STEM Education

https://www.aimspress.com/journal/steme

STEM Education, 6(2): 194–229
DOI:10.3934/steme.2026009
Received: 07 October 2025
Revised: 09 February 2026
Accepted: 23 February 2026
Published: 05 March 2026

Research article

AI leads, humans lead, or collaborate? Empirical findings and the SAGE roadmap for embedding GenAI in systems analysis and design education

Mahmoud Elkhodr*and Ergun Gide

School of Engineering and Technology, Central Queensland University, Sydney, Australia

* **Correspondence:** Email: m.elkhodr@cqu.edu.au.

Academic Editor: Sasha Nikolic

Abstract: Generative AI orchestration competencies, the abilities to evaluate, adapt, and integrate AI outputs with human judgment, remain inadequately addressed by general AI literacy frameworks. This paper applies and extends the SAGE (Structured AI-Guided Education) framework to systems analysis and design education through three experimental assessments targeting requirements synthesis, formal modelling correction, and design evaluation. In the Brisbane baseline cohort ($n = 13$ groups), 84% demonstrated selective and justified AI use (Balanced Integrator or Selective Adapter), and no group reached expert-level synthesis; the subsequent cross-campus cohorts (n=5 groups, yielding $N = 18$ template-compliant groups analysed overall out of 21 participating groups) corroborated these patterns. Students consistently recognised accessibility needs when working with user requirements and interface designs, but almost entirely neglected them when producing system architecture diagrams, suggesting that this competency depends on continuous prompting rather than transferring automatically between tasks. When AI was used to generate formal data flow diagrams from the students' specifications, the resulting errors were not random but clustered around system boundary classification, exception handling, and data lifecycle completeness. These findings carry implications for how educators structure AI collaboration in technical curricula, particularly regarding the sequencing of human and AI contributions, the need for stage-specific scaffolding, and the role of explicit justification in developing professional judgment.

Keywords: SAGE, generative AI (GenAI), systems analysis and design education, AI orchestration, accessibility, data flow diagrams (DFD)

1. Introduction

The integration of Generative Artificial Intelligence into higher education has accelerated rapidly [1], yet evidence-based frameworks addressing discipline-specific orchestration competencies (the abilities to evaluate, adapt, and integrate AI-generated outputs with human judgment) remain underdeveloped. Whilst general AI literacy initiatives provide broad guidance on understanding AI capabilities and ethical considerations [2], they inadequately address the unique challenges facing systems analysts who must synthesise stakeholder requirements with AI-generated suggestions, correct algorithmic misinterpretations of business logic in formal models, and evaluate automated design recommendations against contextual constraints. This orchestration demands sophisticated metacognitive capabilities and critical evaluation skills that transcend generic AI literacy instruction [3].

This paper advances a systematic research program investigating GenAI integration in technical education through methodological maturation and cross-domain application. Building on a validated pedagogical architecture established in cybersecurity education [4], which demonstrated effectiveness through qualitative evidence with 105 students across undergraduate and postgraduate cohorts, this work applies and extends the SAGE (Structured AI-Guided Education) framework, established and validated in cybersecurity education [4], to systems analysis and design education. Whilst SAGE's two-stage progression and three-dimensional competency model were demonstrated through qualitative evidence in that foundational study, this work advances the framework through three innovations. First, it embeds measurable metrics into student work (including decision logs, confidence ratings, and written justifications), making previously invisible orchestration processes observable and assessable. Second, it provides standardised components (prompts, templates, rubrics) enabling educators at different campuses and delivery modes within a single Australian university to replicate the approach rather than designing from scratch. Third, it deliberately varies levels of support across tasks as a pedagogical strategy, providing explicit scaffolding in some assignments whilst removing it in others to test whether students internalise skills or remain dependent on continuous prompts.

Systems analysis education presents distinctive pedagogical challenges [5] for GenAI integration that differ fundamentally from cybersecurity contexts. Unlike cybersecurity policy development, where compliance against regulatory frameworks provides clear evaluation criteria, systems analysis involves subjective stakeholder priority negotiations, competing design trade-offs, and context-dependent feasibility judgments. Requirements emerge through iterative stakeholder dialogue rather than objective specification. Formal models demand precise semantic relationships where subtle errors fundamentally alter system behaviour. Interface design balances competing objectives without definitive optimal solutions. These characteristics necessitate pedagogical approaches specifically calibrated to analytical judgment rather than adapted from general frameworks or transferred without modification from other technical domains [6].

SAGE addresses critical gaps in understanding how GenAI integration frameworks must balance general pedagogical principles with domain-specific adaptations. The framework operationalises AI orchestration competency development through three experimental assessments embedded within authentic systems analysis tasks: (i) requirements synthesis, examining reconciliation of human and AI-generated stakeholder needs; (ii) formal modelling correction, investigating knowledge prerequisites for identifying AI translation errors in Data Flow Diagrams; and (iii) design evaluation, testing contextual judgment when universal usability principles conflict with situated constraints. Each assessment functions simultaneously as a pedagogical learning activity, a summative evaluation instrument, and a controlled experiment, generating systematic empirical data through standardised decision matrices, mandatory justifications, and embedded quantitative metrics.

SAGE was implemented across four campuses and delivery modes of an Australian university (Central Queensland University) (Brisbane, Melbourne, Sydney, Online) with 21 student groups participating, yielding 18 template-compliant groups included in the final analysis. Implementation spanned an initial baseline term and a subsequent term across delivery modes, testing framework robustness under varying institutional conditions whilst enabling investigation of theoretical questions about competency transfer, orchestration pattern development, and durability of skills across system abstraction layers. Cross-campus replication corroborated core pedagogical mechanisms whilst revealing adaptations that enhance framework effectiveness.

To this end, this work makes three contributions to systems analysis pedagogy:

- It demonstrates that structured embedding of GenAI into curriculum successfully develops critical AI partnership skills rather than passive acceptance. SAGE trains students to evaluate AI outputs systematically, deciding when to accept suggestions that genuinely improve their work, when to modify recommendations to fit contextual constraints, and when to confidently reject AI contributions that miss stakeholder needs or violate domain principles. This represents a pedagogical shift from treating AI as either forbidden tool or uncritical assistant to teaching AI as collaborative partner requiring human judgment.
- Second, this work replicates SAGE's transferability beyond its original cybersecurity context by demonstrating that its two-stage progression, three competency dimensions (analytical deconstruction, contextual application, reflective synthesis), and ready-to-use protocols including standardised prompts generalise to a domain characterised by subjective stakeholder negotiation and context-dependent trade-offs, validated through replication across diverse student populations and delivery modes.
- Third, this work provides actionable guidance grounded in cross-campus implementation, revealing what works, what proves fragile, and where students struggle. Systems analysis educators gain evidence-based strategies for assessment design (require written justification for every AI decision), scaffolding approaches (embed critical prompts at each development stage because awareness collapses without continuous support), and evaluation protocols (baseline-comparison enabling systematic gap identification).

The rest of the paper is structured as follows. Section 2 positions SAGE within the GenAI pedagogy literature and articulates the cybersecurity-informed rationale for the framework. Section 3 specifies the framework architecture and the three experimental tasks used to operationalise competency development. Section 4 describes the participants and the multi-campus implementation context. Section 5 reports the empirical results. Section 6 discusses the findings in relation to learning theory and orchestration scholarship. Section 7 derives practical implications for educators and curriculum designers. Section 8 outlines the study limitations and identifies directions for future work. Section 9 concludes the paper.

2. Related work

GenAI integration in higher education has generated diverse pedagogical frameworks addressing implementation challenges across disciplines [7, 8]. This section positions SAGE within three research streams: general AI literacy frameworks, domain-specific integration approaches including the cybersecurity foundation upon which SAGE builds methodologically, and assessment-as-research methodologies. We identify gaps in existing literature that SAGE addresses through quantitative instrumentation, cross-campus validation, and systematic investigation of competency transfer.

2.1. General AI literacy frameworks

Several researchers have proposed broad frameworks for AI literacy development applicable across disciplines [9]. Long and Magerko established foundational competencies through Delphi studies, identifying that AI literacy encompasses understanding what AI is, what it can do, how it works, and how it should be used [10]. Ng et al. extended this through emphasis on ethical dimensions and societal impacts alongside technical understanding [11]. In parallel, Bloom-aligned approaches to structuring cognitive progression in AI-enhanced learning environments argue for staged development across remembering/understanding to application and creation [12]. Complementing these academic accounts, recent sector guidance synthesises institutional imperatives for GenAI adoption, foregrounding transparency, assessment redesign, and staff capability-building [13, 14]. Concurrently, review essays have framed LLMs' affordances and risks for learners and teachers, emphasising opportunities for feedback and personalisation alongside concerns about bias, hallucination, and over-reliance [15,16]. Whilst valuable for establishing general education objectives, these frameworks provide limited operational guidance for discipline-specific orchestration where students must critically evaluate AI outputs against domain expertise rather than simply understand AI capabilities conceptually.

2.2. Domain-specific GenAI integration

Recent work has begun addressing discipline-specific integration challenges with varying degrees of methodological rigour. In computing education, studies report both promise and pitfalls, including over-reliance on AI suggestions and shallow understanding of underlying logic [7, 17]. Within requirements engineering, reviews show that ChatGPT outputs often require substantial human refinement, particularly for non-functional requirements and stakeholder-specific concerns [18]. Parallel lines of inquiry in higher education have examined academic integrity under GenAI, arguing for constructive alignment of assessment, explicit disclosure norms, and redesigned tasks rather than detection arms races [19]. These studies highlight discipline-specific challenges but rarely provide systematic frameworks validated through multi-site research.

Elkhodr and Gide [4] developed a comprehensive two-stage pedagogical framework for cybersecurity education, implementing it across undergraduate and postgraduate cohorts. Their framework integrated constructivist learning principles with Bloom's taxonomy through structured progression from scaffolded tutorials to independent application, demonstrating pedagogical effectiveness through qualitative analysis of student reflections, observational data, and policy document assessments. The framework established three core principles: progressive cognitive development through taxonomy levels, structured AI engagement through controlled prompts and evaluation rubrics, and authentic contextualisation grounding activities in professional practice. The three-dimension competency model operationalised critical thinking through analytical deconstruction, contextual application, and reflective synthesis.

Whilst pedagogically effective in cybersecurity contexts, this work left theoretical questions unanswered regarding cross-domain generalisability, quantitative relationships among competency dimensions, and durability of developed skills across different analytical tasks.

2.3. Positioning relative to prior SAGE research

This study builds methodologically on the SAGE framework established in cybersecurity education [4], which demonstrated pedagogical effectiveness through qualitative analysis of student reflections and policy document assessments with 105 students. That foundational work validated SAGE's two-stage progression and three-dimensional competency model but left critical questions unanswered. First, the cybersecurity context involved compliance-oriented tasks where regulatory frameworks provided clear evaluation criteria; whether SAGE transfers to domains requiring subjective stakeholder negotiation and context-dependent trade-offs remained untested. Second, the reliance on qualitative reflection data precluded systematic investigation of competency relationships and developmental trajectories. Third, single-site implementation left generalisability uncertain.

Systems analysis presents fundamentally different pedagogical challenges. Requirements emerge through iterative stakeholder dialogue rather than regulatory specification. Formal models demand precise semantic relationships where subtle errors alter system behaviour. Interface design balances competing objectives without definitive optimal solutions. These characteristics test whether SAGE's orchestration competencies generalise beyond compliance-focused domains to contexts requiring situated judgment.

This study advances the SAGE research program through three innovations. Quantitative instrumentation replaces qualitative reflection analysis with embedded metrics (decision matrices, error classifications, confidence ratings) enabling correlation analysis and pattern identification. Cross-campus validation with 18 groups across four campuses and delivery modes within a single Australian university tests replicability beyond single-site implementation. Controlled experimental manipulation of scaffolding intensity according to the task's (accessibility cued in Tasks 1 and 3, absent in Task 2) enables investigation of competency transfer and durability. The resulting discoveries—the accessibility U-curve revealing layer-dependent competency expression, the three-zone collaboration model differentiating authority distributions, and the systematic AI translation error taxonomy—represent findings that qualitative single-site designs could not generate.

2.4. Theoretical foundation and gaps addressed

SAGE builds methodologically on assessment-as-research and constructive alignment traditions [20], and aligns with critical thinking frameworks emphasising analysis, evaluation, inference, and self-regulation [21, 22]. SAGE operationalises these principles through experimental assessments serving triple purposes: pedagogical skill development, summative evaluation, and empirical research on orchestration patterns. Standardised instruments create observable evidence through structured decision matrices, mandatory justifications, confidence ratings, and error classification taxonomies.

