Future of Education

GenAI4ED

D2.3: The GenAI4ED evaluation framework

Funded by
the European Union

Vision of the project

GenAI4ED is here to empower students, teachers, parents, and guardians with the knowledge and skills to navigate the world of generative AI in education. We want to foster a digital-literate society that can make smart, informed decisions about the role of AI in education.

Funded by
the European Union

Project Title

Project acronym GenAI4ED

Project full title A platform for Assessing and Bridging Generative AI and Human Skills in Secondary Education

Grant agreement No. 101178648

Call HORIZON-CL2-2024-TRANSFORMATIONS-01

Project starting date 1 October 2024

Project duration 36 months

Document details

Work Package WP2

Authors Constantine Dovrolis, Myrianthi Hadjicharalambous, Valeria Filippou, Emmanouil Kritikos (CYI)

Contributors Tonia Galati, Persia Xenopoulou (PGMS)
Dimitra Dimitrakopoulou (EA)
Arja Krauchenberg, Claudio Masotti (EPA)
David Gikas (TURN-KEY.AI)
Dana Muresan, Cristiana Caserta, Simone Favale, Stefania Aceto (WE)

Reviewing partners Wonderful Education (WE), Trilateral Research (TRI), KonnektAble Technologies Limited (KT)

Dissemination level Public

Document history

Date	Version	Description
07-2025	0.1	Outline
08-2025	0.2	Draft
09-2025	0.3	Final

Funded by the European Union

Due date of the deliverable: 30/09/25

Actual submission date: 30/09/25

Executive Summary

This deliverable (D2.3 GenAI4ED Evaluation Framework) introduces the GEARS evaluation framework (**G**roup, **E**valuation criteria, **A**ssessment indicators, **R**eview method, **S**ource & data type), developed specifically for the GenAI4ED project under Task 2.4 - GenAI4ED evaluation framework. GEARS provides a comprehensive methodology for evaluating generative AI (GenAI) tools in educational settings, addressing critical gaps identified in existing evaluation approaches.

In developing the framework, the document reports on a systematic analysis of current evaluation methodologies, complemented by iterative stakeholder engagement, including focus group discussions with students, teachers, parents/guardians, and AI experts. This multidisciplinary consultation process ensured that GEARS reflects stakeholder priorities while maintaining academic rigour and practical applicability.

The deliverable then sets out the structure of GEARS, which establishes 56 evaluation criteria with corresponding Key Performance Indicators (KPIs) distributed across four stakeholder perspectives: students (18 criteria), teachers (12 criteria), parents/guardians (12 criteria) and AI experts (14 criteria). Each group contributes distinct evaluation priorities and access to different types of evidence, capturing the multi-dimensional impact of educational AI. The framework covers key evaluation dimensions, including usability and accessibility, learning effectiveness and support, ethical considerations and trust, technical performance and reliability, and developmental impact. It employs validated measurement approaches where available, while developing context-specific indicators for emerging evaluation dimensions unique to GenAI applications.

A distinctive strength of GEARS is its preservation of diverse stakeholder perspectives rather than aggregating conflicting evaluations into composite scores. This approach makes tensions and trade-offs explicit for informed decision-making while combining quantitative metrics with qualitative insights through mixed-methods design.

The document also explains how GEARS integrates with the GenAI4ED platform architecture via the Criteria Definition Module (T3.4) ensures systematic evidence collection to support pilot study requirements in WP5 including Task 5.6 policy development. GEARS will provide standardised evaluation protocols for consistent assessment across educational contexts. This will generate comparable evidence on GenAI tool effectiveness and support evidence-based policy recommendations at institutional, national, and European levels.

Overall, this deliverable offers a comprehensive overview of how the GEARS framework was conceived, developed, and positioned within the GenAI4ED project. It establishes a robust foundation for responsible GenAI integration in education while linking evaluation design directly to subsequent pilot implementation and policy guidance.

Table of Contents

List of Tables

Table of Figures

Acronyms

Acronym	Full Form
AI	Artificial Intelligence
ChatGPT	Chat Generative Pre-Trained Transformer
D1.1, D2.3	Deliverable 1.1, Deliverable 2.3
EQ	Emotional Quotient
EU	European Union
GEARS	Group, Evaluation criteria, Assessment indicators, Review method, Source & data type
GenAI	Generative Artificial Intelligence
GenAI4ED	Generative AI for Education
GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
MathVista	Mathematics Vision and Language Understanding
MMMU	Massive Multi-discipline Multimodal Understanding
NASA-TLX	NASA Task Load Index
ScienceQA	Science Question Answering
STEM	Science, Technology, Engineering, and Mathematics
T1.1, T1.2, T2.4, T3.4, T5.1, T5.2, T5.3, T5.4, T5.5, T5.6	Task 1.1, Task 1.2, Task 2.4, Task 3.4, Task 5.1, Task 5.2, Task 5.3, Task 5.4, Task 5.5, Task 5.6
UDL	Universal Design for Learning
URL	Uniform Resource Locator
WP1, WP2, WP3, WP5	Work Package 1, Work Package 2, Work Package 3, Work Package 5

1 Introduction

1.1 Background and Motivation

The integration of generative AI (GenAI) tools in education represents a paradigm shift with both significant opportunities and considerable challenges. On the one hand, recent advances in GenAI promise more personalised learning pathways, enhanced teaching support, and new forms of learner engagement (Jurenka, et al., 2024). On the other hand, the educational sector faces unprecedented difficulties in systematically assessing the pedagogical, ethical, and technical implications of these tools (Chu et al., 2025; Bond et al., 2024).

The gap between GenAI's potential and its practical implementation in education arises primarily from two factors: (1) the difficulty of translating pedagogical principles into meaningful AI-driven interactions, and (2) the absence of established frameworks for evaluating the effectiveness, safety, and broader impact of these technologies (Jurenka, et al., 2024). This highlights the urgent need for systematic, evidence-based evaluation approaches that guide the responsible adoption of GenAI in education.

Task 2.4 of the GenAI4ED project responds to this need by developing the GEARS evaluation framework. Designed through systematic research and stakeholder consultation, GEARS establishes comprehensive evaluation criteria, measurable KPIs, and guidance for standardised protocols that enable multi-perspective assessment of GenAI tools in educational contexts.

1.2 Objectives and Scope of Task 2.4

Task 2.4 pursues four key objectives:

- Develop comprehensive evaluation criteria for GenAI tools from multiple stakeholder perspectives (students, educators, parents/guardians, and AI experts).
- Establish measurable KPIs for each evaluation criterion across ethical, psychological, technical, operational, and user acceptance dimensions.
- Design data collection protocols supporting mixed-methods research that integrate quantitative metrics with qualitative stakeholder insights.
- Formulate processes for translating evaluation outcomes into policy recommendations for responsible GenAI adoption in education.

1.2.1 Relation to Other Tasks

Task 2.4, serves as a pivotal connector within the GenAI4ED project, drawing on early research outputs and providing the foundation for subsequent implementation, evaluation, and policy development:

- **Task 1.1** *Systematic Review of GenAI in Education*: Informs the development of evaluation criteria, ensuring they reflect current research, trends, and best practices.
- **Task 1.2** *Taxonomy Development*: Provides the structural basis for defining evidence-based evaluation criteria for GenAI tools and user groups and for pilot design.

Funded by the European Union

- **Task 3.4** *Criteria Definition Module*: Enables iterative refinement of the evaluation framework using real-world usage data from the platform
- **Task 5.2-5.4** *Pilot Studies*: Standardised protocols from T2.4 guide pilot implementations, ensuring consistent and comparable assessments.
- **Task 5.5** *Comprehensive Pilot Evaluation*: The evaluation criteria and KPIs serve as the operational toolkit for systematic pilot assessments.
- **Task 5.6** *Policy Recommendations*: Task 2.4 outputs provide evidence-based insights to inform policy guidance at institutional, national, and European levels.

1.3 Structure of the Document

This deliverable is organised into seven main sections. Following this introduction, Section 2 presents the GEARS framework structure and design rationale. Section 3 details the development methodology, including literature review findings and stakeholder consultation processes. Section 4 presents the comprehensive evaluation criteria organised by stakeholder group, representing the core contribution of Task 2.4. Section 5 examines cross-stakeholder considerations and potential divergences between different evaluation perspectives. Section 6 outlines framework applications in platform integration, pilot studies, and policy development. Finally, Section 7 provides conclusions and identifies next steps for framework implementation. Detailed criteria specifications are provided in Appendices A1-A4 for each stakeholder group.

