Retrieval-Augmented Generation in Financial Derivatives Education: Prototype Development and Student Perspectives

Nur Liyana Mohamed Yousop[1], Wan Mohd Farid Wan Zakaria[2*], Zuraidah Ahmad[3], Noriza Satam[4], Nazrul Hisyam Ab Razak[5]

[1,2,3]Faculty of Business and Management, Universiti Teknologi MARA Johor, Johor Branch, Segamat Campus, Malaysia

[4]Department of Computing and Information Technology, Johor Branch, Tunku Abdul Rahman University of Management and Technology, Malaysia

[5]School of Business and Economics, Universiti Putra Malaysia, Serdang, Malaysia

[*]Corresponding Author

DOI: https://doi.org/10.47772/IJRISS.2026.102100219

Received: 19 October 2026; Accepted: 24 October 2026; Published: 31 October 2026

ABSTRACT

In recent years, generative artificial intelligence (GenAI) technologies have supported learning in higher education by providing accessible explanations, worked examples, and formative feedback. While students benefit from the efficiency of these GenAI tools, their usefulness in financial derivatives education remains challenging. This study explores our students' experiences with AI WAN, a retrieval-augmented generation (RAG) platform that uses approved module materials to provide conceptual explanations, worked calculations, examination-style practice, and feedback. The exploratory qualitative pilot included six undergraduate Malaysian finance students from a cohort of 29. Analysis of their open-ended responses showed that they generally viewed the platform positively. Participants appreciated getting explanations when needed and working through complex calculations step by step at their own pace. They also valued its alignment with course content and highlighted the importance of clear responses, accurate information, and ongoing lecturer guidance. These outcomes provide an initial account of how students perceive curriculum-grounded RAG as a learning aid in financial derivatives education. However, the small sample and reliance on students' experiences mean that the findings reflect perceived usefulness rather than evidence of output accuracy, improved learning, or continued use. Further investigation is needed to assess these aspects and determine whether AI WAN helps students develop independent financial reasoning and reach informed derivatives decisions.

Keywords: Financial derivatives education, Retrieval-augmented generation, Curriculum-grounded learning, Digital learning, Learning prototype development, Student perceptions.

INTRODUCTION

Generative artificial intelligence (GenAI) refers to a broad class of artificial intelligence (AI) systems that can produce new content, including text, images, audio, code, and other forms of digital output (Banh & Strobel, 2023). Within this broader category, large language models (LLMs) constitute a class of generative models designed to process and generate human language. LLMs underpin the language-processing capabilities of many GenAI applications by learning statistical relationships from large volumes of textual data and predicting subsequent tokens based on the earlier context (Luitse & Denkena, 2021). The growing accessibility of LLM-powered GenAI applications has expanded opportunities to integrate them into higher education. In particular, GenAI may facilitate self-directed learning (SDL) by helping students plan their studies, organise academic tasks, obtain explanations, and deepen their understanding of course content (Al-atrash et al., 2026; Zhao & Gu, 2026). It may also support personalised learning by adapting the complexity of learning materials and offering real-time feedback in response to unique learning needs (Owan et al., 2023; Wang et al., 2024). Despite these

potential benefits, GenAI in education also presents important challenges. Overreliance on GenAI may limit the development of analytical thinking and problem-solving skills, while using AI-generated content in assessed work raises concerns about academic integrity (Cotton et al., 2024; Kasneci et al., 2023) Moreover, extensive exposure to AI-generated information may impede meaningful learning when students lack the capacity to critically evaluate, synthesise, and organise the information provided (Gao et al., 2024). As a result, integrating GenAI into learning does not inherently produce favourable learning outcomes. Rather, its educational value likely depends on how effectively students engage with and regulate their use of these technologies.

These concerns are particularly relevant in financial derivatives education, where students are required not only to perform calculations but also to understand the financial reasoning underlying those calculations. Students must identify the relevant financial exposure, determine the appropriate derivative position, apply the relevant contract specifications, and interpret the resulting value in relation to the underlying risk (Miller, 2011). These steps are interdependent, as even an accurate calculation may lead to incorrect conclusions if the derivative position, contract multiplier, or valuation assumptions are inappropriate (Hull, 2022). Successful learning support should guide students through this process and help them explain why a particular method suits the financial problem under consideration. During SDL, students may have to work through these difficulties devoid of immediate lecturer feedback. In these situations, students often cannot tell why their answer is wrong (J. Hattie & Timperley, 2007). The error may lie in how they understood the underlying exposure, in the formula they chose, in how they substituted the values, or in how they interpreted the financial meaning of the result.

Although GenAI can provide immediate explanations and computational assistance, existing studies identify persistent limitations related to the accuracy and contextual relevance of AI-generated responses (Zhao & Gu, 2026). These limitations matter most in course-specific learning environments, where explanations must correspond to prescribed learning objectives, syllabus content, methodological conventions, and assessment requirements. A response may appear linguistically coherent yet still conflict with the conceptual framework or problem-solving approach expected in a particular course (Ji et al., 2023). Moreover, the growing use of GenAI raises broader concerns regarding learner agency, responsibility for learning, and the evolving role of educators. These concerns are particularly noticeable because GenAI still has limited capacity to address the ethical, pedagogical, and affective dimensions of practical learning contexts (Al-atrash et al., 2026).

As LLM use has expanded across domain-specific applications, attention has increasingly focused on mechanisms that enable these models to generate responses grounded in task-specific knowledge. One such way is Retrieval-Augmented Generation (RAG), which enhances LLM-generated responses by retrieving relevant information from external knowledge sources and incorporating it into the generation process (Guu et al., 2020; Iaroshev et al., 2024; Lewis et al., 2021). In this study, we developed AI WAN as a RAG-enabled learning platform aligned with the financial derivatives syllabus. The application integrates a course content library, module-linked explanations, calculation laboratories, simulations, and assessment activities. Its design guides students from conceptual understanding to applying relevant calculations and interpreting results in a financial context. The development of the AI WAN support curriculum is grounded in linking explanations and learning activities to the content, terminology, and conventions of relevant course modules.

In order to test the usefulness of the AI WAN, this study attempts to examine how AI WAN provides curriculum-grounded learning support for financial derivatives education and to explore students' perceptions and experiences of using the platform. Specifically, the study addresses the following research questions.

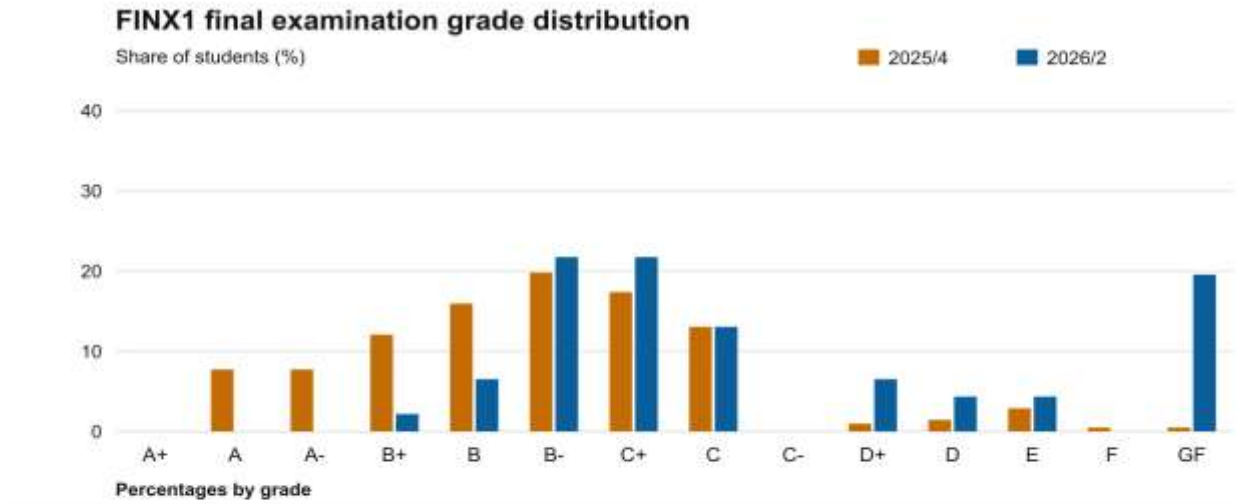
RQ1: How is AI WAN designed to provide curriculum-grounded digital support for the conceptual, quantitative, and decision-making demands of financial derivatives learning?

RQ2: How do participating students describe the prototype's perceived usefulness and their experiences using it?

Finally, this study extends RAG to curriculum-grounded financial derivatives education. It shows how course-specific conceptual, quantitative, and decision-making requirements can inform the design of a structured GenAI learning environment and provides exploratory evidence on students' perceived usefulness and experiences using the platform.

