subsystem.

The aim to introduce human rating was to serve as a cross-check and reliability check mechanism and to contribute to the calculation of the final combined scores which were later used in the PRE-POST calculation.

The raters (teacher01 and teacher02) had access to anonymous responses without revealing the identity of the student. They could see the question, the answer, the soft skill category and the AI-generated score (LLM score). Qualitative comments or verbal observations by the teachers were not possible. The human evaluation was exclusively quantitative to remain comparable to the numerical scores of the system.

Rater UI Operation process

The process included the following steps:

1. Answer loading for evaluation

Rater UI pulls answers from the backend – either from the PRE or POST quiz. Answers are automatically selected based on skill category or question type (open or multiple choice) or assessment status (if not already rated by the rater). Each entry is displayed as a card including:

I. User_id (anonymous identifier)
II. Question_id
III. Category and type
IV. Question text
V. Participant's response

2. View AI automatic score

The AI-generated score (LLM score) is displayed on the card on a scale of 0-10. This is provided only as an indicative reference point. The rater is free to agree or disagree as their score is recorded independently.

3. Enter human rating (slider)

Each rater (e.g. teacher01) rates the answer using a 0-10 slider. There is no comment field as the human intervention is numerical to be comparable to the AI rating. After selecting a value, the rater presses Queue/Submit and the rating is stored as a record in the human_ratings table with the fields: answer_id, rater_id and score.

4. Calculate final score (final_score)

After the human evaluation by both raters is complete, the backend automatically calculates the final score of each answer. It is a two-step calculation:

I. Average human evaluation (human_avg) is based on the principle of multivariate evaluation and combination of sources of assessment (ensemble scoring):

$$human\_avg = \frac{score_{teacher01} + score_{teacher02}}{N}$$

Figure 4.3 – Calculation of the human average score (human_avg)

where N is the number of available scores.

These final scores are the ones used in the analysis of the study results as they incorporate both human- and AI-generated ratings.

II. Combination with AI score:

$$final\_score = 0.6 \cdot human\_avg + 0.4 \cdot llm\_score$$

Figure 4.4 – Calculation of the final score (final_score)

The final score (final_score) combines the average human evaluation (human_avg) and the AI-based score (llm_score), weighting them at 60% and 40%, respectively. This hybrid design reflects established automated scoring practice, where model outputs complement human judgment and are tuned to maximize human–machine agreement rather than replace human raters [100], [101]. The 60/40 split is adopted as a conservative blending choice that prioritizes human evaluation while still capturing the model's contribution, consistent with prior hybrid consensus formulations using a 60–40 human/ML balance [102]. All values are stored in the final_score database attribute.

Table 2. Scoring Variables Recorded for the Hybrid Human–AI Evaluation Process

llm_score
teacher01
teacher02
human_avg
final_score
phase (pre/post)

5. Summary

The Rater UI evaluation served as the final data validation phase. Double scoring (AI and humans) in all phases of the experiment ensured that the PRE-POST comparison was based on verified and weighted results enhancing the reliability and validity of the research conclusions of this study.

4.3.8 Ethical issues and anonymity

In this thesis, research was implemented according to the principles of ethics and personal data protection (General Data Protection Regulation-GDPR). At no stage of the process personal data were collected or stored. Recruitment and identification of participants were performed through distribution of anonymous tokens which were created using the SoftSkills Tokens Utility and stored in the PostgreSQL database with secure encryption. Participation was completely voluntary without academic or professional benefits. All records (tokens, answers, scores, ratings) were stored anonymously and used exclusively for research purposes.

4.3.9 Summary

The experimental process of the SoftSkills AI system combined technological, pedagogical and research elements. The PRE-POST structure allowed the monitoring of the progress while the human evaluation through Rater UI provided a mechanism for verifying the AI-generated scoring. The integration of ethical practices and anonymity mechanisms enhanced the validity of the study ensuring the results are reliable, reproducible and research-based.

4.4 System Architecture

4.4.1 Architecture overview

The SoftSkills AI system was developed with a modular and scalable three-level architecture, designed to enable the operation of its connected subsystems. This architecture ensures clear separation of roles, code reuse and easy creation of future extension, as well as secure data management and anonymous collection.

The structure is organized in three operating levels (tiers):

1. Frontend layer:

It includes two independent interfaces:

I. The SoftSkills User Quiz which allows students to participate in the PRE and POST phases,
II. The Rater UI which is assigned to the human raters (instructors) who have been selected to rate the participants' answers

2. Application Layer/Backend:

Represented by the SoftSkills Bot subsystem which functions as an API server. It processes all client requests, implements automatic evaluation via LLM + GLMP + Fuzzy Logic, stores the results in the database and coordinates communication with the Rater UI. The implementation is based on the Python FastAPI framework.

3. Data Layer:

Here, all data of the answers, evaluations, scores and tokens are stored in the PostgreSQL database. Data management is done with the SQLModel which integrates the advantages of Pydantic (validation) and SQLAlchemy (ORM). This layer also includes the Tokens Utility, a standalone Python script that generates random anonymous PRE and POST links for each participant.

Data flow in the system

The operation of the system follows a circular data flow where each subsystem collaborates with the others through RESTful APIs.

This multi-level scheme ensures data isolation per phase and user (via token) in combination with transparency in the evaluation flow so that each stage (AI and Human) is distinct. Finally, each module operates autonomously and can be re-executed at any time.

Based on this structure, the following sections (4.4.2 – 4.4.5) present the subsystems in detail, accompanied by representative code snippets that illustrate their internal operation.

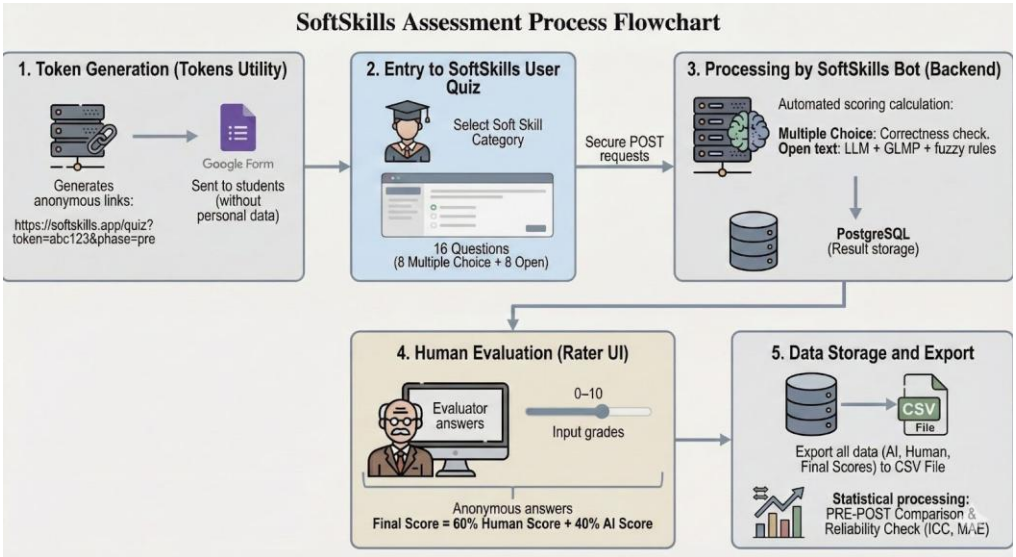

Figure 4.5 – Overview of the SoftSkills Assessment Process (Author's Own)

4.4.2 Frontend: Soft Skills User Quiz

The SoftSkills User Quiz application was developed with React and Vite, utilizing JavaScript and CSS to create a lightweight, fast and scalable user interface. Its architecture is based on the use of React Hooks for state management and React Router DOM for navigation between individual quiz screens. Communication with the backend is done via the native Fetch API, while the export of results is performed with a custom CSV Exporter function. The final application is compiled in static form (static build) via Vite which allows for fast loading and each deployment on any web server.

Table 3. Frontend System Components and Their Functional Roles

Component	Technology	Role
Build tool	Vite	Optimized React application development
Frontend Framework	React(javascript library)	Create dynamic and resuable UI
HTTP client	Fetch API	Send requests to backend
State management	React Hooks	Manage question state and scores
Data export	CSV Export Utility	Download results to a local file

4.4.2.1 Participant initialization: token, attempt and PRE/POST phase

At the core of the quiz logic is the correct identification of the participant and the phase they are in (PRE or POST). This is implemented in src/main.js through a series of helper functions and IIFE (Immediately Invoked Function Expressions) which read the token and attempt parameters from the URL, create an anonymous identifier for each user based on their token and define the phases of the quiz that the user is in (PRE/POST).

```
// --- Study token & attempt helpers ---
function getQueryParam(name) {
  const m = new RegExp(`[?&]${name}=([^&#]*)`).exec(window.location.search);
  return m ? decodeURIComponent(m[1].replace(/\+/g, ' ')) : null;
}
      Konstantinos Smaragdas, 2 weeks ago • Initial commit …
const STUDY_TOKEN = (() => {
  const fromUrl = getQueryParam('token');
  const fromLS  = localStorage.getItem('study_token');
  const tok = fromUrl || fromLS || '';
  if (tok) localStorage.setItem('study_token', tok);
  return tok;
})();

const ATTEMPT_NO = (() => {
  const raw = getQueryParam('attempt');
  const n = parseInt(raw || '1', 10);
  return (n === 1 || n === 2) ? n : 1;
})();

// Δημιουργία ανώνυμου χρήστη + φάση PRE/POST
;(function seedUserFromTokenAndAttempt(){
  if (typeof STUDY_TOKEN !== 'undefined' && STUDY_TOKEN && STUDY_TOKEN.trim()) {
    const anonId = `stu_${STUDY_TOKEN.trim()}`;
    localStorage.setItem(LS.USER_ID, anonId);
    localStorage.setItem('QUIZ_USER', anonId);
  }
  const phase = (ATTEMPT_NO === 1 ? 'PRE' : 'POST');
  localStorage.setItem('QUIZ_PHASE', phase);
})();
```

Figure 4.6 – Functionality of the anonymous identifier and evaluation phase initialization mechanism

Detailed explanation:

1. getQueryParam(name)

It is a helper function that reads parameters from the URL (e.g. ?token=...&attempt=1). It uses a regular expression to find the value of the name argument (e.g. token, attempt) and returns it decoded.

2. STUDY_TOKEN (IIFE)

The function first reads the token from the URL (fromUrl) and then if it is not found it tries to get it from localStorage (fromLS). If it is found in either of the two, it stores it in localStorage with the study_token key ensuring that even if the user closes and reopens the page with the same browser, the token remains available without the need for re-registration.

3. ATTEMPT_NO (IIFE)

This function is one of the core mechanisms of the quiz since it determines the phase of the user participation. In practice, it reads the attempt parameter from the URL, converts it to an integer and limits the results to two values: 1 or 2. In this way, the system recognizes the phase of the test, encoding the logic: attempt=1 corresponds to PRE test, while attempt=2 to POST test. This procedure ensures that the quiz knows which phase the participant is in without any further action from the user.

4. seedUserFromTokenAndAttempt()

This function is responsible for creating and storing an anonymous user identifier based on the token. If there is an active STUDY_TOKEN, the system creates a unique code of type stu_<token> and stores it locally in localStorage with the keys LS.USER_ID and QUIZ_USER. At the same time, it identifies the quiz phase, storing it as QUIZ_PHASE in the same storage space where the value is PRE when ATTEMPT_NO equals 1 and POST when it equals 2. This information becomes accessible by all other quiz subsystems (e.g. loading questions, sending answers, summary page) ensuring consistent behavior in each phase.

4.4.2.2 SoftSkills User Quiz operation logic

The SoftSkills User Quiz is based on an adaptive flow that guides the student from the initial category selection to the completion of the 16 questions. The system is designed to provide a personalized assessment experience, track the user's progress and dynamically record performance data without an interruption or loss of session. The quiz architecture is based on a set of state variables and control functions that lead to the formation of the sequence of questions, the calculation of the initial level and the determination of the order of appearance of the remaining skills.

At the beginning, the list of available soft skills categories and the main variables that control the internal flow of the quiz are defined:

The ALL_CATEGORIES table contains the four soft skills that are assessed and the remaining variables determine the state of the quiz:

- START_CATEGORY stores the category from which the participant chose to start
- BRANCHED indicates whether "branching" has already been done after the first four questions
- LEVEL stores the assigned level based on the average performance of the user on the initial questions
- FINISHED marks the completion of the test

After the user selects the first category, the startNewQuiz() function is executed which prepares the environment, resets progress and retrieves the initial set of questions from the backend. This function is the starting point for each new attempt and initializes the quiz flow. It resets any previous state, reads the category the student has chosen and asks the backend (loadFour(START_CATEGORY)) to return four questions related to that skill. These questions are stored locally in the BUNDLE and are immediately displayed in the user interface via renderCurrent(). When the quiz starts, the participant sees the message “Starting with [Category] (4 questions)” on the screen while their progress is saved in localStorage so that they can continue in case of interruption. This design ensures stability in the flow and prevents data loss.

```
/* ================== Adaptive state ================== */
const ALL_CATEGORIES = ['Communication', 'Teamwork', 'Leadership', 'Problem Solving'];
let START_CATEGORY = null;
let BRANCHED = false;
let LEVEL = null;
let FINISHED = false;
```

Figure 4.7 – Initialization of the adaptive assessment state

After answering the first four questions, the system proceeds to the critical phase “branching”. At this stage, the average of the scores of the first four questions is calculated, the user’s level (LEVEL) is determined and the questions from the remaining categories are automatically loaded. This logic is captured as follows:

```
// === Branch μετά την 4η ===
if (!BRANCHED && CUR === 3) {
  const first4 = RESULTS.slice(0, 4).map(r => (typeof r.score === 'number' ? r.score : 0));
  const avg = first4.length ? (first4.reduce((a, b) => a + b, 0) / first4.length) : 0;
  LEVEL = bandFromAvg(avg);
  BRANCHED = true;

  const others = ALL_CATEGORIES.filter(c => c !== START_CATEGORY);
  const batches = [];
  for (const cat of others) batches.push(await loadFour(cat));
  BUNDLE = [...BUNDLE, ...batches.flat()];

  const lvl = $('#levelBadge');
  if (lvl) {
    lvl.textContent = `LEVEL: ${LEVEL.toUpperCase()}`;
    lvl.dataset.level = LEVEL;
  }

  stopTimer();
  startTimer(DEFAULTS.secondsPerQuestion);
  saveProgress();
}
```

Figure 4.8 – Adaptive logic for determining user level after the first four questions

The if (!BRANCHED && CUR === 3) condition is activated once the user completes the fourth question ensuring that the “branching” process will only be executed once. The system extracts the scores of the first four questions from the RESULTS table, calculates the average and determines the level of the user with the help of the bandFromAvg() function which categorizes the performance into ‘low’, ‘mid’ or ‘high’.

After the level is calculated, the BRANCHED variable is set to TRUE and the quiz automatically loads the rest of the questions for the next skill. The ALL_CATEGORIES.filter(c => c !== START_CATEGORY) command ensures that the original category is not repeated while the new questions are integrated into the BUNDLE so that the user continues without interruption with the next blocks of 12 questions.

Finally, the level is updated in the graphical environment through the #levelBadge element which dynamically changes to display “LEVEL: LOW”, “LEVEL: MID” or “LEVEL: HIGH”. This is immediate feedback on their initial performance and is linked to the learning material that will be provided at the end of the assessment.