Existing literature exhibits three critical gaps SAGE addresses. First, general AI literacy frameworks emphasise conceptual understanding without operational mechanisms for embedding competency development within discipline-specific professional practice, leaving educators without concrete protocols or validated assessment designs [11, 12]. Second, domain-specific studies rarely provide systematic frameworks validated through multi-site empirical research, with single-implementation case studies demonstrating feasibility but leaving generalisability uncertain [7,17–19]. Third, assessment methodologies seldom serve simultaneous pedagogical and research purposes, missing opportunities for evidence-based practice improvement [20].

SAGE contributes complete architectural specification with standardised protocols enabling replication, validated through cross-campus implementation with 18 groups across four campuses and delivery modes within a single Australian university demonstrating core pattern robustness. The accessibility U-curve revealing layer-dependent competency expression and the three-zone orchestration model differentiating authority distributions represent theoretical advancement, patterns unlikely to surface in single-site qualitative implementations but emerging through quantitatively instrumented, cross-campus investigation [23].

3. SAGE framework and experimental design

This study adapts and extends the SAGE (Structured AI-Guided Education) framework, established in cybersecurity education [4], to the systems analysis and design context. SAGE operationalises AI orchestration competency development through a two-stage progression: scaffolded tutorials (Weeks 1–6) developing critical evaluation through repeated generation–analysis–synthesis–reflection cycles, followed by experimental assessments (Weeks 8, 10, 12) serving triple purposes—pedagogical skill development, summative evaluation, and empirical research through standardised prompts, structured decision matrices, and embedded metrics. The competency model encompasses three dimensions (Analytical Deconstruction, Contextual Application, and Reflective Synthesis). Full definitions are provided in Appendix A.

The experimental work consisted of three assessments that instantiate SAGE competencies through domain-specific protocols targeting requirements synthesis, formal modelling correction, and design evaluation. Each assessment embeds controlled experiments within authentic systems analysis tasks, generating systematic data through standardised measurement instruments.

Experiment 1 (Requirements Synthesis):

A three-part workflow comprising a human-only baseline, subsequent AI generation using a standardised prompt, and final synthesis into a prioritised backlog is used to examine whether students can identify gaps in AI outputs whilst productively leveraging them. Synthesis is operationalised through Accept/Modify/Reject decisions with brief justifications, and primary measures include Synthesis Evidence, Domain Knowledge, and Accessibility Awareness.

Experiment 2 (DFD Translation and Correction):

A structured specification is first produced by students, after which AI is prompted to generate a Level-0 DFD component listing; students then identify and correct errors using a formal correction log. Key metrics include Error Detection Rate, Correction Accuracy, and the distribution of error categories (Structural, Completeness, Semantic, Notation, Scope), with cross-cutting analysis of boundary reasoning, state management, and exception handling.

Experiment 3 (Design Evaluation):

Students create low-fidelity wireframes for an elderly-focused kiosk context, obtain an AI heuristic evaluation aligned to Nielsen's principles, and complete a decision matrix that records whether AI suggestions are implemented, modified, or rejected with contextual justification. Evaluation focuses on context sensitivity, trade-off recognition, user empathy, and professional judgement.

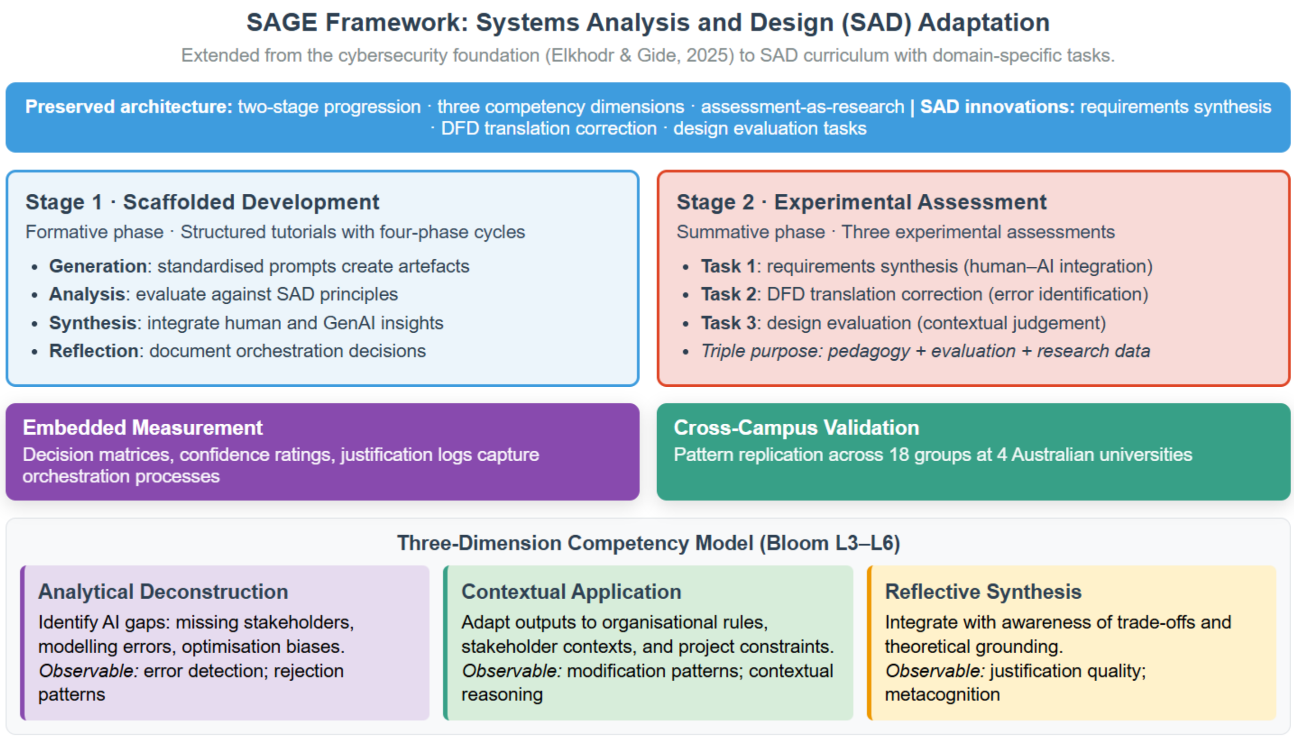

Figure 1. Architecture of SAGE adapted for Systems Analysis and Design, which formalises AI–human orchestration.

Figure 1 presents the SAGE architecture adapted for Systems Analysis and Design, elevating the framework from tool integration to a competency-centred AI–human orchestration model. The adaptation operationalises analytical deconstruction, contextual application, and reflective synthesis, and introduces quantitative instrumentation (decision matrices, confidence/justification logs), controlled scaffolding across tasks, and an assessment-as-research design corroborated through cross-campus replication. SAD-specific protocols (requirements synthesis, DFD translation and correction, design evaluation) serve as contexts for observing orchestration, rather than as mere implementation artefacts. Full protocols and details for all three experiments are provided in Appendix B.

3.1. Cross-task scaffolding manipulation

The three tasks implement controlled experimental variation in accessibility scaffolding to test competency durability. Task 1 explicitly cues accessibility ("elderly users and Indigenous customers" in the standardised AI prompt, accessibility awareness as a measured competency dimension). Task 2 provides zero accessibility scaffolding (no mentions in Process Description Template fields, DFD creation requirements, translation prompts, or reflection questions), testing whether students spontaneously maintain the accessibility focus demonstrated in Task 1 when environmental cues are absent. Task 3 reintroduces accessibility pervasively (target user specifications requiring elderly focus, design justification requirements addressing elderly needs, AI evaluation prompt cueing elderly impact assessment).

This within-subjects manipulation enables investigation of whether accessibility competency represents durable internalised expertise transferring automatically across tasks and abstraction layers, or stimulus-dependent awareness requiring continuous environmental cueing. The experimental design tests two competing hypotheses: (1) competency developed through Task 1 scaffolding transfers robustly to subsequent analytical tasks regardless of explicit prompting, or (2) competency expression remains fragile and prompt-dependent, suppressing when scaffolding is removed before recovering when reintroduced. Evidence supporting the second hypothesis would indicate that accessibility awareness operates differently from other analytical competencies, with implications for curriculum design requiring layer-specific scaffolding rather than assuming transfer from requirements to architecture to interface.

4. Participants and implementation contexts

The study involved a Brisbane baseline cohort (13 groups) during 2025 Term 2, followed by cross-campus validation cohorts in Melbourne, Sydney, and Online in the subsequent term, yielding an initial pool of 21 student groups. Of these, 18 groups produced template-compliant submissions and were included in the final analysis, whilst 3 groups were excluded due to template non-compliance. Brisbane served as the baseline cohort with 13 groups of four students (N=52) enrolled in COIT20248: Information Systems Analysis and Design, a postgraduate course at Central Queensland University. Two additional Brisbane groups were excluded from analysis due to non-compliance with structured response templates, precluding systematic comparison. The cohort comprised primarily international students (about 90%) from the Indian subcontinent, bringing diverse cultural perspectives to accessibility considerations within the GreenHarvest Smart Kiosk case study. Implementation proceeded through scaffolded tutorials (Weeks 1–6), followed by three experimental assessments scheduled for Weeks 8, 10, and 12 of the 12-week term.

Table 1. Cross-campus sample characteristics.

Campus	Groups Included	Groups Excluded	Primary Contribution	Marks (out of 30)
Brisbane	13	2 (template non-compliance)	Baseline quantitative patterns; full statistical analysis	17–28/30
Melbourne	2	0	DFD scaffolding bundle (Melbourne-A); saturation confirmation (Melbourne-B)	22, 27/30
Sydney	2	0	Privacy-aware patterns, payment rigour, unified accessibility controls	19, 26/30
Online	1	1 (template non-compliance)	Saturation confirmation; micro-refinements	24/30

Cross-campus validation data were subsequently collected from Melbourne (two groups), Sydney (two groups), and Online (two groups, with one subsequently excluded) cohorts in the subsequent term. In total, 21 groups participated; 18 groups produced template-compliant submissions and were included in analysis; 3 groups were excluded due to template non-compliance. Table 1 summarises sample characteristics and primary contributions. Melbourne-A uniquely contributed architectural scaffolding innovations (DFD pre-specification bundle reducing AI translation errors), whilst other cohorts confirmed thematic saturation of Brisbane baseline patterns including orchestration distributions, accessibility U-curve replication, and AI systematic failure taxonomies.

4.1. Analytical approach

Analysis employed concurrent mixed methods integrating quantitative pattern recognition with qualitative thematic coding. Descriptive statistics characterised decision distributions, means, and standard deviations across embedded metrics. Spearman rank correlation coefficients examined relationships among ordinal competency codes (Synthesis Evidence, Domain Knowledge, Accessibility Awareness, non-functional requirements (NFR) coverage, Justification Quality), with significance thresholds $p < 0.01$ for primary relationships. Orchestration patterns (Passive Acceptor, Selective Adapter, Balanced Integrator, Critical Synthesizer) were treated as descriptive taxonomy rather than correlational variables, with cluster profiles presenting mean competency scores without computing spurious correlations embedding the construction rule.

Qualitative analysis was conducted by the first author through iterative thematic coding of justification texts and reflection responses. A two-layer coding scheme was employed: (i) *a priori* codes operationalising the study's theoretical constructs (Analytical Deconstruction, Contextual Application, Reflective Synthesis and associated rubric dimensions), and (ii) emergent codes capturing recurring, data-driven patterns not fully specified by the initial framework. Emergent codes (e.g., boundary reasoning competence, exception awareness, context calibration, trade-off recognition) were retained only when they met three criteria: (a) recurrence across multiple groups and tasks rather than isolated observations; (b) grounding in observable, domain-specific behaviours (e.g., systematic misclassification of internal versus external entities; omission of timeout and rollback flows; explicit articulation of accessibility-feature trade-offs) rather than broad interpretive labels; and (c) improved explanatory power in distinguishing high- versus low-quality orchestration beyond what the primary rubric codes captured.

Task 2 error analysis employed dual coding: primary categories (Structural, Completeness, Semantic, Notation, Scope) assigned as single labels with distributions summing to 100%, and cross-cutting themes (boundary reasoning, state management) coded as orthogonal overlays. This separation preserves mutually exclusive error counts for quantitative distributions whilst allowing theoretically meaningful co-occurrence patterns (e.g., a structural error that also reflects boundary reasoning failure) to be analysed qualitatively without conflating measurement levels.

