ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2747

When AI Mentorship Scaffolds or Substitutes: Iterative Design and Evaluation of a Hybrid Validation System for Deep Tech Entrepreneurship Education

Carlos Isaacs Bornand[1]; Erkko Autio[2]; Maria Elizete Kunkel[3]; Luiz Galvão[3]; Ignacio Zambrano[4]

[1,3]*Universidad de la Frontera, Chile. carlos.isaacs@ufrontera.cl;*

[2]*Imperial College London, United Kingdom. erkko.autio@imperial.ac.uk*

[3]*Universidade Federal de São Paulo, Brasil. elizete.kunkel@unifesp.br; legmartins@unifesp.br*

[4]*XpertIA Spa, Chile. izambrano@xpertialab.cl*

Abstract *Deep technology entrepreneurship education requires developing independent judgment under genuine uncertainty, a process that AI tools may either support or undermine depending on their architecture. Three structural limitations characterize current AI tools in this domain: feedback disconnected from validated evidence standards, absence of session memory, and reliance on pre-training data rather than domain-specific entrepreneurial precedents. This paper reports the first Design Science Research (DSR) cycle of ProjectIA, a hybrid system combining eight structured validation frameworks operationalizing the Technology Startup Roadmap (TRL 1–5) with an AI platform constrained to Socratic interrogation and evidence-grounded feedback. Evaluation with 13 technology professionals across 16 weeks tested three hypotheses: H1, that frameworks surface risks professional experience alone cannot detect; H2, that Socratic AI design is perceived as complementary to human mentoring; and H3, that LLM fidelity failures are the primary trust barrier. H1 and H2 were confirmed: frameworks achieved unanimous utility scores (5.0/5.0) and the AI platform scored 4.69/5.0 on complementarity. H3 was confirmed and refined: hallucination control (3.23/5.0) and session memory (3.31/5.0) produced expertise-asymmetric consequences, recoverable for experts but potentially corrupting for novices. Two exploratory patterns require Cycle 2 testing: a scaffolding-substitution risk in less advanced participant groups, and a Technology-Solution Fit gap in eight of thirteen participants. Whether the system develops or substitutes for independent judgment remains the primary open question for Cycle 2.*

Keywords. Entrepreneurship Education; AI Mentoring; Scaffolding-Substitution Boundary; Automation Bias; Design Science Research; Technology Entrepreneurship.

I. Introduction

Every engineering program developing deep technology ventures faces the same structural problem: most universities employ two to five technology transfer professionals supporting twenty to thirty active projects annually, and deep tech mentor pools are shallower still, because the combination of technical domain expertise and entrepreneurial experience required rarely exists in a single individual [7]. That gap between what rigorous early-stage validation demands and what human mentorship supply can provide is precisely what AI tools are being asked to fill.

Early evidence suggests AI tools can fill part of that gap, at least in terms of engagement and perceived usefulness [8], [9], [12]. The harder question, and the one this study is designed to surface, is whether they fill the right part. Technology entrepreneurship education is not about conveying known content; it is about developing independent judgment under conditions of genuine uncertainty, the capacity to reason through incomplete information, test falsifiable assumptions, and revise mental models when evidence demands revision [10], [11]. A tool that relieves the learner of that effortful reasoning may produce outputs that look like entrepreneurial thinking without the cognitive process that should generate them.

That distinction, between scaffolding judgment and substituting for it, has received almost no attention in how AI mentoring tools are currently evaluated. User satisfaction surveys and usability ratings, the dominant instruments in this literature [12], yield similar scores whether a student is developing judgment or delegating it. Ji et al. [13] and Thus et al. [14] both document this dissociation in learning technology contexts: satisfaction and engagement metrics do not reliably track the cognitive outcomes the tools are supposed to produce. Risko and Gilbert [15] describe the underlying mechanism as cognitive offloading. Not all offloading is equivalent: offloading memory or calculation preserves the cognitive activity through which competence develops, while offloading the reasoning process itself removes that activity entirely. When the delegated task is one through which entrepreneurial judgment develops, offloading it to an LLM is not a neutral convenience.