4.4.2.3 Answer recording flow (open and multiple choice)

```
$('#btnPrev')?.addEventListener('click', async (e)=>{
  e.preventDefault(); e.stopPropagation();
  if (CUR > 0){
    const q = BUNDLE[CUR];
    if (q.type === 'open') q.answer = ($('#answer')?.value || '').trim();
    else {
      const sel = document.querySelector('input[name="mcOpt"]:checked');
      q.selected_id = sel ? sel.value : (q.selected_id || null);
    }
    saveProgress();
    await animateQuestionSwap(CUR - 1);
  }
});
      Konstantinos Smaragdas, 2 weeks ago • Initial commit
/* “Επόμενο” όταν δεν είναι finish */
$('#btnNext')?.addEventListener('click', async (e)=>{
  const next = $('#btnNext');
  if (next?.dataset.role === 'finish') return;
  e.preventDefault(); e.stopPropagation();

  if (CUR === BUNDLE.length - 1 && !BRANCHED){
    alert('Κάνε πρώτα Βαθμολόγηση στην 4η ερώτηση για να συνεχίσουμε με τις προσαρμοστικές ερωτήσεις.');
    return;
  }
  if (CUR < BUNDLE.length - 1){
    const q = BUNDLE[CUR];
    if (q.type === 'open') q.answer = ($('#answer')?.value || '').trim();
    else {
      const sel = document.querySelector('input[name="mcOpt"]:checked');
      q.selected_id = sel ? sel.value : (q.selected_id || null);
    }
    saveProgress();
    await animateQuestionSwap(CUR + 1);
  }
});
```

Figure 4.9 – Navigation and answer storage functions in the quiz

Before the user moves to the next or chooses to go back to the previous question, the application records the current answer explicitly distinguishing between open-ended and multiple-choice questions. This logic is implemented in the “Previous” and “Next” buttons. At this point, the BUNDLE array acts as a

“carrier” of all the quiz questions along with the fields associated with the answers. For each current question q, if it is an open-ended question (q.type === ‘open’) the system reads the content of the text field (#answer), “cleans” it with trim() and stores it in q.answer. If it is a multiple-choice question, it looks for the selected radio button (input[name=”mcOpt”]:checked) and stores its value in q.selected_id. If no option has been selected, it keeps the previous value or null.

Immediately after, saveProgress() is called which writes the current state (questions, answers, CUR position, PRE/POST phase) to localStorage. In this way, each answer (open/multiple choice) is bound to the question within the BUNDLE and safely stored before any page or question change.

4.4.2.4. Question distribution mechanism

```
}
/* ================== Data loaders ================== */
async function loadFour(category){
  const base = ensurePrefix(getAPIBase().trim());

  // Phase & Attempt (από το state του UI / URL)      Konstantinos Smaragdas, 2 weeks ago *
  const phase   = (localStorage.getItem('QUIZ_PHASE') || 'PRE').trim(); // "PRE" | "POST"
  const attempt = (typeof ATTEMPT_NO !== 'undefined' ? ATTEMPT_NO : 1); // 1 | 2

 const url = joinUrl(
  base,
  `/questions/bundle` +
  `?category=${encodeURIComponent(category)}` +
  `&n_open=2&n_mc=2` +
  `&phase=${encodeURIComponent(phase)}` +
  `&attempt=${encodeURIComponent(attempt)}` +
  (DEBUG_SHOW_CORRECT ? `&include_correct=true` : ``)
);

  const data = await fetchJSON(url);

  // open: [{id, text}]
  const openQs = (data.open || []).slice(0, 2).map(q => ({
    id: q.id,
    text: String(q.text || ''),
    type: 'open',
    category
  }));

  const mcQs = (data.mc || []).slice(0, 2).map(q => {
    let options = [];
    if (Array.isArray(q.options)) {
      options = q.options.map(o => ({ id: String(o.id), text: String(o.text) }));
    } else if (Array.isArray(q.choices)) {
      options = q.choices.map((t, i) => ({ id: String(i), text: String(t) }));
    }
```

Figure 4.10 – Loading and preprocessing of questions by skill category

As mentioned above, the quiz structure consists of sixteen questions in total (8 MC and 8 OE) which are retrieved from a question bank created for the purposes of this thesis.

The selection of 2 questions of each type per skill (2 MC and 2 OE) is integrated into the system flow ensuring a balanced distribution as shown in the code snippet above.

The loadFour() function calls the backend endpoint /questions/bundle requesting two MC and 2 OE questions for each skill category. These questions are checked for missing or non-valid fields and converted to a uniform format (type: 'mc' or type: 'open') before being integrated into the quiz BUNDLE. In this way, the distribution of questions does not depend on external factors or random selection but is guaranteed to ensure consistency of the methodology and equal evaluation conditions between participants. The assessment process (e.g. calls to the backend for open answers, scoring, GLMP, fuzzy rules, etc.) is described in detail in the next chapter where the backend mechanism and the logic of automatic scoring are examined.

4.4.3. Backend: SoftSkills Bot

4.4.3.1. Technological background and general architecture

```python
def get_engine():
    global _engine
    if _engine is None:
        # Παίρνουμε URL από .env ή πέφτουμε σε local SQLite για dev
        db_url = getattr(settings, "DATABASE_URL", "sqlite:///./softskills.db")

        # Για PostgreSQL (Neon), αποφεύγουμε broken connections
        connect_args = {}
        if db_url.startswith("sqlite"):
            connect_args["check_same_thread"] = False

        _engine = create_engine(
            db_url,      Konstantinos Smaragdas, 2 weeks ago • Initial commit …
            echo=False,
            pool_pre_ping=True,  # αποφεύγει broken connections
            connect_args=connect_args,
        )

    return _engine


def init_db() -> None:
    """
    Δημιουργεί όλους τους πίνακες αν δεν υπάρχουν.
    Τρέχει στην εκκίνηση (π.χ. μέσα στο on_startup).
    """
    engine = get_engine()
    SQLModel.metadata.create_all(engine)


def get_session() -> Generator[Session, None, None]:
    """
    FastAPI dependency για injection μέσω Depends(get_session)
    """
    with Session(get_engine()) as session:
        yield session
```

Figure 4.11 – Configuration and initialization of the SoftSkills Bot database

The SoftSkills Bot backend system is the brain, the central logic of the evaluation and data management of the application. It is implemented in Python using the FastAPI web framework and the SQLModel library (on top of SQLAlchemy) for communication with the database. Its purpose is to cover the requests from the SoftSkills User Quiz and the Rater UI, to automatically rate the answers, apply GLMP/fuzzy rules, store the results of the assessments and provide the necessary data for analysis and reporting.

The initialization of the database and the connection is done centrally in the app/core/db.py file. In this script, a common engine is defined for the entire application as well as a dependency function get_session() that is used by endpoints to obtain a ready Session.

In this way, the backend server is connected to a single PostgreSQL database (or alternatively SQLite for local testing) and all models (e.g. Answer, Evaluation, HumanRating) are written to the metadata so that the corresponding tables are automatically created. Using get_session() as dependency with Depends(get_session) on the FastAPI endpoints ensures that each HTTP request handles its own transaction with the database.

The central definition of the application is located in app/main.py. In this script, the FastAPI app is created, the necessary settings (title, version, root_path) are applied and the routers corresponding to the main subsystems are added: questions, scoring, GLMP/fuzzy assessment, reports and rater interfaces.

```python
# --- Routers ---
from app.routers.questions import router as questions_router
from app.routers.score import router as score_router
from app.routers.diag import router as diag_router
from app.routers import (
    rater_final,
    glmp,
    coach,
    report,
    rules,
    diagnostics,
    rater_calibrate,
)
from app.routers.rater_simple import router as rater_simple_router

API_PREFIX = "/api/softskills"
ROOT_PATH = os.getenv("FASTAPI_ROOT_PATH", "")
BUILD_TAG = os.getenv("BUILD_TAG", "dev")            You, 19 minutes ago • Uncommi

app = FastAPI(
    title=getattr(settings, "PROJECT_NAME", "softskills-bot"),
    version=getattr(settings, "VERSION", "1.0.0"),
    root_path=ROOT_PATH,
)

app.include_router(glmp.router,            prefix=API_PREFIX)
app.include_router(coach.router,           prefix=API_PREFIX)
app.include_router(rules.router,           prefix=API_PREFIX)
app.include_router(rater_calibrate.router, prefix=API_PREFIX)
app.include_router(report.router,          prefix=API_PREFIX)
app.include_router(diagnostics.router)
app.include_router(rater_final.router,     prefix=API_PREFIX)
app.include_router(questions_router,       prefix=API_PREFIX)
app.include_router(score_router,           prefix=API_PREFIX)
app.include_router(rater_simple_router,    prefix=API_PREFIX)
app.include_router(diag_router, prefix="/api/softskills")
```

Figure 4.12 – Main FastAPI architecture and registration of system routers

The use of the common API/PREFIX = "/api/softskills" provides a consistent namespace for most system functions, such as:

Table 4. Core Backend API Endpoints and Their Functional Roles

/api/softskills/questions/bundle	Provides question blocks to the User Quiz
/api/softskills/score-open and /api/softskills/score-mc	For automatic rating of OE and MC questions
/api/softskills/glmp/*	For application of fuzzy rules and GLMP
/api/softskills/report*	For exporting reports and data for statistical analysis

The backend is structured in a way to accept requests from the SoftSkills User Quiz (questions, answers, scoring), applies heuristic and LLM/GLMP logic, updates and reads from the database, provides the Rater UI and reports with all necessary data for human evaluation, comparative analysis and drawing conclusions. In other words, it is responsible for all the logic of the system and for every action that accompanies it in order to implement the desired result for the developer.

4.4.3.2. LLM Coach Engine – Role and operating logic

The logic of the SoftSkills User Quiz including the automatic assessment is based on artificial intelligence. The LLM Coach Engine mechanism aims to interpret, evaluate by recognizing cognitive and emotional patterns but also to provide feedback on the answers simulating the logic of a human evaluator. Its operation is based on a LLM which is guided by carefully designed prompts that convey clear pedagogical instructions.

The process of interaction with the language model begins with the careful construction of two types of prompts which determine the behavior of the system and the evaluation framework of each answer. The first prompt refers to the system (system prompt) and defines the general role of the model as an 'education coach' by setting the rules of evaluation. It specifies that the model's response must be provided exclusively in valid JSON format with specific fields such as numerical score, short feedback sentences and criteria corresponding to cognitive and emotional aspects of soft skills. It also describes the 0-10 scoring scale and gives examples of how a "weak", "satisfactory" or "excellent" answer should be interpreted. The second prompt refers to the user (user prompt) providing the model with the actual data to be evaluated, namely the answer text, the category of the skill to which it belongs and the context of the question. This combination ensures that the model answer is not abstract but places its judgement in the right educational context considering the content and intention of the student. This approach enhances the reliability of the assessment as the LLM is not based exclusively on statistical language patterns but on an integrative process guided by pedagogical criteria

```
SYSTEM_OPEN = (
    "Είσαι ένας έμπειρος coach ανάπτυξης Soft Skills σε περιβάλλοντα επαγγελματικής εκπαίδευσης.\n"
    "ΠΡΕΠΕΙ να επιστρέφεις ΜΟΝΟ έγκυρο JSON (χωρίς κείμενο γύρω-γύρω), με ΑΚΡΙΒΩΣ τα πεδία που ζητούνται.\n"
    "\n"
    "ΚΡΙΤΗΡΙΑ ΒΑΘΜΟΛΟΓΗΣΗΣ (score 0..10):\n"
    "- 0–2: Η απάντηση είναι πολύ φτωχή ή σχεδόν άδεια: εκτός θέματος, μονολεκτική ή δεν απαντά καθόλου στο ζητούμενο.\n"
    "- 3–4: Η απάντηση είναι πολύ γενική ή θα μπορούσε να ταιριάζει σε πολλές ερωτήσεις, με ελάχιστα συγκεκριμένα σημεία.\n"
    "- 5–7: Καλή και σχετική απάντηση. Δείχνει κατανόηση του ζητήματος, περιέχει κάποιες συγκεκριμένες ιδέες/ενέργειες, αλλά δεν είναι πλήρως ανεπτυγμένη.\n"
    "- 8–10: Πολύ καλή/εξαιρετική, στοχευμένη απάντηση, με συγκεκριμένες ενέργειες/παραδείγματα που δείχνουν βαθιά κατανόηση και εφαρμόσιμα βήματα.\n"
    "\n"
    "ΟΔΗΓΙΕΣ ΚΛΙΜΑΚΑΣ:\n"
    "- Αν η απάντηση είναι κατανοητή, σχετική με την ερώτηση και περιέχει έστω μερικά πρακτικά σημεία, συνήθως ανήκει στην περιοχή 5–7.\n"
    "- Κράτα τις βαθμολογίες 0–4 για περιπτώσεις όπου η απάντηση είναι πολύ γενική, εκτός θέματος ή σχεδόν άδεια.\n"
    "- Χρησιμοποίησε 8–10 μόνο όταν η απάντηση είναι πραγματικά στοχευμένη, με σαφή δομή και συγκεκριμένα παραδείγματα/ενέργειες.\n"
    "\n"
    "Αν η απάντηση είναι αδύναμη ή εκτός θέματος, βαθμολόγησε χαμηλά αλλά ΔΩΣΕ ΣΥΓΚΕΚΡΙΜΕΝΟ coaching στα πεδία change/action/drill."
)
```

Figure 4.13 – Internal LLM instruction system (prompt) for Soft Skills response evaluation

The selection of the GPT-4o-mini model as the core of the LLM Coach Engine was based on technical and pedagogical criteria. Compared to larger models such as GPT-4 or Claude-3, GPT-4o-mini provides almost equivalent quality of language understanding and emotion recognition with significantly reduced response time and computational cost [103].Research has shown that smaller versions of LLMs such as GPT-4o-mini or Gemini Nano can yield equally accurate scores in educational assessment tasks when combined with robust prompt engineering. In addition, GPT-4o-mini is distinguished by its stability in producing JSON outputs and its ability to maintain high accuracy in semantic interpretation even for very short responses, a feature particularly useful in education environments with concise textual inputs [49], [104].

Overall, the LLM Coach Engine acts as a "digital mentor" that analyzes the quality of the response before any quantitative calculation. The combination of carefully structured prompts, the use of GTP-4o-mini model and the fallback mechanism ensures that the system remains intelligent and reliable. Furthermore, the LLM output feeds directly into the backend scoring mechanism completing the response-analysis-feedback cycle of the SoftSkills Bot

4.4.3.2.1. Data flow and automatic response evaluation

The data flow to the backend during the assessment of an answer starts from the moment the student submits an answer to the SoftSkills User Quiz and is completed when a complete evaluation package (score, feedback, criteria) is returned to the frontend and at the same time the answer is stored in the database. In practice, each OE or MC answer is converted into a JSON request that corresponds to the Pydantic models ScoreOpenRequest or MCPayload and is sent to the HTTP endpoints/score-open and /score-mc, respectively.