4.2. Ethical considerations

The study operated within normal curriculum delivery, with students informed that aggregated, anonymised assessment data might inform pedagogical research. No additional burden beyond standard coursework was imposed. All data were de-identified before analysis, with group codes replacing student identifiers. Participation in research analysis was voluntary without grade impact. Assessments benefited student learning regardless of research outcomes, ensuring ethical alignment between educational and research objectives. Students retained work ownership, with institutional rights to use anonymised data for quality assurance and research consistent with educational quality frameworks.

5. Results

5.1. Orchestration pattern distribution and competency clustering

Analysis of the Brisbane cohort ($n = 13$) revealed that orchestration patterns clustered around intermediate sophistication levels rather than distributing across the full continuum. Balanced Integrator behaviour, characterised by systematic synthesis of human and AI contributions with explicit trade-off reasoning, emerged in six groups (46%). Selective Adapter patterns, reflecting targeted acceptance of AI outputs with domain-based filtering, appeared in five groups (38%). Passive Acceptor behaviour, indicating minimal critical evaluation of AI suggestions, was observed in only two groups (15%). Notably, this occurred despite explicit assessment guidance instructing students to actively identify gaps, errors, and context misalignments in AI outputs rather than treating AI responses as authoritative. The presence of a small Passive Acceptor subgroup therefore likely reflects a combination of factors such as low confidence in domain knowledge, time pressure leading to “default acceptance,” or misunderstanding the task as primarily documenting AI suggestions rather than critiquing them.

Critically, no groups achieved Critical Synthesizer status, the highest orchestration category requiring evidence of proactive gap identification, systematic bias correction, and generative synthesis beyond mere combination. We conceptualise Critical Synthesizer as an expert-level capability akin to professional systems analysts who can reframe problems, surface latent stakeholder needs, and generate novel directions beyond any single input source. Given the one-semester timeframe and the participants’ developmental stage, we did not expect most groups to consistently demonstrate this level; therefore, 0% achievement should be interpreted as a normal limitation rather than a pedagogical failure. Table 2 reports the distribution of patterns alongside mean competency scores for each category.

Cluster profiles demonstrated systematic competency differentiation across orchestration categories. Passive Acceptor groups exhibited mean justification quality scores of 0.5 on the zero-to-three rubric, with synthesis evidence absent (mean 0.0), domain knowledge weakly applied (mean 0.5), accessibility awareness minimal (mean 0.0), and non-functional requirements coverage absent (mean 0.0). Selective Adapter groups showed marked improvement, with mean justification quality of 2.0, basic synthesis evidence (mean 1.0), strong domain application (mean 2.0), explicit accessibility awareness (mean 2.0), and partial non-functional coverage (mean 1.0). Balanced Integrator groups achieved the highest competency levels, with mean justification quality of 2.67, advanced synthesis evidence (mean 2.0), strong domain grounding (mean 2.0), explicit accessibility considerations (mean 2.0), and comprehensive non-functional breadth approaching the upper threshold (mean 1.5).

Spearman rank correlations among independent competency codes revealed substantial coupling effects, suggesting that orchestration capabilities develop as integrated clusters rather than isolated skills. Justification quality correlated strongly with synthesis evidence ($\rho = 0.84$, $p = 0.0004$), indicating that groups capable of articulating sophisticated reasoning also demonstrated ability to integrate multiple requirement sources coherently. Synthesis evidence correlated substantially with non-functional requirements coverage ($\rho = 0.78$, $p = 0.0019$), suggesting that groups able to synthesize human and AI contributions also attended to broader system qualities beyond functional scope. Justification quality and non-functional coverage showed a moderate positive association ($\rho = 0.69$, $p = 0.0088$). Domain knowledge and accessibility awareness exhibited near-perfect correlation ($\rho = 0.997$, $p < 0.0001$), indicating that groups demonstrating strong domain grounding simultaneously showed heightened sensitivity to elderly and Indigenous user needs within the GreenHarvest case context. This very high correlation should be interpreted cautiously given the small sample ($n = 13$). It may also reflect partial construct overlap in our coding (i.e., groups demonstrating applied domain grounding often articulated user-group implications in the same justification texts), rather than a stable underlying relationship. Replication with larger cohorts and independently-validated measures is needed to test whether this association generalises. The corresponding pairwise associations are reported in Table 3.

Table 2. Orchestration pattern distribution and competency cluster profiles (Brisbane $n = 13$). Acronyms: MJQ = Mean Justification Quality (0–3), MSE = Mean Synthesis Evidence (0–2), MDK = Mean Domain Knowledge (0–2), MAA = Mean Accessibility Awareness (0–2), MNFR = Mean NFR Coverage (0–2).

Pattern	**n**	**%**	**MJQ**	**MSE**	**MDK**	**MAA**	**MNFR**
Critical Synthesizer	0	0	—	—	—	—	—
Balanced Integrator	6	46	2.67	2.00	2.00	2.00	1.50
Selective Adapter	5	38	2.00	1.00	2.00	2.00	1.00
Passive Acceptor	2	15	0.50	0.00	0.50	0.00	0.00

Distribution analysis of specific competency dimensions revealed additional patterns. Accessibility awareness showed explicit recognition in 85% of Brisbane groups during Task 1, with only 8% demonstrating implicit recognition and 8% showing absence of accessibility consideration. Non-functional requirements coverage demonstrated partial breadth in 69% of groups, with comprehensive coverage achieved by 15% and minimal coverage observed in 15%. Synthesis evidence appeared at advanced levels in 38% of groups, basic levels in 46%, and remained absent in 15%. Domain knowledge application was strong across the cohort, with 85% demonstrating applied domain reasoning, whilst only 8% showed generic understanding and 8% exhibited lacking domain grounding. Justification quality followed a normal-like distribution, with 31% providing strong context-specific reasoning, 54% adequate explanations, and 15% weak or missing justifications.

5.2. The accessibility U-curve: Layer-dependent competency expression

A striking non-monotonic pattern emerged in accessibility awareness across the three experimental tasks, revealing that competency expression depends critically on system abstraction layer. Task 1 (Requirements Synthesis) demonstrated explicit accessibility consideration in 84% of Brisbane groups, operationalised through user stories directly addressing elderly checkout processes, Indigenous language preferences, or low-literacy interaction patterns. This high awareness reflected both the explicit prompt scaffolding (“accessibility needs for elderly users and Indigenous customers”) and the natural alignment between requirements language and user-facing concerns.

Task 2 (DFD Translation and Correction) showed dramatic accessibility suppression, with only 10% of groups incorporating accessibility-relevant architectural elements into their corrected data flow diagrams. Operational measurement at this layer required architectural recognition through accessibility profile attributes in data stores, error assistance pathways reflecting cognitive limitations, or explicit read/write semantics for consent and preference management. The Process Description Template provided no explicit accessibility scaffolding, emphasising functional decomposition through triggers, preconditions, and exception handling without prompting consideration of user attribute modelling or accessibility-specific data flows. Groups treating DFDs as "functionally neutral plumbing" externalised no accessibility considerations at the architectural layer despite having demonstrated strong accessibility awareness in requirements just two weeks prior.

Task 3 (Design Evaluation) evidenced robust accessibility recovery, with 90% of groups justifying heuristic evaluation responses through explicit reference to elderly cognitive needs, physical limitations, or Indigenous cultural considerations. Operational measurement required at least two AI feedback decisions (Implement or Modify actions) explicitly motivated by elderly or Indigenous user impacts. The design task naturally invited accessibility reasoning through requirements targeting elderly users (65+) with low digital literacy, standing interaction contexts, and time pressure constraints. The rebound to 90% awareness approached the Task 1 baseline, suggesting that accessibility competency remained latent during Task 2 rather than degrading permanently.

Table 3. Spearman rank correlations among competency codes (Brisbane n = 13).

Relationship	ρ	p-value	Interpretation
Justification Quality–Synthesis Evidence	0.84	0.0004	Strong positive: reasoning quality couples with integration ability
Synthesis Evidence–NFR Coverage	0.78	0.0019	Strong positive: synthesis capability extends to non-functional breadth
Justification Quality–NFR Coverage	0.69	0.0088	Moderate positive: reasoning quality associates with system quality awareness
Domain Knowledge–Accessibility Awareness	0.997	< 0.0001	Near-perfect: domain grounding inseparable from user sensitivity in this cohort

This U-curve pattern (85% to 10% to 90%) reveals that students conceptualise accessibility primarily at interaction layers where user characteristics directly influence interface decisions, but struggle to embed accessibility systematically into system architecture where it manifests through data modelling, state management, and exception pathways. The architectural trough occurred despite students possessing demonstrable accessibility awareness, indicating a translation failure rather than conceptual deficiency. This finding suggests that inclusive design education must explicitly scaffold accessibility at each abstraction level, as competencies do not automatically transfer from requirements to architecture to interface without layer-specific pedagogical intervention.

5.3. Task 2 translation errors: Systematic AI limitations and student corrections

Analysis of 104 DFD corrections across 13 Brisbane groups revealed systematic patterns in AI translation failures when converting structured natural language process descriptions into formal data flow diagrams. Primary error categories showed relatively even distribution across fundamental DFD knowledge domains. Structural errors, involving incorrect component relationships or malformed diagram topology, comprised 30% of corrections. Completeness errors, reflecting missing processes, data flows, or exception pathways, accounted for 27% of corrections. Semantic errors, indicating conceptual misunderstandings of process meaning or data flow directionality, constituted 26% of corrections. Notation errors, involving improper symbology or numbering violations, represented 15% of corrections. Scope errors, where AI introduced out-of-specification elements, appeared minimally at 2% of corrections. The taxonomy, prevalence, and illustrative examples are consolidated in Table 4, while Figure 2 visualises primary category shares and Figure 3 highlights cross-cutting struggle themes.

Table 4. Task 2 error taxonomy and cross-cutting struggle themes (Brisbane n = 13, 104 corrections).

Primary Error Category	**% of Corrections**	**Example**
Structural	30%	Incorrect process decomposition; unbalanced context diagram
Completeness	27%	Missing confirmation flows; absent audit trail pathways
Semantic	26%	Payment confirmation to Gateway instead of Member; login as top-level process
Notation	15%	Improper numbering; incorrect symbology
Scope	2%	Out-of-specification features introduced
Cross-Cutting Theme	**% of Groups Struggling**	**Example**
Boundary reasoning	55%	Internal staff functions marked as external entities; data flow termination errors
State management (CRUD: create, read, update, delete)	45%	Read-only inventory store without write-back; missing bidirectional flows
Exception handling	55%	Absent timeout pathways; missing error recovery mechanisms
Accessibility modelling	≤10%	No accessibility profile attributes or assistance pathways in architecture

Note: Primary categories sum to 100%; cross-cutting themes are orthogonal overlays and do not sum.

Cross-cutting thematic analysis revealed that specific knowledge prerequisites posed systematic challenges independent of primary error categories. Boundary reasoning failures appeared in corrections from 55% of groups, operationalised through entity misclassifications (marking internal staff functions as external entities), incorrect data flow termination points (directing outputs to wrong external actors), or confusion between system scope and environmental context. State management omissions emerged in 45% of groups, manifested through missing bidirectional data flows (particularly read/write semantics for data stores), absent confirmation or receipt pathways, or incomplete order lifecycle representations. Exception handling gaps appeared in corrections from 55% of groups, evidenced through missing error flows, absent timeout pathways, or incomplete failure recovery mechanisms.

Specific recurring errors illustrated these cross-cutting themes. Payment confirmation flows frequently terminated at Payment Gateway external entities rather than returning receipts to Member entities, exemplifying both boundary reasoning failure (misidentifying the proper sink for confirmation data) and state management gaps (incomplete transactional closure). Inventory data stores commonly appeared as read-only without corresponding write-back flows following order placement, demonstrating state management blindness regarding CRUD (create, read, update, delete) operation completeness. Login processes frequently appeared as top-level numbered processes (1.0 Login) rather than as precondition states, revealing confusion between process decomposition and state prerequisites. Audit trails and receipt generation pathways were systematically omitted despite explicit mention in process descriptions, indicating AI difficulty with implicit data lifecycle requirements.

All Brisbane groups successfully identified multiple AI errors, with 100% detection of at least one systematic failure. Correction complexity distributed across simple (40%), moderate (35%), and complex (25%) classifications based on rubric assessment of knowledge depth required. Groups demonstrating strong boundary reasoning competence in corrections showed moderate positive correlation with Task 3 contextual rejection confidence ($\rho = 0.71$). Interpret cautiously given small sample size. This suggests that formal modelling rigour developed in Task 2 supported sophisticated situated judgment in Task 3. However, accessibility awareness at the architectural layer showed no significant correlation with Task 1 or Task 3 accessibility measures, reinforcing the layer-dependent competency expression pattern. While primary error categories (Figure 2) reveal relatively even distribution across AI translation failures, cross-cutting thematic analysis (Figure 3) identifies specific knowledge domains where struggles concentrate, with boundary reasoning and exception handling each challenging over half of all groups.