2 Development Methodology

The GEARS framework was developed through a systematic two-phase approach combining evidence-based literature analysis with iterative stakeholder consultation. This methodology ensured the framework reflects both current research findings and practical stakeholder experiences across diverse educational contexts.

2.1 Literature Review and Evidence Base

2.1.1 Review Context and Scope

The framework development began with a targeted literature review conducted under Task 2.4, building upon the comprehensive overview provided in Deliverable D1.1: Comprehensive Review on GenAI Integration in Education. While D1.1 established the broader context of GenAI in education, the Task 2.4 review focused specifically on current evaluation practices and methodological gaps in existing approaches.

The rapid integration of generative artificial intelligence (GenAI) in education has prompted an urgent need for rigorous, inclusive, and pedagogically sound evaluation frameworks. GenAI tools such as ChatGPT, GitHub Copilot, and other large language models offer unprecedented capabilities for content generation, assessment, and personalisation. However, their implementation in educational settings raises complex questions about accuracy, ethics, stakeholder alignment, and long-term learning outcomes. This review synthesises recent empirical, theoretical, and policy literature (2019–2025), focusing on this timeframe to capture the period of rapid GenAI development and educational integration, particularly following the public release of ChatGPT in late 2022 and subsequent widespread adoption of GenAI tools in educational contexts. It reveals that although the field exhibits methodological innovation and technical sophistication, critical gaps remain in the integration of pedagogical goals, stakeholder values, and robust validation mechanisms.

2.1.2 Methodological Trends and Current Evaluation Practices

Recent evaluations demonstrate growing methodological breadth and technical depth. However, Bond et al.'s (2024) meta-systematic review of 66 evidence syntheses reveals that existing evaluation approaches often lack comprehensive frameworks that integrate technical performance with pedagogical effectiveness and stakeholder perspectives. Similarly, González-Calatayud et al. (2021) review of AI in student assessment demonstrates how current literature focuses on the technical capabilities while lacking systematic evaluation of educational impact and user experience.

Beyond these technical applications, recent studies have evaluated the impact of GenAI tools on student learning outcomes and user experience. Wang and Fan (2025) synthesised 51 studies on ChatGPT, reporting significant positive effects on learning performance, learning perception, and higher-order thinking skills, moderated by course type, learning model, and duration of use. Additional research points to improvements in

Funded by the European Union

student productivity, enhanced efficiency in writing and research tasks, and reductions in teacher workload when GenAI tools are integrated into instruction.

These findings underline that evaluation frameworks need to capture both technical and pedagogical dimensions, including learning outcomes, usability, and stakeholder experience.

2.2 Key Findings and Framework Implications

The literature analysis revealed three critical limitations in current GenAI evaluation approaches that directly informed GEARS design:

Methodological Inconsistencies and Design Limitations

Despite growing interest in GenAI, evaluations often suffer from inconsistent designs, and limited triangulation. Xia et al. (2024) observe the most empirical studies emerged in 2023, undermining generalisability and long-term reliability. Bond et al. (2024) report nearly half of systematic reviews failed to conduct quality assessments, with unclear data coding and inadequate methodological transparency. These design inconsistencies and transparency issues highlighted the need for robust, reproducible, and context-sensitive practices with standardised quality assessment procedures.

Stakeholder Exclusion and Under-representation

Stakeholder perspectives are frequently underrepresented in GenAI-related discussion in education. O'Dea et al. (2024) show that students prioritise usability and practical applications, such as support with writing, multimedia, and problem-solving, over institutional performance metrics, highlighting a gap between institutional priorities and learner needs. report that 68% of students express concerns about bias in GenAI outputs, and over 72% routinely cross-check information, demonstrating cautious but sophisticated evaluation strategies. These findings underline the importance of incorporating student perspectives into policy and practice. Yet other key stakeholders, including parents, AI experts, and employers, are rarely consulted, despite their potential to contribute valuable insights into the ethical, pedagogical, and technical implications of GenAI adoption. Khlaif et al. (2025) reveal institutional misalignment: educators navigate AI integration independently, while students face inequities linked to digital access and AI familiarity. Zawacki-Richter et al. (2019) highlight disciplinary imbalance, with 62% of AI in education research from STEM/Computer Science fields and fewer than 9% from education departments, leading to underdeveloped pedagogical and ethical dimensions in evaluations.

Technical-Pedagogical Misalignment

Evaluation frameworks often prioritise technical performance (accuracy, efficiency) while neglecting pedagogical relevance. Xia et al. (2024) note GenAI feedback can misalign with instructor expectations due to limited curricular awareness. Research emphasises that students, teachers, and parents perceive GenAI differently, affecting evaluation outcomes. Johnson et al. (2024), surveying 2,555 university students, find that students' ethical understandings of AI usage often diverge from institutional policies.

Karran et al. (2025), with over 1,000 participants, demonstrate distinct priorities: students prefer low AI agency and strong privacy, parents worry about AI grading, and teachers value explainability. Zastudil et al. (2024) confirm these findings qualitatively: students focus on personalised learning and efficiency, while instructors emphasise curriculum coherence and long-term development.

These insights established the theoretical foundation for GEARS criteria. They emphasise that effective evaluation must go beyond technical metrics to include trust, transparency, fairness, and pedagogical alignment as core dimensions. At the same time, they strengthen the case for participatory and context-sensitive approaches, ensuring that diverse stakeholders and educational settings are meaningfully represented, an emphasis directly embedded in Task 2.4.

2.3 Stakeholder Consultation Process

The literature findings were complemented by an iterative consultation process involving diverse stakeholders, ensuring the criteria reflected both theoretical robustness and practical realities from multiple educational perspectives.

2.3.1 Consortium draft criteria development

Following the literature review, consortium partners involved in Task 2.4 developed an initial set of draft evaluation criteria. This development process built upon both the systematic analysis of existing evaluation frameworks identified in the literature and partners' collective expertise in educational technology, AI applications, and pedagogical assessment. The draft criteria table served as the foundation for subsequent stakeholder consultation, providing a structured starting point that could be refined, expanded, or modified based on stakeholder feedback and practical implementation considerations.

2.3.2 Participant Groups

Four stakeholder groups contributed distinct insights to framework development:

- **Students** (n=2) were represented through survey responses and feedback collected in Task 1.3, focusing on usability, learning impact, and engagement with GenAI tools.
- **Teachers** (n=11) provided classroom experience insights, highlighting potential uses, challenges, and risks of GenAI tools.
- **Parents/guardians** (n=11) offered perspectives on GenAI impact on students' learning experiences, wellbeing, and ethical concerns.
- **AI experts** (n=9) provided technological and ethical insights, with some reflecting on educational applications.

2.3.3 Consultation methodology

Multiple focus groups and workshops engaged participants directly with draft evaluation criteria through structured sessions. Each session followed a standardised protocol:

Funded by the European Union

Session Structure
Facilitators provided GenAI4ED project context, explained Task 2.4 objectives, and clarified expected consultation outcomes. Participants then reviewed presented criteria through guided discussion formats.

Feedback Collection
Participants assessed criteria relevance, clarity, completeness, and practical applicability. They identified missing criteria, unnecessary elements, and contextual adaptation requirements. Sessions lasted 30-60 minutes, enabling comprehensive discussion and iterative feedback.

Integration Process
Feedback was systematically documented and incorporated into successive framework versions. This iterative approach integrated technical, educational, psychological, and ethical expertise, ensuring criteria achieved both theoretical validity and practical applicability.

The consultation process ensured GEARS development reflected diverse stakeholder priorities while maintaining methodological consistency across educational contexts and tool applications.

Funded by
the European Union

3 The GEARS Framework

The GEARS evaluation framework provides a systematic methodology for assessing GenAI tools in educational contexts through multi-stakeholder perspectives. The framework addresses current evaluation limitations by preserving distinct stakeholder viewpoints rather than aggregating conflicting assessments into potentially misleading composite scores.