In the case of OE questions, the /score-open endpoint acts as the basic evaluation node. Upon input of a request, the backend first calculates a baseline score via the _rubric_open_heuristic_0_10 function. This function analyzes the text of the answer (length, presence of keywords, structure in "steps", use of examples, etc.) and returns an initial overall score of 0-10 along with individual criteria such as "clarity", "relevance", "structure" and "examples". At the same time, the heuristic_open_feedback function produces a first, simple coaching feedback (strong points and gaps) which can be used as backup feedback in case the call to the LLM fails. This heuristic stage is not the main evaluation strategy but acts as a safety valve so that no answer is left without a numerical score.

```
Konstantinos Smaragdas, 2 weeks ago | 1 author (Konstantinos Smaragdas)
class ScoreOpenRequest(BaseModelConfig):
    category: str      Konstantinos Smaragdas, 2 weeks ago • Initial commit …
    question_id: str
    text: str
    user_id: Optional[str] = None


Konstantinos Smaragdas, 2 weeks ago | 1 author (Konstantinos Smaragdas)
class ScoreOpenResponse(BaseModelConfig):
    text: str
    category: str
    question_id: str
    score: float
    feedback: Dict[str, Any]
    model: str = "heuristic"
    answer_id: Optional[str] = None
    interaction_id: Optional[str] = None
    criteria: Optional[List[Dict[str, Any]]] = None
```

Figure 4.14 – Definition of input and output data models for open-ended response evaluation

Next step is the LLM Coach Engine activation. If the system has an LLM available (_HAVE_LLM=True or force_llm=True), the response is forwarded to the llm_coach_open function which calls the GPT-4o-mini model with the prompts described in the previous subsection. The LLM returns a structured JSON with a total score, four coaching fields (keep, change, action, drill) and a list of criteria. If the output is considered valid (i.e. it does not contain an error and is not an "empty" payload), the backend mechanism combines this assessment with the heuristic baseline.

A blended score is first calculated where the LLM participates with a weight of 70% and the heuristic estimate with 30%. If the LLM has returned detailed criteria, these are passed then through the _weighted_from_criteria function which applies category-dependent weights to the individual criteria (different for each soft skill) to generate a more refined, criterion-weighted score. The resulting score is calibrated per category (_calibrate_category_score) where a nonlinear transformation function (with a gamma parameter per category) is applied to reinforce truly high performances and discourage marginal ones.

All the above results (final score, detailed criteria, coaching and information on whether LLM was used) are stored in the database via the _dynamic_insert and _upsert_answers_and_llm helper functions. In the interaction table, an entry is created where the category, question type, answer text and time are recorded as well as entries of the automatic scores and final assessments. At the same time, the endpoint returns to the frontend a ScoreOpenResponse object which includes the final grade 0-10, the structured feedback and the criteria so that they are immediately displayed on the feedback page.

In case the call to LLM fails or the response is deemed unusable (e.g. invalid JSON or completely empty coaching fields), /score-open fully activates the heuristic fallback mechanism. In this path, the final score is derived exclusively from _rubric_open_heuristi_0_10 while the feedback fields are filled in by heuristic_open_feedback. The answer is normally stored in the same database tables and the system continues to operate without interruption, a critical feature for the functional stability of the application in real conditions.

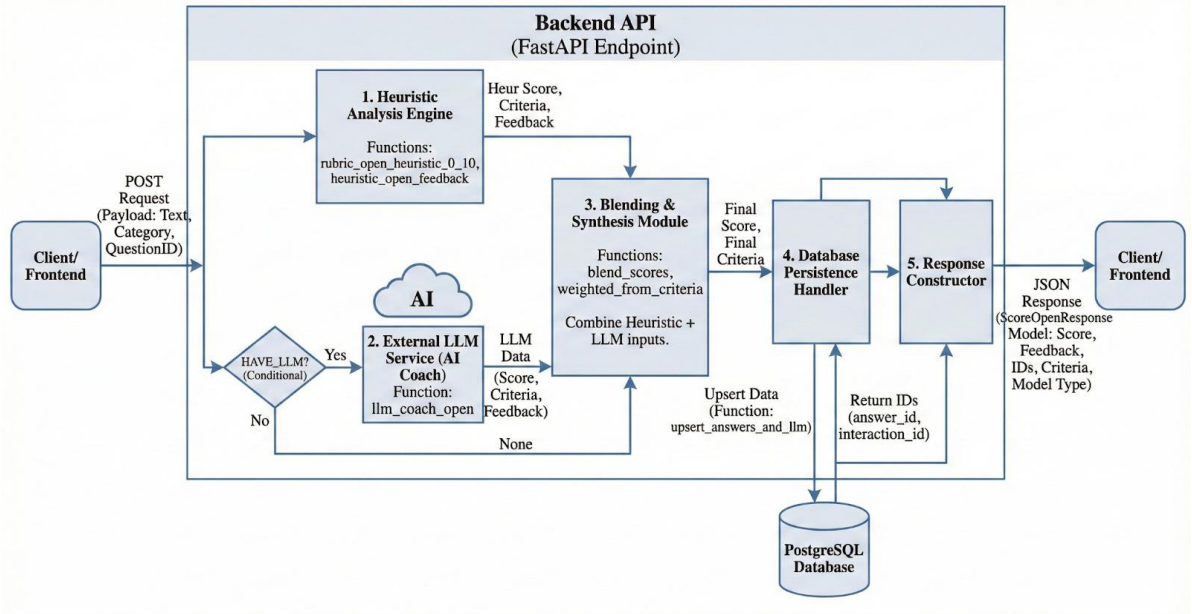

Figure 4.15 – Lifecycle of an Open-Ended Answer Scoring Request in the API (/score-open)

The /score-mc endpoint manages the MC questions, respectively. The backend receives via MCPayload the skill category, the question text, the list of options, the user's choice and the correct_id (where available). If the correct_id has not been given, the system retrieves it from the database or from the question metadata. An initial score (correct/incorrect answer) and a simple coaching message are then generated. If the use of LLM for MC is active (via llm_coach_mc), the model can generate additional verbal feedback, justification of correct or incorrect and support softer evaluation where appropriate (e.g. partial success on similar options). The final result is condensed into ScoreMCResponse which contains the auto_score, whether the answer is correct, a confidence level and the coaching elements that will be displayed to the user.

```
Konstantinos Smaragdas, 2 weeks ago | 1 author (Konstantinos Smaragdas)
class MCPayload(BaseModelConfig):
    category: str
    question_id: str
    user_id: Optional[str] = None
    question_text: str
    selected_id: str
    correct_id: Optional[str] = None
    options: List[Dict[str, Any]]  # [{id, text}, ...]
      Konstantinos Smaragdas, 2 weeks ago • Initial commit …
Konstantinos Smaragdas, 2 weeks ago | 1 author (Konstantinos Smaragdas)
class ScoreMCResponse(BaseModelConfig):
    answer_id: Optional[str] = None
    interaction_id: Optional[str] = None
    source: str
    model_name: str
    llm_used: bool
    correct: Optional[bool] = None
    auto_score: float
    confidence: float
    feedback: Any  # μπορεί να είναι str ή dict
    coaching: Dict[str, Any]
    criteria: Optional[List[Dict[str, Any]]] = None
```

Figure 4.16 – Definition of input and output data models for multiple-choice response evaluation

Finally, for cases where the rating has already been calculated by the GLMP/fuzzy logic engine, the data flow passes through the endpoint/score-open-from-glmp. In this case, the backend does not call the LLM again or re-apply calculation. It simply receives an already calculated score 0-10 from the GLMP engine, stores it in the same tables (interaction, autorating, answers) and updates the relevant fields in a way compatible with all other flows. In this way, SoftSkills Bot unifies all the rating sources (LLM, heuristic rules, GLMP/fuzzy) in a coherent pipeline maintaining a common data structure and full traceability for each answer.

4.4.3.3. Score evaluation and processing mechanisms

The automatic evaluation performed by the backend is not based on a simple score from the LLM. On the contrary, it consists of a set of complementary mechanisms that work together to produce a final, pedagogical consistent and statistically usable score. In this context, there are three levels of processing: the heuristic rule mechanism, the GLMP logic model that utilizes the LLM criteria and finally the normalization through fuzzy logic and calibration per skill category.

4.4.3.3.1. Heuristic Engine - Rules and fallback logic

The heuristic mechanism is the core, rule-based basis of the evaluation and has two roles: an initial, quick assessment of the quality of an answer and an alternative mechanism in cases where the call to the LLM does not yield a usable result. Its logic is implemented through functions such as _rubric_open_heuristic_0_10 and heuristic_open_feedback which apply a set of rules per soft skill category.

At the operational level, elements such as clarity of formulation, correspondence to the question, the existence of structure (e.g. description of steps or actions), reference to specific examples and the presence of empathy or collaborative attitude (where relevant) are examined for each OE answer. Each of these aspects is mapped to a criterion with a score of 0-10 and the overall heuristic score is obtained as an average of the individual criteria. At the same time, a simple verbal comment is created that points out to the users what they do well and where they fall short but without the complexity and depth of the LLM.

This mechanism is used in all cases as a baseline that is a starting point providing a first impression of an answer. If the LLM returns a valid and sufficient evaluation, the heuristic score is combined with the LLM score in a blended result. On the contrary, if the LLM fails or its output is deemed useless (e.g. empty fields, invalid JSON) then the heuristic evaluation takes on the role of the final score. In this way, the system remains fully functional even in conditions with interruptions in the AI service or unexpected errors, while the use of a rule-based mechanism adds an element of repeatability and transparency.

4.4.3.3.2. GLMP Engine - Enriched Evaluation Logic

Beyond the basic numerical score returned by the LLM, the assessment of the soft skills requires a more complex mechanism that can interpret and weigh the qualitative dimensions of the answers. For this purpose, the GLMP (Guided Linguistic Modeling Process) Engine has been developed in the backend as an intermediate layer that converts the analytical information produced by the LLM (such as the individual criteria “clarity”, “relevance”, “empathy”, “specificity” and “actionability”) into a more stable, interpretable and comparable numerical result.

```
_WEIGHTS_BY_CATEGORY = {
    "Communication":   {"relevance": 2, "clarity": 2, "empathy": 3, "actionability": 2, "specificity": 1},
    "Teamwork":        {"relevance": 2, "clarity": 1, "empathy": 3, "actionability": 2, "specificity": 2},
    "Leadership":      {"relevance": 2, "clarity": 2, "empathy": 2, "actionability": 3, "specificity": 1},
    "Problem Solving": {"relevance": 2, "clarity": 1, "empathy": 1, "actionability": 3, "specificity": 3},
}
```

Figure 4.17 – Weight coefficients of evaluation criteria per Soft Skill

The GLMP operation is based on the principle that each soft skill consists of multiple behavioral and cognitive components which are not of equal importance. For example, in communication empathy and clarity have greater weight while in leadership action and taking initiative have priority. These weights are defined in the backend through the _WEIGHTS_BY_CATEGORY table which is the theoretical model behind the application:

When LLM returns its results in JSON format, each response is accompanied by a “criteria” field, i.e. a set of micro-scores per individual dimension. GLMP utilizes this data and calculates a weighted average where the weight of each criterion depends on the skill being evaluated. This operation is performed through the weighted_from_criteria() function which is location in the file app/core/score_glmp.py.

This function calculates the GLMP score for each response which is a numerically smoothed average where each individual dimension contributes according to its weight. If the LLM does not return sufficient data (e.g. the criteria field is missing), the process is automatically skipped and the simple result of the LLM or the heuristic fallback is used. When data is complete, the GLMP incorporates the information and yields a final score with greater reliability and consistency.

```
def _weighted_from_criteria(criteria: list, category: str) -> float | None:
    """Compute weighted score 0..10 from criteria list like [{'name': 'Clarity','score': 7},...]"""
    if not isinstance(criteria, list) or not criteria:
        return None
    weights = _WEIGHTS_BY_CATEGORY.get(_norm_cat(category), _WEIGHTS_BY_CATEGORY.get(category, {}))      You, 1 second a
    acc = 0.0
    den = 0.0
    for c in criteria:
        try:
            name = str(c.get("name", "")).strip().lower()
            score = float(c.get("score", 0))
        except Exception:
            continue
        key = None
        if "clarity" in name:
            key = "clarity"
        elif "relevance" in name or "σχετικ" in name:
            key = "relevance"
        elif "empathy" in name or "ενσυνα" in name:
            key = "empathy"
        elif "action" in name or "πρακτικ" in name:
            key = "actionability"
        elif "specific" in name or "παράδειγ" in name or "τεκμηρ" in name:
            key = "specificity"
        if key and key in weights:
            w = float(weights[key])
            acc += max(0.0, min(10.0, score)) * w
            den += w
    if den <= 0:
        return None
    return acc / den
```

Figure 4.18 – Computation of weighted score based on evaluation criteria

In practice, GLMP acts as a stabilization filter over the results of the LLM. Its goal is not to replace the model but to ensure that the score remains pedagogically valid and consistent even if two responses have

different style or length. This process reduces statistical "noise" and enhances the reliability of comparisons between skills and between PRE-POST phases.

Finally, GLMP can also operate autonomously as in the endpoint /score-open-from-glmp where the responses are evaluated solely based on the criteria already been calculated. In these cases, the result is stored as a "GLMP score" on a scale of 0-10 and is used directly in statistical analysis.

4.4.3.3.3. Fuzzy Logic and score calibration per category

The final stage of the assessment process before storing the results in the database and presenting them to the user involves normalizing and calibrating the scores through a fuzzy normalization approach. This stage was necessary because of the nature of soft skills which are not evenly distributed on a linear scale. Small differences in the score range of 7-8 may indicate substantial progress in behavior while large differences between scores 1-3 may still represent low progress. For this reason, the use of numerical means does not realistically capture the true progress of the participant.

To address this issue, the system implements a final transformation stage inspired by the principles of fuzzy logic and pedagogical calibration. At this stage, the score resulting from the combination of the heuristic mechanism, LLM and GLMP goes through a nonlinear normalization process that adjusts the value depending on the soft skill category. This logic is implemented in the backend in the _calibrate_category_score() function as shown in the following excerpts:

```
_GAMMA_BY_CATEGORY = {
    "Communication": 1.10,
    "Teamwork": 1.15,
    "Leadership": 1.15,
    "Problem Solving": 1.20,
}
```

```
def _calibrate_category_score(score_0_10: float, category: str) -> float:      You, 2 weeks ago • fixing the rater score
    try:
        s = max(0.0, min(10.0, float(score_0_10)))
    except Exception:
        s = 0.0
    gamma = _GAMMA_BY_CATEGORY.get(_norm_cat(category), 1.15)
    return round(((s / 10.0) ** gamma) * 10.0, 2)
```

Figure 4.19 – Category-specific score calibration mechanism using gamma-parameter

In the above implementation, each skill category has its own γ (gamma) factor which controls the degree of "skewness" of the transformation curve. When gamma is greater than 1, the scale becomes "stricter" meaning higher actual performance is required to assign a high score. For example, this is true in the Leadership ($\gamma = 1.15$) or Problem Solving ($\gamma = 1.20$) categories where it is considered pedagogically more difficult to reach an "excellent" skill level. In contrast, in more adaptive skills such as Communication or Teamwork, the coefficient remains close to 1.1-1.15 allowing for a more "lenient" scale that more easily recognizes improvement in everyday interaction.

This mechanism converts linear scores into fuzzy-normalized values mapping numerical differences into "linguistic categories" (low, moderate, high performance). In this way, an increase from 6 to 8 represents a clear move from moderate to high proficiency, while an increase from 2 to 4 remains within the "insufficient" range which better reflects the reality of soft skills training. The result is a scale that approximates the human perception of progress and provides consistent comparisons between different skills and phases (PRE-POST).

The application of fuzzy logic not only changes the score but also enhances the interpretability of the data. Passing all final scores through a common calibrator keeps the result comparable regardless of its source (LLM, GLMP or heuristic fallback). Thus, the database contains a single, standardized scoring format (0-10) allowing for reliable statistical analysis, progress measurement and assessment of the agreement between AI and human rating.