5.4. Task 3 design evaluation: Contextual calibration and professional judgment

The design evaluation experiment revealed sophisticated contextual reasoning that distinguished universal usability principles from situated implementation constraints. Across 130 discrete AI feedback responses from 13 Brisbane groups, acceptance patterns clustered by heuristic category. Visibility and status feedback suggestions (Nielsen H1) achieved 95% implementation or modification rates, as groups recognised the universal value of progress indicators, confirmation messages, and error state visibility regardless of specific context. User control and freedom principles (H3), particularly undo functions and navigation reversibility, achieved 92% acceptance rates. These findings indicate strong alignment between AI-generated universal usability guidance and student recognition of foundational interface requirements. Disposition of AI heuristic suggestions by Nielsen category is summarised in Table 5.

Conversely, context-specific AI suggestions faced systematic rejection based on situated reasoning about the GreenHarvest kiosk environment. Channel modality mismatches achieved 100% rejection rates, as students identified that live chat support proved inappropriate for unstaffed kiosks, hover states violated touch-only interaction constraints, and voice-first interfaces conflicted with noisy supermarket environments. Cognitive load management decisions showed 85% rejection or substantial modification

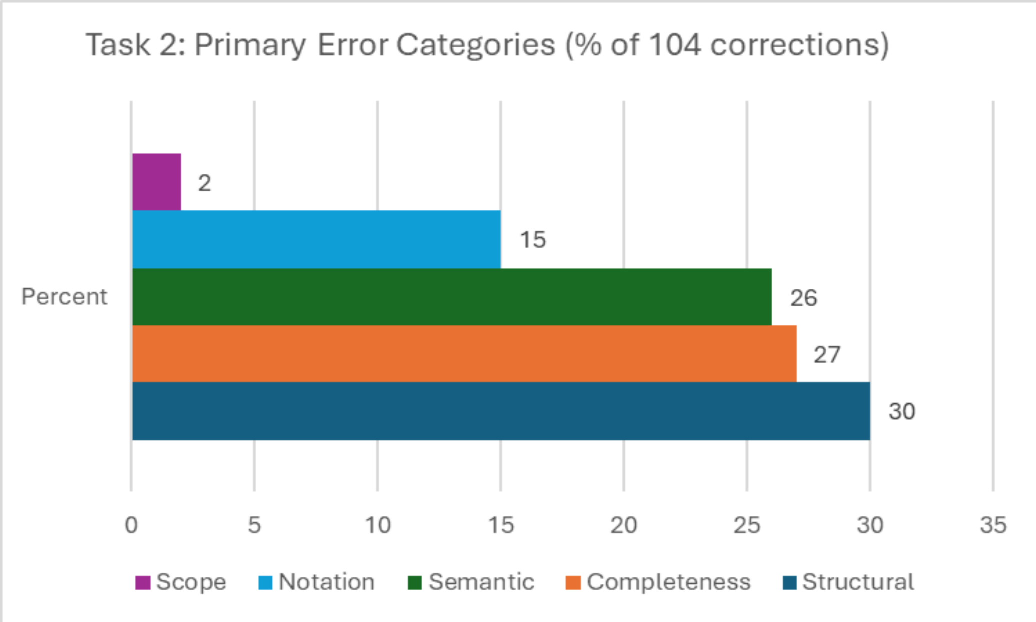

Figure 2. Task 2 DFD translation error distribution, showing a relatively even spread across structural (30%), completeness (27%), semantic (26%), and notation (15%) errors (scope errors: 2%).

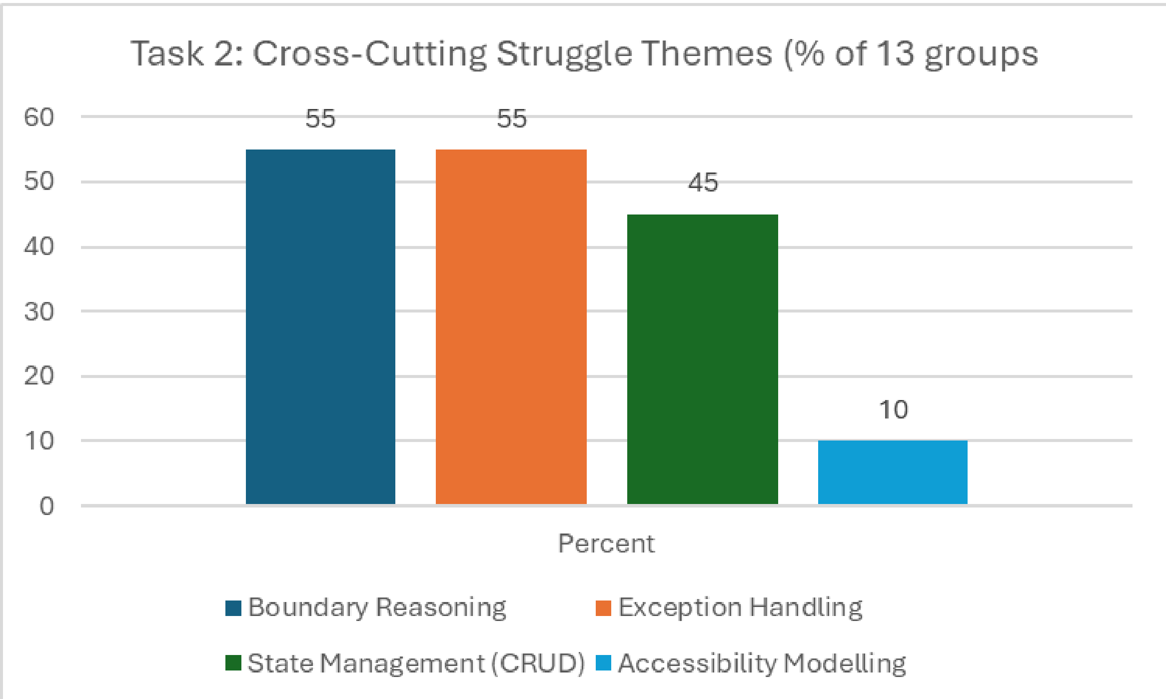

Figure 3. Cross-cutting struggle themes in Task 2, highlighting that boundary reasoning and exception handling challenged 55% of groups, with state management (CRUD) difficulties evident in 45%.

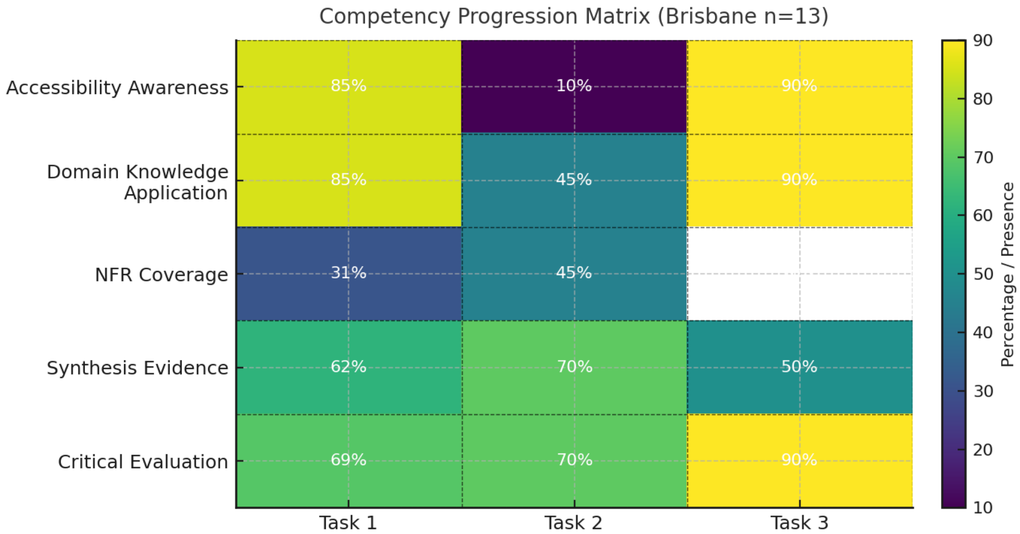

Figure 4. Competency Progression Matrix showing that accessibility awareness drops from 85% (requirements) to 10% (diagrams) to 90% (interface), indicating that students require stage-specific reminders to sustain accessibility thinking throughout the lifecycle.

rates when AI suggested dense information grids, multiple concurrent warning dialogs, or reduced visual contrast, which students recognised as overwhelming for elderly users, inducing alarm fatigue, or unsuitable for bright retail lighting conditions respectively. Redundancy detection drove 95% rejection of suggestions to enlarge already-large touch targets or add buttons when sufficient options existed, demonstrating ability to evaluate current design state against AI feedback validity.

Design choice domains (H2 match to real world, H4/H6 consistency and recognition) operated as collaborative refinement zones with approximately 45% modification rates. Students accepted AI guidance on labelling clarity and recognition-over-recall principles whilst modifying specific icon selections, button placements, and terminology choices to align with elderly user mental models and supermarket task contexts. Aesthetic and legibility suggestions (H8) showed context-dependent acceptance, with 70% implementation when addressing contrast for elderly vision but 40% rejection when AI suggested stylistic elements that sacrificed legibility for visual appeal.

Figure 4 presents competency performance magnitudes across task-layer intersections, with color intensity revealing the dramatic accessibility suppression at architectural layer (10% dark cell surrounded by 85-90% yellow cells), whilst Figure 5 traces developmental trajectories emphasizing distinct transfer patterns: U-curve (accessibility), recovery (domain knowledge), ascending (critical evaluation), and stable (synthesis).

The relationship between AI-assigned severity ratings and student action decisions proved non-linear, evidencing professional judgment development beyond algorithmic deference. Severity-4 (catastrophic) ratings achieved 75% implementation, severity-3 (major) ratings 65% implementation, severity-2 (minor) ratings 70% modification, and severity-1 (cosmetic) ratings 40% rejection. This pattern demonstrates that students exercised contextual trade-offs rather than following severity rankings mechanically. High-severity items were rejected when they conflicted with elderly user needs despite AI categorisation, whilst low-severity items were implemented when they enhanced accessibility beyond AI recognition. Strong correlation between Task 3 contextual calibration

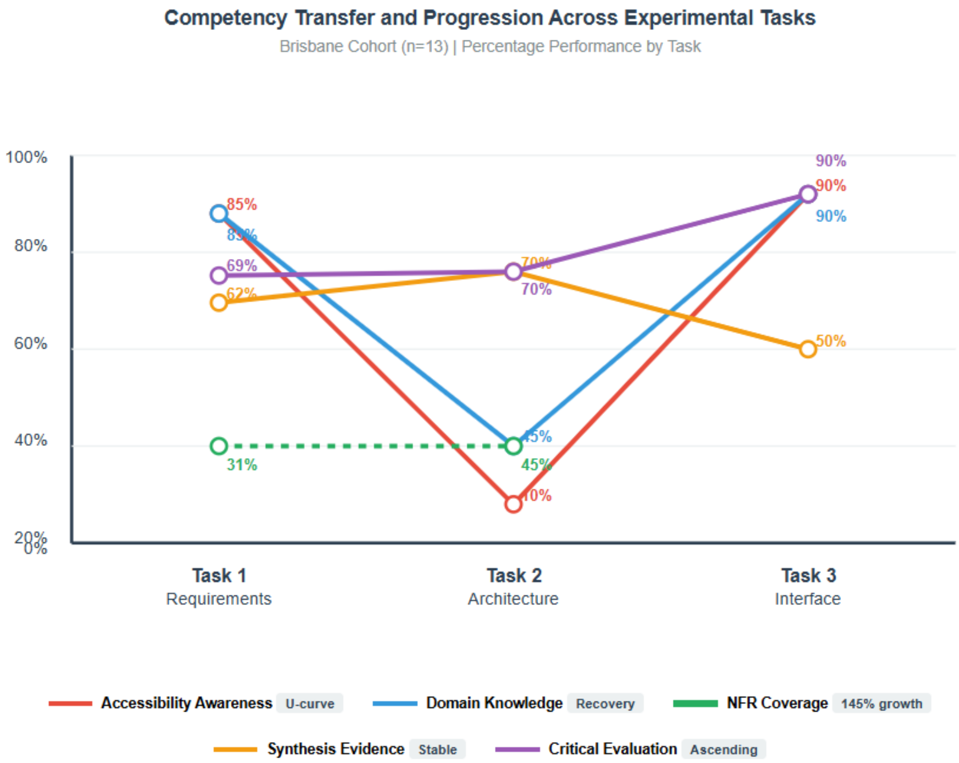

Figure 5. Competency transfer and progression.

sophistication and overall professional readiness assessment (ρ = 0.85, though interpreted cautiously given the small sample size), which suggests that situated judgment in design evaluation serves as a reliable indicator of mature AI orchestration competency.