3.1 Framework Structure and Components

The GEARS framework organises the evaluation process into five interconnected components that systematically address key evaluation questions:

Group Identification

- The stakeholder perspective conducting the evaluation (e.g., students assessing usability, teachers evaluating pedagogical fit, parents focusing on safety)

Evaluation Criteria Establishment

- The specific aspects or dimensions being assessed (e.g., ease of use, learning effectiveness, curriculum alignment, ethical considerations)

Assessment Indicators

- The concrete measures used to evaluate each criterion (e.g., task completion rates, satisfaction scores, learning outcome improvements)

Review Methodology

- The data collection and analysis approaches employed (e.g., surveys, interviews, classroom observations, usage analytics)

Source & Data Type

- The origin and format of evidence gathered (e.g., platform interaction logs, survey responses, interview transcripts, observational notes)

Figure 1: Overview of GEARS Framework components for evaluation of GenAI tools in educational contexts

This structure provides clarity and reproducibility, supporting consistent evaluations across diverse educational settings while remaining adaptable to local contexts.

3.2 Multi-Stakeholder Design Rationale

GEARS recognises that GenAI tools affect different educational groups in distinct ways. Drawing from Holmes et al. (2023), the framework explicitly accommodates diverse perspectives identified in the consultation process. Each stakeholder group contributes distinct evaluation priorities and access to different types of evidence:

Funded by
the European Union

- **Students** assess usability, learning engagement, and personal educational impact from the primary user perspective.
- **Teachers** evaluate pedagogical integration, curriculum alignment, and classroom implementation effectiveness.
- **Parents/guardians** focus on developmental impact, safety considerations, and broader educational value for their children.
- **AI experts** examine technical robustness, algorithmic fairness, privacy protection, and regulatory compliance.

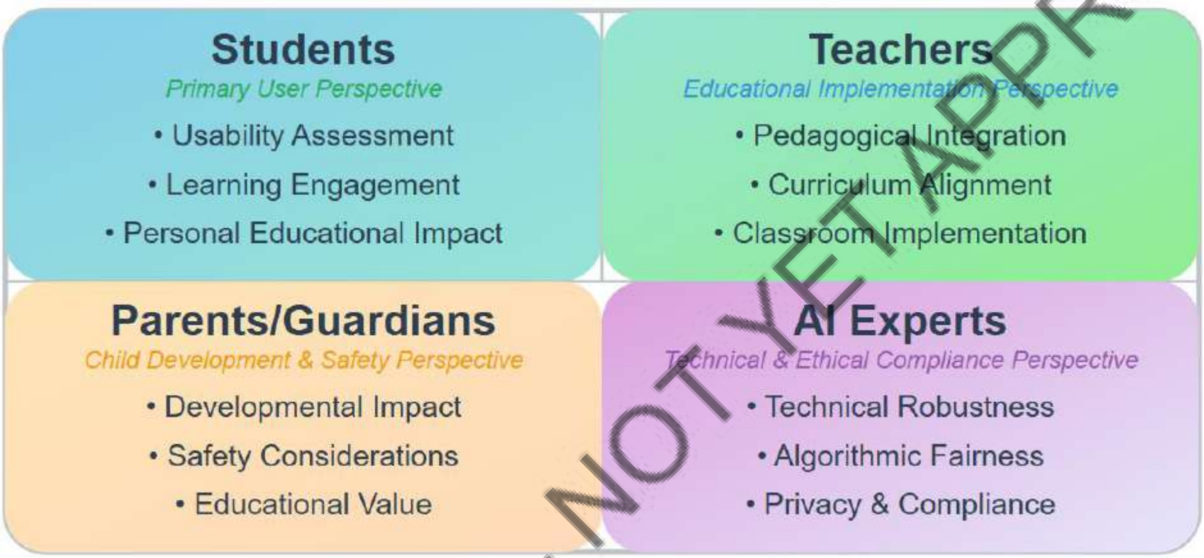

Figure 2: Four key stakeholder groups and their primary evaluation concerns for GenAI tools in education

3.3 Mixed-Methods Design Principles

GEARS employs a mixed-methods evaluation design that systematically integrates quantitative and qualitative data. This approach ensures evaluations capture measurable outcomes while also reflecting lived experiences and diverse stakeholders' priorities.

3.3.1 Data integration approach

Quantitative measures
Usability scores, learning performance indicators, platform analytics, and KPI tracking provide measurable evidence of tool effectiveness and user engagement.

Qualitative inputs
Interviews, focus groups, scenario-based classroom tasks, and open-ended reflections capture nuanced experiences, unexpected outcomes, and contextual factors that numbers alone cannot reveal.

3.3.2 Preserving stakeholder perspectives

Evaluation results are presented separately for each stakeholder group, directly linked to the specific GEARS components relevant to their perspective. This approach preserves interpretability of findings, highlights important differences between stakeholder viewpoints, and ensures results remain both contextually grounded and comparable across studies, enabling informed decision-making about GenAI tool adoption and implementation while respecting the complexity of educational contexts.

4 Evaluation Criteria by Stakeholder Group

This section presents the comprehensive set of 56 evaluation criteria developed through systematic literature analysis and stakeholder consultation, representing the core contribution of Task 2.4. The GEARS framework addresses the critical gaps identified in existing evaluation approaches through deliberate design choices that prioritise stakeholder inclusion, pedagogical relevance, and long-term adaptability.

4.1 Framework Design Principles and Literature Gap Alignment

The criteria systematically address the evaluation limitations identified in current GenAI assessment practices through four key design principles:

Long-term Compatibility and Tool Adaptability

Rather than developing tool-specific metrics, GEARS establishes general evaluation dimensions that remain relevant as new GenAI tools emerge. The criteria focus on fundamental educational considerations, learning effectiveness, safety, pedagogical alignment, and many more, that apply regardless of specific technological implementations, ensuring the framework's sustainability in a rapidly evolving technological landscape.

Multi-Stakeholder Inclusion

The framework directly addresses stakeholder exclusion identified in the literature by systematically incorporating four distinct perspectives. The inclusion of parents/guardians and AI experts alongside traditional student and teacher evaluations ensures comprehensive assessment that captures both immediate usability concerns and broader developmental, ethical, and technical considerations often overlooked in current evaluation approaches.

Pedagogical Primacy

While maintaining technical performance assessment through AI expert criteria, GEARS prioritises pedagogical relevance through dedicated teacher evaluation dimensions and learning-focused student criteria. This design ensures that technical sophistication does not overshadow educational value, addressing the technical-pedagogical misalignment identified in existing frameworks.

4.2 Criteria Development Process

The final 56 criteria emerged through iterative stakeholder consultation where participant priorities directly shaped framework content. During workshop sessions, stakeholders identified evaluation dimensions they considered essential for responsible GenAI adoption in their respective contexts. These suggestions were systematically incorporated into the framework, ensuring criteria reflect authentic stakeholder concerns rather than predetermined research assumptions.

Funded by the European Union

The distribution of criteria across stakeholder groups, 18 for students, 12 for teachers, 12 for parents/guardians, and 14 for AI experts, reflects both the complexity of different stakeholder perspectives and the range of concerns raised during consultation. Students and parents/guardians received more extensive criteria sets due to their comprehensive evaluation needs spanning usability, development, safety, and learning effectiveness. Teachers received a focused set addressing their dual role as users and pedagogical evaluators. AI experts received criteria covering technical robustness and ethical compliance essential for platform integrity.

Given the varied roles and expectations of those affected by GenAI in education, a one-size-fits-all evaluation approach is insufficient. The following sections outline tailored criteria for each stakeholder group, ensuring that diverse priorities such as learning outcomes, pedagogical integrity, ethical compliance, and technical reliability are comprehensively addressed.

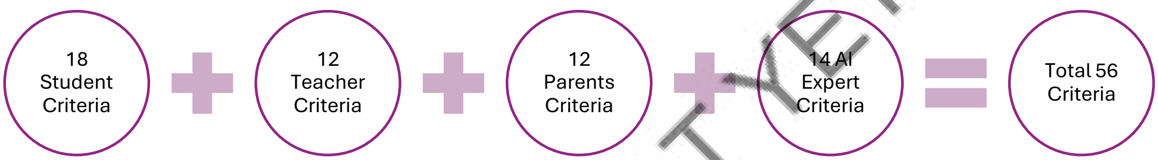

Figure 3: Number of evaluation criteria developed for each stakeholder group in the GEARS framework

4.3 Student Perspective

Students are the primary users of GenAI tools in the educational process, making their evaluation perspective essential for understanding tool effectiveness and educational impact. Through the iterative stakeholder consultation process, students and student representatives identified 18 comprehensive evaluation criteria across five key areas that reflect their priorities and concerns about GenAI integration in learning environments.