4.4.3.3.4. Integration and final score generation

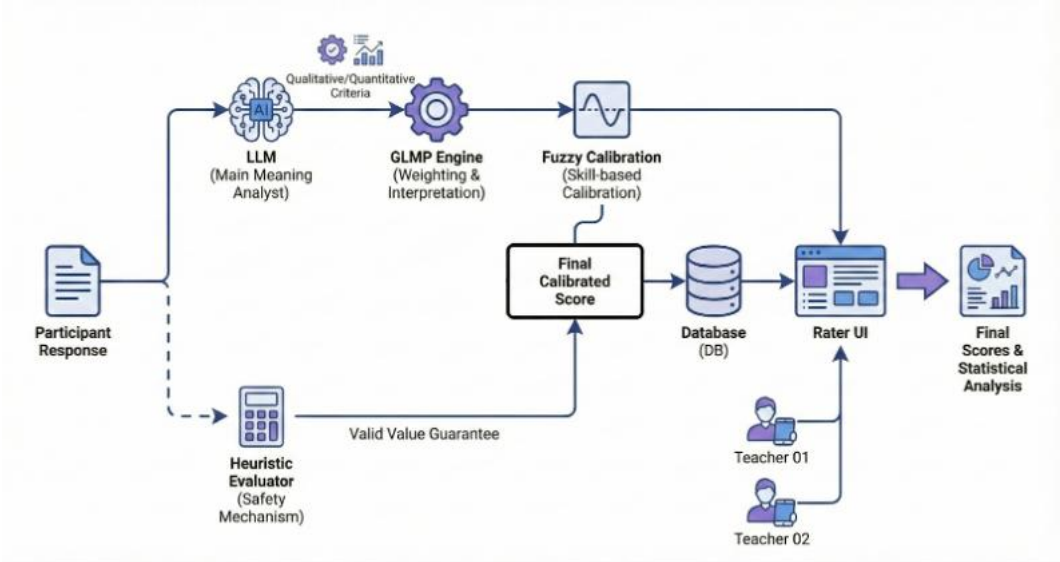

Figure 4.20 – Soft Skills evaluation mechanism using LLM, GLMP, and Fuzzy Calibration

The three mechanisms described above are combined serially to produce the final score for each response. The role of the heuristic mechanism is to ensure that there will always be a valid numerical value even in cases where the call to the LLM fails. The LLM acts as the main analyzers of the meaning of the answers extracting qualitative and quantitative evaluation criteria. The GLMP Engine converts these criteria into a weighted and interpretable form. Finally, the application of fuzzy logic calibrates the result according to the nature of the skill so that the final score reflects more accurately the actual performance of the participant.

This flow ensures that each score stored in the database is the product of a combination of quantitative and qualitative analysis: it starts from the linguistic understanding of the LLM, is enriched by the GLMP that incorporates the relative importance of each dimension and is smoothed through fuzzy calibration. Thus, the results are comparable between participants and between the PRE and POST phases while maintaining pedagogical consistency and stability across the range of skills.

The final calibrated score is the one stored in the database and forwarded to the Rater UI subsystem where it is combined with the human ratings (teacher01 and teacher02) to create the final scores used in the analysis of results and in the statistical assessment of the reliability of the system.

4.4.4. Rater UI - Human rating environment

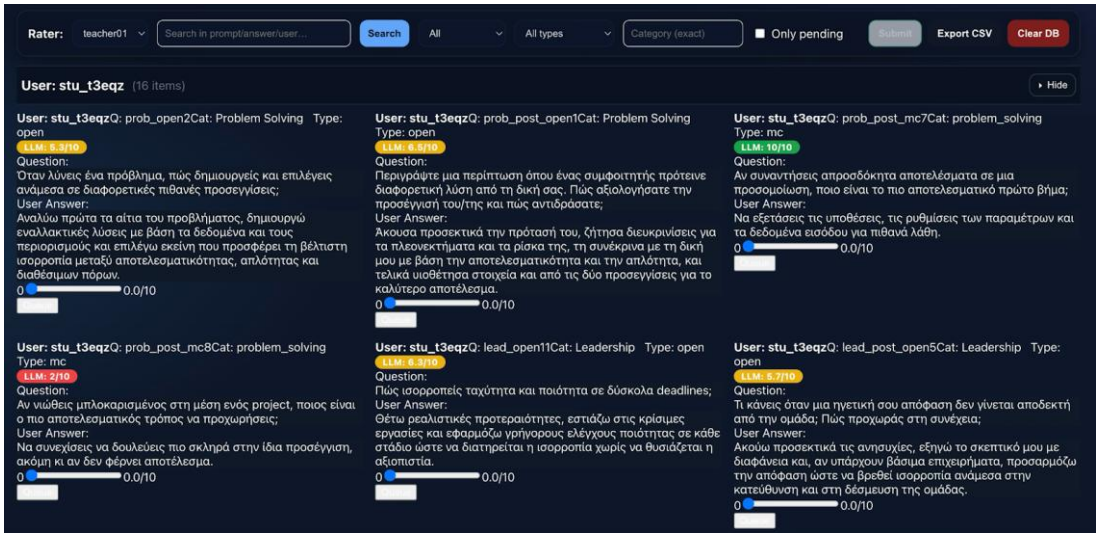

Figure 4.21 – Rater UI interface for manual evaluation and validation of results

Rater UI is the subsystem intended exclusively for teachers and acts as a bridge between the automatic rating system and human judgment. It was developed as a single-page application with React, TypeScript and Vite using simple CSS for the appearance and HashRouter for navigation. The application communicates with the backend through a lightweight client (api.ts) which reads the API address and the access key (API key) from environment variables or from localStorage. In this way, the same UI can be easily connected to both a local development server and a production environment without changes to the code.

The main role of Rater UI is to present teachers with the anonymous responses of students and allow them to assign numerical scores on a scale of 0-10 while having available the AI-generated scores. For each answer, the backend returns an object of type RaterItem containing the answer_id, question_id, anonymous user_id, question type (OE or MC), soft skill category, question and answer text and scores (LLM score, teacher01 or teacher02).

The Rater UI main screen (Figure 3.20) consists of two main areas: the top section with the search filters that each teacher can use and the bottom section with the list of answer cards (AnswerCard). In the first area, each evaluator can filter the large volume of data to facilitate navigation. They can limit the displayed answers based on:

- Keyword in the text of the question or answer,
- Soft skill category (e.g. only Communication or only Leadership),
- Question type (OE or MC),
- Presence or absence of automatic scoring by the LLM,
- “only pending” indication to see only answers that have not yet been scored by the specific evaluator.

At the bottom is the work area with the answer cards. Each card corresponds to a student's answer and includes the question text, the answer text, the category and type of question as well as the initial LLM score on a scale of 0-10. The evaluator selects which answer to grade using a 0-10 slider and when satisfied with the score, presses the corresponding "Queue" button to add the answer to a temporary queue for submission. The evaluator can adjust and check multiple scores before sending them to the backend reducing the risk of errors and allowing for faster batch rating. In addition, the answers are grouped by user and there is a "collapse" option allowing the evaluator to view all answers of a student or focus on individual cases. This is particularly helpful for more consistent assessment as the teacher gets an overview of a participant's profile.

When the evaluator has completed a series of scores, they go to the "Submit" option which sends all queued answers to the backend. When sending, each record includes the answer_id, the rater_id (teacher01 or teacher02) and the final score. Once the sending is complete, the backend updates the database and returns updated values for the scores while the UI clears the queue and refreshes the list highlighting the answers that were successfully graded.

Finally, the Rater UI also offers data management and export functions. The researcher or teacher can export all available assessment data in CSV file format which includes both automatic and human ratings through a button in the top right area of the screen (Figure 4.21). There is also provision for administrative actions such as complete clearing of the database only when a special admin token is given. This function is used exclusively in an experimental/testing context.

4.4.5. Tokens Utility – Anonymous identification and management of PRE/POST links

To ensure that participation in the experiment is simultaneously anonymous, unique (without duplicate participations) and matchable between PRE and POST phases, an auxiliary subsystem in Python was developed which in the context of this thesis is referred to as Tokens Utility.

The basic idea is that each participant is identified by an anonymous token embedded as a parameter in the quiz URL. The Tokens Utility is responsible for creating a pair of links for each future participant. The type of the links is as follows:

- PRE link: ...?token=<unique_code>&attempt=1
- POST link: ...?token=<unique_code>&attempt=2

These links lead to the SoftSkills User Quiz application where the quiz phase is automatically identified by the attempt parameter. The system internally checks the consistency of the token (i.e. that is already registered in the database and has not been used again), preventing the creation of duplicate records or the re-entry of the same participant.

The PRE/POST information is stored exclusively based on the token and is not linked to any other identifier (name, email or device). The result is that the backend can correlate the two phases of the same user (for monitoring progress) without knowing who the participant is.

At the implementation level, the Tokens Utility performs three main functions:

I. Token generation: uses the Python secrets library to generate high-entropy alphanumeric codes (e.g. ab12cd34ef56).
II. Storage: writes the tokens to a CSV file or directly to a database table along with the corresponding phases (attempt=1 and attempt=2).

III. Verification: when the user enters the quiz, the system checks whether the token is valid and associated with a specific phase (PRE or POST).

The choice of this mechanism over a traditional email registration process is based on two main advantages:

- Anonymity and privacy: the tokens are completely pseudonymized and not linked to identifiable data, following the ethical principles set out in this thesis.
- Simplicity and repeatability: the same token can be used automatically in both phases without the need for manual matching or participant tracking.

4.5 Foundation and scientific documentation of the evaluation system

4.5.1. Theoretical basis

The use of AI to assess natural language responses has its roots in the field of Automated Essay Scoring (AES) which has investigated the ability of computer systems to evaluate the quality of written texts in a manner comparable to human judgement [40]. Originally, AES approaches were based on statistical models and surface-level linguistic features, such as sentence length, grammatical complexity and vocabulary variety. However, the rapid development of LLMs and particularly transformer-type architectures has enabled the transition from mechanical feature measurement to semantic understanding and interpretation of content [49], [40], [104].

In this new context, LLMs are no longer limited to measuring the "technical" quality of a text but can assess aspects related to intention, empathy, coherence and communication effectiveness, the elements that define soft skills. Recent studies show that models such as GPT-4.0 achieve high correlation with human scores in a variety of educational scenarios reaching Intraclass Correlation Coefficients (ICC) above 0.85 [49]. This suggests that the use of LLMs can be considered a reliable approach for scoring semi-structured and OE responses given that there is clear guidance of the model through careful prompt engineering and structured output control.

This approach is closely linked to the concept of formative assessment where the sole purpose of evaluation is to facilitate learning through the provision of meaningful feedback [105], [106]. In this context, LLMs can function as "cognitive coaches" providing students with interpretive comments and suggestions for improvement tailored to the content of their response. The value of this approach lies in the accuracy of the scoring and the possibility of automated, personalized feedback which enhances the self-regulation and active learning of the learner.

In the context of this thesis, the GTP-4.0 mini model was chosen which according to the OpenAI technical guide (2024) [103] offers high reliability in JSON-type responses and stable behavior in educational assessment tasks with significantly lower computational cost than larger models of the same family (GPT-4 Turbo, Claude 3). This choice was also based on research data showing that the lighter models of the GPT-4.0 family maintain similar scoring quality in educational tasks given that the instructions and prompts are appropriately designed [49], [104].

To sum up, the theoretical basis of this thesis is based on the coupling of two scientific methods: automatic natural speech assessment (via AES) and pedagogical feedback (formative assessment). The first provides the methodological infrastructure for the quantitative evaluation of the quality of the answers and the second introduces the dimension of learning and personal development. The

combination of these approaches reinforced by the development of LLMs, creates the scientific background on which the automatic soft skills scoring in the present study was based.

4.5.1.1. Justification for the use of GLMP and Fuzzy Logic in Soft Skills Assessment

The integration of GLMP architecture and fuzzy logic in the proposed system was based on a well-established scientific methodology to perform explainable and multidimensional assessment of complex skills. A representative example is the Itzamna Project: a multimodal AI platform for comprehensive transversal skills assessment developed by Guerrero-Sosa et al. (2025) [107]. In their work, they reported that the use of GLMP allows the representation of transversal and soft skills as hierarchical multilevel phenomena, from small quantifiable observations (e.g. clarity or initiative) to synthetic indicators of skills, such as leadership, communication and collaboration. This "granular" modeling logic ensures that the assessment is not limited to a single-dimensional score but considers levels of analysis and weighs relationships between the sub-dimensions of each skill.

GLMP acts as a unifying framework between the numerical data produced by LLM models and the linguistic descriptions used by human raters. The results are translated into interpretable categories allowing the system to combine quantitative consistency with qualitative understanding rather than treating scores as absolute values [107]. In the Itzamna Project, this project is carried out through fuzzy membership functions that have been defined in collaboration with experts in psychology and pedagogy in order to reflect human judgments and avoid arbitrary mathematical delimitations. Similarly, in the proposed system GLMP is used to transform the individual criteria of LLM into a weighted composite score maintaining interpretability and comparability between phases.

The incorporation of fuzzy logic is the second pillar of this methodology. Fuzzy functions allow the system to capture the nonlinearity of human performance [108]. In practice, this means that small improvements in low performance (0-4) do not translate into proportionally equal increases, while high scores (8-10) correspond to increasingly subtle differences in qualitative superiority. As mentioned in subsection 4.4.3.3.3., fuzzy rules act as "calibration filters" transforming raw values into linguistic variables ("insufficient", "satisfactory", "advanced") allowing the system to provide human-readable feedback rather than simply numerical results.

Furthermore, pedagogical validity is of equal importance in this type of research. Systems that utilize fuzzy reasoning and linguistic modelling follow the principles of XAI-ED allowing the learner to understand the reason for receiving a specific score and how it can be improved. This explainability is crucial for building trust in automatic assessment systems and for their ethical legitimacy in educational practice.

Finally, the Itzamna model highlighted the need for cross-cultural and linguistic adaptability while implementing calibration and fairness auditing mechanisms to avoid biases in fuzzy estimates. The same tactic is adopted in the proposed system where the fuzzy rules and GLMP weighting levels are designed to reflect a neutral, multicultural linguistic approach. Overall, the choice of LLM -> GLMP -> Fuzzy logic is methodologically and pedagogically sound combining the algorithmic power of LLMs with the human readability and interpretability of fuzzy systems.

4.6 Data collection and analysis

4.6.1. Data collection

Data collection was designed to ensure completeness, reproducibility and anonymity allowing for accurate statistical processing and comparative analysis between the PRE and POST phases. All data were automatically recorded through the system subsystems (SoftSkills Bot, User Quiz, Rater UI) and stored securely in the PostgreSQL database which acts as a single repository of research records.

The data were organized into four main categories:

1. Participant responses which record results of each test in the PRE and POST phases for both MC and OE questions.
2. Automatic scores (AI-generated scores) derived from the processing of OE answers by the SoftSkills Bot system which combines LLM, GLMP and fuzzy rules.
3. Human ratings collected through the Rater UI environment by two independent raters.
4. Final combined scores which incorporate both AI and human assessments through weighted calculation and form the basis of the final reliability analysis and PRE-POST improvement assessment.

These data were exported in CSV format for statistical processing allowing for:

- Descriptive analysis of performance by skill,
- PRE-POST comparison at participant level,
- Assessment of agreement between AI and human raters.

The statistical analysis tools and methods used to quantitatively investigate the results are presented in the next section.

