education sciences

MDPI

Article

Generating Personalized Viva Questions from Students' Own Assignments with Artificial Intelligence and Bloom's Revised Taxonomy: A Design-Science Study in Higher Education

Md Mainul Islam [1,*], Sraboni Datta [2] and Sutirtha Chowdhury [1]

[1] School of Science, Engineering and Digital Technologies, University of Southern Queensland, Toowoomba, QLD 4350, Australia; sutirthachowdhury1997@gmail.com
[2] Department of Management Information Systems, School of Business and Entrepreneurship, Independent University, Bangladesh, Dhaka 1229, Bangladesh; dattasraboni90@gmail.com
* Correspondence: mainul.islam@unisq.edu.au

Abstract

This study aimed to design and conduct an initial evaluation of a framework that helps lecturers prepare viva questions from each student's own written assignment, as generative AI weakens written work as evidence of understanding. Following design-science research, we developed the Assignment-to-Viva Assessment Framework, with an algorithm turning a submission into ten questions sequenced by intended Bloom's level, and VivaGen-AI, a browser-based tool in which the lecturer reviews every question before use. In an undergraduate materials-engineering course at an Australian university, we analyzed retrospectively 83 interview records from three offerings; in 15 of them, from the latest offering, lecturer-reviewed VivaGen-AI questions were used with students chosen by the lecturer. Instruments were a 32-case software test suite, a 150-mark interview rubric with four understanding bands, and operational criteria for Bloom's levels. On moderated marks, pilot students scored lower than the 68 other records (mean difference 18.6 marks, 95% CI [8.7, 28.5]), but there was no statistically detectable difference from the 16 other interviews of the latest offering (−3.5 marks, 95% CI [−13.7, 6.8]). Lower marks alone do not show a more accurate assessment. We recommend lecturer review of every question, expert validation of question quality, and a randomized comparison before wider adoption.

Keywords: generative artificial intelligence; authentic assessment; oral assessment; viva voce; Bloom's revised taxonomy; automatic question generation; academic integrity; design-science research

Academic Editor: Diego Vergara

Received: 26 September 2026
Revised: 14 November 2026
Accepted: 19 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Generative-AI systems can now produce fluent essays, technical reports, literature reviews, and program code from a brief instruction (Ateeq et al., 2024; Memarian & Doleck, 2026). This has weakened a long-standing assumption in assessment design: that a submitted written artefact is evidence of the student's own understanding. Detection-based responses to this problem are unreliable and systematically biased against particular groups of writers (W. Liang et al., 2023; Perkins et al., 2024), which has shifted scholarly

attention toward redesigning assessment itself around authentic, defensible demonstrations of learning (Perkins et al., 2025).

Oral assessment, and the viva voce in particular, is one of the most consistently AI-resistant formats, because it requires a student to construct and defend an explanation in real time (Curbo et al., 2024). Its main practical limitation is scale: preparing a personalized, cognitively calibrated set of oral questions for every student submission is time-consuming, and this preparation burden is a documented source of educator workload and technostress when institutional support is inadequate (Khlaif et al., 2025). Existing AI-assisted assessment tools mostly address adjacent problems (plagiarism detection, automated grading of closed-response items or generic question banks) rather than generating oral questions from an individual student's own submitted work (Yaacoub et al., 2025; Duong-Trung et al., 2024).

Taken together, these developments create a practical problem for lecturers who wish to use oral assessment at scale, and an open question for research on AI-assisted assessment design. The research problem and its significance are set out below, followed by the objectives and research questions of the study, its scope and the definitions of its key terms.

1.1. Research Problem

Prior studies document three aspects of the problem. First, written work is becoming weaker evidence of a student's own understanding: generative AI can complete many written tasks (Ateeq et al., 2024; Cotton et al., 2024), and AI-text detectors cannot reliably identify AI-generated text. W. Liang et al. (2023) found that seven detectors misclassified essays by non-native English writers as AI-generated at a mean rate of 61%, and Perkins et al. (2024) found that detectors identified unmodified AI-generated text with a mean accuracy of 39.5%, which fell further after simple adversarial edits. Second, oral assessment offers a more secure alternative. Curbo et al. (2024) reported that pairing take-home examinations with a viva was accompanied by a rise in pass rates and was valued by students for learning, and Akimov and Malin (2020) found that an online oral examination provided a valid, authentic assessment that reduced opportunities for cheating, although it raised student anxiety and required careful design to be reliable and fair. Australia's higher education regulator has likewise encouraged interactive oral assessment and the securing of assessment at critical points of a program (Lodge et al., 2023). Third, oral assessment is costly to prepare well: questions must be specific to each submission, and AI assistance in assessment design still requires substantial expert review (May et al., 2025; Khlaif et al., 2025).

The authors' own experience in a third-year undergraduate materials-engineering course, which the first author coordinates and teaches, illustrates the problem. In this course, an individual interview is worth 150 of the 1000 course marks. Preparing questions that are specific to each student's case study takes considerable time, and with up to 46 enrolled students in an offering (2024), this preparation competes with other teaching duties. The course records also show that interview marking practice varied between offerings: marks were awarded in steps of 10 in 2024, only as 120, 130 or 150 in 2025, and in steps of 5 in 2026 (Section 9.1), which suggests that questioning and marking were not standardized across examiners and years. The research problem addressed in this study is therefore how lecturers can prepare individualized, cognitively structured viva questions for every submission at reasonable cost, while keeping the lecturer in control of what is asked and how answers are judged.

1.2. Research Objectives

This paper addresses that gap. It presents the Assignment-to-Viva Assessment Framework (AVAF), which combines authentic assessment principles, Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) and a human-in-the-loop design to convert a single student submission into a ten-question viva, sequenced by intended cognitive level, for lecturer review. AVAF is operationalized as the Assignment-to-Viva Question-Generation Algorithm (AVQGA), and demonstrated in VivaGen-AI, a browser-based reference implementation. Consistent with design-science research (Hevner et al., 2004; Peffers et al., 2007), the paper reports three things clearly and separately, and does not conflate them: a working artefact (framework, algorithm and software), demonstrated on a representative assignment; a real empirical baseline, drawn from existing interview-mark data for three offerings (2024–2026) of one undergraduate course (83 records, 68 of them from interviews with examiner-prepared questions); and a preliminary outcome pilot, comprising the remaining 15 of those records, in which lecturer-reviewed VivaGen-AI questions were used in the regular course interviews of 15 students in the most recent offering of the same course. This preliminary pilot is informative but is not the confirmatory evaluation this paper calls for: although it is drawn from the same course as the baseline, its 15 students were chosen by the lecturer from the 2026 offering rather than at random; its sample is small, and it does not yet include the lecturer/expert rating of question quality or a randomized within-offering comparison, both of which remain specified but not yet executed. The paper therefore does not claim that VivaGen-AI has been shown, under controlled conditions, to improve student outcomes, lecturer workload or assessment validity relative to existing practice; it reports the real evidence gathered so far, at the confidence level that evidence supports, and specifies precisely what remains to establish a stronger claim.

The aim of the study is to design a lecturer-controlled framework for generating personalized viva questions from students' own assignments, and to conduct an initial evaluation of it. The specific objectives are: (O1) to design a framework (AVAF) and algorithm (AVQGA, with its logical decomposition MSVGA) that convert a student's own written submission into a set of ten viva questions sequenced by intended level of Bloom's Revised Taxonomy, for lecturer review; (O2) to implement the framework as a working, browser-based tool (VivaGen-AI) and to test whether it behaves as specified; (O3) to describe how interview marks were distributed in three offerings of the course, and how interview marks relate to performance in the rest of the course; (O4) to compare, in a preliminary pilot, interview outcomes when lecturer-reviewed VivaGen-AI questions were used with those of the other interviews in the three offerings; and (O5) to specify, and where completed to report, a direct expert evaluation of question quality and a randomized design for testing effectiveness.

1.3. Research Questions

The objectives translate into five research questions. RQ1: How can a student's own written submission be converted into a lecturer-reviewable set of ten viva questions sequenced by intended level of Bloom's Revised Taxonomy? (O1; Sections 4, 5 and 7). RQ2: Does the reference implementation behave as specified, and what are its functional limitations? (O2; Sections 6 and 8, Appendix D). RQ3: How were interview marks distributed across understanding bands in three historical offerings of the course, and how are they related to performance in the rest of the course? (O3; Sections 9.1–9.2). RQ4: How did interview outcomes in a small pilot that used lecturer-reviewed VivaGen-AI questions compare with those of the other interviews in the three offerings? (O4; Section 9.3). RQ5: How do experts judge the factual accuracy, source traceability, Bloom alignment, and

suitability for oral delivery of the generated questions? (O5; the protocol is specified in Section 8, and its results will be reported separately).

1.4. Significance of the Study

The paper's contribution is fourfold: (i) AVAF, a theoretical framework that reframes generative AI from a threat to academic integrity into a means of supporting oral checks of students' understanding of their own work; (ii) AVQGA, a computational algorithm, and its logical decomposition into the Multi-Stage Viva Generation Architecture (MSVGA), which turns the framework into a repeatable process; (iii) VivaGen-AI, a working, browser-based, software artefact that is provider-independent at the input/output level; and (iv) an interpretation scheme for interview marks, applied descriptively to 83 historical interview records to provide a reference distribution, not a validation, for a subsequent lecturer-validation and student pilot study.

The study is also significant in practice. For lecturers, it offers a way to prepare questions that are specific to each submission without writing every question by hand, while leaving the questions and the judgement with the lecturer. For institutions, it responds to regulatory guidance in Australia that encourages interactive oral assessment and secured assessment at key points of a program (Lodge et al., 2023), and it states plainly the privacy, cost and validity conditions under which such a tool can be used (Sections 6 and 11). For researchers, it provides a transparent account of the artefact, including the complete prompt, parsing rules and test results (Appendices B and D), and a worked example of why pooled historical comparisons can mislead when baseline offerings differ (Section 9.3).

1.5. Scope and Delimitations

The study is deliberately bounded in four ways. Thematically, it concerns the generation of oral questions from a student's own written submission and their use in a lecturer-conducted viva; it does not address automated grading, detection of AI-generated text, or determination of authorship. Spatially, all empirical data come from one undergraduate engineering course at one Australian university, delivered on two campuses and online. Temporally, the baseline covers the 2024, 2025 and 2026 offerings, each in the same teaching period of the year, and the pilot took place within the 2026 offering. Technically, the evidence concerns one reference build of VivaGen-AI (Section 6), used with the ChatGPT (free version; OpenAI, San Francisco, CA, USA) web interface, and text-based PDF and DOCX submissions. The limitations that affect internal validity, external validity, and the generalizability of the findings are discussed in Section 11, after the results to which they apply.

1.6. Conceptual and Operational Definitions

The key terms are defined below, first conceptually and then operationally, i.e., as they are used and measured in this study.

Oral assessment (viva voce): Conceptually, assessment in which the student's response is given in spoken form, usually in dialogue with one or more examiners (Joughin, 1998). Operationally, a lecturer-conducted interview in which the student answers ten questions about their own submission, with follow-up probes, and is awarded a mark out of 150 using the rubric in Appendix C.

Authentic assessment: Conceptually, assessment that requires students to apply knowledge in realistic tasks and to demonstrate understanding rather than recall (Wiggins, 1990; Villarroel et al., 2018). Operationally, assessment in which students explain, justify, and defend the methods, calculations, and conclusions of their own submitted work.

Bloom's Revised Taxonomy: Conceptually, a framework that classifies cognitive processes into six levels (Remember, Understand, Apply, Analyze, Evaluate, and Create) and knowledge into four types (Anderson & Krathwohl, 2001; Krathwohl, 2002). Operationally, the *intended level* of a question is the level the prompt requests for its position in the sequence (Appendix B.1), and its *demonstrated level* is the level assigned by raters using the criteria in Appendix A (Table A1).

Difficulty label: The fixed Easy, Medium, or Hard label attached to each position in the ten-item sequence. It describes intended cognitive demand and is not a measure of empirical item difficulty (Section 5).

Generative AI and large language model (LLM): Conceptually, AI systems that produce new text in response to an instruction; an LLM is a model trained on large text corpora to generate language. Operationally, the ChatGPT web interface, into which the lecturer pastes the prompt built by VivaGen-AI (Section 6).

Lecturer review (human-in-the-loop): Conceptually, a design in which people retain control over consequential decisions made with AI support (Shneiderman, 2022). Operationally, the lecturer reads every generated question in Teacher Mode before the viva and decides whether to use it, request a revised set, or change it.

Interview mark and understanding band: The interview mark is the mark out of 150 awarded by the examiner. The understanding band is one of four categories (Strong Understanding, Good Understanding, Partial Understanding and Further Verification Needed) obtained retrospectively from the moderated interview mark using the level boundaries of the course rubric (120, 90 and 60 marks; Section 6). The raw mark is the mark awarded by the examiner at the interview; the moderated mark is the mark recorded in the course gradebook after moderation by the course coordinator (Section 8). Unless stated otherwise, interview marks are moderated marks.

Academic integrity, understanding and authorship: Academic integrity refers to honest and responsible conduct in academic work. This study does not measure integrity or authorship. It measures demonstrated understanding of one's own submission in an interview, which is one source of evidence that institutions may use alongside others.

Course, offering, cohort and trimester: The course is a third-year undergraduate materials-engineering course, which, in line with the conditions of the ethics approval, is not named; an *offering* is a single delivery of the course in a given year and teaching period (for example, the 2024 offering); a *cohort* is the group of students enrolled in one offering; and *trimester* refers only to the teaching period. The historical baseline therefore comprises three offerings (2024, 2025 and 2026) of one course, and the pilot comprises the regular course interviews of 15 students in the 2026 offering in which lecturer-reviewed VivaGen-AI questions were used. The 83 analyzed baseline records therefore comprise the 15 pilot records and 68 other records, and the 31 analyzed records of the 2026 offering comprise the 15 pilot interviews and 16 other interviews.

Intended, implemented and evaluated functions: A function of the platform is *intended* if it is part of the AVAF design, *implemented* if it is present in the reference build, and *evaluated* if it has been tested with results reported in this paper. Section 6 maps each function to its status.

2. Relevant Prior Studies

2.1. Generative AI, Academic Integrity and the Case for Redesigned Assessment

Recent reviews agree that generative AI represents a qualitative change in the assessment problem, not merely an extension of existing plagiarism concerns, and that most institutional academic-integrity policy predates this shift (Ateeq et al., 2024; Cotton et al., 2024; Logachev, 2026; Perkins et al., 2025). Purely technical countermeasures are weak: AI-text detectors are demonstrably unreliable and biased against non-native English

writers (W. Liang et al., 2023), a finding echoed in experimental work and in systematic reviews (Perkins et al., 2024; Blessing, 2026). The scholarly consensus has consequently moved toward redesigning assessment around authentic, process-visible demonstrations of learning rather than relying on detection or prohibition (Ateeq et al., 2024; Memarian & Doleck, 2026; Dwivedi et al., 2023).

Several recent frameworks operationalize this shift. Perkins et al. (2025) revise the Artificial Intelligence Assessment Scale (AIAS) so that an AI-free task requires deliberate justification rather than being the default, while cautioning that blanket bans can disadvantage students with disabilities. May et al. (2025) report a design-based research process for building AI-supported assessment items, in which subject-matter experts still flagged three quarters of AI-drafted items as requiring revision, underscoring that AI assistance in assessment design still depends on sustained expert oversight. Related work on experiential learning for authentic assessment (Salinas-Navarro et al., 2024) similarly argues that generative AI should support the design of authentic, constructively aligned learning and assessment, rather than serving only to draft text or answer questions, which motivates assessment formats in which students must explain and apply their own work.

2.2. Bloom's Revised Taxonomy and Automatic Question Generation

Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) remains the dominant cognitive framework used to keep redesigned assessment from collapsing into recall. Empirical audits of existing question banks consistently find lower-order items far outnumbering higher-order ones, including at university level (Belkahla & Azmi, 2026), and conceptual analyses of assessment in the generative-AI era argue that the higher levels of the taxonomy need to be deliberately reframed and targeted (Lubbe et al., 2025). Work that explicitly aligns automatic question generation (AQG) with Bloom's levels reports measurable gains over unconstrained prompting; for example, Duong-Trung et al.'s (2024) BloomLLM, fine-tuned on cognitively labelled questions, which students preferred to questions produced by general-purpose prompting of GPT-4. An earlier systematic review of automatic question generation found little work on generating questions of controlled difficulty (Kurdi et al., 2020), which is why this study treats cognitive level and difficulty as separate properties (Section 5). Even so, transformer-based generators continue to struggle specifically at the higher cognitive levels, producing acceptable factual items but weaker Analyze-, Evaluate-, and Create-level questions (Timosh et al., 2025), and Yaacoub et al. (2025) found that automated classification of AI-generated questions by Bloom's level was less accurate at the higher levels of the taxonomy than at the lower ones. Across this literature, a further, more specific gap recurs: almost none of the reviewed AQG systems generate questions from an individual student's own submitted assignment, as distinct from a generic textbook or item bank (Hwang et al., 2024). Recent viva-oriented systems that do use the student's own submission are discussed in Section 2.3.

2.3. Oral Assessment and Human-Centered, Multi-Agent AI Architectures

Oral and viva voce assessment is repeatedly identified in the literature as one of the more AI-resistant assessment formats, because it requires real-time construction and defense of an explanation (Curbo et al., 2024). At the same time, the preparation burden this format places on lecturers is a documented driver of workload and technostress when institutional support is inadequate (Khlaif et al., 2025), and reviews of AI-assisted assessment design converge on the conclusion that such tools succeed only where teacher readiness, infrastructure and human oversight are preserved (J. Liang et al., 2025; Deepshikha, 2025).

2.4. Commentary on Relevant Prior Studies

This human-centered requirement is echoed in the human-centered AI literature more broadly (Shneiderman, 2022; UNESCO, 2023), and in work on high-stakes assessment outside education that cautions against uncritical confidence in generative-AI-based assessment (Hogan & Corăbian, 2025). It also motivates the use of decomposed, multi-agent architectures for assessment-related reasoning, since staged processing with defined checkpoints is easier to audit than a single undifferentiated model call. Hou et al. (2025) report a directly relevant precedent: a multi-agent LLM architecture for conversation-based assessment that assigns distinct student-facing (expert, peer) and behind-the-scenes (formative assessor, summative assessor) roles to separate reasoning stages, coordinated by a non-LLM controller, an architectural pattern closely related to the staged decomposition described for MSVGA in Section 5 of this paper, though implemented there as genuinely separate agent calls rather than within one prompt.

Three recent systems are closer to the present work, because they generate oral questions from an individual student's own submission. AutoViva (Cao & Zahid, 2025) is a Moodle plugin that calls a commercial LLM when a student submits coursework and generates viva questions tailored to the coursework specification and to that student's submission; its evaluation reports student acceptance and system scalability. Church et al. (2025) describe a viva voce simulator in which an LLM examiner questions the student about an uploaded written submission, opening with a question on a specific detail of the text and following with probes that ask why and how, and then passes the transcript to a human examiner for the final judgement; it was demonstrated as a proof of concept on a single essay. Ipeirotis and Rizakos (2026) report a voice-agent oral examination in an undergraduate course, in which the agent discusses each student's project and a case study, and a panel of LLMs grades the transcript against a rubric, with human audit of flagged cases, across two cohorts.