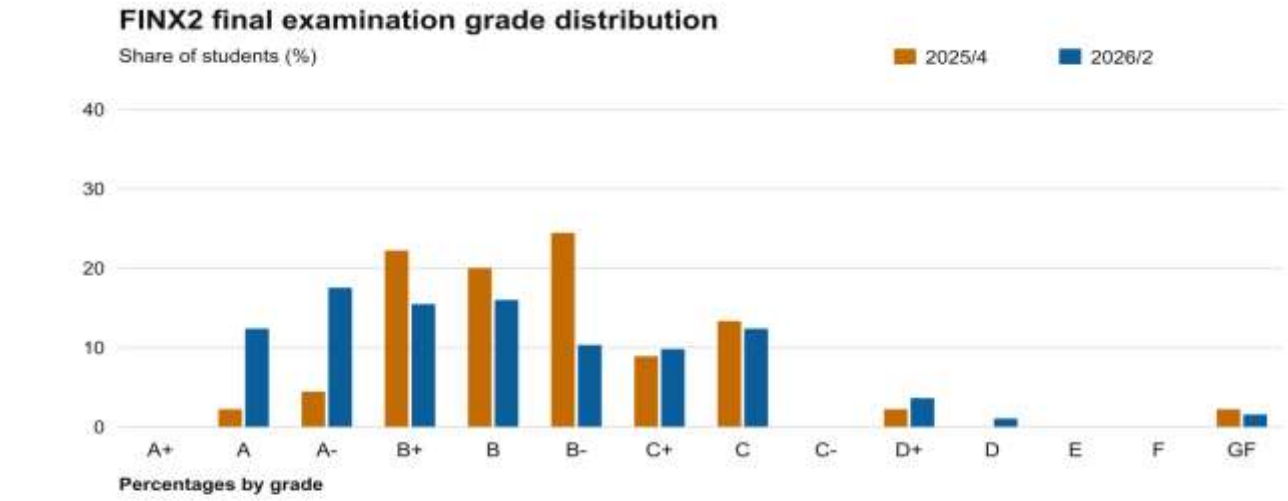
Instructional Setting and Motivation

To understand why AI WAN was built and why RAG was chosen, we first need to look at the instructional setting that led to its development. One main issue was the variation in students' final examination performance. Figures 1 to 3 present the distribution of results for FINX1, FINX2, and INVX1 across the respective semesters. The patterns reflect considerable differences in attainment across courses and cohorts. For FINX1, the cohort mean declined from 2.71 in Semester 2025/4 to 1.78 in Semester 2026/2, while the proportion of students receiving A or A− grades fell from approximately 15.4% to zero. FINX2 followed a different pattern, with the cohort mean increasing from 2.75 to 2.94. For INVX1, more students were concentrated in the middle and lower passing grade categories. Overall, these results suggest that student attainment remained uneven and that some students continued to struggle to understand and apply derivatives-related knowledge.



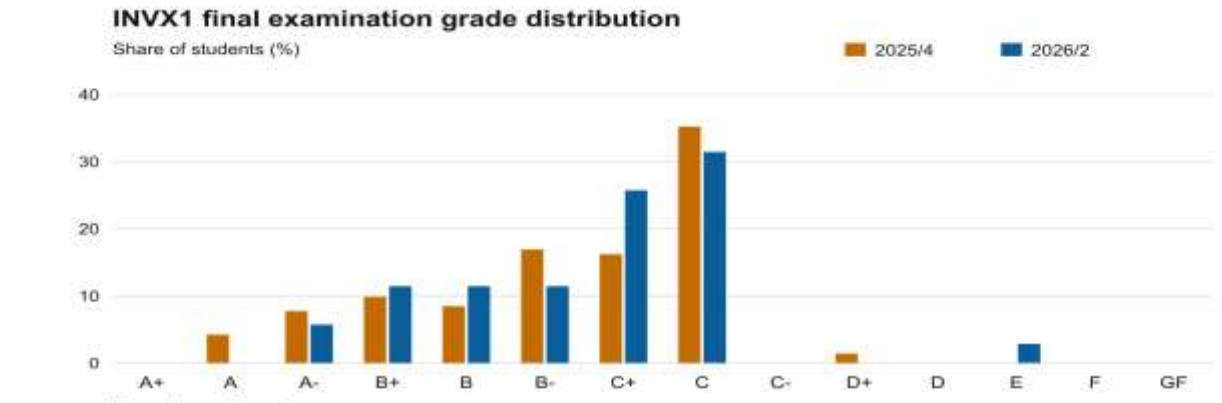
Grade	A+	A	A-	B+	B	B-	C+	C	C-	D+	D	E	F	GF
2025/4	0.0	7.7	7.7	12.1	15.9	19.8	17.4	13.0	0.0	1.0	1.4	2.9	0.5	0.5
2026/2	0.0	0.0	0.0	2.2	6.5	21.7	21.7	13.0	0.0	6.5	4.3	4.3	0.0	19.6

Figure 1 Final examination grade distribution for FINX1



Grade	A+	A	A-	B+	B	B-	C+	C	C-	D+	D	E	F	GF
2025/4	0.0	2.2	4.4	22.2	20.0	24.4	8.9	13.3	0.0	2.2	0.0	0.0	0.0	2.2
2026/2	0.0	12.4	17.5	15.5	16.0	10.3	9.8	12.4	0.0	3.6	1.0	0.0	0.0	1.5

Figure 2 Final examination grade distribution for FINX2



Percentages by grade

Grade	A+	A	A-	B+	B	B-	C+	C	C-	D+	D	E	F	GF
2025/4	0.0	4.2	7.7	9.9	8.5	16.9	16.2	35.2	0.0	1.4	0.0	0.0	0.0	0.0
2026/2	0.0	0.0	5.7	11.4	11.4	11.4	25.7	31.4	0.0	0.0	0.0	2.9	0.0	0.0

Figure 3 Final examination grade distribution for INVX1

Our lecturer's feedback shed further light on these difficulties. The problem went beyond recalling formulas or choosing the right model. Some students produced plausible-looking answers that rested on flawed assumptions, or applied a familiar model without asking whether it suited the problem. As a result, a correct calculation could easily be mistaken for genuine understanding. Students might reproduce the required steps correctly while still misinterpreting the context, overlooking key assumptions, or applying a model inappropriately. The central pedagogical concern was therefore not only whether students could perform a calculation, but whether they understood why a particular method was appropriate and under what conditions it should be applied. This distinction between procedural competence and conceptual understanding became an important consideration in the design of AI WAN.

AI WAN and Retrieval-Augmented Generation

AI WAN is a RAG-enabled instructional prototype designed to provide syllabus-linked explanations and structured support for financial derivatives problems. Its tutoring service uses retrieved curriculum content to contextualise responses from a connected large language model, alongside programmed calculation and assessment tools. The design aims to address difficulties in understanding, applying, and interpreting derivatives concepts, as reflected in the instructional setting. Two instructional requirements shaped the prototype as follows.

1. First, the learning resources and explanations needed to remain closely connected to the relevant course content. The associated module is displayed so that students and lecturers can see where an explanation comes from and refer back to the relevant material when checking it.

2. Second, the learning activities had to make students' reasoning more visible. Instead of focusing only on whether students reach the correct numerical answer, the system encourages them to explain the decisions they make at each stage of the problem. In a hedging task, for example, students may need to identify the underlying exposure and time horizon, justify the position taken, work out the appropriate number of contracts, and consider the risk that remains once the hedge is in place. Looking only at the final answer reveals little about how students interpreted the problem or whether they understood the assumptions behind their decisions. AI WAN therefore uses structured prompts, worked steps, and targeted feedback to help students link financial concepts to calculations and to the decisions that follow.

These features are intended as instructional support rather than as evidence that the system provides adaptive tutoring or enables students to work independently. Whether students actually use the prompts and feedback to check their assumptions, identify errors, and approach subsequent problems without assistance remains an empirical question.

Design Science and Interpretation of Reception

A design science perspective offers a useful way to link an instructional problem to the development and evaluation of an artefact (Hevner et al., 2004; Peffers et al., 2007). In this study, the process involved identifying the problem, developing the prototype, reviewing its implemented functions, and conducting an initial student pilot. Design science also asks researchers to be explicit about the knowledge an artefact contributes, for instance by stating design principles that link features to their intended effects (Gregor et al., 2020). The main features of AI WAN were chosen with this in mind. Source-linked retrieval is intended to keep explanations tied to course content and to limit unsupported output (Ji et al., 2023; Lewis et al., 2021), while structured prompts, worked steps, and targeted feedback are intended to direct attention to students' reasoning rather than to correctness alone (Wisniewski et al., 2020). These are treated as design rationales, not as tested claims.

The evaluation remains preliminary. Evidence that the platform functions as intended, along with positive student responses, cannot, by itself, demonstrate that the design improves educational outcomes. Recent studies of generative AI tutoring show why this caution matters, since perceived helpfulness and actual learning can diverge. Unrestricted AI assistance has been found to improve performance during practice but to harm later unassisted performance (Bastani et al., 2025), whereas tutoring designed around sound pedagogical principles has produced learning gains in controlled comparisons (Kestin et al., 2025). This echoes the lecturer's concern that a successful calculation can be mistaken for real understanding.