Justification analysis revealed that accessibility considerations functioned as a primary decision filter across all Task 3 responses. Suggestions explicitly benefiting elderly users achieved 90% implementation or modification rates, compared to 60% for generic usability improvements and 30% for suggestions lacking accessibility justification. This filtering pattern indicates successful internalisation of the elderly-first design principle established through case study framing and explicit task requirements. Groups consistently articulated trade-offs between universal design patterns and specific cognitive, physical, or cultural needs of the target user population, demonstrating ability to calibrate general principles against situated constraints.

5.5. Cross-campus triangulation and thematic saturation

Validation data from Melbourne, Sydney, and Online cohorts corroborated Brisbane quantitative patterns whilst contributing targeted pedagogical insights. Melbourne-A (highest-performing group, 27/30 marks) uniquely contributed a DFD scaffolding bundle comprising an external entity list with explicit role descriptions, a one-page input-output balancing table for context diagram validation, a mini data dictionary specifying attribute names and directionality, and a CRUD matrix for core data stores (Customer Relationship Management, Orders, Inventory). This scaffolding approach measurably reduced translation errors through systematic pre-specification of DFD components before

AI generation, addressing the boundary reasoning and state management gaps identified in Brisbane data. The scaffolding bundle represents a substantive pedagogical enhancement to Task 2 protocol for future implementations.

Sydney cohorts provided nuanced elaborations on established themes. Sydney-A (lower performance, 19/30 marks) contributed privacy-aware session restoration patterns for kiosks, specifying short-lived cart persistence with personal data masking until explicit user confirmation. Payment handling rigour appeared through tokenised draft-order-to-receipt sequencing with explicit retry and exception loops. Accessibility innovations included focus-order specifications with skip-to-content patterns and unified accessibility control panels co-locating text size, contrast, and voice guidance for improved discoverability. Sydney-B (higher performance, 26/30 marks) elevated notification services to first-class architectural actors with explicit confirmation routing, implemented crash-tolerant ordering through pending-order-on-crash reconciliation patterns, and framed voice input risks with severity-4 classifications requiring confirm-before-execute dialogs for elderly and low-literacy users.

Melbourne-B and Online-A cohorts confirmed thematic saturation without introducing novel patterns. Both groups demonstrated standard orchestration distributions (Balanced/Selective dominance, no Critical Synthesizers), accepted H1/H3 universal principles at high rates, moderated icon and labelling consistency suggestions, and corrected systematic DFD boundary and exception errors consistent with Brisbane taxonomy. Online-A contributed micro-refinements including grid-level product availability display justified through recognition-over-recall principles and icon semantics moderation (rejecting ambiguous staff-badge icons in favour of text labels for kiosks), but these represented tidy instantiations of established patterns rather than new thematic categories.

Table 5. Disposition of AI heuristic suggestions by Nielsen category (Task 3).

Nielsen Heuristic Category	Implement %	Modify %	Reject %	Dominant Zone	Interpretation
H1 Visibility of Status	70%	25%	5%	AI Leads	Universal principle; high alignment
H3 User Control & Freedom	68%	24%	8%	AI Leads	Universal principle; high acceptance
H8 Aesthetic & Minimalist (contrast)	55%	30%	15%	Collaborative	Context dependent; elderly contrast needs
H2 Match Real World	20%	40%	40%	Collaborative	Design choice requiring cultural calibration
H4/H6 Consistency & Recognition	25%	45%	30%	Collaborative	Trade-offs between consistency and simplicity
Context-Specific (channel, modality)	5%	5%	90%	Human Leads	Physical environment expertise essential

Across all campuses, the accessibility U-curve pattern replicated consistently. Requirements specifications showed 75% to 85% explicit accessibility awareness, architectural diagrams suppressed accessibility to 10% to 15% visibility, and interface evaluations recovered to 85% to 95% accessibility-motivated decisions. Orchestration pattern distributions remained stable, with combined Balanced and

Selective categories comprising 70% to 80% of groups and Critical Synthesizer status unachieved in any cohort. Task 2 error taxonomy distributions (Structural 28-32%, Completeness 25-29%, Semantic 24-28%, Notation 13-17%) and cross-cutting struggle themes (boundary reasoning 50-60%, state management 40-50%) showed consistency within measurement tolerance. These replication patterns provide confidence that Brisbane baseline findings generalise beyond single-campus or single-term implementations, whilst Melbourne-A's scaffolding contribution offers a validated pedagogical enhancement for addressing systematic translation difficulties.

6. Discussion

6.1. The three-zone model of human-AI collaboration in systems analysis

The Task 3 results reveal a systematic pattern of competency-dependent collaboration that challenges simplistic narratives of AI augmentation in professional education. Three distinct zones emerged, each characterised by differential human–AI authority. In the *universal principles zone*, students appropriately deferred to AI on foundational usability heuristics. In the *collaborative refinement zone*, genuine co-creation occurred where neither human nor AI perspective dominated. In the *contextual constraints zone*, students asserted human expertise regarding situated factors that AI could not assess.

This three-zone pattern validates SAGE's fundamental premise that AI orchestration competency requires graduated judgment rather than binary acceptance or rejection. The high rejection rate in the contextual constraints zone represents pedagogical success rather than AI limitation, evidencing development of professional confidence to override algorithmic suggestions when human expertise proves superior.

However, the absence of Critical Synthesiser patterns across all cohorts indicates a competency ceiling that the current framework does not penetrate. No groups demonstrated proactive gap identification, systematic bias correction, or generative synthesis beyond combination of human and AI contributions. Whether this ceiling reflects developmental constraints requiring extended practice, task design limitations, or fundamental pedagogical gaps in scaffolding expert synthesis remains an empirical question warranting future investigation.

6.2. Layer-dependent competency expression: The accessibility paradox

The accessibility U-curve reflects scaffold-dependence of competency rather than cognitive inability to apply accessibility reasoning across abstraction layers. The experimental design deliberately manipulated scaffolding intensity: Task 1 explicitly cued accessibility, Task 2 provided none, and Task 3 reintroduced cues. The resulting pattern demonstrates that accessibility competency remains stimulus-dependent despite successful initial development. Students who successfully specified accessible requirements when prompted did not spontaneously apply this reasoning to architectural decisions, yet immediately reactivated accessibility thinking when Task 3 made elderly user needs salient.

Cross-campus replication strengthens confidence in scaffold-dependence as a robust pattern rather than a Brisbane-specific artefact. Several complementary factors may reinforce this phenomenon. Abstraction layer characteristics determine accessibility salience through differential alignment with user-facing concerns—requirements and interfaces operate at interaction layers where user characteristics directly influence decisions, whilst architecture emphasises data flows and process decomposition where user characteristics appear less immediately relevant. Additionally, formal modelling notations evolved primarily for functional representation, with accessibility absent from notation standards and textbooks, leaving students without architectural vocabulary for accessibility translation.

The Melbourne-A scaffolding bundle provides preliminary evidence that architectural accessibility gaps can be addressed through targeted intervention. Future framework iterations might test whether embedding accessibility-specific prompts within process description templates successfully maintains visibility at the architectural layer.

6.3. Systematic AI limitations and student correction patterns

Task 2 error taxonomy reveals that AI translation failures cluster in predictable knowledge domains rather than being distributed randomly. Structural and completeness errors jointly constitute the majority of corrections, indicating systematic weakness in compositional reasoning and systems thinking. AI demonstrates facility with local transformations but struggles with global constraints requiring holistic system understanding: balancing decomposition levels, ensuring data flow closure, and maintaining consistent abstraction.

The cross-cutting struggle themes reinforce this pattern. Boundary reasoning failures manifested through entity misclassifications and scope confusion. State management omissions emerged as missing bidirectional data flows and incomplete lifecycle representations. Exception handling gaps appeared through missing error flows and failure recovery mechanisms. These systematic failures occurred despite students providing structured, detailed process descriptions, indicating AI knowledge limitations rather than inadequate input specification.

The structured template approach proved effective for enabling student identification of these failures, validating SAGE's pedagogical strategy of providing scaffolded comparison opportunities. Students who completed templates thoroughly possessed clear mental models enabling systematic gap identification. However, the finding that error recognition proves easier than sophisticated correction indicates that the latter requires deeper formal modelling knowledge regarding decomposition principles and semantic coherence.

6.4. Competency coupling effects

The strong correlations among independent competency codes demonstrate that orchestration capabilities develop as integrated clusters rather than isolated skills. Groups capable of articulating sophisticated reasoning simultaneously demonstrated ability to integrate multiple requirement sources coherently. Groups able to synthesise human and AI contributions also attended to broader system qualities beyond immediate functional scope. The near-perfect correlation between domain knowledge and accessibility awareness indicates that deep engagement with case context simultaneously develops both domain expertise and stakeholder sensitivity.

This coupling pattern suggests that AI orchestration pedagogy should target competency constellations rather than discrete skills, as improvements in critical evaluation, synthesis, and domain application appears to be mutually reinforcing. The practical implication for assessment design is that evaluating orchestration competency requires examining multiple evidence streams rather than single indicators, as dissociations between articulating one's reasoning articulation and integration of skills provide diagnostic information about specific competency gaps.

6.5. Cross-task competency progression

The three-task sequence reveals differential developmental trajectories for distinct competency dimensions. Critical evaluation capability showed ascending development—students grew progressively more confident in challenging AI outputs as they accumulated experience and familiarity with systematic AI failures. Domain application showed a recovery trajectory rather than linear progression, paralleling the accessibility U-curve and suggesting that architectural tasks generally proved more challenging for translating conceptual understanding into formal structural representations.

These divergent trajectories indicate that competency development is non-uniform and domain-dependent, requiring differentiated pedagogical strategies. Critical evaluation benefited from progressive challenge and cumulative experience, suggesting that graduated difficulty effectively scaffolds judgment. Domain application required layer-specific support, with architectural tasks demanding explicit instruction on boundary reasoning and scope determination. These findings challenge one-size-fits-all approaches to AI orchestration education, indicating that effective pedagogy must adapt scaffolding intensity according to the task's characteristics and abstraction layer demands.

6.6. Theoretical interpretation through Bloom's revised taxonomy

The empirical findings align productively with Bloom's revised taxonomy [24], illuminating both competency progression and the absence of Critical Synthesiser orchestration. The three-task sequence maps onto ascending cognitive levels: Task 1 engaged Apply and Analyse processes as students evaluated AI-generated requirements against stakeholder needs and identified contextual misalignments; Task 2 elevated demands to Evaluate as students judged AI architectural representations against formal modelling principles; Task 3 sustained evaluation whilst approaching Create through contextual calibration of AI feedback against situated constraints.

The absence of Critical Synthesiser patterns finds theoretical explanation within this framework. Critical Synthesiser status required Create-level cognition: proactive gap identification generating novel requirements beyond either source, and generative synthesis constructing innovative solutions rather than combining existing contributions. That 84% of groups demonstrated sophisticated Analyse and Evaluate capabilities yet none achieved Create-level orchestration suggests a developmental threshold that single-term interventions may not penetrate [25]. Future iterations might introduce explicit Create-focused activities, such as requiring students to identify what *neither* human nor AI considered.

The competency coupling effects also find theoretical resonance. The strong correlation between justification quality and synthesis evidence suggests that metacognitive articulation and analytical integration develop interdependently, consistent with views positioning metacognition as operating across cognitive levels [26]. Similarly, the near-perfect domain knowledge and accessibility awareness correlation reflects integrated development through contextual application of abstract principles.