This study's theoretical contribution is the introduction of the scaffolding-substitution boundary as a testable design constraint, not a philosophical concern, grounded in an identifiable data signature. Prior scaffolding literature [3], [4] identifies the principle that effective support must fade as independence develops; this cycle surfaces

exploratory patterns consistent with the conditions under which that fading fails to occur in AI-augmented entrepreneurship education. That construct is not claimed as resolved here; it is operationalized as a measurable design requirement that Cycle 2 must test through a within-subject comparison of supported and unassisted performance.

This article reports the first DSR cycle of ProjectIA, a hybrid digital mentorship system integrating Miro-based validation frameworks, prescribed field work, and an AI platform delivering Socratic feedback. Three hypotheses guided the evaluation: H1, that explicit frameworks surface risks experienced professionals miss without structured guidance; H2, that a Socratic AI design is perceived as complementary to human mentoring rather than substitutive; and H3, that fidelity failures represent the primary trust barrier. All three were tested with 13 technology professionals across 16 weeks.

The structure of this paper is as follows: Section II presents the theoretical framework; Section III describes the system architecture; Section IV details the evaluation methodology; Section V reports the results; Section VI discusses their implications; Section VII presents conclusions and Cycle 2 requirements.

II. Theoretical Framework

A. Deep Tech Validation and the Limits of Lean Startup

Technology ventures carry two distinct risks that most startups do not face simultaneously: market risk, whether customers will pay at viable unit economics, and technical risk, whether the proposed solution can be engineered to function reliably at acceptable cost [19]. These two risk streams are not independent; progress on one frequently gates or enables progress on the other, yet they operate on fundamentally different timescales and require fundamentally different validation methods. Progress on technical risk is assessed through Technology Readiness Levels (TRL), a nine-point scale originally developed by NASA to standardize the assessment of technology maturity from basic research through proven operational deployment [18].

The methodology most widely deployed in entrepreneurship programs to navigate this dual uncertainty, the lean startup framework, was designed for contexts where a minimum viable product can be built and deployed within weeks [20], [21]. That assumption fails for deep tech ventures: a proof-of-concept demonstrating that a novel biosensor, photonic chip, or AI diagnostic algorithm functions under controlled laboratory conditions may require six to eighteen months of sustained R&D before a customer-facing prototype exists [19]. Deploying lean startup heuristics in this context produces two systematic distortions. First, it pressures teams to declare customer validation before the technology can deliver on any validated promise. Second, it treats market risk as the primary uncertainty when technical risk may be the binding constraint at early TRL stages. A framework designed for deep tech ventures must therefore manage both risk streams in parallel, sequencing validation activities according to what is technically demonstrable at each stage of maturity.

The Technology Startup Roadmap [1] addresses this gap by organizing validation across four phases that integrate technical maturation with market validation from TRL 1 through TRL 9. The present study operationalizes and evaluates Phases 1 and 2, which together cover the foundational arc from initial challenge identification through early product-market signals, corresponding to TRL 1 through TRL 5.

Phase 1 (TRL 1 through 3) targets the validation activities that must occur before any product development begins. It opens with Challenge-Solution Fit (CSF), a construct the Roadmap introduces as a necessary antecedent to the lean startup's Problem-Solution Fit (PSF). The distinction is consequential: PSF demands a precisely defined customer segment and a specific pain point, both of which are premature claims when the technology is still at laboratory stage. CSF instead requires evidence that a high-level challenge is real and significant for a broadly defined stakeholder audience, and that the proposed solution concept is plausible in that context [1]. Sufficient evidence for CSF includes rich qualitative data from ten to twenty stakeholder discussions consistently identifying a substantial and recurring challenge, supported by secondary research confirming the challenge's broader relevance. Phase 1 also specifies Proof-of-Concept (PoC) validation, requiring documented laboratory evidence that the core technology functions as theorized (TRL 1 through 3), and an Early IP Strategy milestone, requiring that potentially patentable inventions be identified, freedom-to-operate assessed, and an initial protection strategy documented before incorporation decisions are made.