The Information Systems Success Model and the expectation-confirmation model of information systems continuance offer concepts for interpreting students' responses to the platform. Delone and Mclean (2003) distinguish dimensions such as system quality, information quality, satisfaction, and net benefits, and the model has been applied to e-learning contexts (Al-Fraihat et al., 2020). Bhattacherjee (2001) links expectation confirmation to perceived usefulness, satisfaction, and intention to continue using a system. In this study, these concepts help organise the discussion of what students valued about AI WAN. They are not treated as measured variables, nor are the relationships proposed by these models presented as having been supported through the pilot.

METHODOLOGY

This study used an exploratory qualitative pilot design comprising two parts, namely a review of the deployed prototype and student reflections on their experience of using AI WAN. The prototype review examined the platform's resources and functions and assessed how they matched the instructional needs identified earlier in the study. The review covered curriculum materials, calculation laboratories, assessment tools, source visibility, accessibility features, learning sequences, concept explanations, formula entries, worked examples, quiz items, tutorial cases, error-checking parameters, glossary terms, and contract presets. The examination-practice and error-checking functions draw on examination-style questions and reference answers prepared by the authors with reference to five years of past semester examination questions. The original questions and official marking schemes are confidential and are not reproduced in the platform. Instead, questions of the closest equivalent type, grounded in the prescribed syllabus, were written and supplied to the system. The aim of the review was to document what had been developed and how the different features were intended to support learning. It was not a formal technical evaluation. The platform's outputs were not marked against the official marking schemes, and no standardised test set, independent reference answers, evaluator ratings, or fixed performance criteria were used. Therefore, including a feature only confirms it was available in the prototype and does not show that its accuracy or performance, including the correctness of the automated checking of student answers, was independently tested across different learning tasks. The architecture description also draws on the retained application source listing and the authors' development records. The listing reflects a subsequent expansion of curriculum resources. The core retrieval and response workflow was reported to remain unchanged, while the resource inventory and student reflections reported here refer to the earlier pilot version.

The student component involved six undergraduate finance students from a FINX2 cohort of 29 students at a Malaysian public university. All participants had used AI WAN before taking part in the study, so they could comment on their experience using the platform. However, their views should not be treated as representative of the whole cohort, especially students who had not used the platform or had stopped using it. Participants were

invited by the module lecturer to record short video reflections about their experience with AI WAN. The recordings lasted between about two and four minutes.

The number of participants reflects the exploratory purpose of the pilot rather than an attempt to estimate cohort-level outcomes. In qualitative research, sample adequacy is judged by the information power of the data rather than by a fixed numerical threshold, and fewer participants are required when the study aim is narrow, when participants have direct experience of the phenomenon, and when the analysis is descriptive rather than theory-testing (Malterud et al., 2016). The present study meets these conditions, since its aim was confined to documenting the early reception of the platform among students who had used it. Six reflections were therefore considered sufficient to identify the features that students noticed and to inform the instruments for subsequent evaluation, but not to establish how widely these views are held in the cohort. A pilot of this kind also corresponds to the formative evaluation stage of design science research, in which early evidence guides refinement of the artefact before summative testing (Venable et al., 2016).

There was no fixed interview schedule and no follow-up questioning. This gave students the freedom to describe their experience in their own way, although the short recordings limited the opportunity to explore their comments in detail. Some experiences may also have remained unmentioned. For this reason, the absence of a particular issue in a student's reflection was not taken to mean that the issue had not occurred. The lecturer's involvement in inviting students to participate may also have influenced how openly some students expressed criticism. Participants were informed about the study's purpose, provided informed consent, and were assured that their responses would remain confidential and would not affect their grades. The researchers transcribed and analysed the video reflections using a thematic approach (Braun & Clarke, 2006).

The thematic analysis followed a descriptive sequence. The reflections were recorded in Malay language and transcribed verbatim before analysis. First, the six video-reflection transcripts were read repeatedly to establish familiarity with each account. Second, statements referring to a particular platform feature or learning experience were identified and assigned initial codes. Third, related codes were organised into reporting themes concerning step-by-step explanation, interface organisation and usability, perceived content quality, and revision support or future use. These groupings were then reviewed against the transcripts to check that each reflected what participant had said. Finally, these themes were interpreted in relation to the instructional design and the information-systems concepts introduced earlier in the study. This sequence describes how the six accounts were summarised and does not constitute a quantitative validation of the constructs.

One researcher conducted the coding. As no independent second coder was involved, an inter-rater agreement coefficient was neither calculated nor claimed. This is consistent with reflexive thematic analysis, in which coding is treated as an interpretive activity rather than as a measurement process whose reliability depends on agreement between coders (Braun & Clarke, 2019). To support credibility, a second author reviewed the candidate themes against the transcripts, and differences in interpretation were resolved through discussion. The coding framework, including the definition used to assign statements to each category, is reported in Appendix D. The reported participant counts indicate how many of the six students mentioned a category. Because an account may refer to several features, these categories are not mutually exclusive and the counts do not estimate prevalence in the full cohort. The prototype review documents functions intended to support learning, whereas the reflections document students' perceptions. Neither component establishes technical accuracy or educational effectiveness. The pilot did not include a control group, pre-test or post-test measures, independent answer checking, or longer-term observation of student use.

Prototype Design and Reported Implementation

Curriculum resources

AI WAN is a browser-based progressive web application deployed through Netlify. The authors adapted the prototype to the prescribed financial derivatives syllabus using their own course notes and lecture slides. During the pilot, 29 curriculum modules covered markets, participants, exposures, contract specifications, futures and options positions, profit and loss, margin, hedging, valuation, arbitrage, and spreads. Malaysian derivatives examples, including FKLI and FCPO, contextualised terminology and worked problems. However, the presence

of contract presets does not establish that all specifications were current or independently verified. This would require a dated validation record.

Table 1 summarises the reported resource inventory. These figures indicate the amount of content available in the prototype, but they do not provide evidence about its accuracy, instructional quality, or how extensively students used the resources

Table 1 Reported prototype resource inventory

Resource	Quantity	Resource	Quantity
Curriculum modules	29	Formative assessment items	189
Concept sections	255	Tutorial cases	43
Formula entries	125	Error-checking parameters	160
Worked examples	63	Glossary terms	59
Interactive calculation laboratories	18	Global derivatives contract presets	129

Technical architecture and curriculum-grounded responses

AI WAN combines curriculum-based retrieval with language-model response generation. Its instructional knowledge base was developed from the prescribed financial derivatives syllabus, the authors' course notes, and lecture slides, organised into 29 modules during the pilot. The authors' examination-style questions and reference answers, prepared with reference to five years of past semester papers, were stored separately as application content for assessment activities. Because official examination questions and marking schemes are confidential, these materials comprise equivalent questions rather than reproductions. The examined tutoring retrieval routine does not index the assessment question bank as part of its instructional evidence collection. Figure 4 summarises the retrieval and response pipeline, which is organised conceptually into four processing stages.

Retrieval in AI WAN is lexical rather than embedding-based. In the preparation stage, curriculum materials are structured into labelled instructional items, including module overviews, learning outcomes, key terms, lesson sections, formulae, worked examples, calculation laboratories, and glossary entries. These materials are packaged with the application and made available within the student's browser. No embedding model or vector database is used in the examined retrieval mechanism. In the retrieval stage, a client-side grounding routine normalises the question, identifies relevant terms, and ranks instructional items using weighted lexical matches. The later implementation selects up to seven evidence items in the main tutoring workflow, while the voice-assisted route requests up to six. Selecting a module increases the ranking of its content without necessarily excluding evidence from other modules. For an unchanged question, content collection, and retrieval configuration, the selection process is deterministic, making the grounding step inspectable and repeatable. However, lexical matching may overlook relevant material expressed using different terminology, and retrieval relevance alone does not establish the accuracy of the resulting explanation.

In the generation stage, the student's question, retrieved curriculum evidence, selected tutoring mode, recent conversation history, and learning preferences are transmitted to a server-side Netlify Function, which constructs the request and forwards it to the connected language model. This arrangement keeps the language-model service's access key outside the browser. For curriculum-grounded questions, the later implementation includes an additional language-model review that compares the generated explanation with the retrieved evidence and may revise the response or identify insufficient support. For time-sensitive questions involving market information, contract specifications, or regulatory requirements, a separate verification function uses model-assisted web searches restricted to approved official domains before the main response is generated. The resulting evidence status, confidence category, and supporting references are supplied to the response-generation process. These confidence categories are system-generated assessments rather than independently calibrated measures of accuracy. In the presentation stage, responses and their associated source and verification metadata are retained in the locally stored conversation. The later application provides an expandable panel for reviewing selected

curriculum references and displays verification status, confidence information, and official-source links where available, although its detailed verification-evidence inspector is disabled. Students may flag responses with conflicting, insufficient, or unverified evidence for subsequent review; these flags are stored locally rather than automatically transmitted to a lecturer. Source references and automated review mechanisms are therefore treated as instructional safeguards rather than independent confirmation of answer faithfulness or accuracy.