SAGE Framework Implementation Guide: Systems Analysis Education

Evidence Strength: Strong (≥70%) Moderate (45-69%) Emerging

Task Type	Validated Competencies	Evidence	Critical Implementation Guidance
Task 1 Requirements Synthesis	• 69% non-passive orchestration (Balanced/Selective) • 85% explicit accessibility when prompted • ρ=0.84 justification-synthesis coupling • 62% achieved human-AI integration	STRONG	• Require 8 manual stories BEFORE AI interaction • Standardise prompts; explicitly cue accessibility, NFRs • Mandate 50-75 word justifications per decision • Template compliance critical for assessment
Task 2 DFD Translation	• 100% error detection with templates • 70% moderate-complex corrections • 55% boundary reasoning failures • 10% accessibility without scaffolding	MODERATE	• Structured template enables 100% error detection • AI fails: boundaries (55%), state (45%), exceptions (55%) • Add scaffolding: entity list, I/O table, CRUD matrix • Accessibility collapses to 10% without prompting
Task 3 Design Evaluation	• 95% accept universal principles (H1, H3) • 90% reject context-inappropriate AI • 90% accessibility recovery with re-cueing • ρ=0.85 context calibration-readiness	STRONG	• Mandate 4-3-3 distribution (implement-modify-reject) • 90% rejection rate = success (confident override) • Re-cueing recovers accessibility awareness • Students distinguish principles from constraints
Cross-Task Patterns	• 0% Critical Synthesizer (competency ceiling) • Accessibility U-curve: 85% → 10% → 90% • ρ=0.997 domain-accessibility coupling • Replicated across 4 campuses (n=19 groups)	EMERGING	• Expert orchestration needs advanced scaffolding • Accessibility is stimulus-dependent: scaffold every layer • Competencies cluster: justification↔synthesis↔NFR • Single-exposure instruction insufficient for retention

Validation: Brisbane n=13 groups; cross-campus triangulation (Melbourne, Sydney, Online) | **Framework Success:** 90% contextual rejection demonstrates professional judgment

Figure 6. Variation matrix.

7. Implications

Figure 6 provides at-a-glance synthesis of validated competencies, evidence strength, and critical implementation requirements across all three experimental tasks, enabling educators to rapidly assess the framework's applicability to their institutional contexts. The subsequent sections detail the pedagogical reasoning underlying these recommendations, curriculum design implications for inclusive systems analysis education, and enhancement opportunities for future implementations.

7.1. Pedagogical recommendations for AI orchestration education

The SAGE framework's empirical validation generates actionable recommendations for educators designing AI orchestration curricula in systems analysis or related technical domains.

7.1.1. Structured documentation and baseline comparison enable systematic evaluation

Mandatory decision matrices with embedded justifications, confidence ratings, and source attributions proved essential for generating observable evidence of orchestration processes. That all groups detected at least one AI translation error in Task 2 validates requiring students to create structured baseline artifacts before AI interaction, enabling systematic comparison between human specifications and AI outputs. Educators should structure AI collaboration as controlled experiments: students generate baseline expectations, document them explicitly using standardised prompts, then evaluate AI outputs against baselines. This approach develops judgment competencies rather than merely procedural AI interaction skills. Mandated response distributions (4–3–3 in Task 3) prevented passive acceptance whilst forcing discriminative judgment, though distributions must align with the task's characteristics rather than arbitrary quotas.This aligns with constructive alignment and formative-assessment principles that integrate assessment with learning activities [20].

7.1.2. Scaffolding must be layer-specific and continuously present

The accessibility U-curve demonstrates that competencies developed at one abstraction layer do not transfer automatically without explicit scaffolding at each layer. For prompt-dependent competencies, scaffolding should be built into task templates rather than provided only through briefing instructions. The Melbourne-A DFD scaffolding bundle provides a model: require pre-specification of system components with explicit enumeration of user characteristics, data attributes, and operation types before AI translation. Educators should identify which competencies require continuous environmental support (accessibility awareness) versus which transfer reliably once established (synthesis capability).

7.1.3. Graduated challenge supports progressive development

The three-task progression from synthesis through correction to rejection provided scaffolded increase in metacognitive demand. Task 1 established baseline orchestration in supportive context, Task 2 challenged students to identify systematic AI failures requiring formal modelling knowledge, and Task 3 required confident override based on situated judgment. Educators should sequence AI collaboration tasks to develop progressively sophisticated capabilities rather than expecting expert-level judgment immediately.

7.2. Curriculum design implications for inclusive systems analysis

Current approaches addressing accessibility primarily at requirements or interface stages may produce systems with accessibility conceptualised as add-on features rather than embedded architectural qualities. The dramatic suppression from 85% awareness (requirements) to 10% (architecture) despite maintaining group composition suggests systems analysis education lacks established approaches for teaching architectural-layer accessibility.

Curriculum reform should ensure accessibility receives explicit attention at each development phase: requirements (user characteristics, functional needs), architecture (user profile data stores, adaptive pathways), database design (preference attributes, consent management), algorithm design (bias detection, fairness constraints), and interface design (interaction modalities, cognitive load). Each phase requires distinct pedagogy because accessibility manifests differently at each layer. Formal modelling notations (data flow diagrams (DFDs), entity–relationship diagrams (ERDs), and the Unified Modeling Language (UML)) evolved primarily for functional representation, with accessibility often absent from notation standards and textbooks. Curriculum development should establish standard patterns for representing accessibility architecturally: user capability attributes in entity models, adaptive process branches in activity diagrams, accessibility service layers in architectural views, assistive data flows in DFDs. Providing exemplar models would create currently lacking reference points. Embedding accessibility across artefacts is consistent with accessibility standards that emphasise end-to-end consideration (e.g., the Web Content Accessibility Guidelines (WCAG) [27]).

7.3. Framework enhancements and future directions

Several enhancement opportunities emerge from identified limitations. The absence of Critical Synthesiser patterns indicates current scaffolding does not support expert-level orchestration. Future iterations might introduce advanced tasks requiring proactive gap identification, systematic bias detection, or generative synthesis creating novel requirements beyond either source. Whether structured practice can penetrate this competency ceiling or whether expert orchestration requires different pedagogical approaches (apprenticeship, critique, reflective practice) warrants investigation.

The assessment-as-research design proved effective but creates sustainability challenges. Structured templates and mandatory documentation substantially increase marking time compared to traditional assessments. Institutions should budget appropriate resources and consider semi-automated metric processing whilst preserving qualitative feedback, or alternate between research-grade data collection and streamlined practitioner versions maintaining pedagogical benefits.

Cross-campus validation revealed value in multi-site implementation for quality assurance and innovation discovery. The Melbourne-A contribution emerged from the students' initiative, indicating diverse contexts generate valuable adaptations. Communities of practice among AI orchestration educators could share protocols, student innovations, systematic failure patterns, and pedagogical refinements, accelerating the framework's evolution whilst distributing development effort.

8. Limitations

This study has five main limitations:

- **Small sample:** The core quantitative analysis uses a small cohort (Brisbane $n = 13$ groups), so effect sizes and correlations should be interpreted cautiously.
- **Participant profile:** the cohort was predominantly international postgraduate students (over 90% from the Indian subcontinent), which may limit generalisability to other demographics and program levels.
- **Single-coder analysis:** Qualitative coding was conducted by a single researcher without a formal inter-rater reliability assessment. The primary evidentiary basis rests on quantitative metrics embedded within the assessment instruments (decision matrices, error counts, acceptance rates), with qualitative coding serving an interpretive function to explain observed patterns; nevertheless, future studies should use multi-coder designs to strengthen confirmability of emergent patterns.
- **Model specificity:** Findings reflect the behaviour of the generative AI systems used during the study period (e.g., ChatGPT accessed via our standard student access channels, during the study terms). Behaviour may differ under subsequent model updates or alternative LLMs.
- **NFR scope:** the experimental manipulation and interpretation emphasised accessibility; we did not systematically test whether other non-functional concerns (e.g., security, privacy, performance, maintainability, ethics) exhibit similar layer-dependent patterns.

A further methodological consideration is the absence of a control group comparing SAGE against AI-free instruction. This design was deliberate. The research question was not whether AI improves learning compared with traditional pedagogy, but how to develop critical orchestration competencies given that generative AI is already pervasive in student practice [19]. Students use AI tools regardless of curricular acknowledgment; the pedagogical challenge is transforming unstructured use into disciplined professional judgment. Our contribution demonstrates that structured integration produces observable orchestration competencies and has identified patterns informing curriculum design. Whether SAGE outcomes exceed alternative pedagogies remains an empirical question for future research.

Accordingly, the results should be read as exploratory but reproducible patterns under defined conditions, motivating replications with larger and more diverse cohorts, multi-coder qualitative designs, and broader non-functional coverage.

9. Conclusions

This paper presented SAGE, a pedagogy for embedding generative AI within the systems analysis and design curriculum to strengthen job-ready orchestration competencies. The framework, adapted and extended from a validated cybersecurity architecture, rests on three ideas: staged scaffolding for progressive judgment, assessment-as-research instrumentation, and multi-layer competency operationalisation across requirements, architecture, and interface work.

In the Brisbane baseline cohort ($n = 13$ groups), orchestration behaviours concentrated at intermediate levels (Balanced Integrator 46%, Selective Adapter 38%), with no group reaching expert synthesis. Cross-campus cohorts (n=5 groups, yielding $N = 18$ template-compliant groups analysed overall out of 21 participating groups) corroborated the same qualitative distribution (Balanced/Selective dominance; no Critical Synthesizers) without materially altering the baseline interpretation.

In summary, this experiment reveals that students tend to plateau at a mid-level of orchestration. That means they can merge and tidy AI output, but rarely challenge its framing, or propose genuinely new directions. More broadly, the findings reinforce that effective AI literacy in professional education is not primarily about prompt engineering or output optimisation; it is about developing disciplined critical evaluation and the confidence to override AI when context demands. Designing for orchestration therefore requires assessment tasks that make reasoning visible and reward justified acceptance, modification, and rejection.

Accessibility shows up in requirements and screen designs, then disappears in architecture unless we make it explicit at that layer. AI prose-to-DFD translations often fail at the system boundary and across the data lifecycle, for example misclassifying internal actors as externals, omitting write-backs to stores, and dropping confirmations or error and timeout paths. In practice, the fix is to add one task that rewards gap finding and bias checks, embed accessibility fields in DFD and CRUD templates, and require a short pre-translation scaffold with an entity list, input–output balance, and a CRUD matrix.

Author contributions

Conceptualization: M. Elkhodr. Methodology: M. Elkhodr. Investigation: M. Elkhodr. Data curation: M. Elkhodr. Formal analysis: M. Elkhodr. Visualization: M. Elkhodr. Writing—original draft: M. Elkhodr. Writing—review & editing: M. Elkhodr, E. Gide. Supervision/mentorship: E. Gide. Project administration: M. Elkhodr. Resources: M. Elkhodr, E. Gide. Funding acquisition: Not applicable.

Use of Generative-AI tools declaration

The authors used GenAI tools (OpenAI ChatGPT and Anthropic Claude) to assist with language editing, LaTeX formatting, reference organization, and copy-editing. These tools were not used to generate research ideas, design the methodology, perform the analysis, or make interpretive claims. All study design decisions, data coding, analysis, and conclusions are the authors' own work. The authors reviewed, verified, and take full responsibility for all content.

Acknowledgments

No external funding was received for this work. The authors thank the participating student cohorts for permitting the use of de-identified coursework samples as part of routine teaching evaluation.

Conflict of interest

The authors declare that there are no conflicts of interest related to this work.

Prof. Ergun Gide is an editorial board member for STEM Education and was not involved in the editorial review or the decision to publish this article.

Ethics declaration

This study met Central Queensland University's (CQU) Human Research Ethics Procedure (2024) definition of ”exempt from ethical review”, namely research that carries a lower risk to the participants or the community and satisfies one or more of the conditions outlined in the National Statement on Ethical Conduct in Human Research (Sections 5.1.15–5.1.18). The study involved the analysis of de-identified student coursework produced as part of routine teaching activities, with no additional burden imposed beyond standard assessment requirements. CQU Human Research Ethics Procedure: https://www.cqu.edu.au/about-us/structure/governance/policy/view-current-policy?DocumentId=3127

References

1. Crompton, H. and Burke, D., Artificial intelligence in higher education: the state of the field. *International journal of educational technology in higher education*, 2023, 20(1): 1–22. https://doi.org/10.1186/s41239-023-00392-8
2. Khreisat, M.N., Khilani, D., Rusho, M.A., Karkkulainen, E.A., Tabuena, A.C. and Uberas, A.D., Ethical implications of ai integration in educational decision making: Systematic review. *Educational Administration: Theory and Practice*, 2024, 30(5): 8521–8527. https://doi.org/10.53555/kuey.v30i5.4406
3. Levin, I., Marom, M. and Kojukhov, A., Rethinking ai in education: Highlighting the metacognitive challenge. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 2025, 16(1, Sup. 1): 250–263.