4.6.2. Data analysis

The aim of the data analysis is to provide additional evidence on the reliability, consistency and effectiveness of the SoftSkills AI system in the assessment of soft skills. The process follows the established standards of educational statistical analysis [109], [110] combining descriptive and inductive methods. It was carried out in a Python 3.11 computing environment using specific libraries such as pandas, numpy, scipy and statsmodels. The following data analysis steps were performed:

1) Descriptive Statistics

Aim was to capture the main characteristics of the scores and to form an overview of the performance of the participants. In addition, descriptive statistics help in the detection of extreme values (outliers) or possible asymmetries. For each skill category (Communication, Teamwork, Leadership, Problem Solving) and for each phase of the experiment (PRE and POST), the following were calculated:

- mean (Mean),
- standard deviation (SD),
- median (Median),
- range of variation (Min-Max),
- distribution of scores through graphical representations (e.g. histograms and boxplots).

At this stage, the normality of the PRE-POST differences was also checked with normality tests (e.g. Shapiro-Wilk) in order to decide on the appropriate inductive method [109].

2) PRE-POST change analysis (before-after comparison)

To investigate the change in performance between the PRE and POST phase, paired-samples t-tests were applied. This method is appropriate when the variables are continuous and normally distributed where the skill scores are expressed on a scale of 0-10 and come from the same participants at two points in time [109]. The paired-samples t-test examines whether the mean score in the POST phase differs statistically significantly from the mean score in the PRE phase for each soft skills category. The statistical significance level used in this study was $\alpha = 0.05$ which is an acceptable threshold in social and educational research [109], [110]. In cases where data distribution was not normal, non-parametric tests were used such as the Wilcoxon signed-rank test.

To assess the practical significance of the improvement beyond statistics, the Effect Size was calculated using Cohen's d [110]. This index is defined as:

$$d = \frac{\bar{X}_{\text{POST}} - \bar{X}_{\text{PRE}}}{s_d}$$

Figure 4.22 – Calculation of effect size (Cohen's d) between PRE and POST phases

where s_d is the standard deviation of the differences between PRE and POST scores. According to Cohen (2018), values of $d = 0.2$ correspond to a small effect, $d = 0.5$ to a medium effect and $d \geq 0.8$ to a large effect.

The combined use of paired-samples t-test (or Wilcoxon signed-rank test) and Cohen's d allows for the comprehensive assessment of both the statistical significance and pedagogical effectiveness of the system. This approach is in line with the standards of modern statistical analysis in educational research as documented in the international literature [109], [110], [111].

3) Reliability Analysis

This type of analysis concerns the evaluation of the reliability and consistency of the results produced by the SoftSkills AI system compared to the rating of human raters. Reliability assessment is a crucial step in verifying the stability and objectivity of measurements especially in systems based on AI [43], [66].

Since the scores of the SoftSkills AI system are recorded on a continuous numerical scale (0-10), the analysis focused on two indicators suitable for continuous data:

I. Intraclass Correlation Coefficient (ICC)

ICC is used to measure the degree of agreement between continuous quantitative measurements. In this work, the ICC(2,k) variant was applied which is considered appropriate when there are two or more raters, the raters are considered a random sample from a wider population (random professors of the department in our case) and the measurements represent the average score of all raters [66].

The ICC value ranges between 0 and 1 and is interpreted as follows (Koo & Li, 2016):

- < 0.50: low agreement,
- $0.50 - 0.75$: moderate,
- $0.75 - 0.90$: good,
- > 0.90: excellent reliability.

The use of ICC allows examining the extent to which AI replicates the logic of human evaluators in the overall scoring of soft skills.

II. Mean Absolute Error (MAE)

The MAE index was also calculated which measures the average numerical deviation between the AI score and the average of the human ratings. It is defined mathematically as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| AI_i - H_i \right|$$

Figure 4.23 – Calculation of the Mean Absolute Error (MAE)

where AI_i is the AI system score and H_i is the average of the human evaluators' scores for the same sample. Small MAE values indicate high consistency and low variance which enhances the statistical reliability and stability of the automated assessment [112].

The combination of these metrics allows for a comprehensive assessment of the reliability of the SoftSkills AI system as it controls both the correlation of scores (via ICC) and the true numerical deviation (via MAE).

Table 5. Statistical Methods and Reliability Metrics Used for System Evaluation.

Method/Index	Aim	Reference
Descriptive Statistics (Mean, SD, Median, Range)	Description of score distribution and detection of outliers	[109]
Paired-Samples t-Test	Test for difference in mean values PRE-POST in the same sample	[109], [110]
Cohen's d	Estimate of size and practical significance of change	[110], [111]
Intraclass Correlation Coefficient (ICC(2,k))	Measure of agreement between continuous AI-Human ratings	[66], [43]
Mean Absolute Error (MAE)	Calculate numerical deviation of AI-Human scores	[112]

4.7 Limitations and study reliability

Assessment of soft skills using AI is a field that combines elements of psychometrics, linguistics and machine learning which includes certain limitations. Although the study was designed to minimize sources of errors and ensure research validity, there are factors that might affect the accuracy and generalizability of the results. These limitations fall into four main categories:

Chapter 1: Model Bias (LLM Bias)

LLMs, such as the GPT-4.0 mini used here, are trained on huge and heterogeneous natural language datasets which inevitably include biases that reflect the social, cultural and linguistic specificities of their sources [113], [114]. This means that the model's performance can vary depending on:

- the style and linguistic complexity of the response,
- the cultural context or emotional tone of the text,
- the gender or social context implied in the wording.

For example, an LLM may positively evaluate a "decisive" response to a leadership question but underestimate responses that emphasize cooperation due to underlying associations it has been trained with. The use of GLMP and fuzzy rules in the backend aims to limit these effects by introducing hierarchical and qualitatively guided weighting in the evaluation criteria. In this way, interpretability and controlled adaptability are achieved by the system.

Chapter 2: Human rater subjectivity

Human raters may be influenced by personal attitudes, experiences or perceptions even in environments with clear evaluation criteria [115], [116]. Such rater variability can significantly affect the consistency of the measurements. For this reason, the experiment involved two independent raters working through the Rater UI while their agreement with the SoftSkills AI system was examined with appropriate metrics (ICC(2,k) and MAE). The use of these indicators ensured the measurement of consistency and the identification of potential deviations between human and automatic judgment allowing the GLMP to "adapt" to real pedagogical expectations.

Chapter 3: Sample size limitations

Sample size directly affects the statistical power and generalizability of the findings of a study [110]. Although the present study gathered a sufficient number of participants (N = 15), larger samples would allow for a more detailed investigation of performance by skill, controlling for potential gender, age or cognitive background differences and increased reliability in AI-Human comparisons.

At this stage, the goal was experimental confirmation of the methodological architecture and can be considered preliminary. Future studies may focus on expanding the sample and cross-validating the results in different educational settings.

Chapter 4: Response variability in the POST phase

During the POST phase, participants respond after a period of coaching and feedback reflection. However, variability in duration, level of participation or intrinsic motivation may affect the consistency of responses as has been previously reported in research on engagement and cognitive presence in distance learning environments [117]. A number of participants may invest more time or reflect more deeply which leads to improved responses, not necessarily due to learning but due to the different cognitive context.

4.8 Chapter summary

In this chapter, the methodologies utilized for the development, design and evaluation of the SoftSkills AI project were presented in detail. Initially, the research approach and the experimental protocol were described which was based on the PRE-POST measurement logic with the participation of students and young professionals who were assessed through a combination of multiple-choice and open-ended questions. Then, the architecture of the system which integrates LLMs, the GLMP model and fuzzy logic mechanisms to normalize the results and ensure interpretability, was analyzed in detail. This was followed by a description of the data collection and analysis procedures where statistical methods used in this thesis were presented. In addition, the limitations of the study and error compensation strategies were examined, such as the calibration process, the use of multiple raters and the use of interpretable AI mechanisms. The results of the data analysis, the statistical correlations between human and automatic evaluation and the findings regarding the improvement of the participants' performances between PRE and POST phases are presented in the next chapter.

Chapter 5: Discussion, conclusions and future work

5.1 Summary and main conclusions

Downstream analysis of the collected data was performed using Python (Python 3.11) in Visual Code Studio. The file exported from Rater UI (csv format) was used as input and included assessment records per participant, the time phase of measurement (PRE/POST) and the skill category. All necessary format conversions (separators, numerical formats) were applied to ensure correct operations. To enable valid pairing of before and after measurements, identification variables were standardized. The phase (PRE or POST) was defined based on the user identifier where POST phase records were identified by the suffix "POST". To keep the same participant as one entity across time points, a fixed participant-id was created by removing the suffix from the user ID. In this way, each participant is assigned a unique id and can be linked to the respective observations and allow the following within-subjects statistical tests.

To determine the appropriate level of analysis, we had to take into account the fact that there were more than one record per participant and skill suggesting that each skill is assessed through multiple items. To avoid the risk that the number of records per skill would give disproportionate weight to skills with more items, average values of the scores were calculated at each level (participant/skill/phase). This resulted in a skill-level score unique per individual and time phase which can be interpreted as a summary assessment of the performance on the corresponding skill in that phase.

The total number of participants was 15 and the total number of records was 120 (15 participants x 2 phases x 4 skills) which confirms that the dataset is balanced and can be used for paired statistical comparisons (PRE/POST coupling).

Pre-processing was followed by descriptive statistical analysis to get an overview of the data on the performance per skill and phase. Indices of central tendency (Mean, Median) and data dispersion (SD, Min-Max) were calculated for each category and time phase (Table 6).

Table 6. Descriptive statistics (Mean ± SD) of LLM-based scores (0–10) for PRE and POST assessment across four soft-skill categories (n=15 matched participants).

Skill	PRE Mean ± SD	POST Mean ± SD	Mean (POST–PRE)
Communication	7.432 ± 1.040	7.827 ± 0.621	+0.395
Teamwork	7.392 ± 1.014	7.810 ± 0.404	+0.418
Leadership	7.185 ± 0.967	7.708 ± 0.595	+0.523
Problem Solving	6.765 ± 1.283	7.147 ± 0.818	+0.382

The mean was defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Figure 5.1 – Mean (Sample Average)

where x_i is the skill-level score of the i-th participant and n is the number of participants (here $n = 15$).

The standard deviation is calculated as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Figure 5.2 – Sample Standard Deviation (SD)

and it expresses how dispersed the values are around the mean.

The results of Table 6 show an increase in the average scores from PRE to POST in all skills: Communication from $M = 7.432$ $(SD = 1.040)$ to $M = 7.872$ $(SD = 0.621)$, Teamwork from $M = 7.392$ $(SD = 1.014)$ to $M = 7.810$ $(SD = 0.404)$, Leadership from $M = 7.185$ $(SD = 0.967)$ to $M = 7.708$ $(SD = 0.595)$ and Problem Solving from $M = 6.765$ $(SD = 1.283)$ to $M = 7.147$ $(SD = 0.818)$. In addition to the increase in mean values, it is worth noting that in all categories a simultaneous decrease in dispersion was observed in the POST phase. The decrease in SD suggests that participants show more coherent or less dispersed performance in the second measurement. This can be interpreted as a tendency to converge around a higher score level. These differences are captured in Figure 5.3 which presents the average scores per skill for PRE and POST together with the corresponding standard deviations $(Mean \pm SD)$.

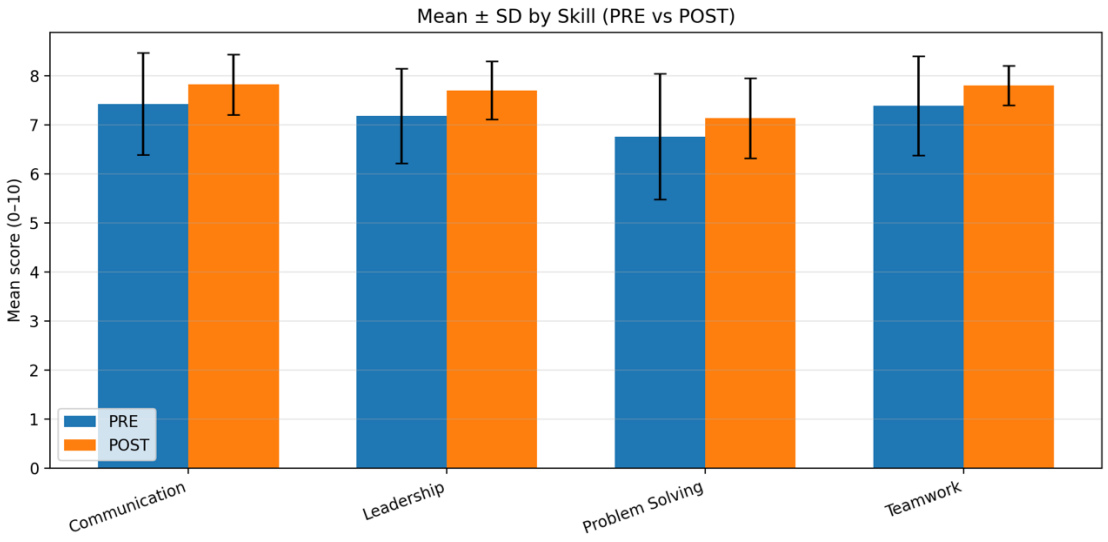

Figure 5.3 – Mean (± SD) LLM-based scores (0–10) for PRE and POST across the four soft-skill categories.

To assess whether the observed PRE-POST changes are statistically significant at the same level, inferential analysis was performed. Due to the fact that this is a within-subjects design (paired- the same participant was measured twice), the analysis was based on differences per participant because the groups are not independent. For each participant i and each skill, the difference, $d_i = \text{POST}_i - \text{PRE}_i$ was calculated, where PRE_i and POST_i are the skill-level scores of the same individual in the two time phases.

From the set of differences, the mean difference is obtained:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

Figure 5.4 – Mean of Paired Differences

as well as the standard deviation of the differences:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

Figure 5.5 – Standard Deviation of Paired Differences

Normality tests were applied to the distribution of the aforementioned differences d_i which is the main assumption for applying the paired t-test. To test for distribution normality, the Shapiro-Wilk test was used as a standard test for small sampling size. The results of the test are presented in Table 7. In all categories, the Shapiro-Wilk p-values were greater than $p > 0.05$ which means that there is no evidence of significant deviation from the normality of the differences and points towards a normal distribution of the values. Based on the above, a parametric paired t-test was then applied for each skill with a null hypothesis that the mean difference is zero ($\bar{d} = 0$).

The t-statistic is calculated as follows:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Figure 5.6 – Paired-Samples t-Statistic

Where the denominator of t is the standard error of the mean difference. The degrees of freedom are $df = n - 1$, here $df = 14$. The results of Table 7 show that the mean differences were positive in all skills (Communication $MD = 0.395$, Teamwork $MD = 0.418$, Leadership $MD = 0.523$, Problem Solving $MD = 0.382$). However, the corresponding p-values did not reach the statistical significance level at $\alpha = 0.05$ when the skills were examined individually.

Table 7. Paired-samples t-tests comparing PRE and POST LLM-based scores (0–10) by skill category. Effect sizes are reported as Cohen's d for paired samples, with 95% confidence intervals for the mean difference.