The response workflow also includes a local fallback mechanism. When the connected language-model service is unavailable, a module-based tutoring routine provides instructional guidance using predefined explanations and relevant curriculum excerpts, including definitions, formulae, worked examples, and common errors where available. Questions requiring current information are handled more cautiously when official verification cannot be completed. Numerical outputs in the interactive calculation laboratories are produced through programmed calculation functions rather than language-model generation. The architecture therefore distinguishes language-model explanations, locally constructed instructional responses, and deterministic numerical calculations. The pilot was documented as incorporating four tutoring modes within four learning stages, as illustrated in Figure 5. Subsequent development expanded the tutoring interface, with the later examined version providing six response approaches.

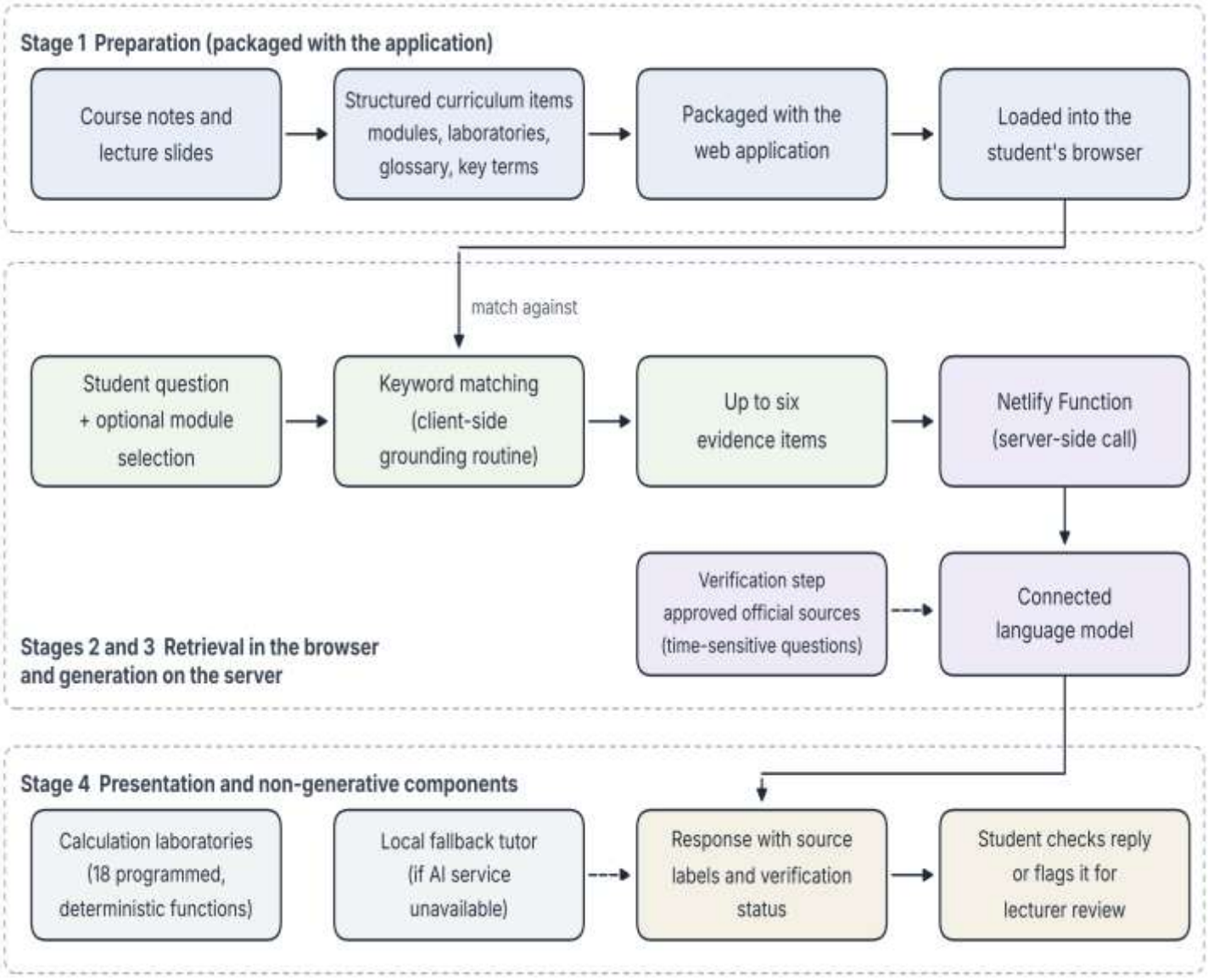


Figure 4 Retrieval and response pipeline of AI WAN

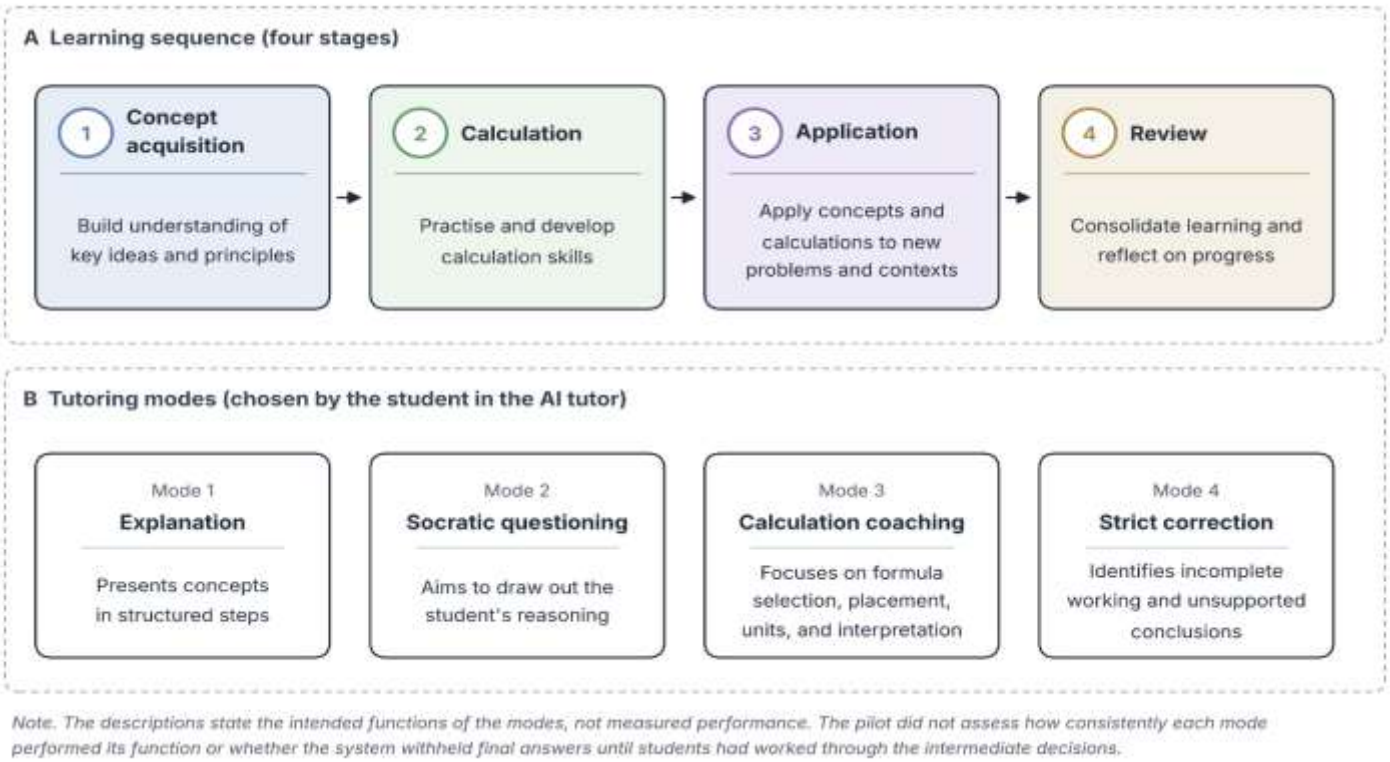


Figure 5 AI WAN Tutoring Modes and Learning Sequence

Netlify provides the deployment configuration for the web application and hosts the server-side functions connecting AI WAN to the language-model service. The technical implementation was cross-checked against the complete source archive of version 12.10.7, a later development version than the version 11.3.1 listing previously used to describe the pilot. The later archive includes the grounding routine, browser-side request handlers, server-side generation and verification functions, and deployment configuration. These files permit inspection of the lexical matching rules, evidence selection, prompt construction, response-generation workflow, and automated review procedures. However, the deployed environment variables, actual production model configuration, and runtime interaction records were not available for independent examination. Furthermore, the later implementation contains developments that cannot automatically be attributed to the pilot without corresponding historical source evidence. The examined workflow is therefore reproducible at the architectural and source-code levels for the later version, whereas exact reproduction of pilot behaviour cannot be established from that version alone. Individual language-model responses may also vary between runs and model versions. Although automated software tests and source-based review mechanisms are incorporated into the application, these do not constitute an independent benchmark of retrieval relevance, answer faithfulness, or response accuracy. No such independent evaluation was undertaken as part of this study.