Phase 2 (TRL 4 through 5) targets early product development and initial market signals, building on the validated challenge concept and feasible technology from Phase 1. It begins with Problem-Solution Fit, requiring twenty to fifty customer discovery interviews with a defined segment and convergence on a specific and recurring

pain point. Following PSF, the Roadmap introduces Technology-Solution Fit (TSF), a validation point absent from lean startup methodology that bridges the gap between a technically feasible PoC and a solution users actually find valuable: whether the specific technology implementation effectively delivers the core solution benefits from the user's perspective, validated through low-fidelity prototypes and early user testing. Successful TSF corresponds to TRL 4. Phase 2 concludes with Initial Demand Validation and MVP Launch at TRL 5. What distinguishes this Roadmap from both lean startup heuristics and generic TRL frameworks is its specification, at each milestone, of the evidence that must exist and the experiments that must generate it before progression is warranted [1]. These sufficiency criteria transform validation from a judgment call into an auditable standard. Evaluating whether a participant's evidence genuinely meets those criteria requires a mentor capable of assessing both methodological rigor and domain-specific plausibility, a combination that, as Siegel and Wright [7] document, cannot be provided at scale by university technology transfer offices.

B. Cognitive Load Theory and Productive Difficulty

Cognitive Load Theory (CLT) distinguishes extraneous load, unnecessary processing that good design should eliminate, from germane load, the effortful schema-building activity through which durable learning occurs [2]. Bjork and Bjork [22] document this empirically: tasks learners find most comfortable are systematically those that consolidate least. A system that generates structured answers before a learner engages with the underlying problem does not reduce extraneous load; it eliminates germane load, and with it, the cognitive activity through which entrepreneurial judgment develops.

C. Scaffolding, Fading, and the Substitution Risk

Vygotsky's zone of proximal development [4] establishes that effective support must be calibrated to what the learner cannot yet do independently and must withdraw as independence develops. Bruner [3] operationalizes this as scaffolding; Sweller et al. [2] formalize the withdrawal mechanism as the guidance-fading effect. A scaffold that does not fade has become a prosthetic. St-Jean and Audet [23] found directive advice increases novice dependency; St-Jean and Tremblay [24] tracked this longitudinally, finding that mentees reporting the highest satisfaction showed the most fragile outcomes once support ended. Expert entrepreneurial judgment develops through analogical practice, matching current situations against repertoires of prior cases [5], [6]; when an AI system generates the analogy and delivers the recommendation, the practice building the learner's repertoire does not occur. Current AI systems produce the linguistic outputs of pattern-matching expertise without the underlying structural comparison [25]. The scaffolding-substitution boundary is therefore a design constraint, not a philosophical concern. Whether AI-augmented entrepreneurship education systems cross that boundary requires a comparison of supported and unassisted performance that this cycle was not designed to deliver; establishing that measurement design is the primary theoretical task this study sets for Cycle 2.

D. Iterative DSR and Staged Hypothesis Testing

Design Science Research is an inherently iterative methodology [16], [17]. Each cycle's failures are as theoretically informative as its successes, because failures identify specific design properties requiring refinement. This study's three hypotheses follow from the theoretical gaps identified above and are operationalized in Sections IV and V.

III. ProjectIA: A Hybrid Miro-Guided Field and AI Platform System

A. System Architecture: Three Interdependent Layers

ProjectIA operates as a hybrid system in which three layers work in sequence. The first layer is the structured instrument: a set of Miro-based validation canvases that translate the Roadmap [1] into operational instructions, specifying which field activities to conduct, which questions to ask stakeholders, which experimental protocols to run, and what evidence thresholds must be met before progression is warranted. The second layer is the field work itself: the customer interviews, laboratory experiments, IP searches, and stakeholder engagements that entrepreneurs conduct in the real world against a live innovation project, and whose results are documented back into the canvas. The canvas structures the inquiry; the field work generates the evidence. Without the canvas instructions the field work has no structure; without the field work the canvas has no evidence to evaluate. The third layer is the AI platform, which receives structured evidence from the completed canvas and provides continuous Socratic feedback, gap detection, and formal proofpoint evaluation.