Relative to these systems, the present work makes three narrower additions. First, AVAF fixes the cognitive structure of the viva in advance: ten items are allocated across the six levels of Bloom's Revised Taxonomy, from recall to justification and defense, whereas the systems above generate open or adaptive questions without explicit cognitive targeting. Second, AVAF requires the lecturer to review and approve every question before it reaches a student, and the lecturer, not an AI examiner, conducts the live viva. In the published descriptions of AutoViva, questions are generated automatically at submission, and no lecturer approval step is reported; in the other two systems, human judgement enters after the AI-led examination, through transcript review or audit. Third, each generated item carries a model answer, a reference to the section of the submission it was drawn from, and a follow-up probe, so that the examiner can check a live answer against the student's own work. These are design differences. The present paper does not show that they produce better questions or more valid judgements than the alternatives; that is a question for the evaluation specified in Section 8. Table 1 summarizes these differences. No study identified in this review combines these three features, and this combination is the specific gap AVAF, AVQGA and VivaGen-AI are designed to address.

Table 1. Comparison of AVAF/VivaGen-AI with the closest systems that generate oral questions from an individual student's own submission.

System	Input Material	Submission-Specific	Cognitive Targeting	Who Conducts the Oral Assessment	Human Review	Evaluation Reported

AutoViva (Cao & Zahid, 2025)	Coursework specification and the student's submission (Moodle plugin)	Yes	Not reported	Questions generated automatically at submission	No lecturer approval step reported	Student acceptance; scalability
Viva voce simulator (Church et al., 2025)	Uploaded written submission	Yes	Open question on a specific detail, then "why/how" probes; no taxonomy	LLM examiner	Human examiner reviews the transcript afterwards	Proof of concept on one essay
Voice-AI oral examination (Ipeirotis & Rizakos, 2026)	Student's course project and course case studies	Yes (project)	Rubric dimensions; no taxonomy reported	AI voice agent	Human audit of flagged gradings by an LLM panel	Two cohorts (36 and 37 students): grading reliability, student survey
VivaGen-AI (this study)	Student's PDF/DOCX submission (first 12,000 characters)	Yes	Fixed ten-item allocation across six Bloom levels (intended)	Lecturer	Lecturer reviews every question before use	Demonstration; descriptive baseline; non-randomized pilot within routine assessment (N = 15); expert validation planned

Entries for other systems are taken from their published descriptions; "not reported" means the feature is not described there, not that it is absent.

Table 2 summarizes the key empirical studies reviewed in this section, giving for each its objective, location, method, sample, instrument, main findings and recommendations, as reported by its authors.

Table 2. Summary of key empirical studies reviewed in Section 2.

Study (Location)	**Objective**	**Method and Instrument**	**Sample**	**Key Findings**	**Recommendation(s)**
W. Liang et al. (2023), USA	Test whether GPT detectors misclassify non-native English writing	Experimental evaluation of detectors; seven publicly available GPT detectors	91 TOEFL essays by non-native writers; 88 US Grade 8 essays	Mean false-positive rate of 61.2% for TOEFL essays; US essays classified almost perfectly	Avoid detectors in evaluative or educational settings, especially for non-native writers
Perkins et al. (2024), tests of online tools (no geographic setting)	Test whether simple adversarial edits defeat AI-text detectors	Experimental; seven detectors, including Turnitin and GPTZero	114 texts (15 AI-generated, 89 modified, 10 human-written); 805 tests	Mean accuracy 39.5% on unmodified AI text, falling by about 17 points after modification	Do not use detectors to decide integrity cases; redesign assessment
Curbo et al. (2024), Sweden	Compare in-person examination with take-home examination plus viva	Comparison of cohorts; course grades and course evaluation survey	45 students (2019); 127 students (2020–2022), biomedical laboratory science	Pass rate rose from 27% to 63%; viva valued for learning, though stressful for some	Authors conclude that take-home examinations with viva follow-up support deeper learning

May et al. (2025), USA	Show how ChatGPT can support test development and gather content validity evidence	Design-based research; expert review with nominal group technique; test review questionnaire (4 yes/no criteria and comments)	12 AI-drafted multiple-choice items; 5 subject-matter experts	9 of 12 items (75%) revised or deleted after expert consensus	Keep humans in the loop; have content experts and psychometricians review AI-drafted items
Duong-Trung et al. (2024), Germany	Generate Bloom-aligned questions linked across levels	Supervised fine-tuning of GPT-3.5-Turbo (BloomLLM); student evaluation; questionnaire comparing BloomLLM and GPT-4 question sets	1026 lecturer-written questions (29 topics); 46 master's students as raters	Students preferred BloomLLM questions to GPT-4 questions	Fine-tune on expert, curriculum-specific data and keep the model updated
Yaacoub et al. (2025), France and Lebanon	Test whether Bloom's levels of AI-generated questions can be classified automatically	Supervised classification and linguistic analysis; four classifiers, including DistilBERT; readability and lexical measures	3691 AI-generated computer-science questions labelled by level	Simpler classifiers least accurate at higher levels; DistilBERT 91% validation accuracy	Integrate Bloom's Taxonomy into AI question tools; use advanced classifiers
Belkahla and Azmi (2026), Morocco	Analyze the cognitive demand of summative examination questions	Content analysis with the two-dimensional Bloom framework; coding scheme based on Bloom's Revised Taxonomy	162 questions from one university English department	Understand–Conceptual most common (21.0%); higher-order and metacognitive items scarce	Scaffold examinations from foundational to higher-order tasks
Cao and Zahid (2025), UK	Automate viva questions to authenticate coursework	System development and evaluation; student acceptance measures; response time and throughput	Students (number not stated in the abstract); load tests	High student acceptance; stable performance as concurrent users increased	Not stated in the published abstract
Church et al. (2025), USA	Simulate a viva on a student's written submission with an LLM	Proof-of-concept demonstration; LLM examiner (Gemini 2.5 Flash) producing a structured assessment	One undergraduate essay	Feasible, but the examiner was overly affirming and cited page numbers it could not see	Further testing; the human examiner keeps the final judgement
Ipeirotis and Rizakos (2026), USA	Scale personalized oral examinations with a voice-AI agent	Deployment in two course cohorts; voice-agent examination; three-LLM grading panel with rubric; student survey	36 and 37 undergraduate students	Grader agreement α = 0.86–0.90 after deliberation; 66–70% felt the exam tested understanding	Keep mandatory human audit; validate against expert judgement

Entries are taken from the published articles or, where only the abstract was accessible, from the published abstract. Numbers of items are written as reported.

Three points of agreement and three points of difference emerge from these studies. The present study agrees with the detector studies (W. Liang et al., 2023; Perkins et al., 2024) that the response to generative AI should be a change in assessment design rather than reliance on detection; with the oral-assessment studies (Curbo et al., 2024; Akimov & Malin, 2020) that a viva can provide evidence of understanding that a written submission alone cannot; and with May et al. (2025) and Duong-Trung et al. (2024) that AI-generated

assessment items need expert review and explicit cognitive targeting. It differs, first, in its input: unlike question-generation studies that work from textbooks, item banks, or course topics (Duong-Trung et al., 2024; Yaacoub et al., 2025), it generates questions from each student's own submission. Second, it differs in who conducts the viva: unlike AutoViva, the viva simulator and the voice-agent examination (Cao & Zahid, 2025; Church et al., 2025; Ipeirotis & Rizakos, 2026), it leaves both the questioning and the judgement to the lecturer. Third, it differs in its evidence: it reports a field setting in an engineering course, with historical baselines, a test suite and the complete prompt, rather than a single demonstration or a classifier study. What distinguishes the study is therefore the combination of these features (Table 1), not any single one, and its evidence on effectiveness remains preliminary (Section 9).

3. Research Methodology

3.1. Research Design

This study follows a design-science research (DSR) methodology (Hevner et al., 2004; Peffers et al., 2007), which is appropriate where the research goal is both to explain a problem and to build and evaluate a working artefact. The study develops and evaluates three related artefacts: (1) AVAF, a conceptual framework combining authentic assessment, Bloom's Revised Taxonomy and human-centered AI; (2) AVQGA, the computational algorithm that operationalizes AVAF; and (3) VivaGen-AI, the browser-based software that implements AVQGA. A pragmatic research philosophy underlies this combination of theory-building, artefact construction and empirical evaluation within one study (Creswell & Creswell, 2018).

Figure 1 sets out the following workflow: literature review and problem identification; requirements analysis; framework development (AVAF); algorithm development (AVQGA/MSVGA); software development (VivaGen-AI); demonstration on a representative assignment; and evaluation. Consistent with DSR practice, evaluation in this study combines three distinct kinds of evidence rather than one composite claim: software testing, confirming that the platform's own code behaves as specified, including its documented limitations (text extraction, prompt construction, rejection of malformed or placeholder replies, and storage and export of records; the software does not enforce exactly ten questions or check their labels); an empirical baseline, drawn from real institutional interview data for three offerings, 68 of whose 83 records come from interviews with examiner-prepared questions; and a preliminary outcome pilot, in which VivaGen-AI was used for real interviews with a small sample of students. A randomized, within-offering comparison and a lecturer/expert validation of question quality (Section 8) remain to be executed. These are reported separately throughout this paper and should not be conflated: the baseline in Sections 9.1 and 9.2 characterizes interview marks in the three offerings, and the pilot in Section 9.3 is a preliminary, same-course but non-randomized comparison with a group of students chosen by the lecturer, not a controlled measurement of VivaGen-AI's effect relative to the existing process. The paper is therefore primarily a design-science artefact paper: its contribution is the framework, the algorithm and the working tool, together with evidence of their feasibility in routine assessment. Neither the baseline nor the pilot is offered as evidence of the validity, reliability, diagnostic accuracy or comparative effectiveness of the generated questions.

In terms of the knowledge contribution framework of Gregor and Hevner (2013), the study offers an improvement: a new solution to a known problem. Its evaluation follows the Framework for Evaluation in Design Science (Venable et al., 2016). Formative and artificial evaluation (iterative refinement, the demonstration and the functional test suite) precedes naturalistic evaluation with real users (the pilot), and summative naturalistic evaluation (the randomized within-offering comparison) is planned (Section 8).

https://doi.org/10.3390/educsci16122007

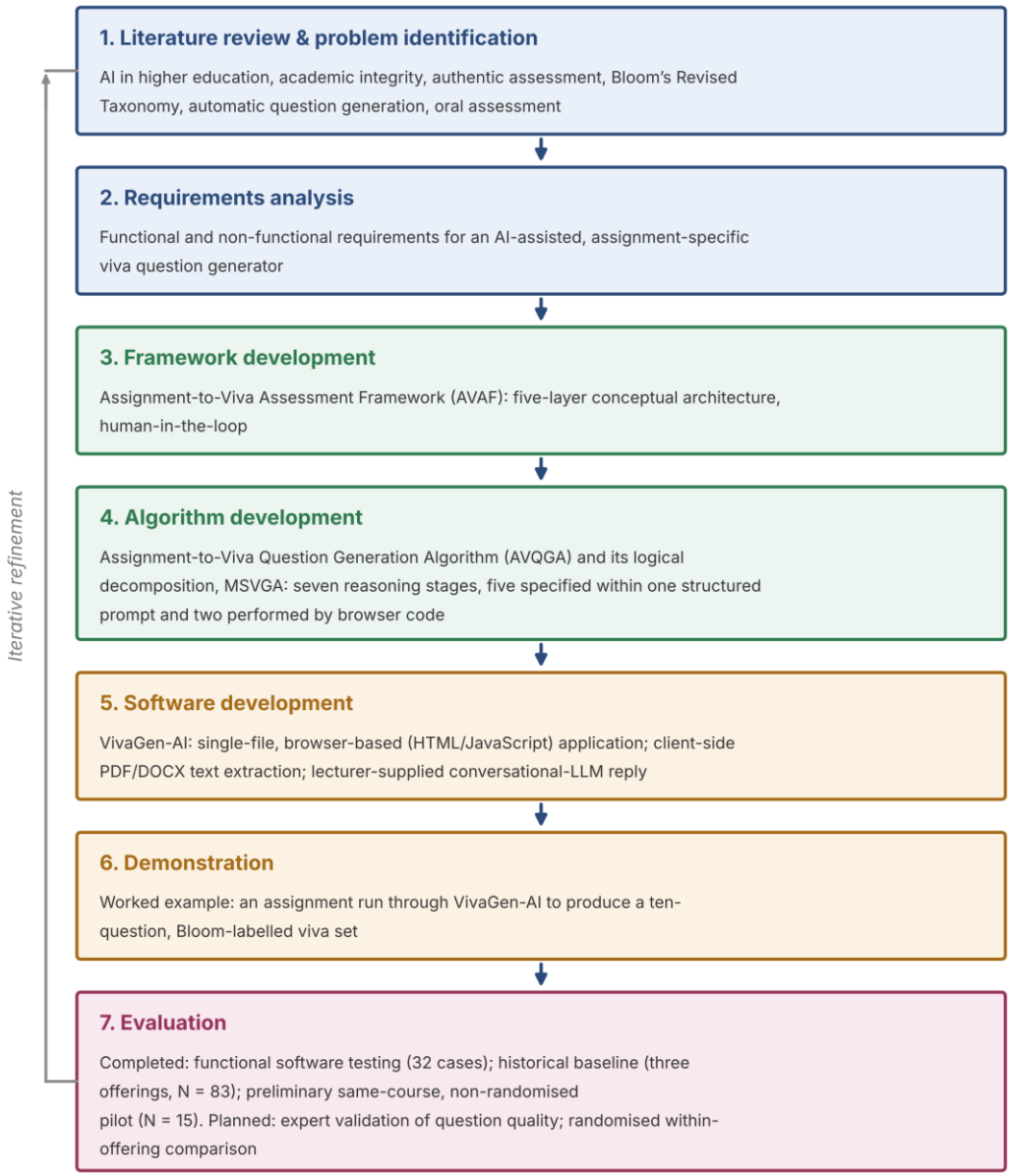


Figure 1. Design-Science Research workflow followed in this study. Stage 7 separates the evaluations completed in this paper from those still planned.

3.2. Population and Sample

The population of interest is undergraduate students enrolled in a third-year materials-engineering course at the University of Southern Queensland, delivered on two campuses and online, together with the lecturers who examine them. Four samples were drawn. (1) Historical baseline: all 96 student records from the 2024, 2025 and 2026 offerings were examined (a census of available records, not a sample), and the 83 records with an interview mark were analyzed; the 13 exclusions are reported in Section 9.1. (2) Pilot: in the 2026 offering, lecturer-reviewed VivaGen-AI questions were used in the regular course interviews of 15 students, chosen by the lecturer on the basis of their marks in earlier assessment components to include both higher- and lower-performing students; these 15 records are among the 83 analyzed baseline records (31 of them in 2026), were analyzed retrospectively, and the pilot group is therefore neither random nor representative of the whole cohort. (3) Demonstration: one de-identified group submission was selected as representative of the case-study assignment. (4) Expert validation (planned): 8–10 de-identified submissions, selected to cover the range of topics and marks, are to be

rated by a panel of at least five experts (Section 8). The software test suite is not a sample of people; its 32 cases were designed to cover every processing rule of the platform.

3.3. Research Instruments

Table 3 lists the seven instruments used in the study, with the purpose, initial form and number of items of each, how each was developed and reviewed, and the validity and reliability evidence available for each. The primary instrument, the question-generation prompt, is the artefact described in Sections 4–7.

Table 3. Research instruments: purpose, form, development, and validity and reliability evidence.

Instrument	Purpose	Form and Number of Items	Development and Expert Review	Validity and Reliability Evidence	Reported in
1. Question-generation prompt (AVQGA, run through VivaGen-AI)	Generate ten viva questions from one submission	Structured prompt; ten items per submission in a fixed sequence of Bloom's levels (Appendix B.1)	Developed iteratively by the authors over successive builds (Figure 1); the ten-item sequence was fixed from the first build; no external expert review before the pilot	Content: items specified against operational criteria (Table A1). Construct: authors' check found 7 of 10 demonstration items at the intended level (Table A2); independent expert classification planned (Section 8). Reliability: consistency across repeated generations not assessed	Sections 4–7; Appendixes A and B
2. Interview marking rubric	Award the interview mark	The course's existing interview rubric: four levels, each with a mark range and descriptor; 150 marks (Appendix C)	Existing course rubric, used unchanged; not separately reviewed for this study	Reliability: marks moderated by the course coordinator, who knew the question source; no double-marking; marking steps differed across offerings (Section 9.1)	Appendix C; Section 9
3. Four-band interpretation scheme	Convert an interview mark into an understanding band	Four bands with fixed thresholds (Section 6)	Level boundaries of the existing course rubric (80%, 60% and 40% of the interview mark), adopted by the authors with four band labels and applied retrospectively	Not validated; correlation with non-interview course marks r = 0.17 and inconsistent across offerings (Section 9.2)	Section 6
4. Operational criteria for Bloom's levels	Classify each question by cognitive level	Six levels, each with a criterion and typical stems (Table A1)	Derived from Anderson and Krathwohl (2001) and Krathwohl (2002); no external review before use	Applied by the authors to the demonstration set; inter-rater agreement to be reported with the expert validation (Section 8)	Appendix A
5. Expert rating form	Rate the quality of generated question sets	Six criteria rated on five-point scales, plus classification of the Bloom's level of each of the ten items	Criteria drawn from the question-generation evaluation literature (Kurdi et al., 2020); draft form, to be reviewed by the expert panel before use	Content validity: item- and scale-level content validity indices; reliability: Fleiss' κ for level classification and the intraclass correlation for ratings	Section 8

6. Functional test suite	Check that the software behaves as specified	32 test cases, each with a predefined expected result	Cases derived from each processing rule in the source code (Appendix B.3)	Covers every processing rule; two independent runs gave identical results	Section 8; Appendix D
7. Lecturer observation notes	Record qualitative observations in pilot interviews	Unstructured notes	Not applicable	Illustrative only; no reliability evidence	Section 9.3

3.4. Validity and Reliability

Content validity: The items of the question-generation prompt and the operational criteria for Bloom's levels were specified from the taxonomy itself (Anderson & Krathwohl, 2001; Krathwohl, 2002) and from the literature reviewed in Section 2. The expert rating form and, through it, the generated question sets are assessed by a panel that includes specialists in curriculum and instruction, measurement and evaluation, artificial intelligence in education and the subject discipline (Section 8). Each expert rates the relevance of every criterion and item, and content validity is summarized by the item-level content validity index (I-CVI) and the scale-level index averaged across items (S-CVI/Ave). With five or fewer experts, an item is retained only if all experts rate it relevant (I-CVI = 1.00), and with six to ten experts, an I-CVI of at least 0.78 and an S-CVI/Ave of at least 0.90 are required (Lynn, 1986; Polit & Beck, 2006). This review has not yet been carried out, and the indices will be reported with the expert validation.