This workflow lets a learner consult a module, ask for clarification, work through a calculation laboratory, apply the result to a case or simulation, and then review the feedback. This structure provides a consistent instructional context. However, the pilot did not track whether participants actually followed the full sequence. The calculation laboratories focus on decisions such as evaluating a position, determining the number of contracts required, assessing residual exposure, and interpreting option payoffs. Appendix A lists the 18 laboratories. They are intended to help students connect numerical inputs and outputs to the underlying financial problem. The reported inventory does not indicate which laboratories individual participants used or how they performed on them.

The trading simulation presents a sequence of settlement prices and shows profit or loss, variation margin, account balances, and possible top-up requirements. The applied cases require students to identify an exposure and time horizon, justify a position, show their working, interpret the result, and consider the remaining risk. These requirements make the intended link between calculation and financial decision-making explicit. The reported assessment functions include quizzes, examination practice, correction, oral explanation, and progress assessment. In the retained source listing, objective items in examination practice (multiple-choice, true/false

and numerical questions) are checked by programmed rules against stored reference answers, with a tolerance for numerical responses, and short quiz answers are checked by matching required key terms. Structured examination answers are not marked automatically. Instead, students submit their answers and then review a worked model answer. A separate lecturer-facing page provides a marking rubric, but marks and feedback are entered manually. The stated answer structure asks students to identify the concept, select the relevant formula, show the substitution and units, give a signed result, interpret the answer, and consider any remaining risk or implementation constraint. Feedback is intended to encourage students to revise their work and return to the relevant learning material. However, the reflections do not show whether students actually completed these revisions or whether their subsequent answers improved.

The prototype also provides separate roles for students, lecturers, and administrators. In the current demonstration, the browser stores accounts and learning records. These interfaces illustrate each role's intended responsibilities, but they do not demonstrate secure institutional authentication, authorisation, or centralised record management. The description should therefore be understood as relating to a demonstration environment rather than a validated institutional deployment. The reported design also includes voice, caption, language, and mobile-oriented functions. Their inclusion in the prototype should not be taken as evidence of an accessibility or usability evaluation. The report does not include an accessibility conformance assessment against the Web Content Accessibility Guidelines or comparative testing across devices. Figure 6 summarises how the reported features address the design rationale. It analytically maps the design, rather than presenting additional test results.

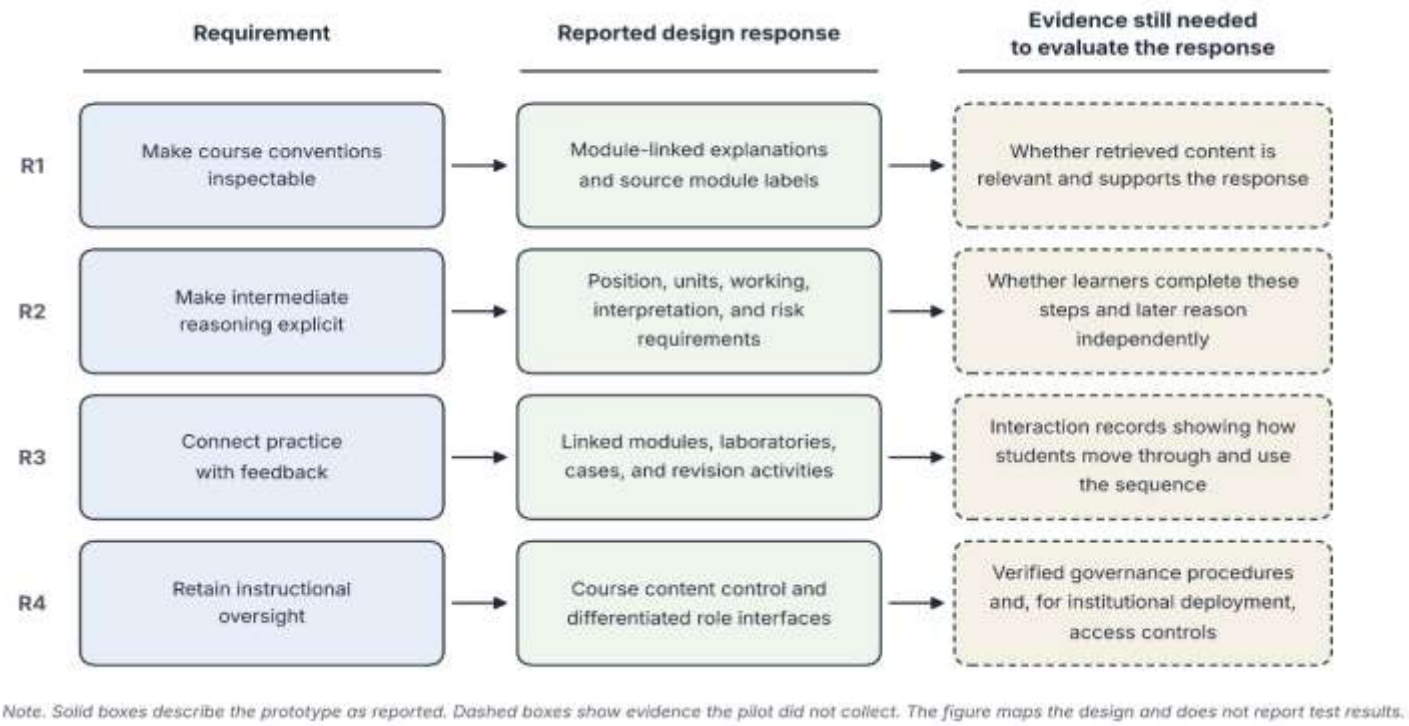


Figure 6 Design requirements and limits of the present evidence

Exploratory student findings

Overall reception

All six participants reported favourable overall views. Their accounts described the platform as a useful supplementary learning resource. This finding concerns a small group of active users who provided feedback, and it cannot be extended to all 29 students or interpreted as a cohort satisfaction rate. Two participants recalled initial scepticism and described a more favourable assessment after use. Their accounts suggest a perceived change in expectations. Because the study did not measure expectations before use, it cannot establish the extent or cause of that change.

Organisation and perceived content quality

Three participants highlighted interface organisation or usability. They valued having topic materials together and described progress assessment as helpful for organising revision. These comments indicate appreciation of the learning environment, rather than performance on a standardised usability assessment. Four participants referred to content clarity plus perceived accuracy. Their accounts valued explanations relevant to futures and options topics and the use of local contract examples. One described using a calculation function to check uncertain working and a quiz to identify a mistake. This illustrates the platform as a study reference, but it does not independently verify the correctness of the platform's answers or the student's diagnosis.

Step-by-step support and the lecturer's role

Four participants highlighted step-by-step explanation. Their accounts emphasised the value of breaking a procedure into understandable stages. Three used a personal-tutor analogy to describe the availability or interpretive role of the system. These categories overlap and should not be combined into a larger participant total. The reported comparison with other tools concerns students' impressions rather than a controlled comparison. The reflections do not show whether AI WAN withheld final answers consistently or whether students completed the intermediate reasoning themselves. Nevertheless, the accounts suggest that explanatory structure was a noticeable aspect of the experience. One participant explicitly affirmed the continuing importance of the lecturer, describing a supplementary role for the platform. This account does not establish how all participants balanced automated assistance with lecturer guidance.

Revision support and future intentions

Four participants reported benefits relating to time efficiency or reduced stress during revision. The accounts described clarifying uncertainties more quickly and feeling less burdened when preparing for examinations. These are perceived benefits, since elapsed task time, cognitive load, and anxiety were not measured independently. Four participants said they would continue using the platform or recommend it to others. These statements indicate favourable prospective attitudes. They do not show subsequent use, retention, or advocacy behaviour. Table 2 retains the original category counts while distinguishing reported experiences from measured outcomes.