4. Elkhodr, M. and Gide, E., The sage framework for developing critical thinking and responsible generative ai use in cybersecurity education. *Discover Education*, 2025, 4(1): 517. https://doi.org/10.70594/brain/16.S1/21
5. Sifakis, J., System design automation: Challenges and limitations. *Proceedings of the IEEE*, 2015, 103(11): 2093–2103. https://doi.org/10.1109/JPROC.2015.2484060
6. Messina, S. and Panciroli, C., Rethinking teaching with genai: Theoretical models and operational tools. *Journal of Inclusive Methodology and Technology in Learning and Teaching*, 2025, 5(1).
7. Qadir, J., Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In *2023 IEEE global engineering education conference (EDUCON)*, 2023, 1–9, IEEE. https://doi.org/10.1109/EDUCON54358.2023.10125121
8. North, M. and Riskas, T., Assessing is learning outcomes effectively in the age of genai. *Issues in Information Systems*, 2025, 26(2): 1–13.
9. Chee, H., Ahn, S. and Lee, J., A competency framework for ai literacy: Variations by different learner groups and an implied learning pathway. *British Journal of Educational Technology*, 2025, 56(5): 2146–2182. https://doi.org/10.1111/bjet.13556
10. Long, D. and Magerko, B., What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, 1–16. https://doi.org/10.1145/3313831.3376727
11. Ng, D.T.K., Leung, J.K.L., Chu, K.W.S. and Qiao, M.S., Ai literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the association for information science and technology*, 2021, 58(1): 504–509. https://doi.org/10.1002/pra2.487
12. Nehru, R., Paredes, S., Roy, S.C., Cuong, T.Q. and Huong, B.T.T., Implementing the revised bloom's taxonomy (2001) in ai-digital and online learning environments: A strategic approach. *Indian Journal of Educational Technology*, 2025, 7(1): 173–189.
13. Holmes, W. and Miao, F., *Guidance for generative AI in education and research*, Unesco Publishing, 2023.
14. Nguyen, K.V., The use of generative ai tools in higher education: Ethical and pedagogical principles. *Journal of Academic Ethics*, 2025, 1–21. https://doi.org/10.2139/ssrn.5003394
15. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al., Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 2023, 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274
16. Fuchs, K., Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse?. In *Frontiers in education*, 2023, 8: 1166682, Frontiers Media SA. https://doi.org/10.3389/feduc.2023.1166682
17. Silva, C.A.G.D., Ramos, F.N., De Moraes, R.V. and Santos, E.L.D., Chatgpt: Challenges and benefits in software programming for higher education. *Sustainability*, 2024, 16(3): 1245. https://doi.org/10.3390/su16031245
18. Marques, N., Silva, R.R. and Bernardino, J., Using chatgpt in software requirements engineering: A comprehensive review. *Future Internet*, 2024, 16(6): 180. https://doi.org/10.3390/fi16060180
19. Cotton, D.R., Cotton, P.A. and Shipway, J.R., Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 2024, 61(2): 228–239. https://doi.org/10.1080/14703297.2023.2190148

20. Biggs, J., Tang, C. and Kennedy, G., *Teaching for quality learning at university 5e*, McGraw-hill education (UK), 2022.
21. Paul, R. and Elder, L., *The miniature guide to critical thinking concepts & tools*, Langara College, 2011.
22. Ennis, R.H., The nature of critical thinking: An outline of critical thinking dispositions and abilities. *University of Illinois*, 2011, 2(4): 1–8.
23. Rittel, H.W. and Webber, M.M., Dilemmas in a general theory of planning. *Policy sciences*, 1973, 4(2): 155–169. https://doi.org/10.1007/BF01405730
24. Anderson, L.W. and Krathwohl, D.R., eds., *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, Longman, 2001.
25. Krathwohl, D.R., A revision of bloom's taxonomy: An overview. *Theory into Practice*, 2002, 41(4): 212–218. https://doi.org/10.1207/s15430421tip4104_2
26. Pintrich, P.R., The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into practice*, 2002, 41(4): 219–225. https://doi.org/10.1207/s15430421tip4104_3
27. W3C Web Accessibility Initiative, Web content accessibility guidelines (wcag) 2.2., 2023. Accessed 2025-10-07. Available from: `https://www.w3.org/TR/WCAG22/`.

A. Appendix A

A.1. The Three-Dimensional Competency Model

The framework operationalises three dimensions of critical thinking adapted from cybersecurity pedagogy for systems analysis contexts, making abstract cognitive capabilities observable through domain-specific behaviours documented in assessment artifacts.

A.1.1. Analytical Deconstruction

This involves systematic identification of gaps, errors, and biases in AI-generated artifacts. In requirements contexts, students identify missing stakeholder perspectives (elderly users, Indigenous customers, accessibility needs) that AI systematically overlooks. In formal modelling, students recognise systematic translation errors including boundary classification failures, state management omissions, and exception handling gaps. In design contexts, students detect optimisation biases where AI recommendations prioritise single dimensions without recognising trade-offs. Observable evidence emerges through error identification rates, rejection patterns, and justification quality articulating systematic AI failure patterns.

A.1.2. Contextual Application

requires adapting generic AI outputs to specific organisational, domain, and stakeholder contexts. In requirements, students modify AI suggestions for GreenHarvest-specific constraints (elderly demographics, fresh produce perishability, regional positioning). In formal modelling, students correct translations for domain-specific business rules (loyalty member versus guest processes, promotional integration, inventory constraints). In design evaluation, students balance recommendations against project constraints (interaction timeframes, elderly cognitive load, environmental conditions).

Observable evidence appears through modification patterns showing sophisticated adaptation and justifications revealing contextual reasoning depth.

A.1.3. Reflective Synthesis

encompasses integrating human and AI contributions whilst maintaining explicit awareness of reasoning processes and trade-offs. In requirements, students justify prioritisation by balancing competing needs using systems analysis principles. In formal modelling, students explain corrections using architectural theory (process mediation, data conservation, CRUD completeness). In design, students articulate trade-offs across stakeholder groups and competing objectives with explicit acknowledgment of who benefits and who compromises. Observable evidence emerges through justification sophistication, synthesis quality ratings, and metacognitive awareness in reflections.

These dimensions operate interdependently: analytical deconstruction enables recognition of when contextual application is necessary; contextual application generates material for reflective synthesis; reflective synthesis deepens future analytical capabilities. Section 4 operationalises these dimensions through measurement frameworks enabling systematic investigation of competency development patterns.

B. Appendix B

B.1. Experiment 1: Requirements Synthesis

This experiment investigates a fundamental question in AI-augmented requirements engineering: can students effectively synthesise human and AI-generated requirements whilst maintaining critical judgment about stakeholder representation, domain appropriateness, and feasibility? Contemporary systems analysts increasingly encounter AI-generated requirement sets that appear comprehensive but may systematically overlook implicit stakeholder needs, organisational constraints, or domain-specific nuances. The experiment tests whether students can identify these gaps whilst leveraging AI's generative capabilities, operationalising synthesis competency through measurable accept/modify/reject decisions.

B.1.1. Task Overview

Students complete a three-phase requirements development workflow for the GreenHarvest Smart Kiosk system (a supermarket self-service kiosk for elderly customers and Indigenous community members). First, they manually author user stories without AI assistance, establishing baseline human requirement quality. Second, they use AI to generate comprehensive requirements using a standardised prompt. Third, they synthesize both sources into a prioritised product backlog, documenting every decision with source attribution, confidence ratings, and justifications. This workflow mirrors professional agile practice where analysts reconcile requirements from multiple sources whilst maintaining traceability.

B.1.2. Detailed Protocol and Phase Aims

Phase 1 - Manual Baseline (Pre-AI): Each of four group members independently writes two user stories (eight total per group) following standard agile format: "As a [stakeholder], I want [functionality], so that [business value]" with 2-3 measurable acceptance criteria and story point estimates (1, 2, 3, 5, or 8).

Aim: Capture students' baseline requirement generation capabilities, analytical priorities, and stakeholder awareness before AI exposure influences their thinking. This phase also documents pre-AI expectations through a structured table where students predict AI's likely contributions and blind spots, enabling later comparison between expectations and actual AI performance.

Phase 2 - AI Generation (Intervention): Groups apply a standardised prompt to ChatGPT requesting comprehensive user stories for all GreenHarvest stakeholders. The prompt explicitly cues consideration of edge cases, non-functional requirements, accessibility needs for elderly users and Indigenous customers, and technical constraints, deliberately testing whether students critically evaluate AI outputs even when prompts appear comprehensive.

Aim: Generate comparable AI outputs across all groups (reduces prompt-quality variation as a confounding factor by using a standardised prompt), whilst creating conditions where students must evaluate AI-generated requirements that mix genuinely useful suggestions with systematic gaps characteristic of current language models.

Phase 3 - Synthesis and Documentation: Groups compile final product backlog of 15+ requirements by evaluating every human and AI story, completing the Product Backlog Synthesis Table (Table 6) for each decision. Required distribution mandates inclusion of all eight human stories, minimum five AI stories (three accepted or modified, two rejected), and at least two hybrid stories merging human and AI ideas. The column definitions and coding intents used in the decision log are summarised in Table 6.

Aim: Force discriminative judgment rather than passive AI acceptance or blanket rejection. The mandated distribution ensures sufficient data for pattern analysis whilst reflecting realistic professional practice where analysts combine multiple requirement sources. Each decision requires a 50-75-word justification, making reasoning processes observable.

Table 6. Product backlog synthesis table structure.

Column	Values	Purpose
Story Title	Brief description	Requirement identification
Source	Human / AI / Hybrid	Attribution tracking enables analysis of source preference patterns
Category	Core / Nice-to-have / Out-of-scope	Tests scope management judgment
Priority	High / Medium / Low	Reveals prioritisation criteria (technical vs user-centred)
Story Points	1, 2, 3, 5, 8	Effort estimation capability
Confidence	1–5 scale	Self-assessment enabling calibration analysis
Decision	Accept / Modify / Merge / Reject	Primary orchestration behaviour
Justification	50–75 words	Reasoning quality and domain knowledge indicator

B.1.3. Measurement and Data Extraction

Completed synthesis tables generate multiple data streams revealing orchestration patterns. Synthesis Evidence is coded as None (no integration, simple copying), Basic (combines sources but minimal reconciliation), or Advanced (sophisticated integration with explicit trade-off reasoning and consistency checking) based on justification content and decision pattern sophistication. Domain

Knowledge is assessed through justification content as Lacking (generic reasoning applicable to any system), Generic (basic retail understanding), or Applied (GreenHarvest-specific insights about elderly customers, Indigenous community needs, fresh produce constraints, or regional supermarket contexts). Accessibility Awareness is rated Absent (no mention of elderly or Indigenous users), Implicit (generic usability language without specific user groups), or Explicit (direct reference to elderly cognitive/physical needs or Indigenous cultural considerations in story titles, acceptance criteria, or justifications).

The distribution of Accept/Modify/Reject decisions across human versus AI sources reveals Orchestration Patterns: Passive Acceptor (predominantly AI acceptance with weak justifications), Selective Adapter (domain-based filtering with differential treatment of technical versus business requirements), Balanced Integrator (systematic synthesis combining strengths from both sources with explicit trade-off reasoning), or Critical Synthesizer (proactive gap identification, systematic bias correction, and generative synthesis producing novel requirements beyond either source). Confidence ratings enable calibration analysis examining whether high-confidence decisions correlate with successful synthesis outcomes. Complete coding rubrics appear in Appendix A. Inter-rater reliability procedures are recommended for future replications.

B.2. Experiment 2: DFD Translation and Correction

This experiment addresses a critical knowledge gap in AI-augmented systems analysis: what formal modelling expertise do students need to identify and correct AI translation errors when converting natural language process descriptions into Data Flow Diagrams? Unlike requirements engineering where evaluation criteria involve subjective stakeholder priorities, formal modelling demands precise semantic relationships and syntactic correctness. AI frequently generates plausible-looking DFDs containing subtle errors in boundary classification, data flow directionality, or process decomposition that fundamentally alter system behaviour. The experiment systematically investigates which types of DFD knowledge prove essential for effective human-AI collaboration in architectural modelling.

B.2.1. Task Overview

Students complete a controlled translation workflow revealing the gap between human conceptual understanding and AI's formal representation. First, they write detailed structured process descriptions for ”Loyalty Member Express Order Processing” without AI involvement, establishing explicit specifications of trigger events, preconditions, process steps, data inputs/outputs, data stores with operation types, external entities, postconditions, and exception handling. Second, they submit these descriptions to AI using a standardised translation prompt requesting Level-0 DFD component listings. Third, they systematically identify and correct AI errors, documenting each correction with error classification, complexity rating, and systems-analysis-based explanations. This workflow tests whether formal modelling knowledge enables error detection that less technically-trained stakeholders would miss.