Construct validity: For the generated questions, the construct of interest is the cognitive level a question demands. It is assessed by comparing the intended level of each item with the level assigned independently by the expert panel, reporting the proportion of agreement and Fleiss' κ across raters (Fleiss, 1971; Section 8). The authors' own check of the demonstration set found 7 of 10 items at the intended level (Appendix A). For interview marks, no independent criterion was available. Their correlation with the rest of the course mark is weak and inconsistent across offerings (Section 9.2), so the four understanding bands should be regarded as unvalidated descriptive categories.

Reliability: Interview marks were moderated by the course coordinator before entry in the gradebook, in all offerings and in the pilot (Section 8), but no double-marking was carried out, so the inter-rater reliability of interview marks could not be estimated; the marking steps used differed across the baseline offerings (Section 9.1), which indicates limited consistency of marking practice across years. Because the language model is stochastic, the same prompt can return a different question set; the consistency of repeated generations was not assessed, and every question set used in the pilot was therefore saved as generated. The functional test suite is deterministic: two independent runs on the same build gave identical results for all 32 cases.

3.5. Research Procedures

The study proceeded in seven steps. (1) Problem identification and requirements: the literature in Section 2 and the authors' teaching experience (Section 1.1) were used to define the requirements for the framework. (2) Design and development: AVAF, AVQGA and VivaGen-AI were designed and refined iteratively (Figure 1). (3) Demonstration: one de-identified submission was processed, and the resulting questions were checked by the authors against the criteria in Appendix A (Section 7). (4) Software testing: the 32-case test suite was run against the reference build (Appendix D). (5) Baseline analysis: the gradebook records for the three baseline offerings were extracted and analyzed using the marks only, without names or identification numbers (Section 9). (6) Pilot: for each pilot student, the lecturer uploaded the student's submission, built the prompt, obtained the

reply from ChatGPT, reviewed the questions, conducted the viva, and awarded a raw mark using the course rubric; the course coordinator then moderated these marks together with those of the other interviews of the offering (Sections 8 and 9.3). (7) Expert validation: the procedure in Section 8 is planned.

Document analysis: Two kinds of documents were analyzed. Students' submissions were analyzed in three stages: the platform extracted the text in the browser; the language model, instructed by the prompt in Appendix B.1, identified the topic, objectives, methods, calculations, tables, findings, and conclusions and drafted the questions; and the course lecturer read each draft against the submission before use. The authors' check of the demonstration questions (Appendix A) was carried out by S.C. The gradebook records were extracted by the course lecturer and analyzed by S.D. All statistical analyses reported in this version were computed in Python (version 3.13; Python Software Foundation, Wilmington, DE, USA) with the SciPy library (version 1.18); all tests were two-sided with α = 0.05 and are treated as exploratory.

4. The Assignment-to-Viva Assessment Framework (AVAF)

4.1. Design Principles

AVAF starts with the observation that a written assignment can no longer, on its own, be trusted as full evidence of a student's independent understanding (Ateeq et al., 2024). This motivates a shift from checking *content* to checking *learning*: judging students on their ability to explain, justify, and defend their own submitted work, consistent with the principles of authentic assessment (Wiggins, 1990; Villarroel et al., 2018) and with calls to redesign assessment so that it preserves meaningful evidence of learning under generative AI (Bearman et al., 2024).

Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) is the framework's cognitive backbone: every candidate question is classified into one of the six taxonomy levels before generation, so that a viva builds from recall toward justification and defense rather than remaining at one cognitive level throughout. Because the educational quality of AI-generated content depends heavily on how the underlying instruction is constructed (Pankaew et al., 2026), AVAF treats prompt design as a distinct step: concepts are extracted and cognitively classified *before* a structured instruction is built, rather than passing the raw assignment to a language model unmediated.

The framework's third design principle is a human-in-the-loop structure: AI performs the computational work of concept extraction, cognitive classification, and first-draft question generation, but the lecturer retains full authority to review, edit, approve, or reject every question before it is used, consistent with human-centered AI principles (Shneiderman, 2022) and with guidance that AI in education should preserve teacher and examiner authority over assessment decisions (UNESCO, 2023). In the current reference build, review and approval are exercised by the lecturer outside the software, as described in Section 6.

4.2. Layered Architecture

AVAF is structured as five loosely coupled layers (Figure 2): Document Ingestion, which extracts machine-readable text from an uploaded PDF or DOCX file (in the current build, extraction yields plain text, so headings and table layout are flattened), with no interpretation performed at this stage; Semantic Analysis, which identifies the assignment's specific claims, methods, calculations and conclusions, i.e., the material a student would reasonably be expected to explain without prompting; Pedagogical Reasoning, which maps this material onto Bloom's Revised Taxonomy and attaches a nominal Easy, Medium or Hard label to each intended cognitive level (the label describes intended

cognitive demand, not empirically calibrated difficulty); Question Generation, which drafts each question, a short model answer, the relevant source section and a follow-up probe, with higher-order items biased toward framing that asks about the student's own reasoning (for example, asking a student to justify a method choice or explain a specific calculation in their own words, rather than to recall a fact), which probes the student's understanding of their own work but does not, on its own, establish authorship; and Human Validation and Delivery, the lecturer approval gate through which every question must pass before it reaches a student, and through which the live viva is delivered.

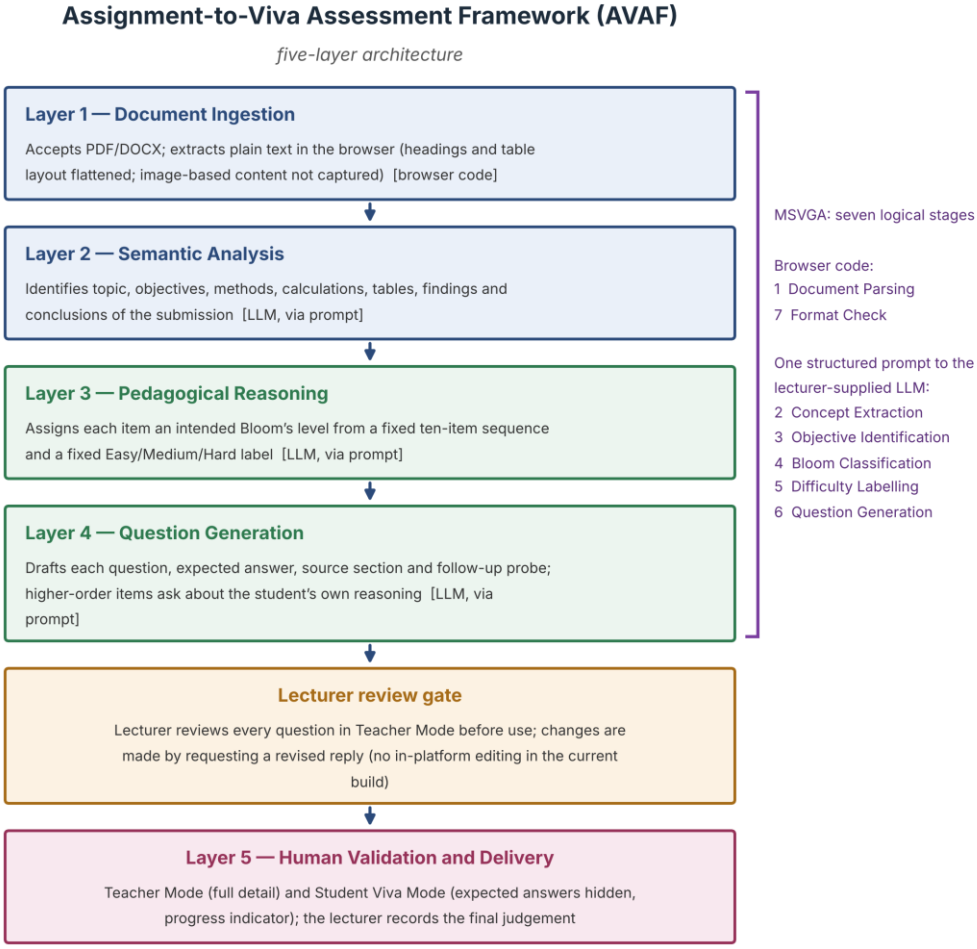


Figure 2. AVAF's five-layer architecture and the seven-stage logical decomposition (MSVGA), of which stages 2–6 are specified within a single structured prompt and stages 1 and 7 are performed by browser code; every question passes the lecturer's review before delivery.

5. AVQGA and MSVGA

The Assignment-to-Viva Question-Generation Algorithm (AVQGA) sets out the computational steps behind AVAF. It takes one assignment as input and produces ten Bloom-labelled questions, sequenced by intended cognitive level, each with a model answer, a source reference and a follow-up probe, as output: (1) ingest the uploaded file and extract its text; (2) validate that sufficient text was extracted to support analysis (at least 50 characters in the current build); (3) extract the assignment's structure (topic, objectives, methods, calculations, tables, findings, conclusions); (4) identify the specific claims and decisions a student should be able to defend without help; (5) classify each candidate item against Bloom's Revised Taxonomy, ensuring coverage across levels rather than concentration at one; (6) sequence the ten questions by intended cognitive level, following the reference build's mapping of Remember to Question 1, Remember/Understand to

Question 2, Understand to Question 3, Apply to Questions 4–5, Analyze to Questions 6 (Medium) and 7 (Hard), Evaluate to Questions 8–9 and Create/Defend to Question 10; (7) generate the natural-language question, model answer, source reference and follow-up probe for each position; (8) bias higher-order items toward the student's own reasoning (for example, "why did you use this method," "how did you calculate this value," "what would you change if you repeated this assignment"), which probes understanding of the student's own reasoning more directly than factual recall does (Villarroel et al., 2018), although it does not by itself establish authorship; (9) constrain tone so that no item implies or states that the student has cheated; (10) check the returned set for placeholder text and, if the check fails, prompt the lecturer to obtain a new reply (in the current build, whether each item can be traced to a specific part of the assignment is judged by the lecturer, not checked by the software); and (11) present the validated set to the lecturer for review before use.

Two distinctions apply to the Bloom labels throughout this paper. The first is between intended and demonstrated alignment. The Bloom level attached to each question is the level the prompt instructs the model to target and that the returned record reports; it is an intended classification. Whether a question actually requires that cognitive process from a student (demonstrated alignment) can only be established by independent classification of the generated questions by expert raters, which is part of the validation protocol in Section 8 and has not yet been completed. The second is between cognitive level and difficulty. Bloom's levels describe the kind of cognitive process a question calls for, not how hard it is: a Remember question about an obscure detail can be harder for a student than an Evaluate question about a central decision. The Easy, Medium and Hard labels in VivaGen-AI are therefore fixed labels attached to each cognitive level by the sequencing rule in Step 6, and they make no claim about empirical item difficulty, which would require item-level response data. The operational criteria used to assign each level, and the authors' check of the demonstration questions against them, are given in Appendix A.

Rationale for the fixed ten-item allocation: The allocation in Step 6 is a design choice, not an empirically optimized one. Ten items were chosen so that a full viva, including follow-up probes, can be completed within a single interview slot. The sequence opens with three lower-order items (Remember and Understand), which let a student settle into the interview on material they should know well, and then allocates seven of the ten items to Apply and above, so that most of the viva concerns the processes in which a student's understanding of their own work is most visible (Section 2.2). A single Create item is placed last because answering it, by proposing a revised method or design, takes longer than the other items. The operational criteria for assigning each level are given in Appendix A (Table A1). Whether a different number or mix of items would work better has not been tested.

The Multi-Stage Viva Generation Architecture (MSVGA) is the name given to this logical decomposition: seven reasoning stages, each with its own defined task: (1) document parsing, (2) concept extraction, (3) objective identification, (4) Bloom classification, (5) difficulty labelling, (6) question generation and (7) format checking. Stage 3 identifies the objectives stated in the assignment itself; the course's intended learning outcomes are not supplied to the model in the current build, so no alignment with learning outcomes is performed. Stage 5 attaches the fixed labels from Step 6 rather than calibrating difficulty. In the current reference implementation, these seven stages are not executed as independent, separately invoked model calls; stages 1 and 7 are performed by browser code, and the platform assembles one structured prompt encoding the instructions and output format for stages 2–6, which the lecturer submits once to a conversational LLM of their choice, and the platform then parses the single returned reply into structured question records (ten are requested, but the parser does not enforce the number or check the labels;

Section 8). The format-checking stage is implemented as the client-side parser described in Appendix B.3; it checks the structure of the reply and rejects placeholder text, but it does not validate factual correctness, source grounding, or cognitive alignment, which rest with the lecturer's review. MSVGA is therefore best understood, and is described here, as a logical decomposition of the reasoning task, five of whose stages are specified as instructions within a single prompt, rather than as a staged, independently executed multi-agent pipeline. The architecture is accordingly named a multi-stage, not a multi-agent, architecture. This distinction matters for two reasons: it sets accurate expectations for readers evaluating the system's technical claims, and it identifies a concrete direction for future work, namely implementing MSVGA's seven stages as genuinely separate, chained model calls with an independent validation pass between each, which the current single-prompt implementation does not yet provide.

AVQGA does not depend on a specific AI vendor for Steps 3–8: the reference implementation is provider-independent at the input/output level, in that it accepts any reply that follows the JSON schema in Appendix B.2, whichever assistant produced it. This is compatibility, not demonstrated performance: only ChatGPT was used in the demonstration and pilot, and the quality of questions produced through other providers has not been tested. In addition, the current build's interface labels (for example, "Open ChatGPT," "Paste ChatGPT's reply") are written around ChatGPT specifically and should be generalized to "your AI assistant" before the platform is presented as provider-independent.

6. System Implementation: The VivaGen-AI Platform

VivaGen-AI is the reference software implementation of AVAF, AVQGA and MSVGA. It is a self-contained, browser-based application (HTML and JavaScript), built to reduce barriers to institutional adoption: it requires no server infrastructure and no software installation, and document parsing for both PDF and DOCX files (via client-side libraries) happens entirely on the user's own device, so an assignment is never transmitted to, or stored by, any third party during parsing. This applies to the parsing step only. The prompt that the platform builds contains text extracted from the assignment, and the lecturer pastes it into an external conversational LLM service, so assignment-derived content does leave the device at the question-generation step. Privacy across the full workflow therefore depends on what text enters the prompt and on the terms of the service used, including whether the institution has an agreement with that provider; depending on institutional policy, a data-sharing agreement or equivalent approval may be required. In the current build, the prompt contains only text extracted from the document itself, limited to its first 12,000 characters; the student name and ID entered in the Upload module are stored locally and are not inserted into the prompt. Identifying information inside the document, such as a cover page or a name in a header, is passed on unless the lecturer removes it before upload. In the pilot, the free version of ChatGPT was used through an individual account, so no institutional agreement or enterprise data controls applied, and the conversations were subject to the provider's standard terms for free accounts. Student names and student numbers were removed from each prompt before it was submitted. This is a limitation of the pilot and the reason for recommendation (6) in Section 13.

The platform has six functional modules (Figure 3). The Upload module accepts a student's assignment together with the student's name and, optionally, a student ID, linking the viva session to a specific student. The Prompt Generation module builds one structured instruction, encoding the Bloom's Taxonomy progression, integrity-related constraints, and required output format, from the assignment's extracted text. The Question-Generation module hands this instruction to a lecturer-supplied conversational LLM: the lecturer copies the prompt to their assistant of choice and pastes the reply back into the

platform, which parses it into structured question records (ten are requested, but the parser does not enforce the number or check the Bloom and difficulty labels; Section 8); this design avoids dependency on a single paid provider, at the cost of a manual copy-and-paste step rather than a fully automated round trip. The Teacher Mode module displays all ten questions with Bloom-level, difficulty, source section, model answer, and follow-up probe, for lecturer review before use. The current build does not provide in-platform editing or removal of individual questions: a lecturer who wants changes asks the LLM for a revised reply, or edits the reply text before pasting it, and approval is the lecturer's decision to proceed to the viva rather than a step recorded by the software. The Student Viva Mode module delivers one question at a time for live use, hiding the model answer and follow-up probe by default and showing progress along the Bloom sequence. The Student Records module saves a record of each viva, comprising the student name and ID, file name, topic, the extracted summary of the submission, the questions with their expected answers and follow-up probes, the date, and the lecturer's final understanding rating. Records can be searched by student name, file name, or topic, deleted individually, and exported as a two-sheet spreadsheet (student records and viva questions).

Reproducibility details: The complete prompt template assembled by the Prompt Generation module, the output schema the parser expects, and the parsing and validation rules applied to the returned reply are reproduced in Appendix B, and the interview marking rubric is given in Appendix C. The platform is a single HTML file (SHA-256: ac0f3f1bf77dbccf419aea0547c95fba4299bea6ab8760238aac9f499ef77a81) that uses three open-source client-side libraries loaded from a public content-delivery network: PDF.js 3.11.174 for PDF text extraction, Mammoth 1.6.0 for DOCX text extraction and SheetJS 0.18.5 for spreadsheet export. The platform does not call a language model itself. The lecturer copies the prompt into the ChatGPT web interface (chatgpt.com) under their own account, and pastes the reply back; the free version of ChatGPT is sufficient, and no subscription or API key is needed. The platform therefore sets no model, version, or sampling parameters: the model is whichever model the lecturer's ChatGPT account is using at the time, and because the web interface does not expose temperature or other sampling settings, the provider's defaults apply. The demonstration and the pilot used the free version of ChatGPT, which applies the provider's default model and does not let the user choose a model or change generation settings; the pilot prompts were run during the 2026 offering. Because the provider updates its models without notice, the same prompt applied to the same assignment can return a different question set. Each parsed question set is therefore saved with the student record at the moment of parsing (see the Student Records module above), and the stored sets used in the pilot have been retained but, because they are derived from students' assignments, cannot be shared under the conditions of the ethics approval (Data Availability Statement).

Data flow: The complete flow of data in the current build is as follows. (1) Upload and extraction: the file is read in the browser, and its text is extracted there; the platform never sends the file anywhere (the three libraries are downloaded from a content-delivery network as code, and no assignment data are sent to it). (2) Prompt construction: the prompt is assembled in the browser from the template in Appendix B.1 and the first 12,000 characters of the extracted text; the student name and ID are not included, but any identifying content within the document text is. (3) External processing: the lecturer copies the prompt into ChatGPT under their own account, so the assignment-derived text is processed, and may be retained, by the provider under that account's terms and settings (for example, chat-history and model-training settings); retention and deletion at this step are controlled through the lecturer's account, not by the platform. (4) Returned output: the lecturer pastes the reply back, and it is parsed in the browser. (5) Browser storage: each accepted reply creates a record containing the student name and ID, file name, extracted

summary, questions, expected answers, follow-up probes, date, and rating. The record is held in the hosting environment's persistent storage or the browser's localStorage until the lecturer deletes it; it is not encrypted, and the platform has no login or access control of its own, so access depends on the security of the device and browser profile. (6) Export: the lecturer can export all records to a spreadsheet file, which then falls under the institution's rules for assessment records. In the pilot, student names and identification numbers did not enter the prompt, because the platform never inserts them, but identifying information inside the submissions was not removed by the platform; the free version of ChatGPT was used, so no institutional agreement or enterprise data controls applied; and the browser records were held on the lecturer's computer.