Table 2 Reported categories in the six student reflections

Category	Participants mentioning the category	Paraphrased account
Overall satisfaction	6 of 6	Favourable assessment of the platform as study support
Perceived usefulness and step-by-step help	4 of 6	Complex procedures presented in more manageable stages
Perceived time efficiency and reduced stress	4 of 6	Perceived quicker clarification and less stressful revision
Continuance and advocacy intentions	4 of 6	Intention to use the resource again or recommend it
Content clarity and perceived accuracy	4 of 6	Explanations and checking functions perceived as relevant and dependable
Perceived system quality and usability	3 of 6	Appreciation of organisation, navigation, and progress functions
Tutor-like support	3 of 6	Perception of accessible explanatory assistance
Revised expectations	2 of 6	Retrospective movement from scepticism towards endorsement

DISCUSSION

The study's main contribution is a context-specific design account that brings curriculum resources and explicit reasoning requirements together in a single prototype. AI WAN is organised around the relationship between

exposure, position, calculation, and interpretation of the result. The study makes this instructional rationale explicit and documents favourable responses from a small group of active users. However, it does not show that the design improves students' reasoning or outperforms existing learning resources. Participants' positive comments about step-by-step explanations, course relevance, and resource organisation align with the design rationale. These responses point to features that warrant closer investigation rather than demonstrating that a particular learning mechanism is operating. The reflections do not show, for example, whether displaying the source material changed how students checked explanations, whether the prompts reduced answer copying, or whether students applied the same reasoning process to unfamiliar problems.

This distinction matters because the platform combines several functions. Students may have valued consolidated notes, calculators, progress assessment, or ready access to explanations independently of its retrieval and generation architecture. The source-level account clarifies the intended interaction between curriculum retrieval, connected language-model responses, and programmed calculations, but the pilot cannot isolate the contribution of any of these components. Nor can it establish that curriculum retrieval improved conceptual understanding or independent financial reasoning. The specific educational value of RAG remains a question for controlled evaluation.

Interpreting Favourable Reception

The reported appreciation of the platform's organisation and content can be discussed in relation to the system-quality and information-quality dimensions of Delone and Mclean (2003) . Comments about usefulness, changed expectations, and intentions to use the platform in future also relate to concepts in Bhattacherjee's (2001) continuance model. These connections offer a useful way to describe the students' accounts, but they do not establish relationships between the constructs. The ten-path adoption model shown in Figure 7 proposes future research and goes beyond what six brief reflections can establish. The current evidence supports a narrower conclusion, namely that some participants linked their positive experiences to an intention to return to the platform. It does not show that information quality led to confirmation, that satisfaction mediated continuance, or that perceived benefits resulted in subsequent behaviour. A separate study would need to examine these relationships.

The uniformly positive responses also need to be considered in light of the recruitment process and study context. Because participants had already used the platform, the sample may have favoured students who found it useful. The lecturer's involvement in inviting students to participate may also have encouraged more appreciative accounts even though participants were told their feedback would not affect their grades. In addition, the short recordings and absence of follow-up questions left limited opportunity for students to discuss reservations or negative experiences. These factors mean that the lack of reported criticism should not be interpreted as evidence that participants had no concerns.

Implications For Teaching Practice

The design offers an approach that instructors could consider when organising supplementary learning resources, which is to link explanations to course material and to ask students to explain what a calculated result means. For example, a derivatives activity could ask students to explain how a selected position relates to an exposure and what risks remain. This is a design implication arising from the prototype, rather than evidence that the approach improves learning. Source display should be understood as an aid to verification rather than a guarantee of quality (Kasneci et al., 2023). Showing the relevant module can help students locate the supporting material, but it does not show that the response represents that material accurately.

Similarly, a numerical checking function may support practice, but a lecturer may still need to review its assumptions and reference answers. The role interfaces illustrate one possible way to organise oversight, but classroom adoption would need to account for the demonstration's actual capabilities. The evidence presented here supports discussion of AI WAN as a supplementary prototype, but it does not support claims that the system is ready for secure institutional deployment or can replace lecturer judgement.

LIMITATIONS

The empirical material consists of six short reflections from active users in a single module at one institution. These accounts offer only a limited view of students' experiences and exclude students who did not use the platform. The small sample and lack of follow-up questions also limit what we can infer about why students valued particular features or whether they experienced any disadvantages. The findings should therefore not be assumed to apply to other cohorts or settings. A single researcher conducted the analysis, so the grouping and interpretation of students' experiences necessarily reflect that researcher's reading of the data. The counts and paraphrases summarise what participants reported, but they do not replace a transparent link between individual extracts, coding decisions, and the interpretations that followed. No claim of thematic saturation is made, consistent with the view that saturation fits poorly with reflexive thematic analysis (Braun & Clarke, 2021).

The prototype review describes available content and functionality, but it does not provide a reproducible benchmark of accuracy. The retained source listing clarifies the functional routing of questions through curriculum retrieval, a connected language-model service, and module-based fallback responses. However, the listing omits the grounding routine, the language-model connection code, and the generation settings, and language-model outputs vary between runs and model versions. These limits exact computational reproducibility. The lexical retrieval method may also miss relevant material when a question is worded differently from the curriculum text, and its retrieval relevance was not measured. Resource counts cannot establish whether explanations, calculations, or contract information are correct. The reference answers used in examination practice were prepared by the authors rather than drawn from the official marking schemes, and no independent reviewer checked them. Because objective items are checked by programmed rules, an error in a stored answer or numerical tolerance would mark student responses incorrectly, and key-term checks of short answers may misjudge correct answers that are worded differently. Neither possibility was evaluated. The confidentiality of the official schemes also means that any later accuracy check against them could not be replicated by other researchers. The study also does not measure retrieval relevance, answer faithfulness, refusal reliability, accessibility, or security. Finally, it includes no independent measures of learning, time, anxiety, or continued use. Positive perceptions may coexist with limited learning or continued reliance on assistance. The evidence cannot establish whether AI WAN improves independent financial reasoning or whether any particular platform component caused the favourable responses.

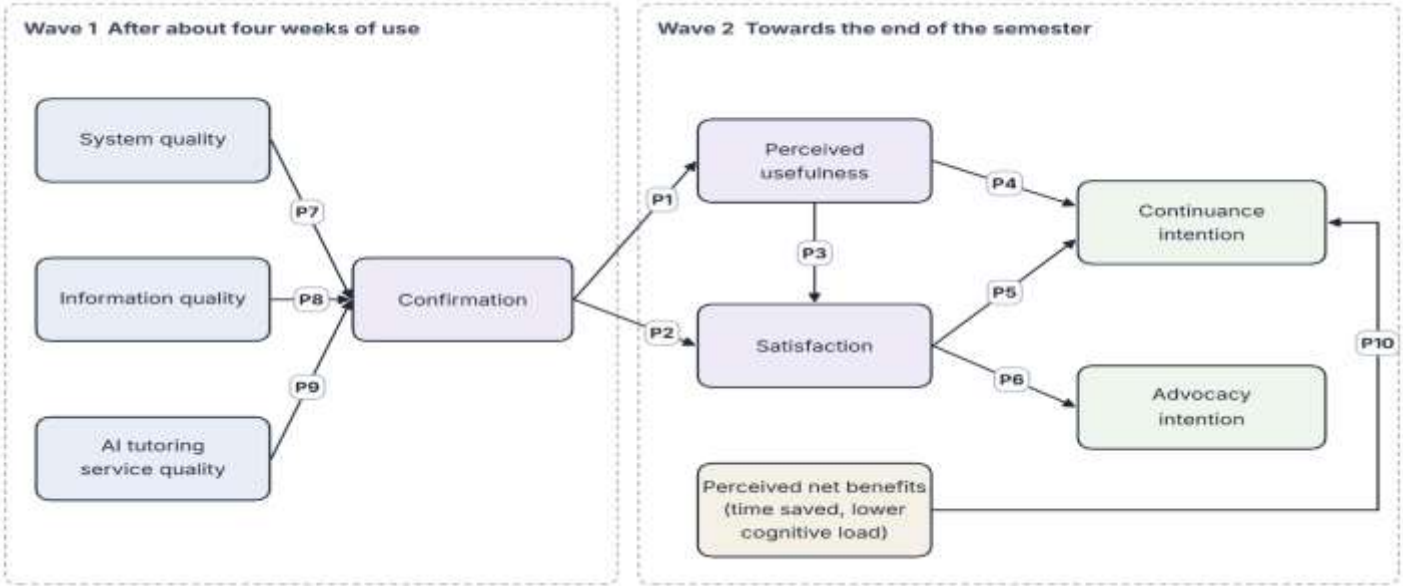
Future research

A subsequent technical evaluation should begin with a pre-specified set of financial derivatives questions and reference answers reviewed by subject specialists. Retrieval relevance, whether generated explanations are supported by the retrieved materials, formula selection, numerical accuracy, and interpretation of financial risk could be assessed separately. Lexical retrieval could also be compared with embedding-based retrieval on the same question set to establish whether the added complexity improves the relevance of the evidence supplied to the language model (Karpukhin et al., 2020). The version of the instructional sources, retrieval configuration, and evaluation rubric should be recorded to improve reproducibility. The authors' examination-style question-and-answer sets could serve as a starting point once their reference answers have been reviewed by subject specialists. These tests are proposed and were not undertaken in the present pilot.