Skill	Mean Diff (MD)	t(14)	p-value	Cohen's d (paired)	95% CI of Diff
Communication	0.395	1.551	0.143	0.400	[-0.151, 0.941]
Teamwork	0.418	1.603	0.131	0.414	[-0.142, 0.978]
Leadership	0.523	1.749	0.102	0.452	[-0.118, 1.165]
Problem Solving	0.382	1.023	0.324	0.264	[-0.419, 1.182]

Cohen's d effect size for paired design was also calculated, which is defined as $d = \frac{\bar{d}}{s_d}$. In this formula, $\bar{d}$ expresses the average size of the change while s_d is the natural variability of the paired differences and d is the change in SD units. The range of Cohen's d values was from 0.246 to 0.452 (Table 7) indicating a small to moderate effect. The presence of positive differences between pre and post phase with a small to moderate effect sizes in all categories reflect a consistent direction of change. Nevertheless, it does not appear to be significant per skill level in this sample.

In addition, to capture the uncertainty of estimates, the 95% confidence intervals (CI) of the mean difference was calculated (Figure 5.7). The confidence interval describes the range within which the true mean difference in the population is expected to lie based on the sample data.

$$CI = \bar{d} \pm t_{0.975,df} \cdot \left(s_d / \sqrt{n} \right)$$

Figure 5.7 – Confidence Interval for the Mean Difference

Finally, to examine the overall change and whether small but consistent improvements in individual skills converge into more detectable change, an overall composite was calculated as the average of the four skills scores per phase for all participants (Table 8). This index functions as a summary index of overall performance irrespective of the skill category and reduces the random variability that might occur when skills are examined separately. Results in Table 8 show an increase in the composite from $PRE\ M = 7.193\ \ (SD = 0.707)$ to $POST\ M = 7.623\ \ (SD = 0.487)$, with a mean difference MD=0.430. The paired comparison for the overall index returned $t(14) = 2.145$ and $p = 0.04997$ while the effect size was $d = 0.553$ (moderate). The 95% CI for the mean difference was [0.000, 0.859] indicating a positive estimated change with a lower bound very close to zero, marginal but consistent shift.

Table 8. Overall composite (mean of four skill scores) PRE vs POST paired comparison (n=15).

Measure	PRE Mean ± SD	POST Mean ± SD	Mean Diff	t(14)	p-value	Cohen's d	95% CI
Overall Composite	7.193 ± 0.707	7.623 ±0.487	0.430	2.145	0.050	0.554	[0.000, 0.859]

The distribution of the overall composite is presented in Figure 5.8 with a boxplot to allow for observations regarding the shift of central tendency and dispersion.

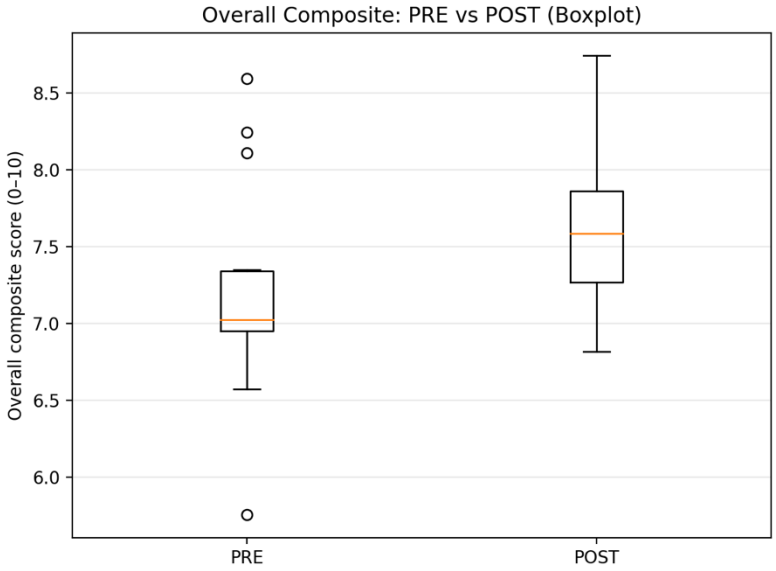

Figure 5.8 – Distribution of the overall composite score (mean across four categories) in PRE and POST (boxplot; n = 15 matched participants).

Interindividual variability and additional transparency at the participant level could be observed at Figure 5.9. The PRE-POST change for each individual separately are presented below, allowing for assessment of whether the overall change is widespread in the sample.

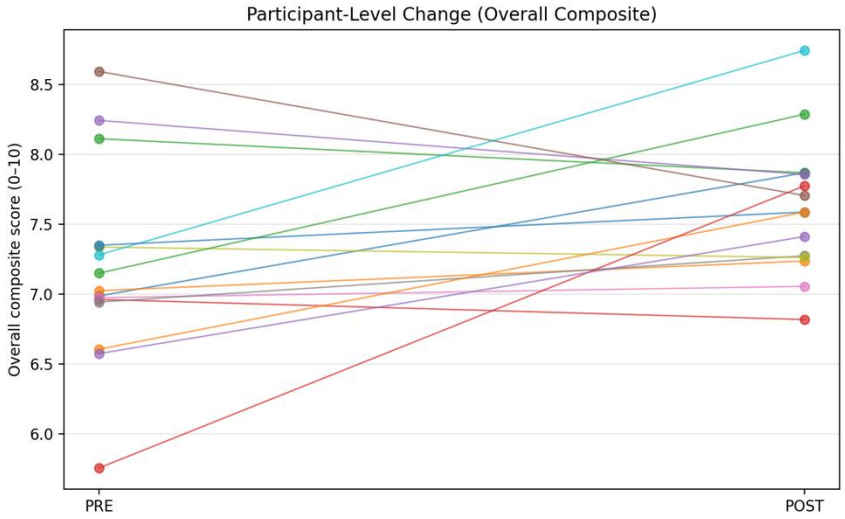

Figure 5.9 – Participant-level change in the overall composite score from PRE to POST (paired line plot; each line represents one participant).

5.2 Reliability and agreement analyses (human raters and model scoring)

Because the evaluation pipeline includes both human ratings and automated scoring, reliability and agreement analyses were conducted to clarify (a) how consistent the two human raters were with each other and (b) how closely the LLM scores aligned with the human reference. These checks are essential in assessment settings, since any interpretation of PRE–POST change depends not only on the magnitude of score differences but also on the stability and reproducibility of the scoring process itself. In practice, this section addresses two complementary questions: first, whether the human scoring can be considered sufficiently consistent to serve as a defensible benchmark; and second, whether the LLM behaves similarly to human evaluators when applied to the same set of responses.

Human–human reliability (Teacher 1 vs Teacher 2)

Inter-rater reliability between Teacher 1 and Teacher 2 was assessed using the Intraclass Correlation Coefficient (ICC). The ICC fits well in this context because it evaluates the agreement between raters on the same targets while considering both correlation and absolute differences in ratings. A two-way random-effects model with absolute agreement was used (ICC (2,*)), which is a common choice when raters are treated as a sample from a broader population of potential raters and when exact agreement in scores (rather than only consistency in rank ordering) is of interest. Two ICC variants were reported: ICC (2,1), which represents the reliability of a single rater, and ICC(2,k), which represents the reliability of the average score when combining the two raters (here k=2). Since the downstream analyses use the human average as a reference score, ICC(2,k) is particularly informative because it quantifies the expected stability of that averaged human benchmark.

To ensure that reliability was assessed at the same analytic level used for inferential testing, ICC was computed on aggregated "targets" rather than on raw item-level rows. Specifically, targets were defined at the participant × phase × skill level, yielding 30 targets per skill (15 participants × 2 phases), and an additional set of 30 targets for the overall composite across skills. The results (Table 9) indicate generally high agreement (common interpretive guidelines classify ICC values above ~ 0.75 as good and above ~ 0.90 as excellent) across skills and at the composite level. This supports the interpretation that the two human evaluators were scoring in a sufficiently consistent way and that their average can be treated as a stable reference point for subsequent comparisons with the LLM.

Table 9. Inter-rater reliability between Teacher 1 and Teacher 2 (ICC; aggregated level, 30 targets).

Scope	Targets	Raters	ICC(2,1)	95% CI low	95% CI high	ICC(2,k)
Communication	30	2	0.7869	0.5979	0.8928	0.8808
Leadership	30	2	0.6983	0.4571	0.8440	0.8224
Problem solving	30	2	0.8338	0.6791	0.9174	0.9093
Teamwork	30	2	0.8026	0.6268	0.9008	0.8905
Overall composite	30	2	0.7975	0.6169	0.8983	0.8874

LLM vs human benchmark (agreement and association)

After establishing the reliability of human scoring, agreement between the LLM and human evaluation was examined at the item level. For this analysis, LLM scores were compared against the human average score (HumanAvg) across all paired observations with non-missing values ($N = 480$). The purpose of this step was to quantify how closely the automated scores track human judgments and to detect any systematic deviations that could affect interpretation.

Agreement was evaluated using complementary metrics that capture different aspects of alignment. Mean Absolute Error (MAE) summarizes the typical absolute deviation between LLM and HumanAvg scores on the 0–10 scale, whereas Root Mean Squared Error (RMSE) shows larger discrepancies more strongly, making it sensitive to occasional large mismatches. In addition, the signed mean error (bias; LLM − HumanAvg) was computed to indicate whether the model tends to systematically over-score or under-score relative to the human benchmark.

Alongside deviation metrics, association measures were used to evaluate whether the LLM preserves the human ordering of responses. Pearson's correlation coefficient (r) was reported as an index of linear association on the original scale, while Spearman's rank correlation (ρ) was additionally included as a robustness check. Spearman's ρ is computed as the Pearson correlation coefficient on ranked values ($\rho = corr(rank(X), rank(Y))$) and is less sensitive to deviations from linearity, outliers, and tied scores that may occur on a discrete rating scale such as 0–10. Corresponding p-values were computed under the null hypothesis of zero association ($H0: r = 0 \; or \; \rho = 0$), as implemented in standard statistical libraries.

As summarized in Table 10, the LLM exhibited very strong alignment with the human benchmark. Absolute deviations were small relative to the scale ($MAE = 0.3563$; $RMSE = 0.6730$), while both Pearson and Spearman correlations were extremely high ($r = 0.9725$; $\rho = 0.9788$), indicating that the model closely follows human judgments and preserves the relative ranking of responses. At the same time, the negative mean error (-0.2417) indicates a modest systematic tendency for the LLM to assign slightly lower scores than the human average. This bias is practically relevant because it suggests that, even under high agreement, calibration or human oversight may be beneficial when LLM outputs are integrated into evaluative workflows.

Table 10. LLM scoring agreement with human evaluation (item-level, N=480).

N (items)	MAE	RMSE	Mean Error	Pearson r	Pearson p	Spearman ρ	Spearman p
480	0.3563	0.6730	-0.2417	0.9725	1.020e-304	0.9788	0.000e+00

Finally, the scatter plot in Figure 5.10 provides an intuitive visualization of the relationship between LLM and HumanAvg. Each point represents one scored response. The dashed line indicates perfect agreement ($y = x$), while the fitted regression line summarizes the relationship in line with the high item-level correlations reported in Table 10.

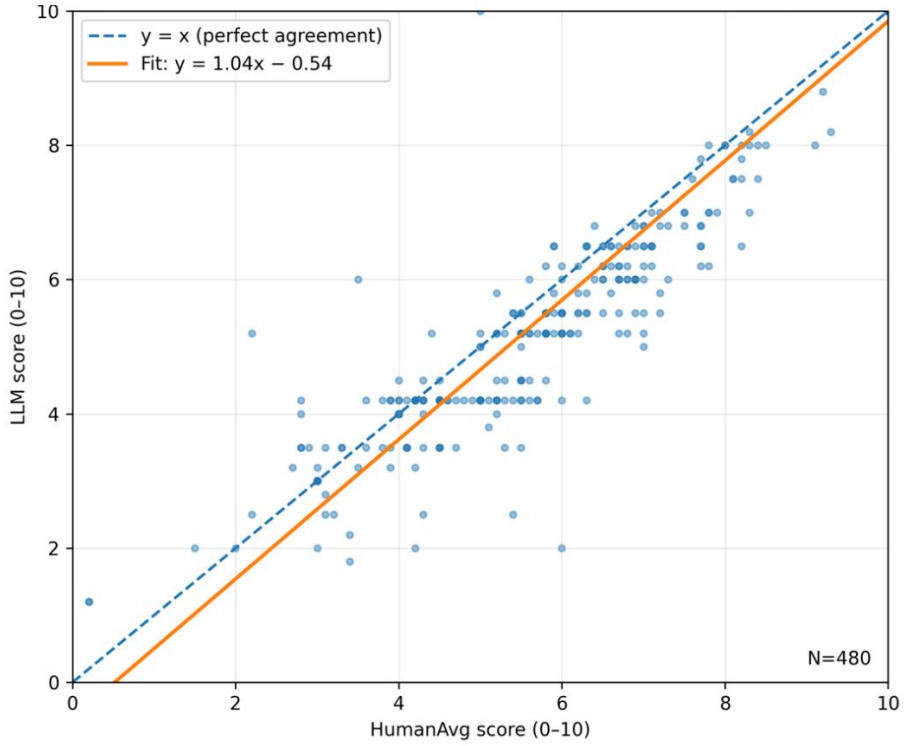

Figure 5.10 – Item-level alignment between LLM and HumanAvg scores (scatter plot; N=480).

5.3 Extended composite analysis across scoring perspectives

Beyond the LLM-only composite reported in Table 8, the overall PRE–POST change was also examined under two additional scoring perspectives: (a) the HumanAvg composite (mean of aggregated human average scores across the four skills), and (b) the FinalScore composite (as defined in the scoring workflow). These composites were computed using the same aggregation logic (participant × phase × skill means, followed by averaging across skills per phase). Paired PRE–POST tests were then conducted at the participant level ($n = 15$) using identical inferential procedures. Results are reported in Table 11, where all three composite versions (LLM, HumanAvg, FinalScore) demonstrate positive mean changes from PRE to POST, with statistically significant paired effects for the human-based composites and moderate effect sizes.

Table 11. Overall composite PRE–POST paired tests across scoring versions (n=15).

Measure	PRE Mean	POST Mean	Mean Diff	Shapiro p	t(14)	p-value	Cohen's d	95% CI
LLM composite	7.1933	7.6229	0.4296	0.9229	2.1451	0.04997	0.5539	[0.0001, 0.8591]
HumanAvg composite	7.3509	7.9529	0.6020	0.9339	2.6649	0.01848	0.6881	[0.1175, 1.0865]
FinalScore composite	7.2930	7.8233	0.5303	0.9523	2.4998	0.02547	0.6455	[0.0753, 0.9853]

5.4 Conclusions

Overall, the results presented in Chapter 5 indicate a consistent PRE–POST upward shift in soft-skill scores, with the direction of change being stable across skills and across scoring perspectives. At the descriptive level, mean values increased from PRE to POST in all four categories (Communication, Teamwork, Leadership, and Problem Solving), while dispersion tended to decrease in the POST phase. This combined pattern suggests both higher average performance and a more concentrated distribution of scores after the second measurement, meaning that participants' outcomes converged around higher levels. When skills were examined separately using paired comparisons, mean differences remained positive in all categories and effect sizes fell in the small-to-moderate range. Although these per-skill tests did not consistently reach the conventional significance threshold, likely due to the rather small sample size, the uniform direction of change across categories supports the interpretation of a coherent improvement tendency rather than isolated, random fluctuations within single skills.

A clearer overall signal emerged when performance was summarized through the composite index (mean across the four skill categories). At this level, the PRE–POST increase became more detectable, which is consistent with the idea that modest improvements distributed across multiple skills can accumulate into a more robust overall shift once random skill-level variability is reduced. Importantly, this conclusion was not dependent on a single scoring source: when the overall composite was computed using the LLM-based scores, the human average scores, and the final combined score, all three versions showed a positive PRE–POST change with comparable magnitude. This convergence across scoring perspectives strengthens the internal consistency of the findings and supports the interpretation that the observed change reflects an underlying pattern in the dataset rather than an artifact of one specific scoring variant.