At the end of a session, the lecturer records one overall judgement from four categories: Strong Understanding, Good Understanding, Partial Understanding, or Further Verification Needed. Where an institution also scores the oral component numerically as part of a wider rubric, for example, an Interview component worth up to 150 marks, the same four categories can be mapped onto that numeric scale (Table 4), giving a consistent qualitative label for any numeric interview mark. In the current build, the platform records only the qualitative judgement; it has no field for a numerical mark. In this study, the interview mark out of 150 was awarded by the examiner during the interview using the course's existing interview rubric (Appendix C), moderated by the course coordinator (Section 8) and recorded in the course gradebook. The cut-points of 80%, 60% and 40% of the interview mark are the level boundaries of that existing rubric, which examiners used in all offerings; the authors adopted the same boundaries for VivaGen-AI, attached the four understanding-band labels to them and applied the bands retrospectively, for analysis only, to the moderated marks (Section 9). The boundaries give three bands of equal width above 40%; as analysis categories, they are provisional, not derived from an institutional standard or a formal standard-setting procedure, and have not been validated. Band-based comparisons (Section 9.3) depend on where these boundaries fall, which is one reason the comparison of mean marks is reported alongside them.

Table 4. Interpretation scheme linking numeric interview marks (out of 150) to the four qualitative understanding categories used by VivaGen-AI.

Interview Mark (Out of 150)	Percentage	Interpretation
120–150	80–100%	Strong Understanding
90–119	60.0–79.3%	Good Understanding
60–89	40.0–59.3%	Partial Understanding
0–59	0.0–39.3%	Further Verification Needed

Interview marks in the baseline records are whole numbers (multiples of 5); percentages are the mark divided by 150, rounded to one decimal place. If fractional marks are awarded, the bands are defined by their lower bounds: Strong Understanding ≥ 120; Good Understanding ≥ 90 and <120; Partial Understanding ≥ 60 and <90; Further Verification Needed < 60.

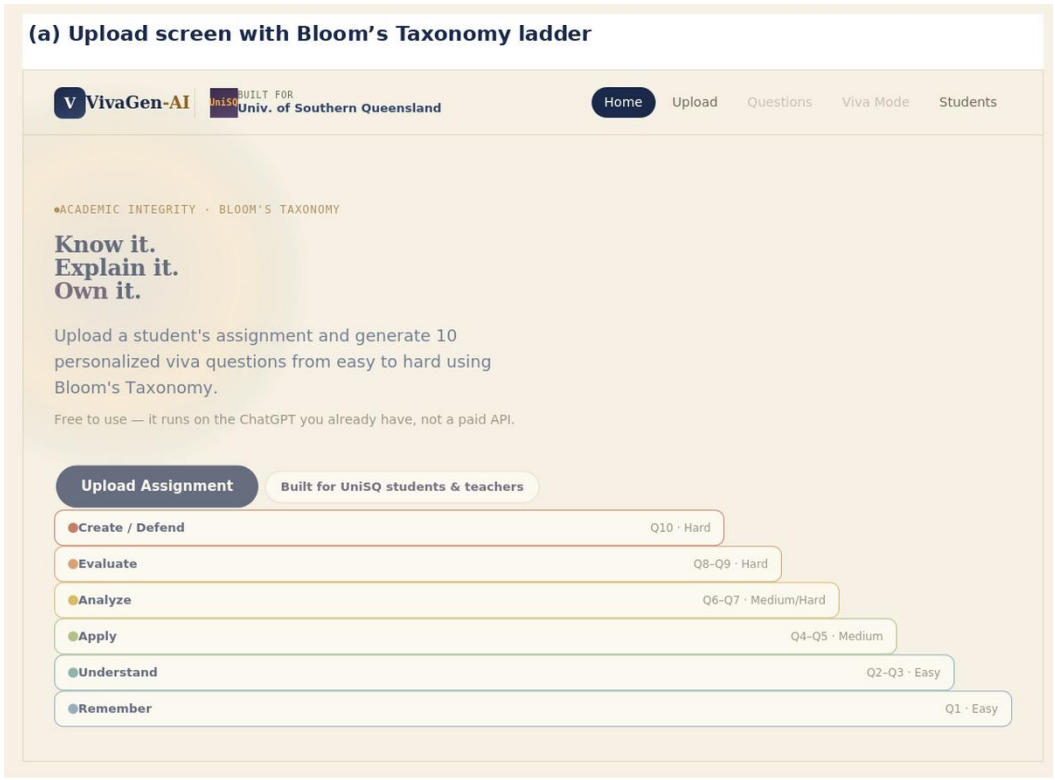
Three design choices distinguish VivaGen-AI from typical AI-assisted assessment tools. First, it relies on a lecturer-supplied LLM reply rather than a paid, built-in subscription, a choice that avoids a subscription fee but does not remove all cost or access barriers, because the lecturer still needs a ChatGPT account (free accounts are subject to usage limits), internet access, and institutional permission to use the service, at the cost of a manual step. Second, document parsing occurs on the user's device, which reduces, but does not remove, a data-privacy concern common to AI-based education tools (Dwivedi et al., 2023); it should be noted, however, that while assignment content never leaves the device during parsing, student name and ID, together with the generated questions and ratings, are stored in the browser, through the hosting environment's persistent key–value storage

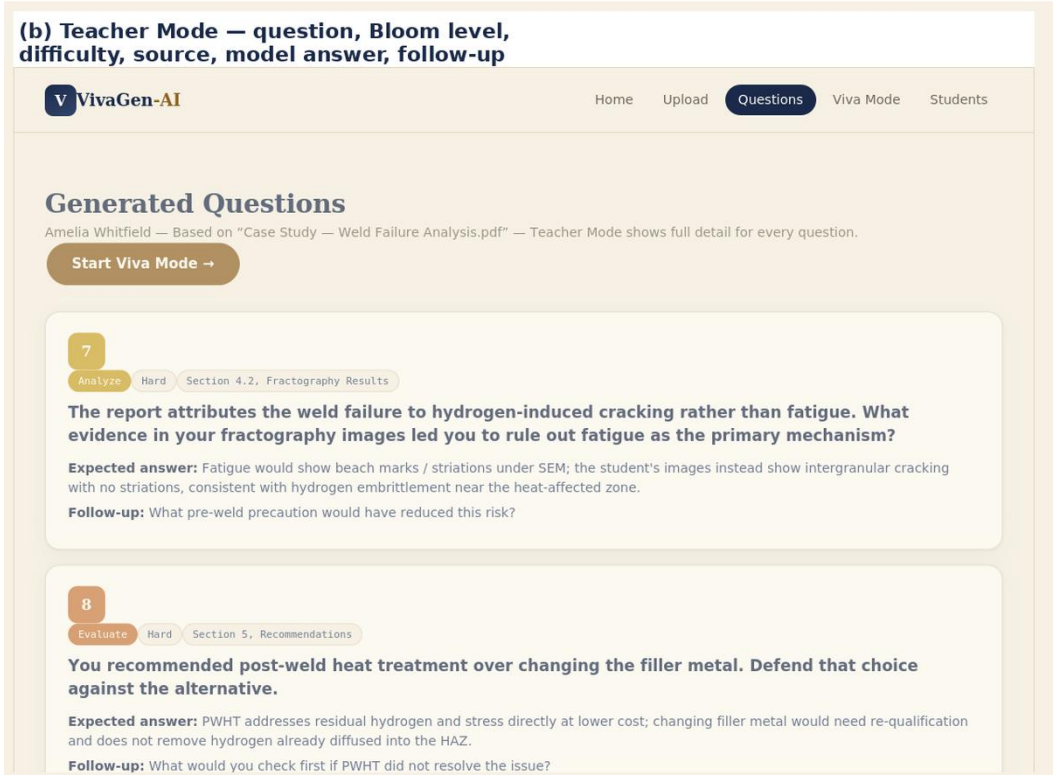
interface where one is available and otherwise the browser's localStorage, and are exportable, and this storage is not encrypted at rest, so institutions should treat the device itself, not just the network, as part of the data-handling boundary. Third, the question-generation instructions include an instruction (AVQGA Step 9) against any content that assumes or states that a student has cheated; this instruction is a safeguard, not a guarantee of the model's output, and the lecturer's review is the control that catches a breach; the system is designed only to support a check of understanding, never to make an accusation.

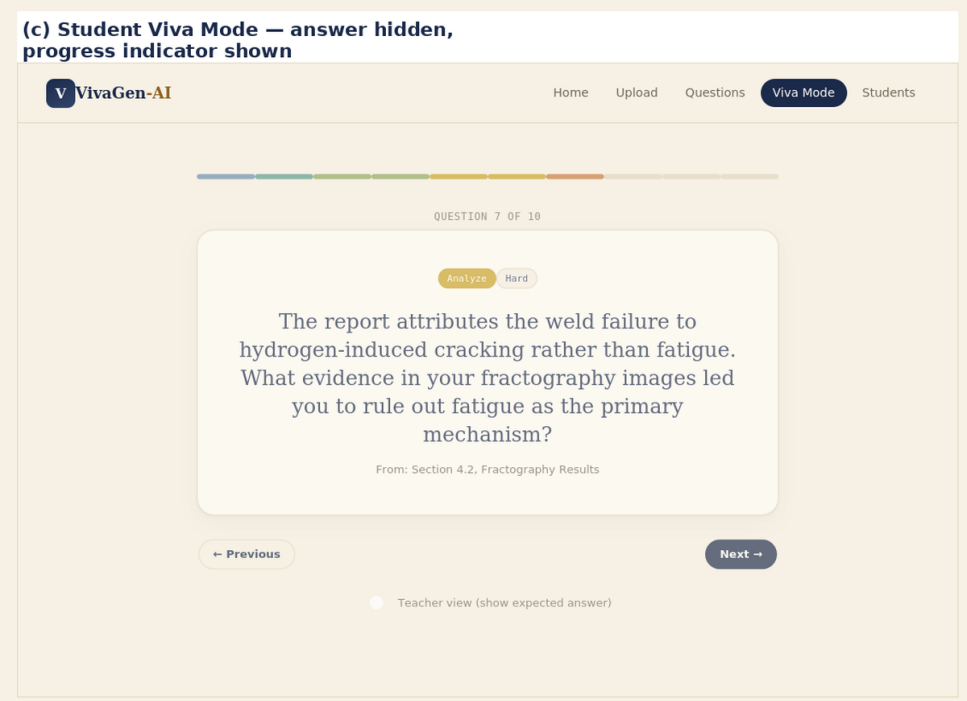
Table 5 separates the operations performed by browser code, the instructions delegated to the language model, the manual actions of the lecturer, and functions that are planned or not implemented.

Table 5. Implementation status of each function of VivaGen-AI: what performs it in the reference build, and whether it is implemented, manual, planned, or not implemented.

Function	Performed by	Status in the Reference Build
File selection and type check (.pdf, .docx)	Browser code	Implemented
Text extraction (PDF.js, Mammoth; at most 30 pages, about 24,000 characters)	Browser code	Implemented; plain text only
Sufficiency check (at least 50 characters)	Browser code	Implemented
Prompt construction (template and first 12,000 characters)	Browser code	Implemented
Sending the prompt and returning the reply	Lecturer (manual copy and paste)	Implemented
Structure extraction and objective identification (MSVGA stages 2–3)	LLM, via prompt	Implemented as instructions only
Bloom-level and difficulty-label assignment, fixed sequence (stages 4–5)	LLM, via prompt	Implemented as instructions only; labels not checked
Question, expected answer, source section and follow-up generation (stage 6)	LLM, via prompt	Implemented as instructions only
Tone constraint: no accusations (AVQGA Step 9)	LLM, via prompt	Instruction only; a safeguard, not a guarantee
Reply parsing and format checks (stage 7)	Browser code	Implemented; format only (Appendix B.3)
Factual-accuracy and source-traceability check	Lecturer	Manual; not supported by the software
Editing, approval or rejection of questions	Lecturer, outside the platform	Manual; no in-platform editing
Live viva delivery (Student Viva Mode)	Browser code and lecturer	Implemented
Final understanding judgement (four categories)	Lecturer	Implemented
Numerical interview mark (out of 150)	Lecturer, outside the platform	Not part of the platform
Record storage, search, deletion and export	Browser code	Implemented; not encrypted
Alignment with course learning outcomes	—	Not implemented
Independent execution and validation of each MSVGA stage	—	Planned
Provider-neutral interface labels	—	Planned
Question quality with providers other than ChatGPT	—	Not evaluated


(a) Upload screen with Bloom's Taxonomy ladder
VivaGen-AI
BUILT FOR
Univ. of Southern Queensland
Home
Upload
Questions
Viva Mode
Students
ACADEMIC INTEGRITY · BLOOM'S TAXONOMY
Know it.
Explain it.
Own it.
Upload a student's assignment and generate 10 personalized viva questions from easy to hard using Bloom's Taxonomy.
Free to use — it runs on the ChatGPT you already have, not a paid API.
Upload Assignment
Built for UniSQ students & teachers
Create / Defend
Q10 · Hard
Evaluate
Q8–Q9 · Hard
Analyze
Q6–Q7 · Medium/Hard
Apply
Q4–Q5 · Medium
Understand
Q2–Q3 · Easy
Remember
Q1 · Easy


(b) Teacher Mode — question, Bloom level, difficulty, source, model answer, follow-up
VivaGen-AI
Home
Upload
Questions
Viva Mode
Students
Generated Questions
Amelia Whitfield — Based on "Case Study — Weld Failure Analysis.pdf" — Teacher Mode shows full detail for every question.
Start Viva Mode →
7
Analyze
Hard
Section 4.2, Fractography Results
The report attributes the weld failure to hydrogen-induced cracking rather than fatigue. What evidence in your fractography images led you to rule out fatigue as the primary mechanism?
Expected answer: Fatigue would show beach marks / striations under SEM; the student's images instead show intergranular cracking with no striations, consistent with hydrogen embrittlement near the heat-affected zone.
Follow-up: What pre-weld precaution would have reduced this risk?
8
Evaluate
Hard
Section 5, Recommendations
You recommended post-weld heat treatment over changing the filler metal. Defend that choice against the alternative.
Expected answer: PWHT addresses residual hydrogen and stress directly at lower cost; changing filler metal would need re-qualification and does not remove hydrogen already diffused into the HAZ.
Follow-up: What would you check first if PWHT did not resolve the issue?



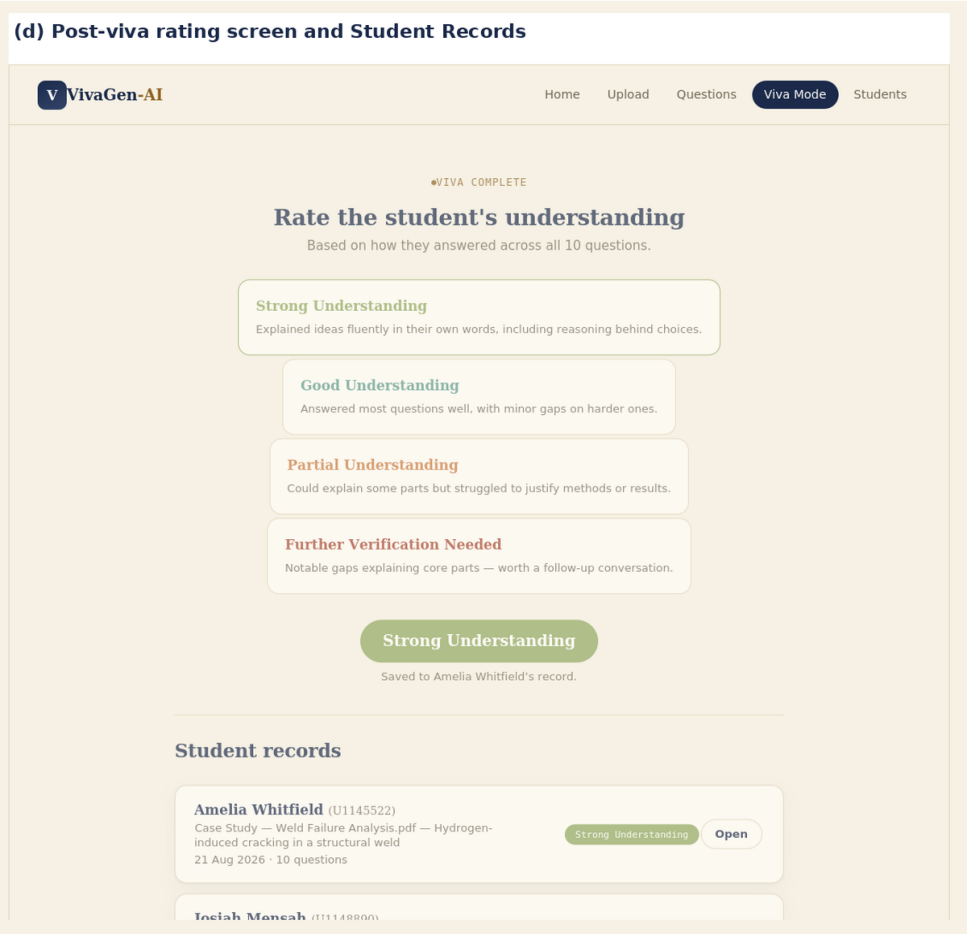


Figure 3. VivaGen-AI interface: (**a**) upload screen with the Bloom's Taxonomy question ladder; (**b**) Teacher Mode showing a generated question card; (**c**) Student Viva Mode with the model answer hidden; (**d**) the post-viva rating screen and Student Records view. The student names and IDs shown are fictitious. The panels are shown at a larger size, (**a**,**b**) above and (**c**,**d**) below. The screenshots were taken from an earlier build whose question ladder in panel (**a**) labels Q1 as Remember

and Q2–Q3 as Understand; the reference build identified in this section labels Q1–Q2 as Remember and Q2–Q3 as Understand. The screens are otherwise the same.

7. Demonstration

Following the demonstration phase of design-science research (Peffers et al., 2007), this section illustrates AVAF, AVQGA and VivaGen-AI on a representative submission from the course: a de-identified failure-analysis case study in which a student group investigated hydrogen-induced cracking (HIC) in a welded steel joint, including a description of the failure mode, the microstructural and hardness evidence examined, the post-weld heat treatment (PWHT) parameters applied, and a set of recommendations for preventing recurrence.

Group-produced submissions: The demonstration submission was produced by a student group. The current build generates questions from the whole submission and has no input for an individual student's declared contribution, so it cannot by itself direct questions at the parts a given student wrote. The demonstration is illustrative only: no student was examined on it, so no individual contribution had to be established. Where group work is examined, the lecturer is responsible for selecting or adapting questions so that a student is not penalized for parts of the work outside their agreed contribution, while questions on the shared method, findings and recommendations continue to assess the intended learning outcomes that all members are expected to meet.

Given this input, the Semantic Analysis layer identified the assignment's core topic (hydrogen-induced cracking in the weld), its stated methodology (microstructural examination, hardness testing and PWHT parameter selection), the specific numerical values used in the student's own analysis, and the assignment's final recommendations. Table 6 presents the resulting ten-question sequence generated by the reference implementation for this submission.