A future quasi-experimental study could compare AI WAN-supported revision with conventional learning resources across comparable financial derivatives tutorial groups. Both groups would study the same syllabus and complete equivalent, unassisted pre-tests, post-tests, and delayed retention assessments. Outcomes would include conceptual understanding, calculation accuracy, selection of appropriate futures or options strategies, and interpretation of residual risk on unfamiliar problems. Independent marking with a common rubric and analysis accounting for baseline attainment would help distinguish performance differences from initial group differences. Random allocation could be adopted where feasible. A third arm using AI WAN with curriculum retrieval disabled would isolate the contribution of retrieval from that of the other platform features, which addresses the question raised in the Discussion about the specific educational value of RAG. This is a proposed design rather than evidence of learning gains.

Testing the proposed adoption model in Figure 7 would require a separate, adequately powered longitudinal study. Future cohorts could complete validated measures of system quality, information quality, AI tutoring service quality, confirmation, perceived usefulness, satisfaction, perceived net benefits, continuance intention, and advocacy intention in two waves. System quality, information quality, AI tutoring service quality, and confirmation would be measured after approximately four weeks of use, and perceived usefulness, satisfaction, perceived net benefits, continuance intention, and advocacy intention towards the end of the semester. Separating the waves in this way reduces common method bias and respects the temporal ordering implied by the expectation-confirmation model (Bhattacherjee, 2001). Items would be adapted from established scales (Al-Fraihat et al., 2020; Bhattacherjee, 2001; Delone & Mclean, 2003). Usage records, where collected ethically, could complement stated intentions with observed engagement. The model would be estimated with partial least squares structural equation modelling, which suits prediction-oriented models with several mediating paths (Hair et al., 2019), after reliability, convergent validity, and discriminant validity have been assessed (Hair et al., 2006). Mediation pathways would be examined through bootstrapped indirect effects (Preacher & Hayes, 2008). Because a single cohort of 29 students cannot support this analysis, data would be pooled across cohorts or partner institutions until the minimum sample indicated by an a priori power analysis is reached. Additional interviews with non-users and less satisfied students would also help identify barriers that the present active-user pilot could not capture.

Figure 7 presents a proposed adoption model for future testing. The arrows represent relationships among system quality, information quality, AI tutoring service quality, confirmation, perceived usefulness, satisfaction, perceived net benefits, continuance intention, and advocacy intention that could be examined. The model draws on the conceptual models discussed above and provides a basis for developing future research questions. However, none of these constructs or proposed relationships were measured or validated in the six-participant pilot. The proposed paths are confirmation to perceived usefulness (P1), confirmation to satisfaction (P2), perceived usefulness to satisfaction (P3), perceived usefulness to continuance intention (P4), satisfaction to continuance intention (P5), satisfaction to advocacy intention (P6), system quality to confirmation (P7), information quality to confirmation (P8), service quality to confirmation (P9), and perceived net benefits to continuance intention (P10).



Note. P1–P10 denote proposed relationships, not estimated effects. The model is retained as a research proposition and should not be interpreted as a result of the exploratory pilot.

Figure 7 The adoption model for future testing

CONCLUSION

AI WAN was developed in response to local concerns about attainment in derivatives education. The prototype is intended to connect course-linked explanations, calculation activities, and decision-centred assessment with

the reasoning required to interpret financial problems. The six participating active users expressed favourable views, particularly about explanatory structure, course relevance, and revision support. The evidence suggests perceived usefulness within this small group and supports a preliminary account of reception. These accounts do not establish technical accuracy, improved attainment, independent mastery, or sustained adoption. The contribution is the description of a context-specific instructional design and the experiences reported by its early users. The attainment graphs explain the motivation for development, while the adoption diagram proposes relationships for future study. Whether AI WAN fulfils its intended learning objectives remains to be evaluated.

ACKNOWLEDGEMENTS

We gratefully acknowledge the anonymous reviewers and the journal editor for their valuable comments and suggestions. The authors also thank colleagues who provided constructive feedback on earlier versions of the manuscript. The authors used ChatGPT (OpenAI) and Claude to refine language, improve clarity, develop images, and enhance readability. The authors critically reviewed and revised all AI-assisted outputs and take full responsibility for the final content of the manuscript.

REFERENCES

1. Al-atrash, A., Hamdan, M., & Dweikat, K. (2026). Qualitative insights into EFL students' use of ChatGPT-assisted autonomous learning for enhancing self-directed learning in Palestinian higher education. Discover Global Society, 4(1). https://doi.org/10.1007/s44282-026-00350-5
2. Al-Fraihat, D., Joy, M., Masa'deh, R., & Sinclair, J. (2020). Evaluating e-learning systems success: An empirical study. Computers in Human Behavior, 102, 67–86. https://doi.org/10.1016/J.CHB.2019.08.004
3. Banh, L., & Strobel, G. (2023). Generative artificial intelligence. Electronic Markets, 33(1). https://doi.org/10.1007/s12525-023-00680-1
4. Bastani, H., Bastani, O., Sungu, A., Ge, H. I., Kabakcı, Ö., Mariman, R., & Brunskill, E. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), Article e2422633122. https://doi.org/10.1073/pnas
5. Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model1. Management Information Systems Quarterly, 25(3), 351–370. https://doi.org/10.2307/3250921
6. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
7. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health, 11(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806
8. Braun, V., & Clarke, V. (2021). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative Research in Sport, Exercise and Health, 13(2), 201–216. https://doi.org/10.1080/2159676X.2019.1704846
9. 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
10. Delone, W. H., & Mclean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. Information Systems Research, Journal of Management Information Systems, 19(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748
11. Gao, Z., Cheah, J.-H., Lim, X.-J., & Luo, X. (2024). Enhancing academic performance of business students using generative AI: An interactive-constructive-active-passive (ICAP) self-determination perspective. International Journal of Management Education, 22(2). https://doi.org/10.1016/j.ijme.2024.100958
12. Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. Journal of the Association for Information Systems, 21(6), 1622–1652. https://doi.org/10.17705/1jais.00649

13. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. Proceedings of the 37th International Conference on Machine Learning, 3929–3938.
14. Hair, J. F., Black, W. C., Hattiesburg, B. J. B., Anderson, R. E., & Cicinnati, R. L. T. (2006). Multivariate data analysis (6th ed.). Pearson Education Inc.
15. Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. European Business Review, 31(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203
16. Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/00346543029848
17. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research 1. MIS Quarterly, 28(1), 1–75. https://doi.org/10.2307/25148625
18. Hull, J. C. (2022). Options, futures, and other derivatives (11th edition). Pearson.
19. Iaroshev, I., Pillai, R., Vaglietti, L., & Hanne, T. (2024). Evaluating retrieval-augmented generation models for financial report question and answering. Applied Sciences Switzerland, 14(20). https://doi.org/10.3390/app14209318
20. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730
21. Karpukhin, V., O˘ Guz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781. https://github.com/facebookresearch/DPR.
22. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274
23. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-97652-6
24. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 9459–9474. http://arxiv.org/abs/2005.11401
25. Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. Big Data and Society, 8(2). https://doi.org/10.1177/20539517211047734
26. Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. Qualitative Health Research, 26(13), 1753–1760. https://doi.org/10.1177/1049732315617444
27. Miller, R. S. (2011). Key issues in derivatives reform. In Derivatives Reform and Regulation. Nova Science Publishers.
28. Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia Journal of Mathematics, Science and Technology Education, 19(8), 1–15. https://doi.org/10.29333/ejmste/13428
29. 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
30. Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

31. 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
32. Wang, N., Wang, X., & Su, Y.-S. (2024). Critical analysis of the technological affordances, challenges and future directions of Generative AI in education: a systematic review. Asia Pacific Journal of Education, 44(1), 139–155. https://doi.org/10.1080/02188791.2024.2305156
33. Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 1–4. https://doi.org/10.3389/fpsyg.2019.03087
34. Zhao, H., & Gu, H. (2026). Thoughtless use of generative artificial intelligence and college students' self-directed learning: a multi-group SEM analysis of gender differences. Scientific Reports, 16(1). https://doi.org/10.1038/s41598-026-54337-y