The student perspective emphasises not just immediate usability and academic impact, but also longer-term developmental considerations including emotional experience, ethical awareness, cognitive development, and learning independence. These criteria assess whether AI tools serve as empowering companions in the learning journey or risk creating dependency, confusion, or undermining authentic learning processes. The full list of student evaluation criteria can be found in Appendix [1], while below is a description of the key dimensions and rationale for their inclusion.

4.3.1 Core Evaluation Areas

Usability (2 criteria) addresses fundamental accessibility and cognitive considerations, emphasising that GenAI tools must be intuitive and reduce rather than increase mental burden. Key criteria include basic accessibility/usability for diverse learners and cognitive load assessment using validated measures like NASA-TLX scoring[1].

Learning Impact (7 criteria) represents the largest evaluation area, reflecting student focus on educational effectiveness across multiple dimensions. Students emphasised the importance of high-quality AI feedback and explanations, study efficiency improvements, and confidence building. Significantly, they highlighted metacognitive awareness, the ability to understand and explain AI-generated content, as crucial for authentic learning. Additional criteria address broader learning outcomes, skill development across multiple areas, and overall learner well-being including emotional, social, and psychological support.

AI Ethics (4 criteria) demonstrates sophisticated understanding of ethical considerations, identifying trust and accuracy, overreliance awareness, ethical understanding of academic honesty and bias issues, and AI transparency comprehension as essential evaluation dimensions. These criteria reflect student awareness that GenAI use involves complex ethical decisions requiring informed judgment.

AI Functionalities (3 criteria) evaluates specific AI capabilities that enhance their learning experience, including personalisation effectiveness, collaboration support, and social presence/peer dialogue facilitation. These criteria assess whether AI tools successfully integrate into collaborative learning environments while maintaining meaningful peer interaction.

Other Critical Dimensions (2 criteria) focuses on engagement and motivation/curiosity/creativity as fundamental measures of AI tool success, emphasising that effective educational AI should inspire and motivate rather than merely complete tasks.

4.3.2 Assessment Methods

The student criteria will be assessed using multiple approaches: thematic qualitative analysis, output/interaction tracking and analytics, quiz and benchmark testing, Likert scale and rubric-based feedback, observation-based evaluation, manual/expert review and rubric scoring, and ethics, policy, and safety evaluation. This methodological diversity ensures comprehensive assessment that captures both quantitative performance measures and qualitative learning experiences.

Complete student criteria specifications in GEARS format are provided in Appendix A.1, including detailed assessment indicators, review methods, and data source specifications for each evaluation criterion.

[1] https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf

Funded by the European Union

4.4 Teacher Perspective

Teachers play a dual role in GenAI integration: they are both end-users evaluating how AI affects their professional experience and evaluators assessing how AI impacts pedagogy, curriculum delivery, and student learning outcomes. This includes evaluating the degree to which specific educational tasks, such as content creation, personalised tutoring, and automatic grading, are enhanced or transformed by AI tools, ultimately providing ranked assessments of which teaching tasks benefit most from GenAI integration. Through iterative stakeholder consultation, teachers identified 12 comprehensive evaluation criteria across five key areas that reflect their unique priorities and concerns about responsible GenAI adoption in educational settings.

The teacher perspective emphasises both immediate practical considerations, such as time efficiency and curriculum alignment, and broader professional concerns including pedagogical soundness, well-being impact, job security, and ethical compliance. These criteria assess whether AI tools enhance rather than undermine teaching quality, support professional growth, and maintain pedagogical integrity while addressing concerns about working conditions and potential job displacement. Particular attention is given to how generative AI can transform core teaching tasks, such as lesson content creation, personalised tutoring, and automated grading, highlighting which aspects of teachers' work stand to benefit most from technological support. The complete set of teacher evaluation criteria is detailed in Appendix [2], with the following section describing the primary categories and their pedagogical significance.

4.4.1 Core Evaluation Areas

Learning Impact (4 criteria) prioritises educational effectiveness and professional alignment as primary evaluation dimensions. Key areas include curriculum alignment and learning objectives assessment, pedagogical soundness evaluation, professional growth measurement, and differentiation/ Universal Design for Learning (UDL) accessibility support. These criteria ensure AI-generated materials match educational standards while supporting diverse learner needs through adaptive materials and inclusive design approaches.

Usability (4 criteria) reflects teacher focus on practical implementation and workplace impact. This includes time efficiency measurement comparing lesson planning duration with and without AI support, productivity assessment evaluating how AI accelerates specific educational tasks (content creation, personalised tutoring, and automatic grading), well-being and workload criteria measure AI's impact on teacher stress, burnout, work-life balance, and concerns about job replacement and changing working conditions. Finally, connection evaluation addressing meaningful interactions facilitated by AI tools.

AI Functionalities (2 criteria) emphasises understanding and effectively utilising AI capabilities. This covers AI literacy support assessment measuring teachers' developing competence with prompts and risk awareness, and data-driven insights evaluation

Funded by the European Union

addressing quality and usefulness of AI-generated reports and analytical tools supporting instructional decision-making.

AI Ethics (1 criterion) addresses regulatory and ethical requirements through ethical and privacy management assessment ensuring GDPR and EU AI Act compliance, including consent management, data anonymisation, and fair, unbiased AI implementation. This reflects teachers' dual responsibility as both professional users and student protectors.

Other Critical Dimensions (1 criterion) identifies collaborative teaching as essential for successful AI integration, evaluating the extent to which AI tools facilitate professional collaboration and co-design among colleagues, recognising that sustainable AI adoption requires community support and shared expertise development.

4.4.2 Assessment Methods

The teacher criteria will be assessed using diverse approaches: manual/expert review and rubric scoring, Likert scale and rubric-based feedback, output/interaction tracking and analytics, and quiz and benchmark testing. This ensures comprehensive assessment capturing both objective performance measures and subjective professional experiences with AI integration, including impacts on teacher wellbeing and professional security.

Complete teacher criteria specifications in GEARS format are provided in Appendix A.2, including detailed assessment indicators, review methods, and data source specifications for each evaluation criterion.

4.5 Parent/Guardian Perspective

Parents and guardians serve as essential gatekeepers for safety, wellbeing, and developmental appropriateness in their children's educational technology use. Through the iterative stakeholder consultation process, parents and guardians identified 12 comprehensive evaluation criteria across five key areas that reflect their unique concerns about GenAI's impact on their children's learning, development, and overall wellbeing.

The parent/guardian perspective emphasises practical oversight concerns and long-term developmental impacts that parents can observe at home and in family settings. These criteria assess whether AI tools support healthy development and authentic learning while maintaining parental control and age-appropriate boundaries, with particular focus on maintaining children's independence, social-emotional growth, and creative thinking. All parent/guardian evaluation criteria are listed in Appendix [3], while the discussion below outlines the main areas of concern and their developmental importance.

4.5.1 Core Evaluation Areas

Funded by the European Union

Safety & Parental Control (3 criteria) prioritises parents' ability to maintain oversight and ensure age-appropriate use through effective parental control, age-appropriate safeguards ensuring content and interactions suitable for developmental stages, and progress monitoring offering quality tracking and reporting features for parent engagement.

Learning Independence & Development (4 criteria) represents the largest evaluation area, reflecting parent concerns about authentic learning and self-reliance. This includes dependency prevention assessing whether children can complete tasks independently without over-relying on AI assistance, knowledge retention measuring whether children learn and remember information rather than just getting quick answers, independent problem-solving evaluating children's ability to approach challenges without immediately turning to AI and critical thinking skills assessment tracking development of analytical and evaluative thinking capabilities.

Social-Emotional Impact (3 criteria) addresses AI's potential effects on children's social and emotional development through emotional intelligence impact tracking changes in emotional recognition and regulation, social interaction quality measuring whether AI use affects children's ability to engage meaningfully with family and peers, and emotional expression assessment evaluating children's ability to understand, express, and regulate emotions appropriately.