Table 6. Illustrative ten-question viva sequence generated for the demonstration weld-failure (hydrogen-induced cracking) submission.

#	Bloom Level (Difficulty)	Question	Source Section
1	Remember (Easy)	State the failure mode you identified for this weld and the primary microstructural feature that indicated it.	Findings
2	Understand (Easy)	In your own words, explain why hydrogen-induced cracking was identified as the failure mechanism rather than another cracking mode.	Microstructural analysis
3	Understand (Easy)	Explain what the hardness values you measured near the weld heat-affected zone indicate about susceptibility to hydrogen cracking.	Microstructural analysis
4	Apply (Medium)	Walk through how you selected the post-weld heat treatment (PWHT) temperature and holding time used in your procedure.	Methodology/calculations
5	Apply (Medium)	Show how you applied the relevant hydrogen diffusion or hardness criterion to justify the chosen PWHT parameters.	Methodology/calculations
6	Analyze (Medium/Hard)	Why did you select this PWHT procedure rather than an alternative approach available to you?	Methodology
7	Analyze (Hard)	Analyze how the assumptions in your hardness or hydrogen assessment could affect the accuracy of your conclusions.	Methodology/calculations

8	Evaluate (Hard)	Evaluate whether your recommended PWHT procedure would fully prevent recurrence of hydrogen-induced cracking, or only partially mitigate the risk.	Recommendations
9	Evaluate (Hard)	Defend your choice of recommended fix against a plausible alternative (for example, a filler-metal change versus a PWHT change).	Recommendations
10	Create/Defend (Hard)	If you repeated this failure analysis, what would you change in your methodology, and why?	Whole submission

Each question in Table 6 is paired, in the platform, with a short model answer keyed to the specific table or paragraph it was drawn from, so that the lecturer can quickly check the student's live answer against the submitted work. This worked example illustrates the framework's central claim: a ten-question viva, sequenced by intended cognitive level, can be produced from one assignment submission, with every item linked to a stated section of the student's own material, intended to support a check of understanding rather than a generic oral quiz. It does not, on its own, demonstrate that lecturers or students judge these questions to be of high quality; that judgement is the subject of the evaluation protocol in Section 8. The Bloom labels in Table 6 are the intended levels returned by the system. The authors' check of each item against the operational criteria in Appendix A (Table A2) found seven of the ten items consistent with their intended level and three (Questions 4, 6 and 10) closer to, or spanning, an adjacent level; it also found one item (Question 5) whose wording hedges between two criteria, suggesting that the system had not identified which criterion the submission used. For two items (Questions 2 and 6), the label returned by the model differs from the label the prompt requests (Appendix B.1), which the parser does not check. The questions are reported as returned by the model, without editing, and their factual accuracy against the source submission has not been independently verified; that verification is part of the expert validation specified in Section 8. In particular, Question 4 asks for a retrospective account of how a procedure was selected, which is not the same performance as applying it, and the combined Create/Defend label on Question 10 merges defending an existing decision (Evaluate) with producing a new solution (Create). This check was made by the authors, not by independent raters, and is reported as a transparency measure rather than as evidence of demonstrated alignment. The filled prompt can be reconstructed from the template in Appendix B.1, but the submission itself is not reproduced and cannot be shared, to protect the students' work and identity. Table 6 (each question, its intended level, and the section of the submission it draws on) and Table A2 (the cognitive process each question requires and the rationale for its classification) together show how Bloom alignment is operationalized item by item. Until the expert validation in Section 8 is completed, Bloom alignment is a design objective of the system, not an independently validated property.

8. Evaluation Methodology

This study evaluates AVAF and VivaGen-AI through five components, reported at different stages of completion and summarized in Table 7. Software testing (completed, functional behavior only) ran 32 scripted test cases against the reference build in a headless browser (Appendix D, Table A4). The cases cover input handling, prompt construction, the page and character limits of text extraction, reply parsing, rendering safety, viva delivery and record storage, and each has an explicit expected result (the acceptance criterion). All 32 cases matched their predefined expected results. Passing the suite therefore shows that the build behaves as its code specifies, including its known limitations, not that it always returns exactly ten complete, correctly labelled questions: five cases document limitations of the current build: replies with fewer than ten questions, with

mislabeled questions or with up to two placeholder items are accepted; trailing text containing braces after the JSON causes a valid reply to be rejected; and each accepted reply creates a new student record. When a reply is rejected, recovery is manual: the lecturer obtains a new reply and pastes it again (Appendix B.3). Because the third-party extraction libraries were replaced by stubs, the suite tests the platform's own logic, not extraction quality on real documents; extraction quality on real scanned documents, equations and tables was not tested systematically, although scanned PDFs without a text layer are rejected by the sufficiency check (case E3). The platform does not itself enforce exactly ten questions or validate Bloom labels, and the interpretation scheme in Table 4 is applied by the examiner, not by the software. Passing these tests verifies the software's functional behavior only; it does not establish the pedagogical quality of the generated questions (their relevance, accuracy or Bloom alignment), which is the subject of the expert validation described below. Historical baseline analysis (completed, reported in Sections 9.1 and 9.2) applies that interpretation scheme descriptively to interview-mark data from three offerings (83 records, comprising 68 interviews with examiner-prepared questions and the 15 pilot interviews), establishing the empirical comparison point for future evaluation of VivaGen-AI itself. Preliminary outcome pilot (completed, reported in Section 9.3) is a retrospective analysis of the 15 regular course interviews in the 2026 offering in which lecturer-reviewed VivaGen-AI questions were used; the students were chosen by the lecturer and not randomly assigned; it reports the resulting understanding-band distribution and interview marks, compared descriptively and statistically with the other baseline records, with the explicit caveat that the absence of random assignment, and the lecturer's choice of the pilot students, mean the observed shift cannot yet be attributed causally to VivaGen-AI itself. Lecturer/expert validation of question quality (specified, not yet executed) remains the step required before any claim about the quality of VivaGen-AI's generated questions, as distinct from the resulting interview outcomes, can be made: it is designed as an exercise with 8–10 assignments, using de-identified real student submissions run through the platform, with a panel of at least five experts, comprising specialists in curriculum and instruction, measurement and evaluation, artificial intelligence in education and the subject discipline (materials engineering), independently rating each generated question set against six criteria, relevance to the assignment, alignment with the intended Bloom's Taxonomy level, clarity, appropriateness of the difficulty progression across the ten items, overall teaching value, and suitability for live oral delivery, each on a five-point scale, with inter-rater agreement reported alongside a qualitative summary of the changes raters would make. Each expert also classifies the intended Bloom's level of every item, and content validity is summarized with item- and scale-level content validity indices (Section 3.4). The fifth component, a prospective, randomized within-offering comparison (for example, randomly allocating consenting students within a single offering of the course to VivaGen-AI-generated or examiner-prepared question sets, with interviewer and marking rubric held constant across arms), also remains to be executed and was selected as the design for directly testing whether VivaGen-AI changes outcomes relative to the existing process while holding course, discipline, trimester and assignment type constant. Randomizing students within one offering was chosen because it removes the differences in cohort, offering, and examiner that limit the pilot (Section 9.3). Two alternatives were considered and not chosen: a crossover design, in which each student would answer both question sets, because answering one set would affect performance on the other; and a comparison with matched historical cohorts, because it would leave differences between examiners and offerings uncontrolled. The expert-validation and randomized-comparison protocols are described as specified in the sense that they were written down before any confirmatory data were collected; they have not been formally

preregistered, and the authors intend to register the randomized-comparison protocol on the Open Science Framework before its data are collected.

Table 7. Evaluation components: purpose, status, and where each is reported.

Component	Purpose	Status	Reported in
1. Software testing	Check that the reference build behaves as specified	Completed (functional; 32 cases)	Section 8; Appendix D
2. Historical baseline	Reference distribution of interview marks in the three offerings (83 records, including the 15 pilot interviews)	Completed (descriptive)	Sections 9.1 and 9.2
3. Preliminary pilot	First use of the workflow in live interviews	Completed (non-randomized; N = 15)	Section 9.3
4. Expert validation of question quality	Direct evaluation of the generated questions	Planned (protocol specified)	Section 8
5. Randomized within-offering comparison	Test of effect relative to the existing process	Planned	Section 8

Pilot timing, participants and assessors: The pilot took place in the 2026 offering, which is also the most recent of the three baseline offerings. For 15 students enrolled in that offering, chosen by the lecturer based on their marks in earlier assessment components so as to include both higher- and lower-performing students, the regular course interview was conducted with lecturer-reviewed VivaGen-AI questions; the interviews of the other students in that offering used examiner-prepared questions. The pilot interviews were part of routine assessment: the examiner's raw marks were moderated before entry in the gradebook, as were all interview marks in the three offerings, and the moderated marks counted towards the course grade in the same way as every other interview mark (see Moderation below). No student was recruited, and no data were collected for the research; the records were analyzed retrospectively after the assessment outcomes had been finalized (see Ethics below). Table 8 sets out how the pilot differs from each baseline offering in academic year, participant selection, assignment type and examiner arrangements. The pilot interviews were conducted and marked by the course lecturer, using the same rubric and four-band scheme as the course interview, whereas the other baseline interviews were marked by course staff, with different examiners in different years (Section 9.1). Three consequences follow. Because the pilot students were selected to span the range of performance, their marks are expected to be more dispersed than those of a whole cohort; because their records are part of the 2026 baseline, the pilot is compared with the other baseline records, not with groups that include it; and because the examiner and the question source differ between the pilot and the other interviews, a difference between them cannot be attributed to the question source alone.

Moderation: In each offering, including 2026, all interview marks were moderated by the course coordinator (the first author) after all interviews of that offering had been completed and before the marks were entered in the gradebook. The coordinator reviewed each raw mark against the level descriptors of the course rubric (Appendix C, Table A3) and the examiner's interview notes, and adjusted marks so that the same standard was applied across the offering. The same procedure was applied to the 15 pilot interviews and to the 16 other interviews of the 2026 offering. Moderation was not blind: the coordinator knew which interviews had used VivaGen-AI questions, and, because the coordinator also examined the pilot interviews, moderation of the pilot marks was a review by the same person rather than an independent check. Moderation narrowed the range of the pilot marks from 47–141 (raw marks) to 75–120. All comparisons in Section 9.3 use the

moderated marks recorded in the gradebook, so that the pilot and the other records are on the same basis; results on the raw marks are reported as a secondary analysis.

Table 8. Description of the three baseline offerings and of the pilot interviews, which form part of the 2026 offering.

Characteristic	2024 Offering	2025 Offering	2026 Offering	Pilot
Course	The course	The course	The course	The course
Offering (year)	2024	2025	2026	2026 (15 interviews within the 2026 offering)
Records available/analyzed	46/37	17/15	33/31 (including the 15 pilot interviews)	15/15 (among the 31 analyzed 2026 records)
Excluded records (reason)	9 (7 completed other components but no interview mark; 2 with no assessment marks)	2 (1 completed other components but no interview mark; 1 quizzes only)	2 (1 completed other components but no interview mark; 1 quizzes only)	0
Mode of enrolment	On campus (two campuses) and online	On campus and online	Online 33	Online (as of 2026)
Course mark composition (out of 1000)	Quizzes 200; literature review 200; moderated case study; interview 150	Quizzes 200; literature review 200; case study 450; interview 150	As 2025	As of 2026, the pilot interview was the course interview
Interview marks recorded (moderated)	Steps of 10 (90–150)	120, 130 or 150 only	Steps of 5 (75–120)	Steps of 5 (75–120) after moderation; examiner's raw marks 47–141
Grading scale	HD/D/C/P/F	HD/D/C/P/F	A/B/C	Not applicable
Disciplinary composition	Undergraduate engineering students	Undergraduate engineering students	Undergraduate engineering students	Undergraduate engineering students
Assignment type	Literature review and case study	Literature review and case study	Literature review and case study	Rail-incident failure-analysis case study submitted for the course
Selection	All students with a recorded interview mark (retrospective)	As 2024	As 2024	Chosen by the lecturer, by marks in earlier assessment components, to include higher- and lower-performing students; analyzed retrospectively
Recruitment	Not applicable	Not applicable	Not applicable	None (routine assessment)
Examiner(s)	Course staff	Course staff	Course staff (16 interviews); course lecturer (15 pilot interviews)	Course lecturer
Question source	Examiner-prepared	Examiner-prepared	Examiner-prepared (16 interviews); VivaGen-AI, lecturer-reviewed (15 pilot interviews)	VivaGen-AI, lecturer-reviewed

Baseline data are from the course gradebook records for each offering (96 records: 83 analyzed, 13 excluded). One student appears in both the 2025 offering (excluded; no interview mark) and the 2026 offering (analyzed), so no student contributes more than one analyzed baseline record. The 15 pilot interviews were regular course interviews of the 2026 offering and are included in its 31 analyzed records, so the 83 analyzed records comprise the 15 pilot records and 68 other records.

Ethics: The study was approved by the University of Southern Queensland Human Research Ethics Committee (approval ETH2026-0687, 3 November 2026) as a

retrospective, secondary analysis of existing assessment records from the three offerings, including the 15 interviews in which VivaGen-AI questions were used. All interviews, including those 15, were conducted as part of the course's routine assessment, before and independently of this research; no students were recruited, no additional data were collected, and the analysis had no bearing on any student's marks. The records were de-identified before analysis and are reported only in aggregate, and, in line with the conditions of the approval, the course and its teaching period are not named. Individual informed consent was not obtained. The application sought a waiver of the requirement for consent for all three data sets, and the Committee approved the study based on that application. Scope and chronology: the interviews in all three offerings, including the 15 VivaGen-AI interviews in 2026, took place as routine course assessment before the ethics application was lodged in September 2026. In those 15 interviews, the lecturer entered text extracted from each student's assignment into the free web version of ChatGPT to draft the questions, with student names and student numbers removed from the prompt (Section 6). This external processing was part of the lecturer's assessment practice, not a research procedure; it was disclosed to the Committee in the application, but it preceded the application and is not covered by the approval, which covers only the retrospective analysis of the de-identified records. Students were not asked to consent to this processing, and no institutional agreement with the provider applied; these are limitations of the pilot (Section 11), and recommendation (6) in Section 13 sets out the safeguards that should apply in future use.

Student safeguards: A Further Verification Needed judgement records that the student could not adequately explain core parts of their own submission in the interview; it is not a finding of academic misconduct, and the platform neither makes nor records such a finding. In the platform, such a judgement is described as a reason for a follow-up conversation with the student, not as an outcome in itself. In the pilot, as in the baseline offerings, the interview mark counted towards the course grade, and students could clarify or contest an interview outcome through the university's standard assessment review procedures. Students with approved assessment adjustments are entitled to them in an interview, as in any other assessment.

9. Results

Data sources and provenance: Three data sets are used. (a) Historical baseline records for the 2024–2026 offerings: the course gradebook records for each offering (96 records; 83 with an interview mark), containing for each student the component marks (quizzes, literature review, case study and interview out of 150), the course total out of 1000 and the grade. Only the marks were used in the analysis; names and identification numbers were not used and are not reported. These records were used for Table 9, Figures 4 and 5, the correlations in Section 9.2 and the year-specific comparisons in Section 9.3, and they have been retained. Because the individual baseline marks are available, the comparisons in Section 9.3 were computed from them; all interview marks in the gradebook are moderated marks. (b) Pilot records: for each of the 15 pilot students, recorded under an anonymous code, the assignment topic, the examiner's raw interview mark and the moderated mark recorded in the gradebook (both out of 150), the understanding band and a short lecturer note, together with the saved question set; all pilot statistics in Section 9.3 were computed from these individual records. Because the pilot interviews were the students' regular course interviews, each pilot student's moderated mark is also one of the 31 analyzed 2026 baseline records. This linkage was established by the course coordinator, who matched each pilot record to the student's entry in the 2026 gradebook. (c) The demonstration submission and its generated question set (Section 7). Because of the

conditions of the ethics approval, these records cannot be shared (see the Data Availability Statement).

9.1. Interview Performance in the Baseline Offerings (2024–2026)

This section applies the interpretation scheme in Table 4 descriptively to real institutional data, establishing an empirical baseline for the outcome pilot reported in Section 9.3. Historical records for the course were examined across three consecutive offerings (2024, 2025, 2026), each in the same teaching period of the year. Each cohort record included an Interview (out of 150) mark, the oral-assessment component of the course, awarded under the course's existing interview process; in 15 of the 31 analyzed 2026 interviews (the pilot, Section 9.3), the questions were VivaGen-AI questions reviewed by the lecturer. Across the three cohorts, 96 student records were available; 83 (86.5%) had a recorded Interview mark and are analyzed below (interview scores: range 75–150, mean = 115.6, SD = 17.8). The remaining 13 students had withdrawn, were marked absent, or had no recorded interview mark, and were excluded. By offering, 9 records were excluded in 2024, 2 in 2025, and 2 in 2026 (Table 8). Of the 13, 9 belonged to students who had completed other assessment components but had no interview mark recorded, and 4 to students with no marks beyond the quizzes or no assessment marks at all, consistent with withdrawal or non-completion. Students without an interview mark had markedly lower course marks than those analyzed (2024: mean 62.3% for the 7 with a course mark, against 78.9%; 2025: 34.6% against 79.0%; 2026: 44.8% against 76.1%), so the excluded records come disproportionately from the lower end of the distribution. Because these students have no observed interview mark, the absence of students in the Further Verification Needed band is a property of the 83 analyzed records only; it should not be read as showing that no student in these offerings would have fallen in that band, and no scores are imputed for the excluded students.

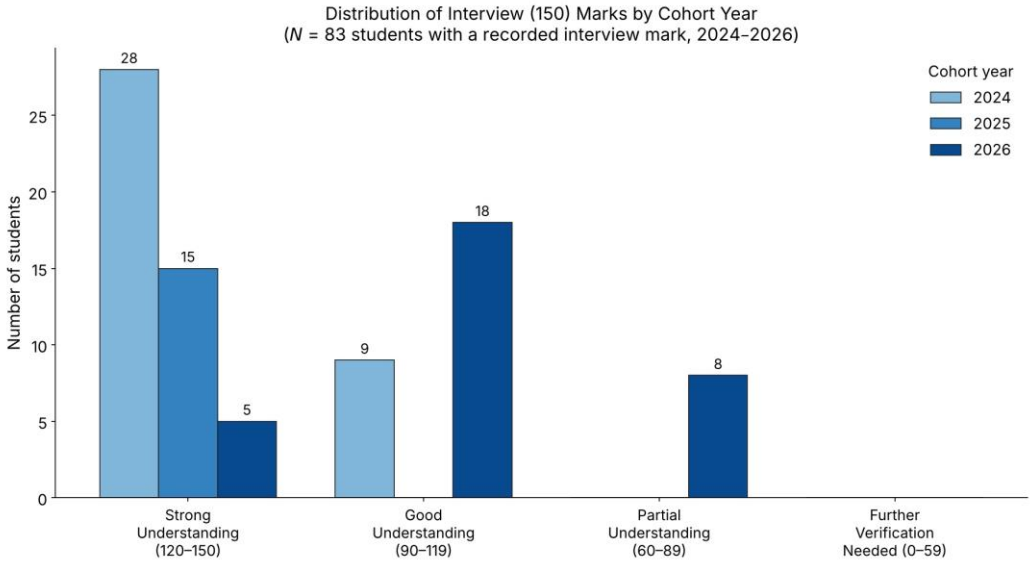


Figure 4. Distribution of interview marks (out of 150) by offering (N = 83).