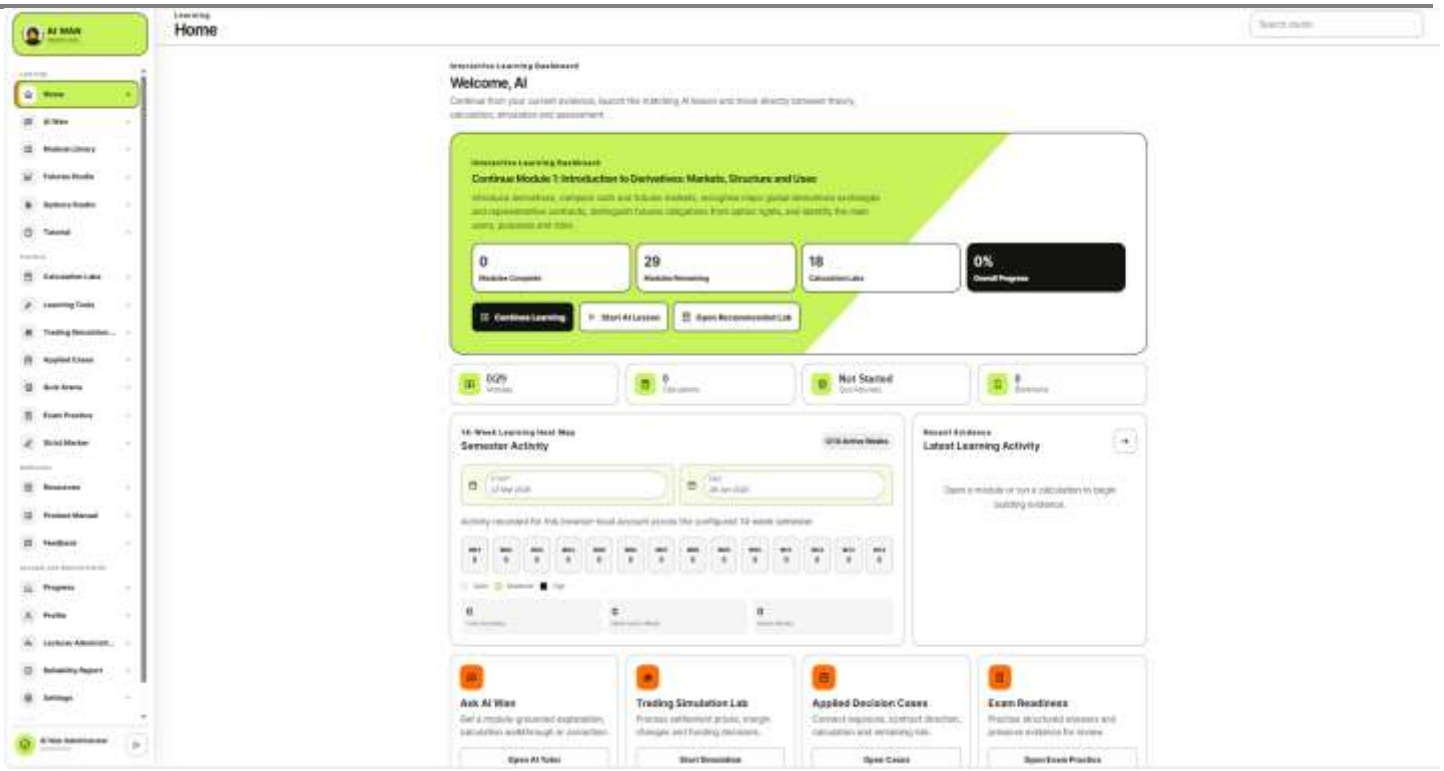
APPENDICES

Appendix A: Reported calculation laboratories

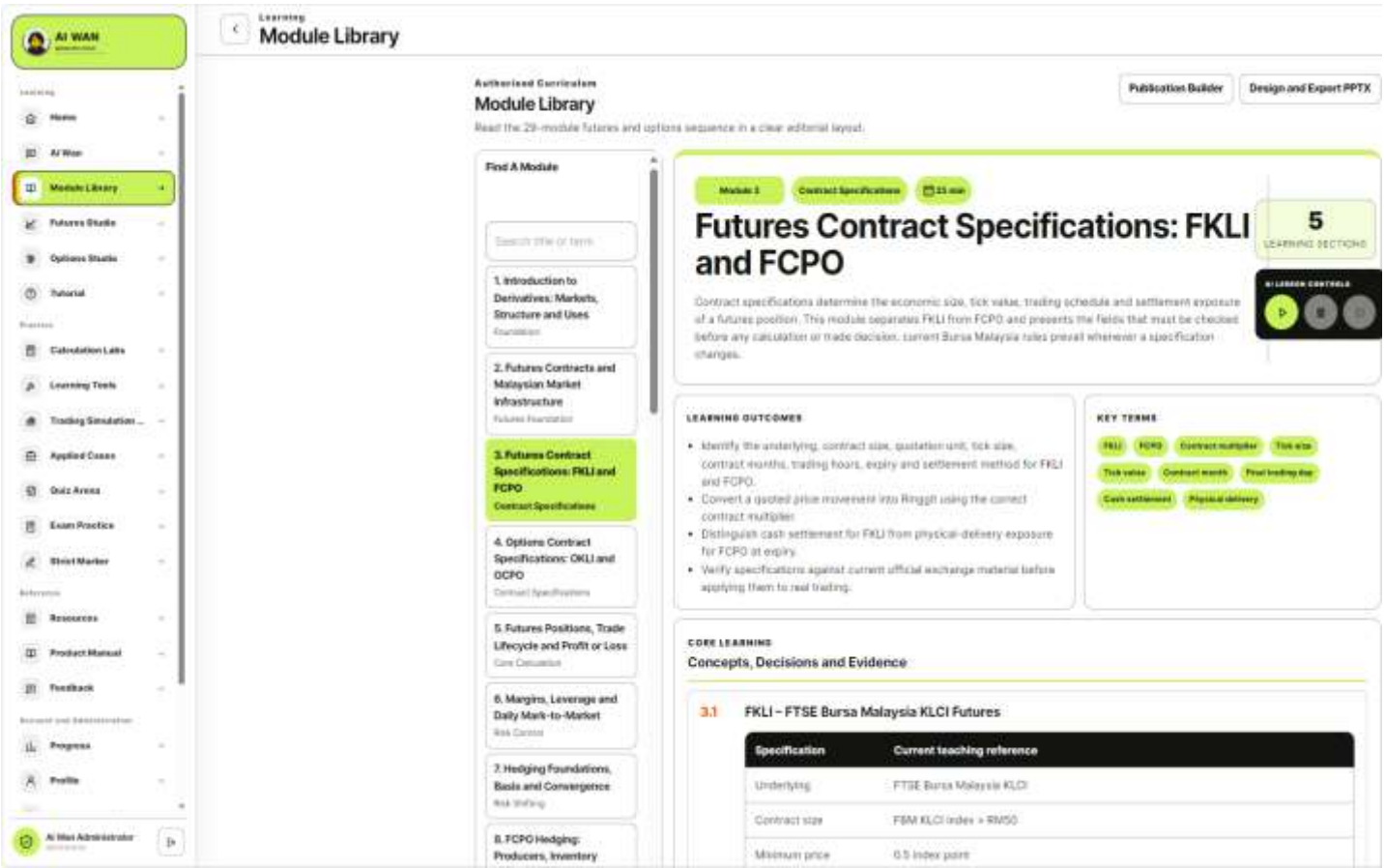
Number	Laboratory	Decision supported
1	Futures profit or loss	Evaluation of a futures position
2	Margin and leverage	Assessment of funding requirements and exposure
3	Daily mark-to-market	Evaluation of settlement flows and margin balances
4	Basis and convergence	Interpretation of the cash–futures relationship
5	Fair futures value	Estimation of theoretical futures value
6	Hedge ratio and contracts	Determination of the required number of contracts
7	Hedge effectiveness and residual exposure	Assessment of remaining risk
8	Target-beta FKLI portfolio hedge	Adjustment of portfolio systematic risk
9	Futures arbitrage investigation	Assessment of mispricing and execution feasibility
10	Calendar spread and commission	Evaluation of a near–far spread position
11	Near–far relative value	Assessment of relative richness or cheapness
12	Global inter-exchange spread	Normalised comparison of contracts across markets
13	Basic option position	Evaluation of a single call or put position
14	Option strategy builder	Construction and interpretation of combined payoffs
15	Black–Scholes–Merton pricing	Estimation of closed-form option value
16	Binomial option pricing model	Estimation of stepwise option value
17	Stock and option hedge	Evaluation of a combined hedge
18	FCPO complete hedge case	Integrated commodity hedging decision

Appendix B: Prototype interfaces

The dashboard and module library illustrate the organisation of the prototype. Interface appearance does not establish response accuracy or learning effectiveness.



Appendix C: AI WAN module library



Appendix D: Coding framework for the student reflections

Category	Definition used to assign statements	Participants
Overall satisfaction	Statements giving an overall evaluation of the platform as a study resource	6 of 6

Perceived usefulness and step-by-step help	Statements describing explanations or procedures presented in stages that the student found helpful	4 of 6
Perceived time efficiency and reduced stress	Statements describing quicker clarification of uncertainties or less pressure during revision	4 of 6
Continuance and advocacy intentions	Statements of intention to use the platform again or to recommend it to others	4 of 6
Content clarity and perceived accuracy	Statements about the clarity, course relevance, or correctness of explanations and checking functions	4 of 6
Perceived system quality and usability	Statements about navigation, the organisation of materials, or progress functions	3 of 6
Tutor-like support	Statements likening the platform to a tutor or describing it as assistance available outside class	3 of 6
Revised expectations	Statements describing a change from initial doubt to a more favourable view after use	2 of 6

Note. A statement could be assigned to more than one category. Counts show the number of participants whose reflection contained at least one statement in the category and do not estimate prevalence in the cohort.