Beyond the magnitude and direction of PRE–POST change, Chapter 5 also provides evidence about the reliability and interpretability of the scoring process itself. Inter-rater reliability between the two human evaluators was high (ICC), indicating that human scoring was sufficiently consistent to serve as a stable reference point. Against this benchmark, item-level agreement between LLM scores and the human average was very strong, with small absolute error on the 0–10 scale and extremely high correlations. At the same time, the signed mean error suggested a modest systematic tendency for the LLM to score slightly lower than the human average, a finding that is practically important because it highlights the value of calibration and human oversight when LLM outputs are used in evaluative workflows. Taken together, the chapter supports two complementary conclusions: first, that the data show an overall PRE–POST improvement tendency most clearly captured at the composite level; and second, that the measurement pipeline is methodologically robust, combining strong human–human agreement with close LLM–human alignment, thereby providing a defensible basis for subsequent interpretation and reporting of the final scoring outcomes.

5.5 Suggestions for future research and improvements

Future work can extend the present study along three complementary directions: (i) methodological strengthening of the research design, (ii) technical refinement and calibration of the scoring pipeline, and (iii) expansion of data modalities and responsible-use governance. First, larger and more diverse samples (across institutions, study levels, and learner profiles) would increase statistical power and enable more granular analyses, such as skill-specific effects, subgroup comparisons, and robustness checks across educational contexts. A longitudinal design with additional measurement points (beyond a single PRE–POST comparison) would further help distinguish short-term fluctuations from more

stable changes that reflect sustained skill development. In parallel, future protocols could incorporate additional human raters and structured rater training in order to further stabilize the human reference standard and examine whether high agreement is maintained under broader rater variability.

Second, at the level of scoring and interpretation, it would be useful to consider adopting an integer-only rating scale (e.g., 0–10 using whole numbers only) rather than decimal scores. In soft-skill assessment, decimal precision may imply a level of measurement accuracy that is difficult to justify consistently, given that these constructs are inferred indirectly and are strongly context-dependent. An integer scale may reduce the appearance of over-precision, support more consistent rubric application across raters, and improve interpretability at the reporting level. At the same time, further improvements can focus on rubric engineering and calibration, including clearer operational definitions, sharper boundaries between adjacent performance levels, and systematic evaluation of scoring stability under controlled variations in prompts or ambiguous responses. Moreover, because LLM-based scoring may exhibit variability, future work should explicitly evaluate LLM stability under small prompt modifications and/or across different execution times (prompt/time stability). Such analyses would document robustness under realistic conditions and inform the degree of standardization and oversight required for dependable use.

Third, an important area for future extension concerns the adoption of multimodal data. In later research stages, incorporating video-based and/or audio-based responses would enable the use of additional indicators that are often critical for soft skills, like facial expressions, gestures, patterns of movement, pacing, pausing, and vocal tone. Such signals could increase ecological validity, as communication, collaboration, and leadership are frequently expressed through non-verbal and paralinguistic cues rather than through written content alone. Additionally, future research that would enhance the criterion validity of the produced scores could look at the relationship between the produced scores and other independent external criteria, such as instructor evaluations of authentic assignments or projects, peer assessment, performance in structured simulations, or validated soft-skill questionnaires. Establishing such links would support the argument that the scores are not only internally coherent but also meaningfully interpretable in real educational and professional settings. Finally, any expansion to richer data modalities and more powerful AI components should be accompanied by stronger consent procedures, data minimization, privacy safeguards, and continuous bias monitoring, given the increased sensitivity associated with behavioral and biometric indicators.

References

[1] M. L. Matteson, L. Anderson, and C. Boyden, "' Soft skills': A phrase in search of meaning," *Portal Libr. Acad.*, vol. 16, no. 1, pp. 71–88, 2016.

[2] M. Cinque and S. Kippels, "Soft Skills in Education: The role of the curriculum, teachers, and assessments," *Reg. Cent. Educ. Plan. Res. Pap. UNESCO*, 2023, Accessed: Jan. 25, 2026. [Online]. Available: https://rcepunesco.ae/en/KnowledgeCorner/ReportsandStudies/ReportsandStudies/05_Soft_Skills_in_Education_RP_EN.pdf

[3] B. Schulz, "The importance of soft skills: Education beyond academic knowledge.," 2008, Accessed: Jan. 25, 2026. [Online]. Available: https://ir.nust.na/bitstream/10628/39/1/The%20Importance%20of%20Soft%20%20Skills-Education%20beyond%20academic%20knowledge.pdf

[4] C. G. P. Berdanier, "A hard stop to the term 'soft skills,'" *J. Eng. Educ.*, vol. 111, no. 1, pp. 14–18, Jan. 2022, doi: 10.1002/jee.20442.

[5] S. Malinen, M. Galster, and A. Mitrovic, "Industry report: Soft skills assessment framework (2SAF)," 2025, Accessed: Jan. 25, 2026. [Online]. Available: https://www.canterbury.ac.nz/content/dam/uoc-main-site/documents/pdfs/research/Soft%20skills%20assessment%20framework_2SAF.pdf

[6] M. Yamaguchi, A. Lobo, and J. Richardson, "Teacher soft skills: Building a student-centred conceptual framework".

[7] Z. S. Bisschoff and L. Massyn, "A conceptual soft skills competency framework for enhancing graduate intern employability," *High. Educ. Ski. Work-Based Learn.*, vol. 15, no. 7, pp. 66–81, Dec. 2025, doi: 10.1108/HESWBL-08-2023-0239.

[8] B. Cimatti, "Definition, development, assessment of soft skills and their role for the quality of organizations and enterprises," *Int. J. Qual. Res.*, vol. 10, no. 1, p. 97, 2016.

[9] A. P. N. De Freitas and R. A. Almendra, "Soft skills in design education, identification, classification, and relations: Proposal of a conceptual map," *Des. Technol. Educ. Int. J.*, vol. 26, no. 3–2, pp. 245–260, 2021.

[10] C. Dede, "Immersive Interfaces for Engagement and Learning," *Science*, vol. 323, no. 5910, pp. 66–69, Jan. 2009, doi: 10.1126/science.1167311.

[11] J. Lamri and T. Lubart, "Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach," *J. Intell.*, vol. 11, no. 6, p. 107, Jun. 2023, doi: 10.3390/jintelligence11060107.

[12] G. Dogara, M. S. B. Saud, Y. B. Kamin, and M. S. B. Nordin, "Project-based learning conceptual framework for integrating soft skills among students of technical colleges," *Ieee Access*, vol. 8, pp. 83718–83727, 2020.

[13] F. Munir, "More than technical experts: Engineering professionals' perspectives on the role of soft skills in their practice," *Ind. High. Educ.*, vol. 36, no. 3, pp. 294–305, Jun. 2022, doi: 10.1177/09504222211034725.

[14] B. Colman and P. Willmot, "How soft are 'soft skills' in the engineering profession?," 2016, Accessed: Jan. 25, 2026. [Online]. Available: https://repository.lboro.ac.uk/articles/How_soft_are_Soft_Skills_in_the_engineering_profession_/9558893/files/17190839.pdf

[15] I. Holik, I. D. Sanda, and G. Molnár, "The necessity of developing soft skills in STEM areas in higher education, with special focus on engineering training," *Athens J. Technol. Eng.*, vol. 10, no. 4, pp. 199–214, 2023.

[16] F. Ahmed, L. Fernando Capretz, S. Bouktif, and P. Campbell, "Soft skills requirements in software development jobs: A cross-cultural empirical study," *J. Syst. Inf. Technol.*, vol. 14, no. 1, pp. 58–81, 2012.

[17] D. González-Morales, L. M. M. De Antonio, and J. L. R. García, "Teaching 'Soft' skills in software engineering," in *2011 ieee global engineering education conference (educon)*, IEEE, 2011, pp. 630–637. Accessed: Jan. 25, 2026. [Online]. Available:

https://ieeexplore.ieee.org/abstract/document/5773204/?casa_token=hLe_Y-eqUbMAAAAA:u2yb4m0yp3OvQZK_-fptrVbrYnlsStOkMi3OoLDW7gJklabeOBJPOuHZOaghErUTC2jlgLl7ntI

[18] B. Aseel, "Professionalism for engineers: Soft Skills in engineering education to prepare for professional life," 2018, Accessed: Jan. 25, 2026. [Online]. Available: https://www.cdio.org/sites/default/files/documents/127_Final_PDF.pdf

[19] D. Jelonek and T. Nitkiewicz, "Soft skills of engineers in view of industry 4.0 challenges," *Qual. Prod. Improv.-QPI*, vol. 2, no. 1, pp. 107–116, 2020.

[20] R. F. D. Laguna, D. L. Aguinaga, D. E. C. Torres, and B. A. L. Cueto, "Soft Skills and the Use of Industry 4.0 as Determinants of Professional Development in Engineering Graduates: A SEM Approach," *Sustain. Futur.*, p. 100742, 2025.

[21] D. B. de Campos, L. M. M. de Resende, and A. B. Fagundes, "The importance of soft skills for the engineering," *Creat. Educ.*, vol. 11, no. 8, pp. 1504–1520, 2020.

[22] D. Konings and M. Legg, "Delivering an effective balance of soft and technical skills within project-based engineering courses," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, 2020, pp. 157–164. Accessed: Jan. 25, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9368493/?casa_token=WUMf3clJ24UAAAAA:Gx-1ty0SlsjeWtr0VEYSUHx_L7-8Ja0jYO_n2Nd3Ds1pN0Pv2mMTc_5wMd1OCpRAd9wE7Uefm9I

[23] J. P. Queiroz-Neto, M. A. Rodrigues, M. S. Pereira, and N. A. Gouvea, "Integrating Project-Based Learning and Design Thinking: An Innovative Approach to Enhancing Hard and Soft Skills in Industrial Robotics Education," in *2024 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2024, pp. 1–8. Accessed: Jan. 26, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10893483/?casa_token=OyTU7Am_SCoAAAAA:qaVAzVHOaiZ2ngAeQmA0LFNifpVQYwXwkiVt2tG28j1ww1LEAV5iolTuoG1U9kvirjHuF-WUe90

[24] I. G. Gerasimova and I. S. Oblova, "Development of Soft Skills with the Purpose of Enhancing Employability of Engineering Students," 2025, Accessed: Jan. 26, 2026. [Online]. Available: https://www.preprints.org/frontend/manuscript/c26aac785c4483fba19e28534ae0bfc0/download_pub

[25] M. Danaher, "Assessment of 21st Century skills in a computing programme during the Covid-19 pandemic," *Glob. J. Eng. Educ.*, vol. 24, no. 3, 2022, Accessed: Jan. 26, 2026. [Online]. Available: http://www.wiete.com.au/journals/GJEE/Publish/vol24no3/03-Danaher-M.pdf

[26] M. Caeiro-Rodriguez *et al.*, "Teaching Soft Skills in Engineering Education: An European Perspective," *IEEE Access*, vol. 9, pp. 29222–29242, 2021, doi: 10.1109/ACCESS.2021.3059516.

[27] U. R. Cukierman and J. M. Palmieri, "Soft skills in engineering education: A practical experience in an undergraduate course," in *2014 International Conference on Interactive Collaborative Learning (ICL)*, IEEE, 2014, pp. 237–242. Accessed: Jan. 26, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7017776/?casa_token=gg5GJ3pzWRkAAAAA:Fe1mf4xoEoUgowEnkc2VyhCWEZalSb8vOxv-5doqJUhy9kQyc7hk6_AiDXic9WmBi4u_ZB4V6mI

[28] M. L. Fioravanti, B. Sena, and E. F. Barbosa, "Assessing the development of soft skills for project management using PBL: a case study," in *2020 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2020, pp. 1–8. Accessed: Jan. 26, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9274099/?casa_token=k-ilzA-t8ZkAAAAA:rnZLtAIBM-qQ-aiIQG73b899Sg8pV3ZVGXMaWN98NGBZpMAEp-Yn6jROVRy0GnbKbGaWI7zwRwQ

[29] V. Caggiano, T. Redomero-Echeverría, J. L. Poza-Lujan, and A. Bellezza, "Soft skills in engineers, a relevant field of research: Exploring and assessing skills in Italian engineering students," *Ing. E Investig.*, vol. 40, no. 2, pp. 81–91, 2020.

[30] G. Zheng, C. Zhang, and L. Li, "Practicing and Evaluating Soft Skills in IT Capstone Projects," in *Proceedings of the 16th Annual Conference on Information Technology Education*, Chicago Illinois USA: ACM, Sep. 2015, pp. 109–113. doi: 10.1145/2808006.2808041.

[31] B. Williamson and R. Eynon, “Historical threads, missing links, and future directions in AI in education,” *Learn. Media Technol.*, vol. 45, no. 3, pp. 223–235, Jul. 2020, doi: 10.1080/17439884.2020.1798995.
[32] L. Chen, P. Chen, and Z. Lin, “Artificial Intelligence in Education: A Review,” *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
[33] I. Roll and R. Wylie, “Evolution and Revolution in Artificial Intelligence in Education,” *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, Jun. 2016, doi: 10.1007/s40593-016-0110-3.
[34] R. Luckin and W. Holmes, “Intelligence unleashed: An argument for AI in education,” 2016, Accessed: Jan. 26, 2026. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/1475756/
[35] K. R. Koedinger, E. Brunskill, R. S. J. D. Baker, E. A. McLaughlin, and J. Stamper, “New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization,” *AI Mag.*, vol. 34, no. 3, pp. 27–41, Sep. 2013, doi: 10.1609/aimag.v34i3.2484.
[36] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – where are the educators?,” *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 39, Dec. 2019, doi: 10.1186/s41239-019-0171-0.
[37] S. Rahimi and V. J. Shute, “Stealth assessment: a theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments,” *Educ. Technol. Res. Dev.*, vol. 72, no. 5, pp. 2417–2441, Oct. 2024, doi: 10.1007/s11423-023-10232-1.
[38] S. Akgun and C. Greenhow, “Artificial intelligence in education: Addressing ethical challenges in K-12 settings,” *AI Ethics*, vol. 2, no. 3, pp. 431–440, Aug. 2022, doi: 10.1007/s43681-021-00096-7.
[39] B. Williamson and R. Eynon, “Historical threads, missing links, and future directions in AI in education,” *Learn. Media Technol.*, vol. 45, no. 3, pp. 223–235, Jul. 2020, doi: 10.1080/17439884.2020.1798995.
[40] M. D. Shermis and J. Burstein, Eds., *Handbook on automated essay evaluation: current applications and new directions*. New York: Routledge, 2013.
[41] Y. Attali and J. Burstein, “AUTOMATED ESSAY SCORING WITH E-RATER® V.2.0,” *ETS Res. Rep. Ser.*, vol. 2004, no. 2, Dec. 2004, doi: 10.1002/j.2333-8504.2004.tb01972.x.
[42] Y. Ren, W. Fan, and J. Wang, “Intelligent text analysis for effective evaluation of english Language teaching based on deep learning,” *Sci. Rep.*, vol. 15, no. 1, p. 28949, Aug. 2025, doi: 10.1038/s41598-025-14320-5.
[43] K. O. McGraw, “Forming Inferences About Some Intraclass Correlation Coefficients”.
[44] V. J. Shute, “Focus on Formative Feedback,” *Rev. Educ. Res.*, vol. 78, no. 1, pp. 153–189, Mar. 2008, doi: 10.3102/0034654307313795.
[45] A. Gandolfi, “GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions,” *Int. J. Artif. Intell. Educ.*, vol. 35, no. 1, pp. 367–397, Mar. 2025, doi: 10.1007/s40593-024-00403-3.
[46] X. Li *et al.*, “A Systematic Study of Code Obfuscation Against LLM-based Vulnerability Detection,” Dec. 18, 2025, *arXiv*: arXiv:2512.16538. doi: 10.48550/arXiv.2512.16538.
[47] E. Kasneci, K. Sessler, F. Fischer, U. Gasser, and G. Groh, “ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education”.
[48] M. Jukiewicz and M. Wyrwa, “Can ChatGPT Replace the Teacher in Assessment? A Review of Research on the Use of Large Language Models in Grading and Providing Feedback,” Sep. 15, 2025, *Social Sciences*. doi: 10.20944/preprints202509.1233.v1.
[49] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability,” *Comput. Educ. Artif. Intell.*, vol. 6, p. 100234, Jun. 2024, doi: 10.1016/j.caeai.2024.100234.
[50] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer, “Is GPT-4 a reliable rater? Evaluating consistency in GPT-4’s text ratings,” *Front. Educ.*, vol. 8, p. 1272229, Dec. 2023, doi: 10.3389/feduc.2023.1272229.
[51] A. Mizumoto and M. Eguchi, “Exploring the potential of using an AI language model for automated essay scoring,” *Res. Methods Appl. Linguist.*, vol. 2, no. 2, p. 100050, Aug. 2023, doi: 10.1016/j.rmal.2023.100050.