Table 9. Distribution of interview marks (out of 150) by offering, classified using the four-band scheme in Table 4.

Understanding Band	2024 (n = 37)	2025 (n = 15)	2026 (n = 31)	Combined (N = 83)
Strong Understanding (120–150)	28 (75.7%)	15 (100.0%)	5 (16.1%)	48 (57.8%)
Good Understanding (90–119)	9 (24.3%)	0 (0.0%)	18 (58.1%)	27 (32.5%)
Partial Understanding (60–89)	0 (0.0%)	0 (0.0%)	8 (25.8%)	8 (9.6%)

Understanding Band	2024 (n = 37)	2025 (n = 15)	2026 (n = 31)	Combined (N = 83)
Further Verification Needed (0–59)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Two observations follow from Table 9. First, across all three cohorts, no student with a recorded interview mark fell into the Further Verification Needed band. Second, the distribution shifts substantially by year: the 2024 and 2025 cohorts sat predominantly in Strong Understanding (75.7% and 100.0%), while the 2026 cohort spread across Strong, Good and Partial (16.1%, 58.1%, 25.8%). Because interview marking across these cohorts was conducted by different staff in different years, this shift most plausibly reflects differences in marking practice, rubric interpretation or cohort composition rather than a single underlying cause; the unusually uniform 2025 result (100% Strong Understanding from a comparatively small cohort of 15 students) is consistent with a single-marker effect in a small sample rather than with a genuine change in cohort ability, though the available data cannot distinguish these explanations definitively. This pattern is one motivation for a structured, criterion-based approach such as AVAF, in which the question set, its cognitive-level mix and its marking guidance are fixed in advance rather than left to the examiner in the moment. The records also show different marking practice across offerings: interview marks were awarded in steps of 10 in 2024 (90–150), only as 120, 130 or 150 in 2025, and in steps of 5 in 2026, when no mark exceeded 120 (Table 8). Applying the four-band scheme to these historical marks is descriptive: it does not validate the meaning of the bands or their thresholds, which are fixed percentage boundaries (Table 4).

9.2. Relationship Between Interview Performance and Overall Course Outcome

A Pearson correlation between interview percentage and overall course percentage, computed across the 83 students with both records, found a statistically significant, moderate positive association, $r(81) = 0.37$, $p < 0.001$ (Figure 5), indicating that interview performance shares approximately 14% of its variance with overall course outcome. Because the interview mark is itself part of the overall course mark, this association is reported as a descriptive result only. It is not independent evidence that the interview measures something distinct from the written and computational components, and it does not validate the four understanding bands.

This coefficient should be read with one caveat. The Interview mark is itself a component of the overall course mark, so this is a part–whole correlation, and the reported association is consequently inflated relative to what a correlation between interview performance and the *non-interview* components of the course would show; the part–whole relationship alone can produce a substantial correlation. If the interview percentage I and the aggregate of the non-interview components N were uncorrelated, and the overall percentage were $O = wI + (1 - w)N$, the correlation between I and O would still be $w \cdot SD(I)/\sqrt{[w^2 \cdot SD(I)^2 + (1 - w)^2 \cdot SD(N)^2]}$. If the overall percentage is such an unscaled weighted sum, N can be recovered for each student from the paired values already used here, without the individual component scores, as $N = (O - wI)/(1 - w)$, where w is the interview's share of the course total. In this course, the total is the unscaled sum of four components (quizzes, 200 marks; literature review, 200; case study, 450, which in 2024 entered the total as a moderated mark; interview, 150), and it is converted to a percentage by dividing by 10. Hence $w = 0.15$, and because the component marks are recorded, N can be computed directly for each student as the non-interview marks expressed as a percentage of the remaining 850; no scaling, rounding, or hurdle rule applied to the total prevents this, and the 2024 moderation affects N but not I. With $SD(I) = 11.9$ and $SD(N) = 9.4$ percentage points, the correlation expected from the part–whole relationship alone is 0.22. The observed correlation between the interview mark and the non-interview aggregate is $r(81) = 0.17$, 95% CI [−0.04, 0.38], $p = 0.11$, and it differs sharply between offerings: $r = -0.03$

in 2024 (n = 37; 95% CI [−0.35, 0.30]), r = −0.39 in 2025 (n = 15; [−0.75, 0.15]) and r = 0.73 in 2026 (n = 31; [0.51, 0.86], $p < 0.001$). The pooled r = 0.37 is therefore largely a part–whole effect, and the records give no consistent evidence across offerings of an association between interview performance and performance in the rest of the course; r = 0.37 is retained only as a descriptive result.

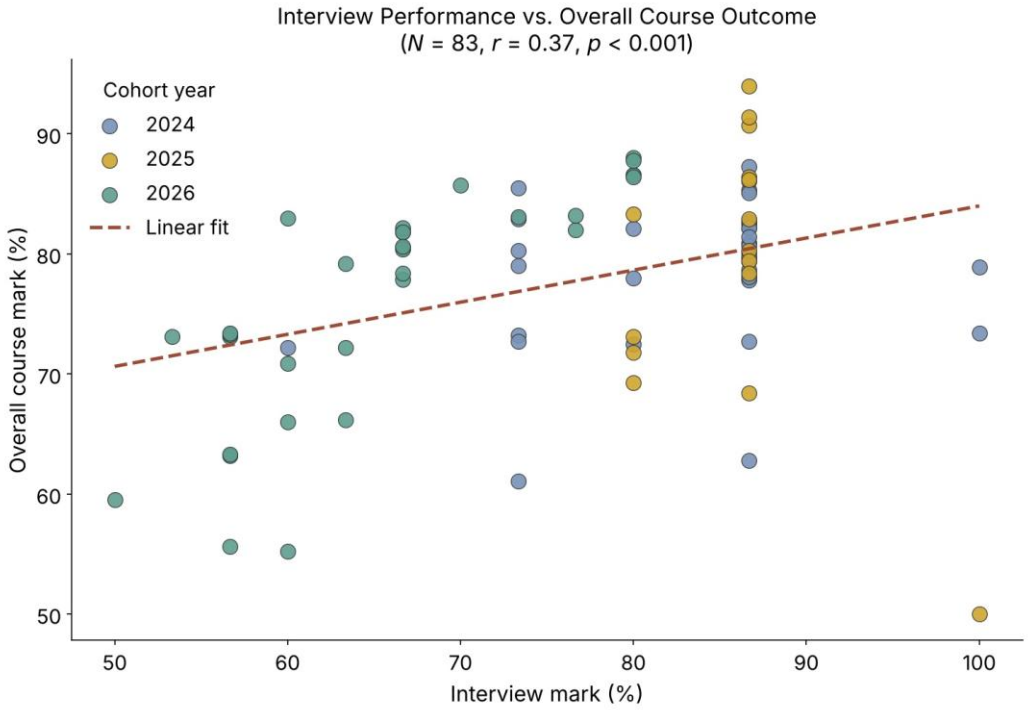


Figure 5. Interview performance (%) plotted against overall course outcome (%) for 83 students in the 2024–2026 offerings.

9.3. Preliminary Outcome Pilot: VivaGen-AI-Supported Interviews (N = 15)

In the 2026 offering, lecturer-reviewed VivaGen-AI questions were used in the regular course interviews of 15 students from the same course as the historical baseline in Section 9.1, chosen by the lecturer to include higher- and lower-performing students (Section 8). Because course and assignment type are held constant, this pilot removes the discipline confound that would apply to a cross-course comparison, though it remains a non-randomized comparison of a group chosen by the lecturer rather than the prospective, randomized within-offering comparison specified in Section 8. The pilot's timing, student selection, and assessor arrangements are reported in Section 8. For each student, the lecturer uploaded the student's own assignment, reviewed and approved the AI-generated ten-question set, conducted the interview in Student Viva Mode, and awarded a raw mark out of 150 using the same course rubric (Appendix C) as for the other interviews; the marks were then moderated (Section 8), and the four understanding bands (Table 4) were applied retrospectively to the moderated marks. Because the pilot records are part of the 2026 baseline, the pilot is compared with the 68 other baseline records; Table 10 and Figure 6 compare the understanding-band distributions of the two groups. Moderated pilot marks ranged from 75 to 120 (mean = 100.3, SD = 16.7; as a percentage: mean = 66.9%, SD = 11.2%), compared with a mean of 119.0 (SD = 16.3; 79.3%) for the other baseline records. Table 11 and Figure 7 summarize these descriptive statistics directly. Figure 8 shows the moderated mark of each pilot student, colored by understanding band, together with the examiner's raw mark; the dashed lines mark the 40%, 60% and 80% thresholds of the Partial, Good and Strong Understanding bands (Table 4).

Table 10. Understanding-band distribution (moderated marks): other baseline records (existing process, N = 68) versus the VivaGen-AI pilot interviews (same course, N = 15).

Understanding Band	Other Baseline Records, Existing Process (N = 68)	VivaGen-AI Pilot, Same Course (N = 15)
Strong Understanding (120–150)	44 (64.7%)	4 (26.7%)
Good Understanding (90–119)	20 (29.4%)	7 (46.7%)
Partial Understanding (60–89)	4 (5.9%)	4 (26.7%)
Further Verification Needed (0–59)	0 (0.0%)	0 (0.0%)

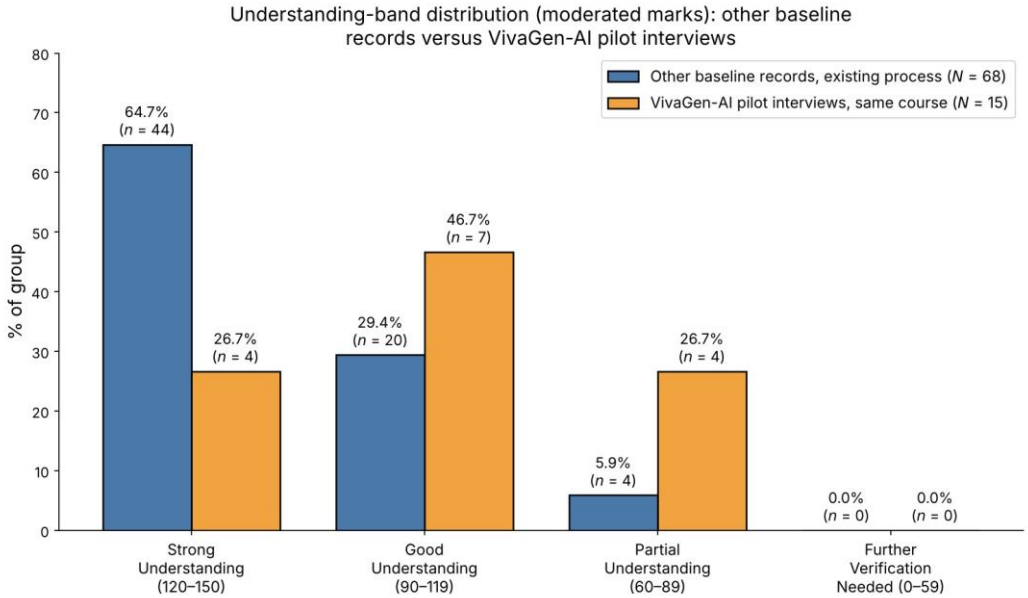


Figure 6. Percentage of students in each understanding band (moderated marks): other baseline records (existing process, N = 68) versus the VivaGen-AI pilot interviews (same course, N = 15); each bar is labeled with the percentage and the number of students (n).

Table 11. Descriptive statistics comparison (moderated marks): interview mark and percentage, other baseline records (N = 68) versus VivaGen-AI pilot interviews (N = 15).

Metric	Other Baseline Records, Existing Process (N = 68)	VivaGen-AI Pilot, Same Course (N = 15)
Interview mark, mean (SD), /150	119.0 (16.3)	100.3 (16.7)
Interview mark, range, /150	75–150	75–120
Interview mark, mean (SD), %	79.3 (10.9)	66.9 (11.2)
95% CI for mean, %	[76.7, 81.9]	[60.7, 73.1]

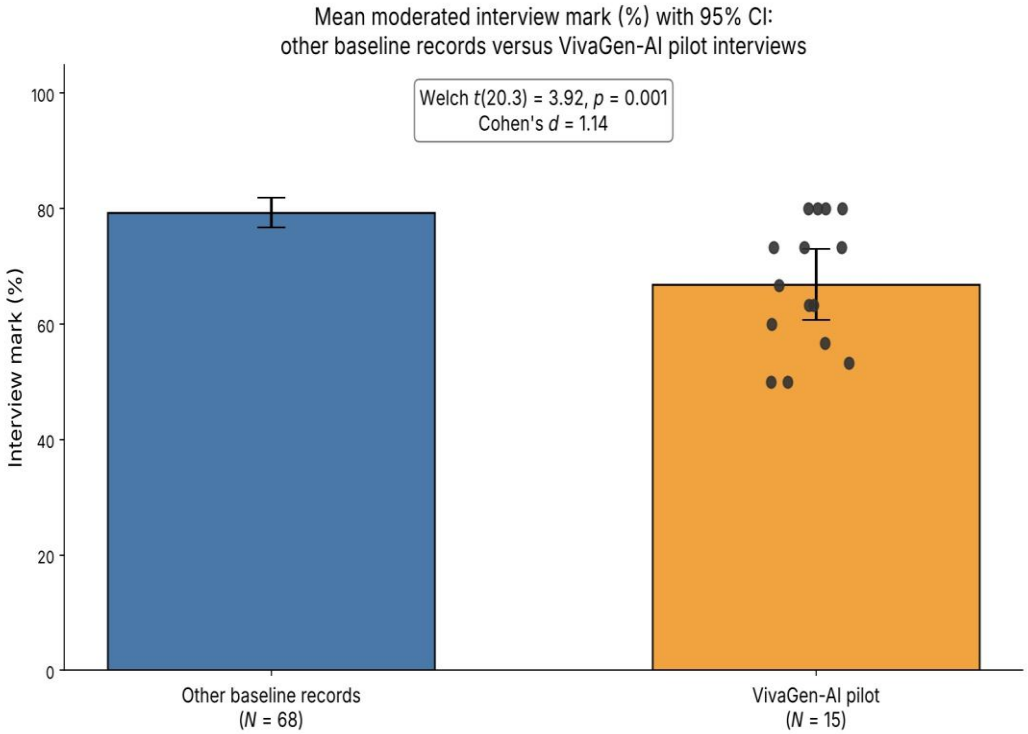


Figure 7. Mean moderated interview mark (%) with 95% confidence intervals, other baseline records versus VivaGen-AI pilot interviews, with individual pilot observations overlaid.

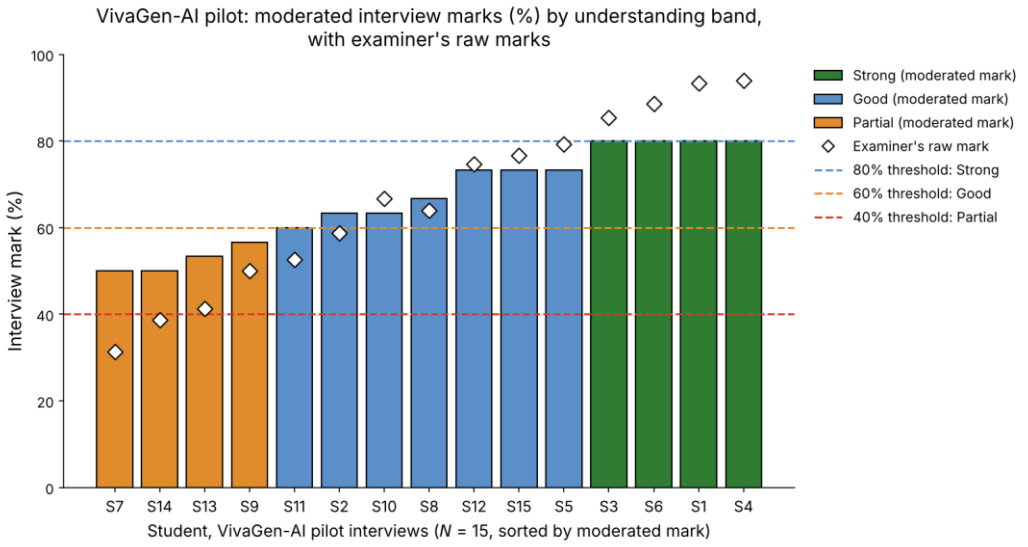


Figure 8. Moderated interview marks (%) for the 15 pilot students, sorted and colored by understanding band, with the examiner's raw marks shown as diamonds.

On moderated marks, the pilot distribution is shifted toward the lower bands relative to the other baseline records: 26.7% of pilot students fell into Partial Understanding or Further Verification Needed, against 5.9% of the other records, and no student in either group fell into Further Verification Needed (two pilot students had raw marks in that range before moderation).

Two kinds of tests were applied to these data, and they address different questions. The first concerns band membership. On the full table of the three occupied bands, an exact test of independence (the Freeman–Halton extension of Fisher's exact test; Freeman

& Halton, 1951) gives $p = 0.008$. Collapsing the table to two categories (Strong/Good versus Partial/Further Verification Needed) separates students whose explanation of their own work the examiner judged largely adequate (60% or more of the interview mark) from those whose explanation was judged incomplete, and so gives a single interpretable effect. Because the grouping was chosen after the data were seen, this comparison is exploratory. Fisher's exact test on the collapsed table gives $p = 0.032$ (sample odds ratio 5.82; conditional maximum-likelihood estimate 5.64, 95% CI [0.91, 35.35]); the interval is wide and includes 1. Fisher's exact test does not require large expected counts, so the collapse is not needed for the test to be valid. A chi-square test is not reported for the full table, because several expected counts were small.

The second test concerns mean marks. A Welch two-sample *t*-test comparing the pilot's moderated marks with those of the other baseline records (mean = 119.0, SD = 16.3, N = 68), computed on marks out of 150 with Welch–Satterthwaite degrees of freedom, gives $t(20.3) = 3.92$, $p = 0.001$. The mean difference (other records minus pilot) is 18.6 marks, 95% CI [8.7, 28.5], and the standardized difference is Cohen's d = 1.14 (the mean difference divided by the pooled SD of 16.4 marks; Hedges' g = 1.13; approximate 95% CI [0.54, 1.71]). On this pooled comparison, the pilot students scored lower, but the cohort-year analysis below shows that the difference arises from the 2024 and 2025 offerings, not from within 2026. Before moderation, the examiner's raw pilot marks were more dispersed (mean 99.5, SD 30.6, range 47–141; 6 of 15 students below 60% of the interview mark), as expected for students selected to include both higher- and lower-performing students. As a secondary analysis, raw marks give a similar pooled difference (19.4 marks, 95% CI [2.1, 36.7]) and again no statistically detectable difference from the other 2026 interviews (−2.7 marks, 95% CI [−20.1, 14.8]), so moderation narrowed the spread of the pilot marks without changing the pattern of results. Table 11 reports the descriptive statistics, and Figure 7 plots the two groups with 95% confidence intervals and the individual pilot observations.