Creative & Cognitive Development (1 criterion) maintains children's original thinking and creative expression through creative independence assessment of whether children maintain original thinking and creative problem-solving capabilities rather than relying on AI-generated ideas.

Family Support & Communication (1 criterion) emphasises receiving actionable guidance for supporting children's learning through actionable insights evaluation assessing the quality and usefulness of suggestions provided to parents for supporting learning at home.

4.5.2 Assessment Methods

The parent/guardian criteria incorporated innovative, family-centred assessment approaches including role-play scenarios with different tool configurations, workshop evaluation of control interface usability, simulated scenarios testing safety responses, memory games and recall exercises, creative challenges and original thinking exercises, and group collaboration exercises. This diverse methodology ensures comprehensive evaluation capturing both immediate safety concerns and long-term developmental impacts that parents can authentically observe and assess.

Complete parent/guardian criteria specifications in GEARS format are provided in Appendix A.3, including detailed assessment indicators, review methods, and data source specifications for each evaluation criterion.

Funded by
the European Union

4.6 AI Expert Perspective

AI experts provide technical and ethical backbone to the evaluation process, focusing not on how tools are used, but on how they are built. Through systematic stakeholder consultation, AI experts identified 14 comprehensive evaluation criteria across five key technical areas that examine algorithmic robustness, model behaviour, fairness, and system-level integrity. Their assessments help ensure that GenAI tools deployed in education are transparent, secure, and fit for purpose.

The AI expert perspective emphasises rigorous technical validation and system-wide integrity assessments that can only be conducted through specialised testing methodologies and expert analysis. These criteria assess whether AI tools meet industry standards for safety, reliability, and ethical deployment while maintaining high performance across diverse educational contexts, with particular focus on preventing harmful outputs, ensuring fairness across demographic groups, and maintaining system stability and data protection. The comprehensive list of technical and ethical evaluation criteria can be found in Appendix [4], with the subsequent description focusing on the core technical and compliance dimensions.

4.6.1 Core Evaluation Areas

Security & Safety Systems (3 criteria) represents the most critical evaluation area, focusing on system resilience against adversarial prompts and jailbreak attempts through stress testing, safety system robustness evaluation using specialised safety benchmarks, and data privacy compliance assessment through comprehensive privacy compliance frameworks to prevent training data leakage.

Model Performance & Consistency (4 criteria) focuses on technical reliability and performance stability through output consistency measuring response reproducibility across sessions using statistical variance analysis, computational efficiency evaluating processing speed and resource optimisation, system reliability assessing model stability in production environments through uptime monitoring, and version consistency measuring performance across model updates through regression testing.

Accuracy & Reasoning Quality (3 criteria) prioritises intellectual integrity of AI-generated content through domain knowledge accuracy using established benchmarks like MMMU[2] (Xiang, et al. 2023) and ScienceQA[3], reasoning quality evaluating step-by-step logical inference through reasoning pathway analysis, and factual accuracy measuring hallucination rates using specialised fact-checking tools.

Fairness & Language Equity (2 criteria) addresses algorithmic bias and linguistic inclusivity through statistical analysis identifying disparities across demographic variables using bias detection datasets, and language performance parity measuring

[2] https://arxiv.org/abs/2311.16502
[3] https://scienceqa.github.io/

performance differences between high-resource and low-resource languages through cross-linguistic evaluation.

Technical Integration & Efficiency (2 criteria) evaluates advanced technical capabilities through prompt efficiency assessing model interpretation and response quality across different prompt styles and complexity levels, and multimodal integration evaluating technical quality of combining text, image, and other modalities through specialised benchmarks.

4.6.2 Assessment Methods

The AI expert criteria included sophisticated technical assessment approaches including automated jailbreak testing and security vulnerability scanning, red-team attack simulations, statistical variance analysis, bias detection workshops using demographic test datasets, cross-linguistic performance comparison across multilingual benchmarks, prompt engineering workshops, performance profiling and resource monitoring, expert domain evaluation sessions, safety system stress testing, fact verification using hallucination detection tools, privacy compliance assessment, reliability testing with failure analysis, version comparison analysis, and multimodal processing evaluation. This comprehensive methodology ensures thorough technical validation capturing both immediate performance metrics and long-term system integrity requirements.

Complete AI expert criteria specifications in GEARS format are provided in Appendix A.4, including detailed assessment indicators, review methods, and data source specifications for each evaluation criterion.

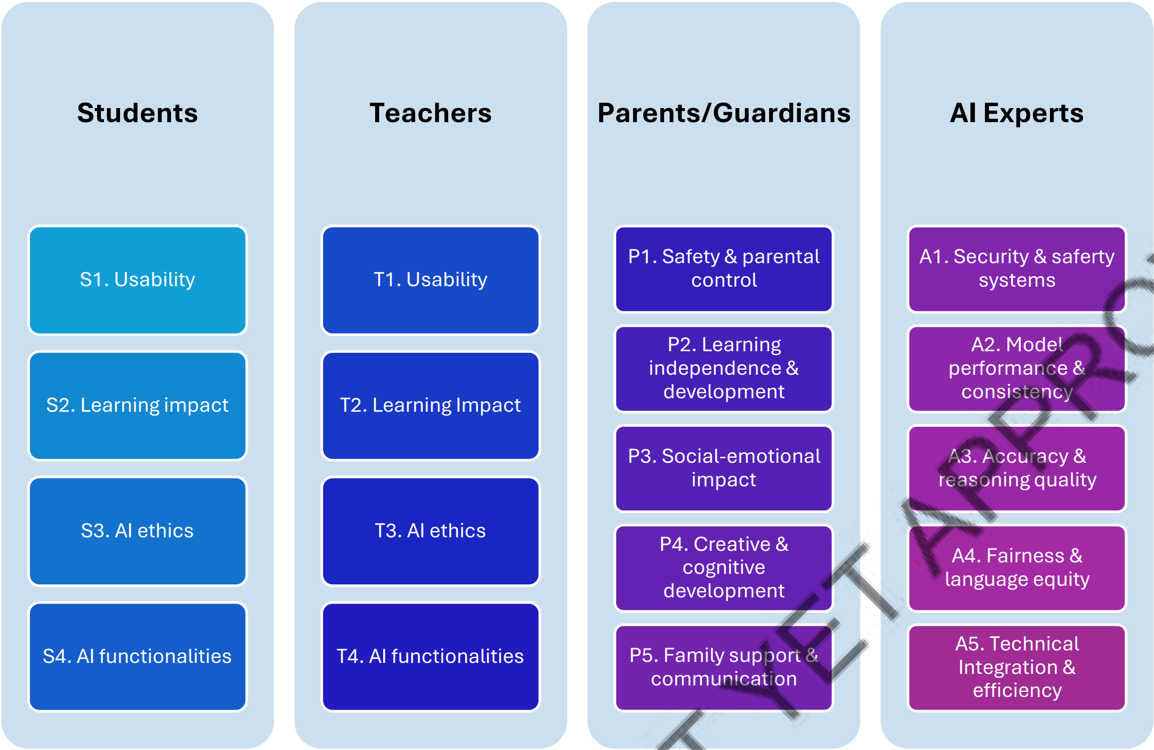

Figure 4: Core evaluation criteria areas identified for the development of GEARS framework. While some criteria appear across multiple stakeholder groups, each stakeholder evaluates these criteria from their unique perspective and context.

5 Cross-Stakeholder Considerations

The GEARS framework incorporates diverse perspectives from four key stakeholder groups: students, teachers, parents/guardians, and AI experts. While this inclusive structure ensures comprehensive evaluation of GenAI tools in education, it also introduces complexity through potentially conflicting assessments of the same tool. Rather than viewing divergence as problematic, GEARS makes these differences explicit to support informed decision-making about tool adoption and implementation.

5.1 Complementary and Diverging Perspectives

Each stakeholder group brings distinct evaluation priorities that complement each other from different perspectives. These differences reflect the real complexity of AI adoption in educational settings and provide valuable insights for implementation decisions.