[52] T. Wolfram, F. C. Tropf, and C. Rahal, “Large Language Models Predict Cognition and Education Close to or Better than Genomics or Expert Assessment,” May 03, 2022, *SocArXiv*. doi: 10.31235/osf.io/a8ht9.
[53] G. Kortemeyer, “Could an artificial-intelligence agent pass an introductory physics course?,” *Phys. Rev. Phys. Educ. Res.*, vol. 19, no. 1, p. 010132, May 2023, doi: 10.1103/PhysRevPhysEducRes.19.010132.
[54] J. Kim and B. N. Vajravelu, “Assessing the Current Limitations of Large Language Models in Advancing Health Care Education,” *JMIR Form. Res.*, vol. 9, pp. e51319–e51319, Jan. 2025, doi: 10.2196/51319.
[55] O. Alkhatib, M. Siddique, M. Qureshi, S. Elsetouhy, and F. Alessi Longa, “Explainable Ai For Automated Assessment: Improving Transparency And Trust In Computer-Assisted Grading Systems,” *Annu. Methodol. Arch. Res. Rev.*, Nov. 2025, doi: 10.63075/qp2e4m49.
[56] K. Swathi, R. Yuvatez, and S. Kovid, “Enhancing Trust in AI-Powered Learning: The Role of Explainable AI in the Education Field”.
[57] European Education and Culture Executive Agency, Ed., *Explainable AI in education: fostering human oversight and shared responsibility: by the European Digital Education Hub’s Squad on artificial intelligence in education*. Luxembourg: Publications Office, 2025. doi: 10.2797/7156884.
[58] G. Türkmen, “The Review of Studies on Explainable Artificial Intelligence in Educational Research,” *J. Educ. Comput. Res.*, vol. 63, no. 2, pp. 277–310, Apr. 2025, doi: 10.1177/07356331241310915.
[59] Z. M. Altukhi and D. S. Pradhan, “Systematic Literature Review: Explainable AI Definitions and Challenges in Education,” 2024.
[60] D. Hooshyar and Y. Yang, “Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction,” *IEEE Access*, vol. 12, pp. 137472–137490, 2024, doi: 10.1109/ACCESS.2024.3463948.
[61] F. S. Mohammed and F. Ozdamli, “A Systematic Literature Review of Soft Skills in Information Technology Education,” *Behav. Sci.*, vol. 14, no. 10, p. 894, Oct. 2024, doi: 10.3390/bs14100894.
[62] O. L. Liu, H. Kell, K. Williams, G. Ling, and M. Sanders, “ETS Skills Taxonomy,” *ChineseEnglish J. Educ. Meas. Eval.*, vol. 4, no. 4, Dec. 2023, doi: 10.59863/NMIE9603.
[63] A. Mukashova *et al.*, “AI-driven framework for automated competency formalization: from professional standards to adaptive learning outcomes,” *Front. Comput. Sci.*, vol. 7, p. 1710358, Dec. 2025, doi: 10.3389/fcomp.2025.1710358.
[64] N. Kerimbayev, K. Adamova, R. Shadiev, and Z. Altinay, “Intelligent educational technologies in individual learning: a systematic literature review,” *Smart Learn. Environ.*, vol. 12, no. 1, p. 1, Jan. 2025, doi: 10.1186/s40561-024-00360-3.
[65] I. Cheong, Y. E. Huh, and S. Puntoni, “Lay beliefs about AI assessment of interpersonal skills in personnel selection,” *Sci. Rep.*, vol. 15, no. 1, p. 25317, Jul. 2025, doi: 10.1038/s41598-025-10358-7.
[66] T. K. Koo and M. Y. Li, “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.
[67] I. Molenaar, “Towards hybrid HUMAN-AI learning technologies,” *Eur. J. Educ.*, vol. 57, no. 4, pp. 632–645, Dec. 2022, doi: 10.1111/ejed.12527.
[68] M. Cukurova, “The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence,” *Br. J. Educ. Technol.*, vol. 56, no. 2, pp. 469–488, Mar. 2025, doi: 10.1111/bjet.13514.
[69] C. DeLuca, L. Volante, and M. Holden, “The AI3 Model: Future Directions for Artificial Intelligence, Assessment Innovation, and Academic Integrity”.
[70] X. Kong, H. Fang, W. Chen, J. Xiao, and M. Zhang, “Examining human–AI collaboration in hybrid intelligence learning environments: insight from the Synergy Degree Model,” *Humanit. Soc. Sci. Commun.*, vol. 12, no. 1, p. 821, Jun. 2025, doi: 10.1057/s41599-025-05097-z.
[71] S. Järvelä, A. Nguyen, and A. Hadwin, “Human and artificial intelligence collaboration for socially shared regulation in learning,” *Br. J. Educ. Technol.*, vol. 54, no. 5, pp. 1057–1076, Sep. 2023, doi: 10.1111/bjet.13325.

[72] D. Hooshyar *et al.*, “Towards responsible AI for education: Hybrid human-AI to confront the Elephant in the room”.
[73] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, “Are We There Yet? - A Systematic Literature Review on Chatbots in Education,” *Front. Artif. Intell.*, vol. 4, p. 654924, Jul. 2021, doi: 10.3389/frai.2021.654924.
[74] C. W. Okonkwo and A. Ade-Ibijola, “Chatbots applications in education: A systematic review,” *Comput. Educ. Artif. Intell.*, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.
[75] D. S. M. Pereira, F. Falcão, L. Costa, B. S. Lunn, J. M. Pêgo, and P. Costa, “Here’s to the future: Conversational agents in higher education- a scoping review,” *Int. J. Educ. Res.*, vol. 122, p. 102233, 2023, doi: 10.1016/j.ijer.2023.102233.
[76] A. S. Alrajhi, “Artificial intelligence pedagogical chatbots as L2 conversational agents,” *Cogent Educ.*, vol. 11, no. 1, p. 2327789, Dec. 2024, doi: 10.1080/2331186X.2024.2327789.
[77] M. Ciaschi and M. Barone, “Exploring the role of Artificial Intelligence in assessing soft skills,” presented at the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, Oct. 2024, pp. 573–578. doi: 10.15439/2024F2063.
[78] Z. Yin, S. V. Chinta, Z. Wang, M. Gonzalez, and W. Zhang, “FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications,” Nov. 02, 2025, *arXiv*: arXiv:2407.18745. doi: 10.48550/arXiv.2407.18745.
[79] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, “The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment,” *J. Univ. Teach. Learn. Pract.*, vol. 21, no. 06, Apr. 2024, doi: 10.53761/q3azde36.
[80] W. Holmes *et al.*, “Ethics of AI in Education: Towards a Community-Wide Framework,” *Int. J. Artif. Intell. Educ.*, vol. 32, no. 3, pp. 504–526, Sep. 2022, doi: 10.1007/s40593-021-00239-1.
[81] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas, “A Multimodal Framework for Explainable Evaluation of Soft Skills in Educational Environments,” May 03, 2025, *arXiv*: arXiv:2505.01794. doi: 10.48550/arXiv.2505.01794.
[82] “AI scoring for international large-scale assessments using a deep learning model and multilingual data,” OECD Education Working Papers 287, Feb. 2023. doi: 10.1787/9918e1fb-en.
[83] A. Marengo, A. Pagano, B. Lund, and V. Santamato, “Research AI: integrating AI and gamification in higher education for e-learning optimization and soft skills assessment through a cross-study synthesis,” *Front. Comput. Sci.*, vol. 7, p. 1587040, Sep. 2025, doi: 10.3389/fcomp.2025.1587040.
[84] “Digital Progress and Trends Report 2025: Strengthening AI Foundations,” 2025.
[85] C. Succi and M. Canovi, “Soft skills to enhance graduate employability: comparing students and employers’ perceptions,” *Stud. High. Educ.*, vol. 45, no. 9, pp. 1834–1847, Sep. 2020, doi: 10.1080/03075079.2019.1585420.
[86] A. Bobbio and A. M. Manganelli, “LEADERSHIP SELF-EFFICACY SCALE. A NEW MULTIDIMENSIONAL INSTRUMENT,” vol. 16, no. 1, 2009.
[87] S. I. M. Bouland-van Dam, J. K. Oostrom, and P. G. W. Jansen, “Development and validation of the leadership learning agility scale,” *Front. Psychol.*, vol. 13, p. 991299, Dec. 2022, doi: 10.3389/fpsyg.2022.991299.
[88] E. Britton, N. Simper, A. Leger, and J. Stephenson, “Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills,” *Assess. Eval. High. Educ.*, vol. 42, no. 3, pp. 378–397, Apr. 2017, doi: 10.1080/02602938.2015.1116497.
[89] M. Hebles, C. Yániz-Álvarez -de-Eulate, and M. Alonso-Dos-Santos, “Teamwork competency scale (TCS) from the individual perspective in university students,” *J. Technol. Sci. Educ.*, vol. 12, no. 2, p. 510, Jul. 2022, doi: 10.3926/jotse.1478.
[90] A. R. Phillips and C. Lambie, “Assessing Civil Engineering Students Perceptions of Their Problem Solving Ability”.
[91] P. P. Heppner and C. H. Petersen, “Problem-Solving Inventory.” Sep. 12, 2011. doi: 10.1037/t04336-000.
[92] N. Kourmousi, E. Amanaki, C. Tzavara, and V. Koutras, “Active Listening Attitude Scale (ALAS): Reliability and Validity in a Nationwide Sample of Greek Educators,” *Soc. Sci.*, vol. 6, no. 1, p. 28, Mar. 2017, doi: 10.3390/socsci6010028.

[93] Y.-C. Huang and S.-H. Lin, “An inventory for assessing interpersonal communication competence of college students,” *Br. J. Guid. Couns.*, vol. 46, no. 4, pp. 385–401, Jul. 2018, doi: 10.1080/03069885.2016.1237614.
[94] OECD, *Innovating Assessments to Measure and Support Complex Skills*. OECD Publishing, 2023. doi: 10.1787/e5f3e341-en.
[95] A. V. Sánchez and M. P. Ruiz, “Competence-based learning - A proposal for the assessment of generic competences”.
[96] M. Forehand, “From Emerging Perspectives on Learning, Teaching and Technology”.
[97] D. R. Krathwohl, “A Revision of Bloom’s Taxonomy: An Overview,” *Theory Pract.*, vol. 41, no. 4, pp. 212–218, Nov. 2002, doi: 10.1207/s15430421tip4104_2.
[98] B. R. Belland, “Instructional Scaffolding: Foundations and Evolving Definition,” in *Instructional Scaffolding in STEM Education*, Cham: Springer International Publishing, 2017, pp. 17–53. doi: 10.1007/978-3-319-02565-0_2.
[99] P. Subban, “Differentiated instruction: A research basis”.
[100] J. Atkinson and D. Palma, “An LLM-based hybrid approach for enhanced automated essay scoring,” *Sci. Rep.*, vol. 15, no. 1, p. 14551, Apr. 2025, doi: 10.1038/s41598-025-87862-3.
[101] H. Chen and B. He, “Automated Essay Scoring by Maximizing Human-Machine Agreement,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 1741–1752. doi: 10.18653/v1/D13-1180.
[102] C. Gupta and V. Gupta, “Hybrid Human–Machine Consensus Framework for SME Technology Selection: Integrating Machine Learning and Planning Poker,” *Systems*, vol. 14, no. 1, p. 42, Dec. 2025, doi: 10.3390/systems14010042.
[103] “GPT-4o mini Model | OpenAI API.” Accessed: Jan. 27, 2026. [Online]. Available: https://platform.openai.com
[104] Y. Liu, H. Qi, and X. Lu, “Enhancing GPT-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures,” *Comput. Assist. Lang. Learn.*, pp. 1–20, Jun. 2025, doi: 10.1080/09588221.2025.2518430.
[105] P. Black and D. Wiliam, “Classroom assessment and pedagogy,” *Assess. Educ. Princ. Policy Pract.*, vol. 25, no. 6, pp. 551–575, Nov. 2018, doi: 10.1080/0969594X.2018.1441807.
[106] D. J. Nicol and D. Macfarlane-Dick, “Formative assessment and self-regulated learning: a model and seven principles of good feedback practice,” *Stud. High. Educ.*, vol. 31, no. 2, pp. 199–218, Apr. 2006, doi: 10.1080/03075070600572090.
[107] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas, “Itzamna: A multimodal artificial intelligence platform for comprehensive transversal skills assessment,” *SoftwareX*, vol. 31, p. 102262, Sep. 2025, doi: 10.1016/j.softx.2025.102262.
[108] L. A. Zadeh, “Fuzzy logic = computing with words,” *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 2, pp. 103–111, Feb. 1996, doi: 10.1109/91.493904.
[109] A. Field, “DISCOVERING STATISTICS USING BM SPSS STATISTICS”.
[110] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Routledge, 2013. doi: 10.4324/9780203771587.
[111] C. O. Fritz, P. E. Morris, and J. J. Richler, “Effect size estimates: Current use, calculations, and interpretation.,” *J. Exp. Psychol. Gen.*, vol. 141, no. 1, pp. 2–18, 2012, doi: 10.1037/a0024338.
[112] C. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, pp. 79–82, 2005, doi: 10.3354/cr030079.
[113] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 2017, doi: 10.1126/science.aal4230.
[114] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[115] W. T. Hoyt and J. N. Melby, “Dependability of Measurement in Counseling Psychology: An Introduction to Generalizability Theory,” *Couns. Psychol.*, vol. 27, no. 3, pp. 325–352, May 1999, doi: 10.1177/0011000099273003.
[116] S. E. Stemler and T. Jessica, “Best Practices in Interrater Reliability Three Common Approaches,” *Best Pract. Quant. Methods*, pp. 29–49, 2008, doi: 10.4135/9781412995627.d5.
[117] G. D. R, “Text-based environment : Computer conferencing in higher education,” *Internet High. Educ*, vol. 2, no. 2, pp. 87–105, 2000.