Cohort-year sensitivity analysis: The pooled baseline combines offerings with very different band distributions (Table 9): all 2025 students and 75.7% of 2024 students were in Strong Understanding, whereas 25.8% of the 2026 cohort were already in Partial Understanding. Because the composition of the pooled comparator affects the pilot comparison, Table 12 repeats the comparison against each baseline year separately; for 2026, the comparator is the 16 other interviews of that offering, which used examiner-prepared questions, so the groups do not overlap. The pilot differs clearly from the 2024 and 2025 offerings, but there is no statistically detectable difference from the other 2026 interviews (collapsed table: 4/15 versus 4/16 in the lower two bands, Fisher's exact $p = 1.000$, odds ratio 1.09, 95% CI [0.16, 7.44]; full table, $p = 0.294$). The 2026 comparison does not show that the two distributions are equivalent: with these sample sizes, its confidence interval is compatible both with no difference and with a substantial one. Moreover, the pilot students were chosen by the lecturer, not allocated at random; a within-offering comparison with random allocation is recommended in Section 13. It does show that the pooled result depends heavily on the 2024 and 2025 cohorts, and that differences between baseline years (Section 9.1) are larger than the difference between the pilot and the other interviews of its own offering. Year-specific comparisons of mean marks, computed from the individual baseline records, show the same pattern more clearly (Table 12): the pilot mean was 24.3 marks lower than in 2024 (95% CI [14.3, 34.2]; Welch $p < 0.001$) and 28.3 marks lower than in 2025 (95% CI [18.4, 38.2]; $p < 0.001$), but 3.5 marks higher than the other 2026 interviews (comparator minus pilot −3.5, 95% CI [−13.7, 6.8]; $p = 0.49$; d = −0.26). The pilot mean (100.3) and the mean of the other 2026 interviews (96.9) are therefore similar, and the pooled mean difference is driven by the 2024 and 2025 offerings. This again does not establish equivalence, because the confidence interval for the 2026 comparison is wide. In any case,

statistical comparisons of this kind cannot remove the underlying lack of comparability between the pilot and baseline cohorts.

Table 12. Cohort-year sensitivity analysis (moderated marks): proportion of students in the two lower understanding bands (Partial Understanding or Further Verification Needed) and mean interview mark, VivaGen-AI pilot versus each baseline comparator.

Comparator	n	Lower Two Bands, n (%)	Fisher 2 × 2: *p*	Odds Ratio [95% CI]	Exact, Full Table: *p*	Mean Mark (SD)	Mean Difference [95% CI]	Welch: *p*
VivaGen-AI pilot (reference)	15	4 (26.7%)	—	—	—	100.3 (16.7)	—	—
Other baseline records, 2024–2026	68	4 (5.9%)	0.032	5.64 [0.91, 35.35]	0.008	119.0 (16.3)	18.6 [8.7, 28.5]	0.001
2024 offering	37	0 (0.0%)	0.005	Not estimable *	<0.001	124.6 (12.2)	24.3 [14.3, 34.2]	<0.001
2025 offering	15	0 (0.0%)	0.100	Not estimable *	<0.001	128.7 (7.4)	28.3 [18.4, 38.2]	<0.001
2026 offering, other interviews	16	4 (25.0%)	1.000	1.09 [0.16, 7.44]	0.294	96.9 (9.6)	−3.5 [−13.7, 6.8]	0.491

Odds ratios are conditional maximum-likelihood estimates (odds of a pilot student falling in the lower two bands relative to the comparator), with exact 95% confidence intervals. The full-table *p*-values are from the Freeman–Halton exact test on the occupied bands. * No comparator student fell in the lower two bands, so the odds ratio is unbounded above (lower 95% limits: 1.85 for 2024; 0.73 for 2025). The 2026 comparator excludes the 15 pilot interviews, so all comparisons are between independent groups. Mean marks are moderated marks out of 150; mean differences are comparator minus pilot, with Welch 95% confidence intervals computed from the individual records. All tests are two-sided and exploratory; no correction for multiple comparisons was applied.

The examining lecturer's notes on the six pilot students whose raw marks fell below 60% of the interview mark (two in the Further Verification Needed range and four in the Partial Understanding range on raw marks; on moderated marks, four were in Partial Understanding and two in Good Understanding) converge on a common pattern: these students could often state what their submission concluded but could not explain the underlying calculation or methodology when asked to justify it, particularly at the Analyze and Evaluate levels of the sequence, whereas students in the Strong Understanding band justified their own calculations without prompting and defended their methodology choices across the full ten-question set. These notes are illustrative only. They were not analyzed systematically, and they cannot distinguish between a sequence that surfaces genuine gaps in understanding and questions that were harder, or less well matched to what students could reasonably be expected to explain, than those used in the baseline years. Nor does difficulty in explaining a submission show that a student did not produce it: the four-band scheme records demonstrated understanding in the interview, not authorship. The examining lecturer recorded a short note for each of the 15 pilot interviews, and the pattern reported here for these six students was identified by the authors from a reading of the notes, without a coding scheme or a second coder.

10. Discussion

The baseline findings bear on one claim made in Section 2. The interview–course correlation (Section 9.2) is a part–whole association and is not used here as evidence that oral assessment captures something distinct from written work. What the baseline does show is that the substantial year-to-year swing in outcomes (Table 9), together with the complete absence of low-scoring outcomes across 83 real interviews (among the students with

a recorded mark, which excludes a disproportionately low-performing group; Section 9.1), is consistent with documented concerns about the reliability of unstructured, marker-dependent oral assessment (Joughin, 1998), which is one motivation for a criterion-based, Bloom-sequenced approach in which every question is classified before the interview takes place rather than improvised by the examiner.

The preliminary pilot in Section 9.3 adds a first, real signal about VivaGen-AI's own use, but it admits more than one explanation, and the paper deliberately does not adjudicate between them. One reading, consistent with the qualitative pattern in the lecturers' notes (which, as noted in Section 9.3, are illustrative only), is that the ten-question, Bloom-sequenced structure surfaces a gap between what a submission states and what its author can independently defend, a gap the existing, unstructured interview process may under-detect. A second, equally available reading is that the shift reflects ordinary cohort-to-cohort variation rather than any effect of VivaGen-AI itself, since the pilot is a small sample of 15 students chosen by the lecturer to include higher- and lower-performing students, whose marks are therefore expected to be more widely spread than those of a whole cohort, and the comparisons with 2024 and 2025 are across different cohorts. The cohort-year sensitivity analysis (Table 12) reinforces this caution: against the other interviews of the 2026 offering, there is no statistically detectable difference in either the band distribution or the mean mark (96.9 versus 100.3), although the wide confidence intervals mean that this is not evidence of equivalence. A third possibility is that the AI-generated questions were, in some cases, harder or less well calibrated than the questions used under the existing process, something the platform's own difficulty labels cannot verify without the lecturer/expert quality rating specified in Section 8. A fourth is examiner behavior: the pilot and baseline interviews were not all marked by the same examiners (Section 8), so differences in marking stringency could produce the same pattern. The data collected so far cannot distinguish these explanations, and the paper does not claim to have done so; distinguishing them is precisely the purpose of the randomized within-offering comparison and the lecturer/expert-validation instrument that remain to be executed.

What the pilot does establish is narrower but real: VivaGen-AI can be used, by a lecturer, to conduct and score genuine oral interviews for real assignment submissions in the course, and in this small pilot the resulting marks were distributed differently from those of the other baseline records. Lower marks do not, by themselves, indicate more accurate or more valid assessment. Whether that difference reflects question difficulty, examiner behavior, cohort composition, assessment quality or other factors is unresolved. It does not establish, and this paper does not claim, that VivaGen-AI improves on, or matches, the course's existing interview process, since the pilot students were chosen by the lecturer rather than at random, were examined by a different examiner, and, for the 2024 and 2025 comparisons, come from a different offering. Closing that specific evidentiary gap requires the randomized within-offering comparison and the lecturer/expert-validation instrument specified in Section 8, neither of which has yet been conducted.

11. Limitations

Internal validity: The pilot was small (N = 15), chosen by the lecturer to include higher- and lower-performing students, and not randomized; the question source is confounded with the examiner and, for the 2024 and 2025 comparisons, with cohort and offering (Section 8; Table 8); moderation narrowed the range of the pilot marks (from 47–141 to 75–120), although results on the raw marks follow the same pattern (Section 9.3); and moderation was carried out by the course coordinator, who also examined the pilot interviews and knew the question source, so it was neither blind nor independent. The baseline offerings differ markedly from one another, so the pooled comparison depends on their composition (Table 12), and 13 baseline records without an interview mark were

excluded. Baseline students without an interview mark had markedly lower course marks than those analyzed (Section 9.1). All statistical comparisons are exploratory, and there is no statistically detectable difference between the pilot and the other interviews of the 2026 offering in either band distribution or mean mark, which does not establish equivalence (Table 12).

Construct validity: The four understanding bands, whose cut-points are those of the existing course rubric and which were applied retrospectively, are unvalidated descriptive categories, and the interview–course correlation is a part–whole association (Section 9.2). The Bloom levels and Easy/Medium/Hard labels of the generated questions are intended classifications; the only check of them reported here was made by the authors (Appendix A), and no item-difficulty data were collected. Because the expert validation has not yet been completed, the quality of the generated questions has not been evaluated directly. The lecturers' qualitative notes were not analyzed systematically. The interview measures demonstrated understanding of the submission, not authorship.

Implementation: MSVGA's stages are realized as instructions within one prompt and two browser-code steps, not as independently executed and validated stages (Section 5). The parser checks format only and accepts replies with fewer than ten questions, unchecked labels, or up to two placeholder items (Appendix D), and the platform offers no in-platform editing. Only the first 12,000 characters of a submission are used, and image-based content, equations stored as images, and table layout are lost in extraction. The model version is not controlled by the platform and may change without notice. In the pilot, assignment text was entered into the free web version of ChatGPT under an individual account, without an institutional agreement and without students being asked to consent to this processing; names and student numbers were removed from the prompt, but the platform does not remove other identifying content in a submission.

External validity and generalizability: The results come from one course, one discipline, one institution, and one LLM service, with small samples and no random sampling of students. They may not generalize to other disciplines, assessment types, student populations, institutions, or models, and the findings should be generalized only to settings that resemble the one studied (Section 1.5).

12. Conclusions

This paper presents the Assignment-to-Viva Assessment Framework (AVAF), a human-centered design that combines authentic assessment principles and Bloom's Revised Taxonomy to generate a personalized oral examination, sequenced by intended cognitive level, from a student's own submitted assignment, subject to lecturer review before use. AVAF is operationalized as the Assignment-to-Viva Question-Generation Algorithm (AVQGA) and demonstrated through VivaGen-AI, a self-contained, browser-based platform that performs client-side document parsing and routes structured prompts to a lecturer-supplied conversational LLM. Applied descriptively to three offerings of historical interview data (N = 83), the interpretation scheme provides a reference distribution for future evaluation; the interview–course correlation (r = 0.37) is largely a part–whole effect, falling to r = 0.17 (p = 0.11) when the interview is removed from the course mark, and is not evidence of the scheme's validity. A preliminary pilot in which VivaGen-AI was used for real interviews in the regular course interviews of 15 students in the 2026 offering shows lower moderated marks than the other baseline records (mean difference 18.6 marks, 95% CI [8.7, 28.5]) but no statistically detectable difference from the other interviews of its own offering (−3.5 marks, 95% CI [−13.7, 6.8]), which does not establish equivalence; the pooled difference reflects differences between offerings. Because the pilot was small, chosen by the lecturer and not randomized, these results cannot show that the questions assess understanding more accurately; question difficulty, examiner differences, and

cohort differences remain open explanations. Lower marks do not establish better assessment, and the understanding bands are unvalidated descriptive categories. The paper's central contribution is this framework, algorithm, architecture and working artefact, together with a real baseline and an early, appropriately caveated outcome signal; it does not yet claim, and should not be read as claiming, that VivaGen-AI improves on existing practice within a matched course, a question the specified randomized within-offering comparison and lecturer/expert-validation protocol are designed to answer. Until the direct evaluation of question quality is reported, what has been demonstrated is limited to three things: the workflow can produce a question set of ten items labelled with intended Bloom's levels from a student's submission, in the specified format, for lecturer review, although the software does not enforce the number of items or check the labels and the alignment of the items with those levels has not been independently validated; a lecturer has used it in live interviews in one course; and the resulting marks were distributed differently from a pooled historical sample, for reasons that remain unresolved.

13. Recommendations

The following recommendations follow directly from the findings; the finding on which each rests is given in brackets.

(1) Lecturers who use VivaGen-AI, or a similar tool, should read every generated question against the submission before use and correct or replace items whose level or content is wrong. [The parser checks format only and accepts incomplete or mislabeled question sets (Appendix D), and 3 of the 10 demonstration questions were not at their intended level (Appendix A).]

(2) Developers should add automatic checks that a reply contains exactly ten questions with the requested labels, in-platform editing of questions, more robust extraction of the reply, and prevention of duplicate records. [Test cases R4, R7, R12, R13 and S6 in Appendix D.]

(3) Course teams that use oral assessment should adopt a common written rubric, brief examiners on it, and have a sample of interviews in each offering double-marked by a second examiner who does not know the question source. [Marking practice and mark distributions differed markedly between the baseline offerings (Section 9.1; Table 9), and in this study moderation was carried out by a single coordinator who knew the question source (Section 8).]

(4) Researchers and institutions should not treat the correlation between interview and course marks as evidence of validity, and should instead relate interview performance to the non-interview components and to independent criteria. [The pooled correlation of 0.37 fell to 0.17 when the interview was removed and differed in sign between offerings (Section 9.2).]

(5) Before any claim of effectiveness is made, a randomized within-offering comparison should be conducted, with the same examiners and rubric in both arms, covariates such as prior performance recorded and the analysis plan registered in advance. Within an offering, interviews that use VivaGen-AI questions should be compared with those that use examiner-prepared questions. [The pilot differed from the other baseline records, but there was no statistically detectable difference from the other interviews of the 2026 offering, and the differences between baseline offerings were larger than that between the pilot and its own offering (Section 9.3; Table 12).]

(6) Institutions should require identifying information to be removed from submissions before upload, use an AI service and account settings that the institution has approved, and tell students before the assessment that text from their assignment will be processed by an external service. [Assignment-derived text leaves the device when the prompt is pasted into the external service (Section 6).]

https://doi.org/10.3390/educsci16122007

(7) For group assignments, the platform should allow each student's declared contribution to be entered, and lecturers should adapt the questions accordingly. [The current build generates questions from the whole submission (Section 7).]

(8) The expert validation of question quality should be completed with a panel that covers curriculum and instruction, measurement and evaluation, artificial intelligence and the subject discipline, before the tool is used for summative decisions. [Question quality has so far been checked only by the authors (Section 7; Appendix A).]

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci16122007/s1, S1, the functional test script and results summarized in Appendix D; S2, the reference build of VivaGen-AI (single HTML file) identified in Section 6.

Author Contributions: Conceptualization, M.M.I.; methodology, M.M.I.; software, S.D.; validation, S.C.; formal analysis, S.D.; investigation, S.C.; data curation, S.D.; writing—original draft preparation, M.M.I. and S.C.; writing—review and editing, S.D. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a University of Southern Queensland (UniSQ) Learning and Teaching Grant (2025).

Institutional Review Board Statement: The study was approved by the University of Southern Queensland Human Research Ethics Committee (approval ETH2026-0687, 3 November 2026) as a retrospective analysis of existing, de-identified assessment records, and was conducted in accordance with the National Statement on Ethical Conduct in Human Research (2023). The approval covers this analysis only; the entry of assignment text into an external AI service during routine assessment in 2026 preceded the application, was disclosed to the Committee and is not covered by the approval (Section 8).

Informed Consent Statement: Individual informed consent was not obtained. The study involved only the secondary analysis of existing assessment records generated during routine course assessment, which were de-identified before analysis and are reported only in aggregate. The University of Southern Queensland Human Research Ethics Committee approved the study (ETH2026-0687, 3 November 2026) on the basis of an application that sought a waiver of the requirement for consent.

Data Availability Statement: (a) Interview-mark records (baseline and pilot) and (b) students' assignments: these are not publicly available and cannot be shared, because the conditions of the ethics approval restrict their use to this project; the cohorts are small, and the records could allow students to be re-identified. Aggregate results are reported in full in the article. (c) Generated question sets: because they are derived from students' assignments, they are subject to the same restriction; the demonstration question set is reported in full in Table 6. (d) Software artifact: the reference build of VivaGen-AI and the functional test script are provided as Supplementary Materials S1 and S2, and the complete prompt template, output schema and parsing rules are reproduced in Appendix B.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AIAS	Artificial Intelligence Assessment Scale
AQG	Automatic question generation
AVAF	Assignment-to-Viva Assessment Framework

AVQGA	Assignment-to-Viva Question-Generation Algorithm
CI	Confidence interval
CVI	Content validity index
DSR	Design-science research
FEDS	Framework for Evaluation in Design-Science
HIC	Hydrogen-induced cracking
JSON	JavaScript Object Notation
LLM	Large language model
MSVGA	Multi-Stage Viva Generation Architecture
OR	Odds ratio
PWHT	Post-weld heat treatment
RQ	Research question
SD	Standard deviation

Appendix A. Operational Criteria for Bloom's Levels and the Authors' Check of the Demonstration Questions

Table A1 gives the operational criteria used to classify an oral question about a student's own submission. Where a question fits two levels, it is classified at the highest cognitive process that a complete answer necessarily requires. Table A2 applies these criteria to the ten demonstration questions in Table 6. The check was made by the authors and was not blinded; it documents intended alignment and is not a substitute for the independent expert classification specified in Section 8. The intended level is the level requested by the prompt (Appendix B.1). For two items, the label returned by the model differs from the requested one: Question 2 was requested as Remember/Understand and returned as Understand, and Question 6 was requested as Analyze (Medium) and returned as Analyze (Medium/Hard).

Table A1. Operational criteria for assigning Bloom's Revised Taxonomy levels to viva questions.

Level	Operational Criterion (Oral Question on the Student's Own Submission)	Typical Stems
Remember	Retrieve a fact, value, or statement that appears in the submission; a complete answer needs recall only.	State, name, identify
Understand	Explain the meaning, cause, or interpretation of content in the submission in the student's own words; no procedure must be carried out.	Explain, interpret, what does X indicate
Apply	Carry out a procedure, calculation, or criterion used in the submission on the submitted values or on a value changed by the examiner.	Show how, calculate, use X to
Analyze	Break down the submission's reasoning: relate parts, identify assumptions, or separate evidence from inference.	How does X affect Y, what assumptions
Evaluate	Make and justify a judgement against a criterion, such as the adequacy of a conclusion or the merits of a choice against an alternative.	Evaluate, justify, defend, is X sufficient
Create	Produce something not in the submission, such as a revised method, design or plan.	Propose, design, what would you change

Table A2. Authors' check of the intended Bloom levels of the ten demonstration questions (Table 6) against the criteria in Table A1.