5.1.1 Efficiency vs. Independence Divergences

Time efficiency conflicts
Teachers prioritise "Time Efficiency" (Appendix A2), measuring minutes saved per lesson and planning time reduction. However, parents emphasise "Dependency Prevention" (Appendix A3), assessing whether children complete tasks independently without direct AI answers. A tool that significantly reduces teacher lesson planning time might simultaneously increase student reliance on AI assistance, creating a tension where teacher productivity gains come at the cost of student independence. Similarly, students value "Study Efficiency" (Appendix A1), reporting time savings and better outcomes, which may conflict with parent concerns about authentic learning processes.

Learning depth tensions
Parents focus on "Knowledge Retention" (Appendix A3), measuring knowledge retained after 1 week, 1 month, and 3 months, while students emphasise immediate "Learning Support and Feedback Quality" (Appendix A1), seeking quick comprehension assistance. A tool providing instant, detailed answers might score highly on student satisfaction but poorly on long-term retention measures valued by parents.

5.1.2 Safety vs. Usability Trade-offs

Control vs. accessibility
Parents prioritise "Parental Control Features" (Appendix A3) and "Age-Appropriate Safeguards," while students focus on "Accessibility/Usability" (Appendix A1), describing tools as "easy to use." Comprehensive parental controls might introduce complexity that students find difficult, creating friction between safety and usability priorities.

Privacy vs. functionality
AI experts evaluate "Data Privacy Compliance" (Appendix A4) through privacy risk scores and compliance assessment, while teachers value "Data-driven Insights" (Appendix A2) requiring dashboard usefulness and analytical capabilities. Strong privacy protections

Funded by the European Union

might limit the data collection necessary for generating useful teacher analytics, creating tension between compliance and functionality.

5.1.3 Quality vs. Engagement Conflicts

Accuracy vs. motivation
AI experts measure "Factual Accuracy" (Appendix A4) through hallucination rates and fact-checking tools, while students assess "Motivation, Curiosity and Creativity" (Appendix A1) through engagement and inspiration scores. A highly accurate but conservative AI tool might score well on expert technical measures but poorly on student engagement metrics if it provides cautious, unimaginative responses.

Pedagogical soundness vs. personalisation
Teachers evaluate "Pedagogical Soundness" (Appendix A2) through expert rubrics for age appropriateness and instructional value, while students value "Personalisation Effectiveness" (Appendix A1) through adaptive learning paths. Pedagogically sound approaches might feel generic to students seeking personalised experiences.

5.1.4 Technical vs. User Experience Perspectives

System reliability vs. user satisfaction
AI experts focus on "System Reliability" (Appendix A4) measuring deployment stability and uptime, while all user groups prioritise immediate functionality. A technically robust system with 99% uptime might still frustrate users during brief outages, creating divergent evaluations of the same tool's quality.

Algorithmic bias vs. user preference
AI experts assess "Algorithmic Bias" (Appendix A4) through demographic fairness scores, while students and teachers may not notice or may even prefer certain biased outputs that align with their expectations or cultural contexts.

5.2 Identifying Tensions Through Mixed-Methods Design

The GEARS framework's mixed methods approach systematically captures these tensions through both quantitative metrics and qualitative insights. Each stakeholder group provides evaluation data through their specific criteria, enabling direct comparison of perspectives on the same tool.

Quantitative divergence identification
Numerical scores across stakeholder groups reveal systematic tensions. For example, a tool might receive high student "Study Efficiency" scores (85% reporting time savings) while showing low parent "Dependency Prevention" ratings (only 45% completing tasks independently). These numerical differences immediately highlight areas requiring attention.

Funded by the European Union

Qualitative context understanding
Open-ended feedback and scenario-based evaluations provide context for quantitative divergences. Student interview transcripts explaining why they value AI assistance can be compared with parent concerns about over-reliance, revealing the underlying reasoning behind conflicting scores.

5.3 Implications for Evaluation and Decision-Making

The GEARS framework enables educational institutions to make informed decisions about GenAI tool implementation by comparing evaluation results across all stakeholder groups. Decisions are made case-by-case, considering the specific educational context and balancing potentially conflicting stakeholder priorities. The following examples demonstrate how this process works in practice.

5.3.1 Context-Specific Tool Adaptation

Primary education scenario
In a primary school setting using GenAI for creative writing support, high student "Engagement" scores (Appendix A1) might coincide with teacher concerns about "Pedagogical Soundness" (Appendix A2) and parent worries about "Creative Independence" (Appendix A3). Implementation decisions could involve restricting AI to brainstorming phases while requiring independent final drafts, addressing multiple stakeholders concerns simultaneously.

Secondary education scenario
For secondary students using GenAI in collaborative research projects, student "Collaboration Support" (Appendix A1) and teacher "Productivity" (Appendix A2) benefits might conflict with AI expert "Security & Attack Resistance" (Appendix A4) concerns about data sharing. Solutions might include implementing secure collaboration protocols that maintain functionality while addressing security requirements.

5.3.2 Evaluation-Based Decision Making

Resource allocation decisions
When stakeholder evaluations diverge, institutions can make informed trade-offs. High AI expert technical scores combined with poor user experience ratings might indicate needs for training and support rather than tool replacement.

Regulatory compliance balancing
Conflicts between AI expert compliance requirements and user functionality preferences help identify areas where additional guidance, training, or technical solutions are needed to satisfy both regulatory and user needs.

Funded by the European Union

5.4 Supporting Holistic and Ethical Evaluation

The GEARS framework's strength lies in preserving rather than resolving stakeholder tensions. By maintaining separate evaluation streams, the framework enables decision-makers to understand why different groups experience the same tool differently and to develop targeted interventions that address multiple perspectives simultaneously.

Transparency in decision-making
Rather than forcing artificial consensus through averaged scores, GEARS provides clear visibility into where stakeholder priorities align and where they conflict, enabling more thoughtful and adaptive implementation strategies.

Ethical complexity acknowledgment
Educational GenAI deployment involves legitimate competing values, efficiency vs. independence, safety vs. accessibility, innovation vs. caution. GEARS supports ethical decision-making by making these trade-offs explicit rather than hidden within composite metrics.

This approach ensures that GenAI tool evaluation and implementation decisions are grounded in comprehensive stakeholder understanding rather than simplified metrics, supporting more effective and ethically aware educational AI integration.

Funded by
the European Union

6 Applications and Future Directions

The GEARS framework serves as a foundational resource that will inform three primary areas of future development within the GenAI4ED project: platform integration, pilot study implementation, and policy development. While specific implementation approaches will be determined in later project phases, this section outlines how the framework will guide their development.

6.1 Framework Applications in Project Development

6.1.1 Platform Integration Through Question Development

The GEARS framework will inform the development of evaluation questions integrated into the GenAI4ED platform through the Criteria Definition Module (T3.4). The 56 evaluation criteria provide a comprehensive foundation for question development, ensuring systematic coverage of usability, learning impact, ethical considerations, and technical performance across all stakeholder groups. Each criterion's specified assessment indicators offer concrete guidance for translating abstract evaluation dimensions into measurable questions.

Questions derived from GEARS criteria will be integrated into user workflows through methods to be determined in later development phases, potentially including pop-up interaction windows or follow-up communications. The framework's mixed-methods design supports both quick quantitative ratings and more detailed qualitative feedback collection.

6.1.2 Pilot Study Development and Implementation

GEARS will guide development of questionnaires and protocols for pilot studies involving teachers and students using various GenAI tools across different educational contexts. The framework ensures consistent evaluation approaches while accommodating diverse pilot objectives and tool functionalities.

Pilot study questionnaires and usage protocols will be developed in later project phases (T5.1) using GEARS criteria as reference points for comprehensive evaluation coverage. The framework's assessment indicators and review methods provide concrete guidance for creating measurable evaluation instruments appropriate for diverse educational contexts. While the complete framework contains 56 criteria, pilot studies will select relevant subsets based on their specific objectives, participant characteristics, and tool functionalities.

6.1.3 Policy Development Guidance (T5.6)

GEARS will inform policy development processes by providing a systematic foundation for understanding GenAI tools' impacts across multiple stakeholder perspectives and evaluation dimensions. The multi-stakeholder design captures implementation complexity, providing policy developers with structured understanding of competing priorities and potential conflicts between different evaluation perspectives.