B.2.2. Detailed Protocol and Phase Aims

Phase 1 - Structured Specification (Baseline):

Students complete the Process Description Template specifying all elements of loyalty member express ordering: what triggers the process (member scans card at kiosk), what must be true beforehand (member has active account, previous orders exist in system), 8-10 numbered process steps with data transformations, all incoming data with sources (member ID from card reader, order history from database), all outgoing data with destinations (order confirmation to member, order details to kitchen management system), data stores accessed with read/write operations, external entities with interaction types, what must be true afterwards (order recorded, inventory updated, payment processed), and 2-3 exception scenarios (payment failure, out-of-stock items, system timeout).

Aim: Create detailed comparison standard enabling systematic evaluation of AI translation accuracy. The structured template ensures students think through all DFD components before seeing AI's interpretation, preventing AI outputs from anchoring student mental models.

Phase 2 - AI Translation (Intervention):

Students submit completed descriptions using standardised prompt requesting Level-0 DFD conversion with components listed in structured format: numbered processes with descriptions, external entities with roles, data stores with contents, and data flows with source-destination-content specifications. The prompt explicitly requests balancing with context diagram.

Aim:

Generate AI translations exhibiting systematic failures observable across groups—enabling investigation of which error types students successfully detect versus overlook, and which formal modelling knowledge proves essential for correction.

Phase 3 - Error Analysis and Correction:

Students identify minimum eight errors in AI-generated DFDs, completing the DFD Correction Log (Table 7) for each. They specify what component AI generated incorrectly, provide corrected version, classify error type, rate correction complexity, and explain why correction is necessary using systems analysis terminology (concepts like balancing, decomposition levels, external entity constraints, data flow closure). The DFD Correction Log fields and their analytic purposes are defined in Table 7.

Aim: Make formal modelling knowledge visible through correction explanations. Students who deeply understand DFD semantics articulate corrections using theoretical principles (e.g., ”external entities cannot directly access data stores; interactions must flow through processes”), whilst those with surface knowledge provide procedural fixes without theoretical grounding. This phase also tests accessibility transfer: does awareness demonstrated in Task 1 requirements spontaneously transfer to architectural modelling without explicit scaffolding? Task 2 deliberately provides no accessibility prompts in templates or reflection questions.

Table 7. DFD correction log structure.

Column	Values	Purpose
Component Type	Process / Entity / Store / Flow	Identifies which DFD element contains the error
AI Output	[Generated version]	Documents AI's incorrect interpretation
Correct Version	[Fixed specification]	Shows the student's correction
Error Category	Structural / Semantic / Notation / Completeness	Enables error taxonomy; reveals systematic AI weaknesses
Complexity	Simple / Moderate / Complex	Indicates knowledge depth required for the correction
Explanation	30–50 words	Reveals understanding through systems analysis reasoning

B.2.3. Measurement and Data Extraction

The correction logs enable multiple analyses revealing formal modelling competency patterns. Error Detection Rate calculates percentage of systematic AI failures successfully identified by each group, indicating observational accuracy and formal notation literacy. Correction Accuracy assesses whether fixes genuinely resolve issues (Perfect), partially address problems (Partial), introduce new errors (Incorrect), or worsen the model (Worse), revealing depth of DFD semantic understanding beyond surface pattern recognition.

Error Distribution analysis across taxonomy categories reveals which AI failure types students recognise versus overlook. Structural errors involve incorrect component relationships or malformed topology (unbalanced decomposition, inappropriate connections between element types). Completeness errors reflect missing processes, flows, or exception pathways (absent confirmation loops, missing audit trails, incomplete state transitions). Semantic errors indicate conceptual misunderstandings of process meaning or data directionality (payment confirmations directed to wrong entities, login treated as process rather than precondition). Notation errors involve improper symbology or numbering conventions (incorrect process numbering, data store labelling violations). Scope errors introduce out-of-specification elements beyond stated requirements.

Cross-cutting thematic analysis examines which knowledge domains prove most challenging independent of primary error category. Boundary Reasoning competence (distinguishing internal processes from external entities, correctly identifying data flow termination points, understanding system scope boundaries) is assessed through corrections involving entity classifications and flow directions. State Management capability (recognising missing bidirectional flows, identifying absent write-back operations, ensuring CRUD operation completeness) is evaluated through corrections addressing data store interactions and transactional closure. Exception Handling sophistication (identifying missing timeout pathways, detecting absent error recovery mechanisms, recognising incomplete failure flows) is measured through corrections adding exception-related DFD elements.

Accessibility Modelling (whether groups spontaneously include accessibility-relevant states, flows, or stores without explicit prompting) is tracked through presence of accessibility profile attributes in data stores, error assistance pathways reflecting cognitive limitations, or explicit read/write semantics for consent and preference management. This measurement tests whether strong accessibility awareness demonstrated in Task 1 requirements transfers to architectural representations, or whether competency expression depends on system abstraction layer with accessibility readily visible at requirements but suppressed at architecture without explicit scaffolding. Detailed error classification criteria and knowledge assessment rubrics appear in Appendix B. Inter-rater reliability procedures are recommended for future replications.

B.3. Experiment 3: Design Evaluation

This experiment investigates how students develop contextual judgment to critically evaluate AI design feedback, specifically examining the tension between universal usability principles and situated constraints. Unlike binary compliance decisions (policy follows regulation or not) or correctness judgments (DFD balances or not), design evaluation requires subjective trade-offs where multiple valid solutions exist. AI design feedback typically applies universal heuristics (Nielsen's usability principles) without recognising contextual factors like physical environment constraints (standing versus seated interaction), demographic-specific needs (elderly users versus general populations), or organisational priorities (accessibility versus feature richness). The experiment tests whether students can confidently override AI recommendations when context demands whilst appropriately accepting universally valid suggestions.

B.3.1. Task Overview

Students complete a design-evaluate-decide workflow mirroring professional user interface (UI) design practice. First, they create low-fidelity wireframes for two GreenHarvest kiosk screens (Welcome/Home and Product Search Results) specifically targeting elderly users (65+) with low digital literacy in standing interaction contexts with 2-3 minute time constraints. They justify design decisions relative to elderly cognitive/physical needs, Indigenous cultural considerations, standing constraints, and supermarket time pressure. Second, they submit wireframes to AI requesting heuristic evaluation using Nielsen's 10 usability principles with explicit attention to elderly impact and busy supermarket context. Third, they critically assess AI feedback, completing evaluation matrix with exactly 10 decisions following mandated 4-3-3 distribution (four implementations, three modifications, three rejections), forcing discrimination between universally valid suggestions and context-inappropriate recommendations.

B.3.2. Detailed Protocol and Phase Aims

Phase 1 - Design Creation and Justification:

Students develop wireframes addressing GreenHarvest's specific constraints: elderly users requiring large touch targets (minimum 44x44 pixels), high contrast for vision limitations, simple navigation without nested menus, visible status feedback, and undo capabilities for error recovery; Indigenous customers potentially requiring language toggles or cultural sensitivity in imagery; standing interaction constraining session duration and favouring recognition over recall (no password entry); busy supermarket environment creating noise interference (visual feedback over audio) and time pressure (progressive disclosure rather than feature completeness). Design justification document (200-250 words) explicitly connects wireframe decisions to these constraints.

Aim: Establish students' independent design reasoning and user empathy before AI evaluation influences their perspective. The justification requirement makes initial design logic transparent, enabling later assessment of whether AI feedback reinforces, challenges, or redirects student thinking.

Phase 2 - AI Heuristic Evaluation (Intervention):

Students submit wireframe descriptions using a standardised prompt requesting evaluation against Nielsen's 10 heuristics (visibility of status, match real world, user control, consistency, error prevention, recognition over recall, flexibility, aesthetic minimalism, error recovery, help documentation). The prompt explicitly cues elderly user impact and supermarket context considerations.

Aim: Generate AI feedback exhibiting systematic patterns across universal principles, design choices, and context-specific recommendations, enabling investigation of whether students appropriately distinguish principle-based feedback from context-blind suggestions and develop graduated authority distribution across these domains.

Phase 3 - Critical Evaluation and Decision Documentation:

Groups complete AI Feedback Evaluation Matrix (Table 8) for exactly 10 AI suggestions with mandated distribution: four implementations (accepting AI feedback as valuable improvement), three modifications (adapting AI suggestion to context constraints), three rejections (overriding AI when recommendation conflicts with elderly needs, standing interaction limits, or supermarket environment realities). Each decision requires 50-75 word justification explicitly addressing elderly impact, contextual appropriateness, and trade-off considerations.Table 8 specifies the AI Feedback Evaluation Matrix used to capture decisions and justifications.

Aim: Develop confident professional judgment to lead human-AI collaboration rather than deferring to algorithmic authority. The mandated rejection quota is pedagogically critical—it forces students to find instances where their contextual expertise supersedes AI's generic recommendations, building confidence in situated judgment. This phase reintroduces accessibility scaffolding through explicit design requirements and evaluation prompts, testing whether awareness suppressed in Task 2 architecture can be reactivated at interface layer.

Table 8. AI feedback evaluation matrix structure.

Column	Values	Purpose
AI Suggestion	15–25 word summary	Documents AI recommendation content
Heuristic	H1–H10	Links to Nielsen's usability principle
AI Severity	1–4 scale	Records AI's importance rating
Agreement	Agree / Partial / Disagree	Student's evaluation position
Action	Implement / Modify / Reject	Decision type revealing authority distribution
Elderly Impact	High / Medium / Low	Context sensitivity indicator
Justification	50–75 words	Reasoning quality and contextual judgment evidence

B.3.3. Measurement and Data Extraction

Completed evaluation matrices reveal sophisticated patterns in human-AI authority negotiation. Context Sensitivity is rated High (justifications explicitly reference elderly users, standing posture, time limits, or supermarket noise), Medium (generic usability reasoning without context-specific factors), Low (purely aesthetic preferences), or None (no justification), enabling assessment of whether students ground decisions in situated analysis versus abstract principles. Trade-off Recognition is classified as Explicit (articulates competing objectives like ”large buttons improve elderly accessibility but reduce information density; prioritising accessibility given target users”), Implicit (acknowledges complexity without detailed reasoning), or Absent (treats decisions as straightforward without recognising tensions), revealing sophistication in handling design's inherently contested nature.

User Empathy is assessed as Strong (demonstrates deep understanding of elderly cognitive load, physical dexterity limits, vision requirements, or Indigenous cultural needs through specific references), Present (acknowledges user characteristics generically), or Lacking (ignores or dismisses elderly-specific considerations), measuring whether students maintain a stakeholder-centred perspective versus defaulting to designer preferences or AI recommendations.

Professional Judgment is evaluated through the confidence and appropriateness of AI overrides. Mature judgment shows confident rejection of context-inappropriate suggestions with sophisticated reasoning articulating why contextual factors supersede universal principles in specific cases. Developing judgment shows some independent thinking with hesitation or over-qualification. Novice judgment defers to AI authority based on severity ratings or algorithmic confidence without critical contextual evaluation.

The matrix enables quantitative analysis of collaboration zones by examining acceptance patterns across heuristic categories. Decisions regarding universal usability principles versus context-specific recommendations reveal whether students appropriately distribute authority, accepting AI guidance where principles genuinely apply, whilst asserting human expertise where situated factors dominate. Analysis investigates whether high AI severity ratings correlate with student implementation decisions, or whether students exercise independent judgment, overriding severity assessments when contextually appropriate.

Author's biography

Dr. Mahmoud Elkhodr is a Senior Lecturer in the School of Engineering and Technology at Central Queensland University, Australia. His research spans cybersecurity, Internet of Things security and management, and AI-enabled digital systems, with a particular focus on privacy-preserving architectures, trustworthy AI, and evidence-based pedagogy for emerging technologies. His recent work advances structured approaches for integrating generative AI into higher education through measurable orchestration competencies, assessment instrumentation, and cross-campus implementation. He has authored and co-authored over 100 peer-reviewed publications and contributes to curriculum innovation that aligns professional practice with rigorous student capability development.

Professor Ergun Gide is a full Professor of Information and Communication Technology at Central Queensland University, Australia, and a nationally recognised recipient of multiple teaching awards. His research and academic leadership encompass intelligent information systems, digital transformation, and technology-enhanced learning, with extensive experience in curriculum design and scalable work-integrated education. He has supervised more than 30 research higher degree graduates and has led institutional initiatives that strengthen learning quality and industry relevance across diverse student cohorts. In this study, he contributed supervisory mentorship and methodological oversight supporting cross-campus implementation and quality assurance.