#	Intended Level	Process Required by the Wording	Outcome	Note
1	Remember	Recall of own stated finding	Consistent	Second clause (linking evidence) adds a light Understand element.

2	Remember/Understand (returned: Understand)	Explain why one mechanism was identified	Consistent with Understand	Comparison with other modes approaches Analyze; returned label differs from requested label.
3	Understand	Interpret measured values	Consistent	—
4	Apply	Retrospective explanation of a selection	Adjacent level (Understand)	Describing how a procedure was selected is not applying it; Question 5 is the true Apply item.
5	Apply	Apply a criterion to justify parameters	Consistent	Wording hedges between two criteria ("hydrogen diffusion or hardness"); possible factual-specificity weakness.
6	Analyze (Medium; returned: Medium/Hard)	Compare chosen procedure with alternatives	Spans Analyze/Evaluate	Returned difficulty label differs from the requested "Medium"; overlaps with Question 9.
7	Analyze	Identify effect of assumptions on conclusions	Consistent	—
8	Evaluate	Judge adequacy of a recommendation	Consistent	—
9	Evaluate	Defend a choice against an alternative	Consistent	Defending an existing decision is Evaluate, not Create.
10	Create/Defend	Propose a revised method	Spans Create/Evaluate	The combined label merges two processes; relabel as Create and require a concrete revised method.

Appendix B. Prompt Template, Output Schema and Parsing Rules

Appendix B.1. Prompt Template

The Prompt Generation module assembles the following prompt, reproduced verbatim from the reference build; only the final block varies between submissions. Line breaks are as in the original.

You are an academic viva assistant. Read the following student's submitted assignment and:

(1) Identify its structure: topic, objectives, methods, calculations, tables, findings, conclusions, and key parts a student should be able to explain.

(2) Generate exactly 10 viva questions that test whether the student understands their own work, progressing from easy to difficult using Bloom's Taxonomy:

Q1: Remember (Easy)
Q2: Remember/Understand (Easy)
Q3: Understand (Easy)
Q4: Apply (Medium)
Q5: Apply (Medium)
Q6: Analyze (Medium)
Q7: Analyze (Hard)
Q8: Evaluate (Hard)
Q9: Evaluate (Hard)
Q10: Create/Defend (Hard)

Favor academic-integrity style questions at the higher levels—for example asking the student why they chose a method, to explain a result in their own words, how they calculated a value, what was hardest about the assignment, what would happen if an assumption changed, or what they'd improve next time. Do not accuse the student of cheating or using AI anywhere—your only job is to help the teacher verify the student's own understanding and ownership of the work.

For each question include: the question, its Bloom's level, difficulty, which part of the assignment it comes from, a short expected answer for the teacher, and one follow-up question.

IMPORTANT: Every value below must be REAL content you write based on the actual assignment text—specific, concrete sentences about this student's actual topic, methods, numbers and findings. Never write "..." or any placeholder text, and never repeat this template back—the shape below is only an example of the JSON structure, not something to copy literally.

Respond with ONLY valid JSON, no markdown fences, no commentary before or after. The structure (fill every field with real content, do not leave any "..."):

{"topic":"the assignment's actual topic in one sentence","objectives":["a real objective from the assignment"],"methods":["a real method used"],"calculations":["a real calculation, or empty array if none"],"tables":["a real table description, or empty array if none"],"findings":["a real finding"],"conclusions":["a real conclusion"],"questions":[{"n":1,"question":"a real question about this specific assignment","bloomLevel":"Remember","difficulty":"Easy","section":"which part it refers to","expectedAnswer":"a real short expected answer","followUp":"a real follow-up question"}]}

The "questions" array must contain exactly 10 filled-in objects like the one above, with "n" running from 1 to 10, matching the levels and difficulties listed earlier in order.

Assignment text:

"""

[ASSIGNMENT TEXT—the first 12,000 characters of the text extracted from the submission]

"""

[Appended only when the extracted text is longer than 12,000 characters:] [Note: assignment text was long, so this is the first portion of it.]

Appendix B.2. Output Schema

The model is asked to return a single JSON object with the keys topic (string), objectives, methods, calculations, tables, findings and conclusions (arrays of strings), and questions (an array of ten objects). Each question object has the keys n (1–10), question, bloomLevel, difficulty, section (the part of the assignment the question refers to), expectedAnswer (shown to the lecturer as the model answer) and followUp. The structural fields are shown to the lecturer in Teacher Mode as a summary of the submission; the question objects become the ten question records.

Appendix B.3. Parsing and Validation Rules

The reference build applies the following rules, in order. (1) Input: only .pdf and .docx files are accepted, and a student name must be entered before a prompt can be prepared. (2) Extraction: PDF text is read from the text layer with PDF.js, from at most the first 30 pages, stopping once more than 24,000 characters have been read; DOCX text is read as raw text with Mammoth. Images, scanned pages, equations or figures stored as images, and text inside figures are not captured; headings and table layout are flattened into running text. (3) Sufficiency: if fewer than 50 characters are extracted, as happens with scanned PDFs, the process stops with an error. (4) Prompt: the first 12,000 characters of the extracted text are inserted into the template in B.1, with a truncation note when the text is longer. (5) Reply: the pasted reply must contain at least 20 characters; Markdown code fences are removed, the text between the first opening brace and the last closing brace is parsed as JSON, and the reply is rejected if it cannot be parsed or contains no non-empty questions array. (6) Placeholder check: the reply is rejected if three or more questions have an empty or ellipsis-only question or expected answer. (7) Normalization: question objects

without question text are discarded, the rest are sorted by n, and at most ten are kept. When a reply is rejected, the platform shows an error message telling the lecturer what to ask ChatGPT for, and the lecturer pastes a new reply; there is no automatic regeneration. These checks concern format only. The parser does not require exactly ten questions, does not check that the Bloom and difficulty labels match the requested sequence, and does not check source references, factual correctness, or cognitive alignment; these rest with the lecturer's review.

Appendix C. Interview Marking Rubric

The interview is marked out of 150 under the course's existing interview rubric, which was used unchanged in the pilot and in the 2024, 2025 and 2026 baseline offerings. The examiner awards a single holistic mark out of 150 for the student's answers to the ten questions and their follow-up probes, guided by the four level descriptors in Table A3. The level descriptors and mark ranges are those of the existing rubric, which examiners used during the interviews in all offerings; the authors adopted them for the rating screen of VivaGen-AI (Figure 3d) and for the interpretation scheme in Table 4, attaching the four understanding-band labels, which were applied retrospectively to the recorded marks for analysis only. The examiner's raw marks were moderated by the course coordinator before entry in the gradebook (Section 8), but no double-marking was carried out, so inter-rater reliability could not be estimated. The understanding band reported in Section 9 was derived from the moderated mark using the boundaries in Table 4.

Table A3. Interview marking rubric: levels, mark ranges and descriptors (out of 150).

Level	Mark Range (Out of 150)	Percentage	Descriptor
Strong Understanding	120–150	80–100%	Explained ideas fluently in their own words, including the reasoning behind their choices.
Good Understanding	90–119	60.0–79.3%	Answered most questions well, with minor gaps on the harder ones.
Partial Understanding	60–89	40.0–59.3%	Could explain some parts but struggled to justify methods or results.
Further Verification Needed	0–59	0.0–39.3%	Notable gaps in explaining core parts of the submission; a follow-up conversation is warranted.

Appendix D. Functional Test Suite

The tests were run on 21 October 2026 against the reference build identified in Section 6 (SHA-256: ac0f3f1bf77dbccf419aea0547c95fba4299bea6ab8760238aac9f499ef77a81), in headless Chromium driven by the Playwright automation library. Each case sets up an input, calls the platform's own functions, and compares the outcome with an expected result defined before the run. The three third-party libraries were replaced by deterministic stubs (a simulated PDF document with a set number of pages and characters per page, a fixed DOCX text, and a spreadsheet writer that records the sheets it receives), so the suite tests the platform's own logic and not the extraction quality of PDF.js or Mammoth on real documents. Cases whose expected result begins "Documented behavior" or "Documented failure mode" pass when the platform behaves as its code specifies; they are listed because they are limitations of the current build. The test script is provided as Supplementary Material S1.

Table A4. Functional test cases, expected and observed results.

ID	Area	Test Case	Expected Result	Observed	Result
I1	Input	DOCX selected, student name blank	Prompt preparation blocked	Prepare button disabled = True	Pass
I2	Input	Unsupported file type (.txt)	Rejected with message	Please upload a PDF or DOCX file	Pass
P1	Prompt	DOCX with student name and ID entered	Prompt contains document text; name and ID not inserted	Document text present; name absent; ID absent	Pass
P2	Prompt	Short document (<12,000 characters)	No truncation note	No truncation note	Pass
P3	Prompt	DOCX of 20,000 characters	First 12,000 characters inserted, truncation note appended	12,000 characters inserted; note present = True	Pass
E1	Extraction	PDF with 40 text pages	Only pages 1–30 read	Last page read = 30; page 31 text present = False	Pass
E2	Extraction	PDF with 10 pages of 5000 characters each	Reading stops once > 24,000 characters	Stopped after page 5 (25,035 characters)	Pass
E3	Extraction	Image-only (scanned) PDF: no text layer	Stopped with error (<50 characters)	Could not read enough text from this file. Try a different file	Pass
E4	Extraction	DOCX yielding 10 characters	Stopped with error (<50 characters)	Could not read enough text from this file. Try a different file	Pass
R1	Reply parsing	Valid JSON, 10 questions	Accepted, 10 records	10 records; page = questions	Pass
R2	Reply parsing	Valid JSON inside Markdown code fences	Fences removed, accepted	10 records	Pass
R3	Reply parsing	JSON with commentary before and after	Commentary ignored, accepted	10 records	Pass
R4	Reply parsing	Only 7 questions returned	Documented behavior: accepted (exactly ten not enforced)	Accepted with 7 records, no warning	Pass
R5	Reply parsing	12 questions returned	First 10 by n kept	10 records kept (n = 1–10)	Pass
R6	Reply parsing	Template echoed back ("…" in every field)	Rejected (placeholder check)	Rejected: template-echo message shown	Pass
R7	Reply parsing	2 of 10 questions are "…" placeholders	Documented behavior: accepted (threshold is 3)	Accepted with 10 records, including 2 placeholder items	Pass
R8	Reply parsing	Truncated reply (invalid JSON)	Rejected with re-copy instruction	Rejected: JSON parse error message shown	Pass
R9	Reply parsing	JSON without a "questions" array	Rejected	Rejected: missing questions list	Pass
R10	Reply parsing	Reply shorter than 20 characters	Rejected	Rejected: reply too short	Pass
R11	Reply parsing	Questions returned out of order	Sorted by n	Order after parsing: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	Pass
R12	Reply parsing	All 10 labelled "Remember"/"Easy"	Documented behavior: accepted (labels not validated)	Accepted; all labels "Remember", no warning	Pass
R13	Reply parsing	Commentary after JSON containing braces	Documented failure mode: rejected	Rejected: last-brace slicing captures trailing text	Pass

R14	Reply parsing	Questions missing expectedAnswer, section, and followUp	Documented behavior: accepted with empty fields	Accepted with 10 records; empty fields	Pass
X1	Rendering	HTML/script markup inside a returned question	Escaped; not executed	Executed = False; image elements created = 0	Pass
V1	Viva mode	Expected answer visibility	Hidden by default; shown when Teacher view is switched on	Hidden by default = True; shown on toggle = True	Pass
V2	Viva mode	Advance through all 10 questions	Rating screen shown	Rating screen shown = True	Pass
S1	Records	Storage when no host storage interface is present	Record saved to localStorage	Saved as plain (unencrypted) JSON with name, ID, questions, expected answers, and rating	Pass
S2	Records	Search by topic, then by student ID	Topic matches; ID not a search field	Topic search: "11 of 11 records"; ID search: "0 of 11 records"	Pass
S6	Records	Several accepted replies in one session (same student)	Documented behavior: one new record per accepted reply	11 records created for one student from 11 accepted replies	Pass
S3	Records	Export	Two-sheet workbook including expected answers	File VivaGen-AI-Student-Records-2…; sheets = ['Student Records', 'Viva Questions']; question rows = 107	Pass
S4	Records	Delete a record	Record removed from storage	1 record removed	Pass
S5	Records	Host persistent key–value storage interface present	Record saved there, not in localStorage	Saved to host storage = True; localStorage used = False	Pass

References

Akimov, A., & Malin, M. (2020). When old becomes new: A case study of oral examination as an online assessment tool. *Assessment & Evaluation in Higher Education*, *45*(8), 1205–1221. https://doi.org/10.1080/02602938.2020.1730301.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete edition). Addison Wesley Longman.

Ateeq, A., Alzoraiki, M., Milhem, M., & Ateeq, R. A. (2024). Artificial intelligence in education: Implications for academic integrity and the shift toward holistic assessment. *Frontiers in Education*, *9*, 1470979. https://doi.org/10.3389/feduc.2024.1470979.

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment and Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321.

Belkahla, S., & Azmi, N. (2026). Cognitive demand in higher education summative assessment: A two-dimensional analysis using Bloom's revised taxonomy. *British Journal of Teacher Education and Pedagogy*, *5*(6), 31–44. https://doi.org/10.32996/bjtep.2026.5.6.3.

Blessing, E. (2026). Artificial intelligence in medical education: A systematic review of teaching, learning, and academic integrity issues. *International Journal of Applied Resilience and Sustainability*, *2*(3), 16–48. https://doi.org/10.70593/deepsci.0203002.

Cao, H., & Zahid, S. (2025). Automated viva voce using generative AI for student coursework authentication. In *Proceedings of the 2025 International Conference on Educational Technology and Artificial Intelligence (ETAIC 2025)*. ACM. https://doi.org/10.1145/3766557.3766564.

Church, I. M., Drake, L., & Harris, M. (2025). Using LLMs to support assessment of student work in higher education: A viva voce simulator. *arXiv*, arXiv:2511.05530. https://doi.org/10.48550/arXiv.2511.05530.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

Curbo, S., Lindkvist, A., & Hultgren, C. (2024). Unlocking learning: Exploring take-home examinations and viva voce examinations in microbiology education for biomedical laboratory science students. *Journal of Microbiology & Biology Education*, *26*(1), e00193-24. https://doi.org/10.1128/jmbe.00193-24.

Deepshikha, D. (2025). A comprehensive review of AI-powered grading and tailored feedback in universities. *Discover Artificial Intelligence*, *5*, 251. https://doi.org/10.1007/s44163-025-00517-0.

Duong-Trung, N., Wang, X., & Kravčík, M. (2024). BloomLLM: Large language models based question generation combining supervised fine-tuning and Bloom's Taxonomy. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education* (EC-TEL 2024, LNCS Vol. 15160, pp. 93–98). Springer. https://doi.org/10.1007/978-3-031-72312-4_11.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, *38*(1–2), 141–149. https://doi.org/10.2307/2332323.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625.

Hogan, N. R., & Corăbian, G. (2025). Generative artificial intelligence in violence risk assessment: Emerging technology and the ethics of the inevitable. *Behavioral Sciences & the Law*, *43*(6), 606–615. https://doi.org/10.1002/bsl.70014.

Hou, X., Forsyth, C., Andrews-Todd, J., Rice, J., Cai, Z., Jiang, Y., Zapata-Rivera, D., & Graesser, A. (2025). An LLM-enhanced multi-agent architecture for conversation-based assessment. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education (AIED 2025)* (Part II, LNCS Vol. 15878, pp. 119–134). Springer. https://doi.org/10.1007/978-3-031-98417-4_9.

Hwang, K., Wang, K., Alomair, M., Choa, F. S., & Chen, L. K. (2024). Towards automated multiple choice question generation and evaluation: Aligning with Bloom's taxonomy. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (AIED 2024, LNCS Vol. 14830). Springer. https://doi.org/10.1007/978-3-031-64299-9_35.

Ipeirotis, P., & Rizakos, K. (2026). Scalable and personalized oral assessments using voice AI. *arXiv*, arXiv:2603.18221. https://doi.org/10.48550/arXiv.2603.18221.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, *23*(4), 367–378. https://doi.org/10.1080/0260293980230404.

Khlaif, Z. N., Hamamra, B., Bensalem, E., Mitwally, M. A. A., & Sanmugam, M. (2025). Factors influencing educators' technostress while using generative AI: A qualitative study. *Technology, Knowledge and Learning*. Advance online publication. https://doi.org/10.1007/s10758-025-09885-z.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y.

Liang, J., Stephens, J. M., & Brown, G. T. L. (2025). A systematic review of the early impact of artificial intelligence on higher education curriculum, instruction, and assessment. *Frontiers in Education*, *10*, 1522841. https://doi.org/10.3389/feduc.2025.1522841.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779.

Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency. Available online: https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf (accessed on 23 October 2026).

Logachev, M. (2026). Artificial intelligence in educational assignments: Issues of academic integrity. *Frontiers in Computer Science*, *8*, 1729059. https://doi.org/10.3389/fcomp.2026.1729059.

Lubbe, A., Marais, E., & Kruger, D. (2025). Cultivating independent thinkers: The triad of artificial intelligence, Bloom's taxonomy and critical thinking in assessment pedagogy. *Education and Information Technologies*, *30*(12), 17589–17622. https://doi.org/10.1007/s10639-025-13476-x.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017.

May, T. A., Provinzano, K., Koskey, K. L. K., Sondergeld, C. J., Stone, G. E., Archer, J. N., & Rimkunas, N. (2025). Expanding the team: Integrating generative artificial intelligence into the assessment development process. *Applied Sciences*, *15*(18), 9976. https://doi.org/10.3390/app15189976.

Memarian, B., & Doleck, T. (2026). A systematic review of academic integrity and misconduct with artificial intelligence in higher education. *SN Computer Science*, *7*, 49. https://doi.org/10.1007/s42979-025-04569-y.

Pankaew, K., Inai, K., & Suprom, S. (2026). Artificial intelligence and the application of prompt engineering in developing digital age learning assessment tools. *Journal of Education and Learning*, *15*(3), 271. https://doi.org/10.5539/jel.v15n3p271.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302.

Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale (AIAS): A refined framework for educational assessment. *Journal of University Teaching and Learning Practice*, *22*(7). https://doi.org/10.53761/rrm4y757.

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education*, *21*(1), 53. https://doi.org/10.1186/s41239-024-00487-w.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147.

Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. *Education Sciences*, *14*(1), 83. https://doi.org/10.3390/educsci14010083.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001.

Timosh, R. R., Joy, E., Evangeline, S., & Aroma, J. (2025). Automated question generation with fine-tuned NLP model based on Revised Bloom's Taxonomy. *SN Computer Science*, *6*, 919. https://doi.org/10.1007/s42979-025-04453-9.

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36.

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment and Evaluation in Higher Education*, *43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, *2*(1), 2. https://doi.org/10.7275/ffb1-mm19.

Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment. *International Journal of Computer Theory and Engineering*, *17*(3), 114–125. https://doi.org/10.7763/IJCTE.2025.V17.1374.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.