Policy development will draw upon evidence generated through GEARS-informed pilot studies and platform implementations, creating a systematic knowledge base for evidence-based governance recommendations. Framework-informed policies will address multiple governance levels, from institutional AI adoption guidelines to national education technology standards and European-level coordination mechanisms.

6.2 Future Directions and Evolution

Framework Refinement Through Implementation

While GEARS remains fixed during the project implementation phase to ensure consistent evaluation, data collected through platform integration and pilot studies will inform future framework development. Usage patterns, stakeholder feedback, and evaluation effectiveness evidence will guide post-project framework evolution.

Research and Development Opportunities

Framework implementation will generate research opportunities including evaluation methodology advancement, stakeholder engagement optimisation, and comparative analysis across educational contexts contributing to broader understanding of educational AI evaluation.

Long-term Impact and Sustainability

GEARS establishes foundations for responsible AI evaluation in education extending beyond immediate project applications. By providing comprehensive, stakeholder-inclusive methodology, the framework contributes to broader European initiatives for evidence-based educational technology governance and responsible AI development in educational contexts.

7 Conclusions and Next Steps

The GEARS framework represents both a significant research contribution and a practical implementation tool for the GenAI4ED project. As a research deliverable, it addresses critical gaps in current GenAI evaluation approaches through systematic literature review and extensive stakeholder consultation, providing a comprehensive, stakeholder-inclusive methodology for evaluating generative AI tools in educational settings.

As a practical tool, GEARS serves three distinct operational functions within the project ecosystem: providing methodological foundation for platform development (WP3) through the Criteria Definition Module (T3.4), offering comprehensive reference sets for context-appropriate pilot study design (WP5), and generating systematic evidence for policy development (Task 5.6) on regulatory compliance, equity considerations, and educational quality.

The framework's mixed methods design preserves stakeholder perspective diversity rather than forcing artificial consensus, enabling informed decision-making about tool selection, adaptation, and implementation. This approach acknowledges that different stakeholder groups possess legitimate but potentially conflicting evaluation priorities, making tensions and trade-offs explicit for contextual interpretation.

GEARS therefore fulfils the dual objectives of advancing research understanding while providing immediate operational value for platform development, pilot study design, and policy recommendation formulation. The framework's modular structure ensures sustained relevance throughout the project lifecycle and supports adaptation for future educational AI evaluation contexts beyond the GenAI4ED project scope.

Through this combination of theoretical thoroughness and practical applicability, GEARS advances the project's goals of responsible innovation, user-centred design, and evidence-based policy development for GenAI integration in European educational systems, establishing a robust foundation for responsible GenAI integration in education while promoting evidence-based innovation in educational technologies.

Funded by the European Union

References

Bond, M., Khosravi, H., De Laat, M. *et al.* A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *Int J Educ Technol High Educ* **21**, 4 (2024). https://doi.org/10.1186/s41239-023-00436-z

Chen, K., Tallant A.C., Selig, I. (2025). Exploring generative AI literacy in higher education: student adoption, interaction, evaluation and ethical perceptions. *Information and Learning Sciences*; 126 (1-2): 132–148. https://doi.org/10.1108/ILS-10-2023-0160

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., ... & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

European Data Protection Board. (2024). EDPB opinion on AI models: GDPR principles support responsible AI. European Data Protection Board. https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. *Applied Sciences*, *11*(12), 5467. https://doi.org/10.3390/app11125467

Holmes, W., Iniesto, F., Anastopoulou, S., & Boticario, J. G. (2023). Stakeholder Perspectives on the Ethics of AI in Distance-Based Higher Education. *The International Review of Research in Open and Distributed Learning*, *24*(2), 96–117. https://doi.org/10.19173/irrodl.v24i2.6089

Johnston, H., Wells, R.F., Shanks, E.M. *et al.* Student perspectives on the use of generative artificial intelligence technologies in higher education. *Int J Educ Integr* **20**, 2 (2024). https://doi.org/10.1007/s40979-024-00149-4

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., ... & Ibrahim, L. (2024). Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Karran, A.J., Charland, P., Trempe-Martineau, J. *et al.* Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. *npj Sci. Learn.* **10**, 44 (2025). https://doi.org/10.1038/s41539-025-00333-2

Khlaif, Z. N., Alkouk, W. A., Salama, N., & Abu Eideh, B. (2025). Redesigning Assessments for AI-Enhanced Learning: A Framework for Educators in the Generative AI Era. *Education Sciences*, *15*(2), 174. https://doi.org/10.3390/educsci15020174

O'Dea, X., Tsz Kit Ng, D., O'Dea, M., & Shkuratskyy, V. (2024). Factors affecting university students' generative AI literacy: Evidence and evaluation in the UK and Hong Kong contexts. *Policy Futures in Education*, *0*(0). https://doi.org/10.1177/14782103241287401

UNESCO. (2021). AI and education: Guidance for policy-makers. Paris: UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000376709

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Wang, J., Fan, W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanit Soc Sci Commun* **12**, 621 (2025). https://doi.org/10.1057/s41599-025-04787-y

Xia, Q., Weng, X., Ouyang, F. *et al.* A scoping review on how generative artificial intelligence transforms assessment in higher education. *Int J Educ Technol High Educ* **21**, 40 (2024). https://doi.org/10.1186/s41239-024-00468-z

Zastudil, C., Rogalska, M., Kapp, C., Vaughn, J., & MacNeil, S. (2023, October). Generative ai in computing education: Perspectives of students and instructors. In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

Zawacki-Richter, O., Marín, V.I., Bond, M. *et al.* Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *Int J Educ Technol High Educ* **16**, 39 (2019). https://doi.org/10.1186/s41239-019-0171-0

Appendices

Appendix A1: Student Evaluation Criteria

Table 1: GEARS framework related to students with 18 evaluation criteria

Group	Evaluation Criteria	Assessment Indicators (KPI)	Review Method	Source & Data Type
Students	Accessibility/Usability (S1)	Thematic ease-of-use score; % of students describing tool as "easy to use"	Thematic Qualitative Analysis	Coded interview transcripts; Survey open-text responses (Qualitative)
	Cognitive Load (S1)	NASA-TLX score ≤ 40/100	Output/Interaction Tracking and Analytics	Micro-survey TLX (Quantitative)
	Learning Support and Feedback Quality (added from Workshop) (S2)	% of students answering comprehension questions correctly; Avg. quiz score ≥ 75% on comprehension questions	Quiz and Benchmark Testing	Student quiz results; AI-generated instructional content (Quantitative)
	Study Efficiency (S2)	% reporting time savings or better outcomes	Output/Interaction Tracking and Analytics	Self-report reflections; Survey open source (Mixed)
	Confidence and Autonomy (S2)	% showing improved confidence	Likert Scale and Rubric-Based Feedback / Thematic Qualitative Analysis	Pre- and post-use confidence and autonomy rating (Quantitative)
	Metacognitive Awareness (S2)	≥ 70% correct explanations	Manual/Expert Review and Rubric Scoring	Student ability to explain AI-generated content (Mixed)
	Learning Outcome (Academic Impact) (S2)	Avg. performance increase; %	Likert Scale and Rubric-Based Feedback	Quiz results; Self-

	reporting improvement		assessment (Mixed)
Learner Skills Development (S2)	% of learners showing improvement in ≥ 2 skill areas ≥ 80%; Avg. score increase ≥ 20% gain	Assessment method not specified	Pre/post skill assessments (Quantitative)
Learner Well-being (S2)	% completing weekly well-being check-ins ≥ 85%; Avg. well-being sentiment score ≥ 3.5/5; % reporting high inclusion ≥ 80%	Assessment method not specified	Regular well-being check-ins; Social inclusion measures (Mixed)
Ethics/Trust and Accuracy (S3)	Avg. trust score; % rating ≥ 4/5	Likert Scale and Rubric-Based Feedback	Trust and reliability assessments (Quantitative)
Overreliance Awareness (S3)	% recognizing when to limit AI use	Observation-Based Evaluation	Behavioural observation during AI use (Mixed)
Ethical Understanding (S3)	% with ≥ 70% on ethical use assessments	Ethics, Policy, and Safety Evaluation	Ethics assessment; Case discussions on academic honesty, bias, diversity (Mixed)
AI Transparency Comprehension (S3)	≥ 80% correct answers in AI identification quiz	Quiz and Benchmark Testing	AI content identification assessments (Quantitative)
Personalisation Effectiveness (S4)	Likert score ≥ 4/5 on personalised adaptive learning paths and feedback	Likert Scale and Rubric-Based Feedback	Personalised learning paths and feedback evaluation (Quantitative)

Collaboration Support (S4)	% reporting improved collaboration	Observation-Based Evaluation	Group work enhancement assessment (Mixed)
Social Presence/Peer Dialogue (S4)	≥ 2 constructive comments per student	Output/Interaction Tracking and Analytics	Peer interaction quality around AI use (Mixed)
Engagement	Average engagement score on Likert scale; % of students rating ≥ 4/5	Likert Scale and Rubric-Based Feedback	Self-reported engagement measures (Quantitative)
Motivation, Curiosity and Creativity	Avg. score ≥ 4/5; % reporting increased curiosity	Likert Scale and Rubric-Based Feedback	Inspiration and idea generation assessment (Quantitative)

Appendix A2: Teacher Evaluation Criteria

Table 2: GEARS framework related to teachers with 12 evaluation criteria

Group	***Evaluation Criteria***	***Assessment Indicators (KPI)***	***Review Method***	***Source & Data Type***
Teachers	Curriculum Alignment and Learning Objectives (T2)	% of lesson plan elements aligned with official curriculum standards	Manual/Expert Review and Rubric Scoring	AI-generated content analysis; Curriculum standards mapping (Mixed)
	Pedagogical Soundness (T2)	Average score on expert evaluation rubric for age appropriateness, learning depth, instructional value	Manual/Expert Review and Rubric Scoring	Expert pedagogical review; Activity quality assessment (Mixed)
	Professional Growth (T2)	% increase in AI tool confidence and self-rated tech integration skill	Likert Scale and Rubric-Based Feedback	Pre/post confidence surveys; Professional competence

			assessment (Quantitative)
Differentiation, UDL and Accessibility (T2)	≥ 1.5 variants per student for adaptive materials	Output/Interaction Tracking and Analytics	Material adaptation tracking; Accessibility feature usage (Quantitative)
Time Efficiency (T1)	Average minutes saved per lesson; % reduction in planning time	Output/Interaction Tracking and Analytics	Time-on-task comparison; Planning duration analysis (Quantitative)
Productivity (T1)	% of tasks completed faster or improved through AI	Output/Interaction Tracking and Analytics	Task completion tracking; Efficiency measures (Quantitative)
Well-being & Workload (T1)	Avg. reduction in stress level; % reporting better work-life balance; Job-security stress level	Likert Scale and Rubric-Based Feedback	Stress/burnout surveys; Work-life balance assessment (Quantitative)
Connection (T1)	% of users interacting with ≥ 2 other users weekly (≥ 70%); % of tasks using integrated workflows (≥ 80%)	Assessment method not specified	User interaction tracking; Workflow integration analysis (Mixed)
AI Literacy Support (T4)	≥ 80% of teachers scoring competent (Likert ≥ 4)	Likert Scale and Rubric-Based Feedback / Quiz and Benchmark Testing	AI competency assessment; Prompt effectiveness evaluation (Mixed)
Data-driven Insights (T4)	Dashboard usefulness score (Likert ≥ 4 and rubric ≥ 80%)	Likert Scale and Rubric-Based Feedback	Dashboard usability surveys; Report quality assessment (Quantitative)

Ethical & Privacy Management (T3)	100% GDPR and EU AI Act compliance requirements met	Manual/Expert Review and Rubric Scoring	Compliance audit; Privacy protection assessment (Quantitative)
Collaborative Teaching	≥ 30% of teachers involved in co-design activities	Output/Interaction Tracking and Analytics / Likert Scale and Rubric-Based Feedback	Collaboration participation tracking; Co-design engagement surveys (Mixed)

Appendix A3: Parent/Guardian Evaluation Criteria

Table 3: GEARS framework related to parents/guardians with 12 evaluation criteria

Group	***Evaluation Criteria***	***Assessment Indicators (KPI)***	***Review Method***	***Source & Data Type***
Parents	Dependency Prevention (P2)	% of tasks completed independently without direct AI answers	Role-play scenarios with different tool configurations	Work independence assessment; Task completion analysis (Mixed)
	Parental Control Features (P1)	Number of available control features / Total desired features	Workshop evaluation of control interface usability	Control interface evaluation; Usability assessment (Mixed)
	Age-Appropriate Safeguards (P1)	Pass/fail rate on age-appropriate safety checks	Simulated scenarios testing safety responses	Safety protocol testing; Age-compliance verification (Quantitative)
	Knowledge Retention (P2)	% knowledge retained after 1 week, 1 month, 3 months	Memory games and recall exercises in workshop setting	Retention testing; Longitudinal memory assessment (Quantitative)
	Critical Thinking Skills (P2)	Score improvement on critical thinking rubrics	Structured debates and analysis exercises	Critical thinking assessment; Analytical skill evaluation (Mixed)

	Emotional Intelligence Impact (P3)	EQ score changes and social interaction quality ratings	Role-playing scenarios and peer interaction exercises	EQ testing; Social interaction assessment (Mixed)
	Independent Problem-Solving (P2)	% of problems solved independently within time limits	Individual challenge rounds without AI tools	Problem-solving assessment; Independence testing (Quantitative)
	Actionable Insights (P5)	Parent satisfaction score with recommendations and implementation rate	Workshop evaluation of recommendation clarity and actionability	Recommendation assessment; Parent satisfaction surveys (Mixed)
	Progress Monitoring (P1)	Parent usage rate of monitoring features and perceived value	Mock parent dashboard evaluation sessions	Dashboard usability testing; Feature usage tracking (Mixed)
	Social Interaction Quality (P3)	Quality ratings of peer interactions and collaboration effectiveness	Group collaboration exercises and social skill demonstrations	Social skills assessment; Peer interaction evaluation (Mixed)
	Emotional Expression (P3)	Improvement in emotional vocabulary and expression complexity	Emotional expression exercises and communication challenges	Emotional development assessment; Expression capability testing (Mixed)
	Creative Independence (P4)	Originality scores and creative process independence ratings	Creative challenges and original thinking exercises	Creativity assessment; Independence evaluation (Mixed)

Appendix A4: AI experts

Table 4: GEARS framework related to AI experts with 14 evaluation criteria

Group	***Evaluation Criteria***	***Assessment Indicators (KPI)***	***Review Method***	***Source & Data Type***
AI Experts	Security & Attack Resistance (A1)	Jailbreak success rate (%)	Stress testing	Adversarial prompts

Output Consistency (A2)	Response Reproducibility Score (0–1)	Variance testing	Same prompts tested multiple times
Algorithmic Bias (A4)	Demographic Fairness Score	Bias detection tools	Bias detection datasets
Language Performance Parity (A4)	Cross-Language Accuracy Drop (%)	Multilingual evaluation	Multilingual benchmarks
Prompt Efficiency (A5)	Prompt-Response Effectiveness Score	Prompt scoring workshop	General prompt datasets & subject-specific prompts
Computational Efficiency (A2)	Tokens/sec, latency, memory use	Performance benchmarking	System performance logs & benchmark metrics
Domain Knowledge Accuracy (A3)	Subject-Specific Error Rate	Domain tests + reviews	Domain benchmarks
Reasoning Quality (A3)	Logical Consistency Score	Reasoning pathway analysis	Reasoning benchmarks
Safety System Robustness (A1)	Harmful Content Detection Rate	Safety system stress testing	Safety benchmarks
Factual Accuracy (A3)	Hallucination Rate (%)	Fact-checking tools	Hallucination benchmarks
Data Privacy Compliance (A1)	Privacy Risk Score	Privacy compliance assessment	Training data audits & privacy compliance documents
System Reliability (A2)	Deployment Stability Score	Reliability testing + failure analysis	System logs
Version Consistency (A2)	Model Update Impact Score	Version comparison analysis	Version history & performance metrics
Multimodal Integration (A5)	Cross-Modal Processing Quality	Cross-modal evaluation + integration testing	Multimodal benchmarks

Funded by the European Union

Project Partners

Funded by
the European Union