ProjectIA's seventeen Miro-based canvases cover the full four-phase roadmap from TRL 1 through TRL 9. This DSR cycle operationalizes and evaluates Phases 1 and 2 (TRL 1 through 5). Phase 1 (TRL 1 to 3) addresses foundational validation: futures analysis and challenge identification, empathic inquiry and Innovation Challenge design, value proposition mapping, Challenge-Solution Fit synthesis, business

model exploration, TRL-grounded technology assessment [18], early IP and Freedom-to-Operate strategy, and Proof-of-Concept experimental design. Phase 2 (TRL 4 to 5) addresses product-market validation: Problem-Solution Fit market assessment, demand intensity testing, complementary asset mapping, go-to-market architecture, the PFTB matrix, NABC investor synthesis, MVP experimental design, scalability planning, and comprehensive IP strategy.

B. Frameworks Evaluated in Cycle 1

Of the seventeen canvases, eight were selected for Cycle 1 evaluation, covering the complete Phase 1 validation arc from challenge identification through investor-ready pitch synthesis, plus Framework 09 as the Phase 2 entry point. The eight frameworks are: (01) Scenarios and Preferred Futures; (02) Innovation Challenge, empathic inquiry protocols generating quantified severity and frequency metrics; (03) Value Proposition, jobs-to-be-done mapping; (04) Concept Development and Validation; (06) Technology Assessment, component-level TRL scoring; (07) Early IP Strategy, covering invention disclosure, Freedom-to-Operate, and protection strategy; (09) Market Assessment, Problem-Solution Fit validation with explicit sufficiency thresholds (severity 3.5/5, minimum eight qualified interviews); and (14) NABC Builder, investor-ready synthesis. The remaining nine canvases were assigned to participant project plans to establish structural continuity into Cycle 2.

C. AI Software Platform: Design and v1.0 Implementation

ProjectIA's AI platform was designed around five principles, implemented in v1.0 at three levels. Fully implemented: the Co-Pilot asks probing questions rather than giving direct answers [2], [22]; three stakeholder personas (Skeptical Investor, Methodology Auditor, Technical Reviewer) provide role-specific evaluation frames [30]; and entrepreneurs can override any AI evaluation [28]. Partially implemented: the constraint preventing unsupported claims was enforced in the Formal Validation Agent but not consistently in the Co-Pilot; and session memory was limited to the immediately preceding session rather than full project history [29]. Not implemented: override logging with written justification [28], deferred to Cycle 2. Four functional components deliver these principles: a structured workspace for proofpoint documentation; a Co-Pilot for gap identification; a Formal Validation Agent rendering VALIDATED, PARTIALLY VALIDATED, or REJECTED with cited rationale; and an Experiment Plan Generator for follow-on field work.

IV. Evaluation Methodology

A. DSR Evaluation Framework

DSR evaluation criteria [16], [17] were operationalized across five dimensions: Validity, Utility, Usability, Quality, and Generalizability. Methods triangulated quantitative surveys (Likert scales; 32 items for frameworks, 28 for software), open-ended survey responses (qualitative items embedded in each instrument, all 13 participants), and artifact analytics (Miro time logs per framework, software interaction logs capturing Co-Pilot queries, Validation Agent submissions, and hallucination incidents).

B. Participants

Participants were recruited from a postgraduate Deep Technology Management course at a public research university in Brazil (UNIFESP). All 13 enrolled participants completed all three evaluation instruments (N=13, no withdrawals). The cohort comprised professionals with active careers in technology-intensive sectors, enrolled at master's or doctoral level, and working on innovation projects either as founders, as employees within technology companies, or as researchers advancing applied innovations within institutional contexts. Eligibility required an active technology innovation project at TRL 1 to 5 and a 16-week commitment applying the frameworks to that live project. Recruiting experienced technology professionals rather than novice students was a deliberate methodological choice: when practitioners with active industry experience report usability barriers, those barriers reflect genuine structural constraints in the system rather than gaps in user preparation. For analysis purposes, participants were organized into three sectoral cohorts based on their project domain and TRL stage. Table I presents the formal characterization of each group.

TABLE I
Participant Group Characterization (N=13)

Group	n	Domains	TRL Range
1: AI / GovTech	5	Artificial Intelligence, Gov. Tech.	TRL 4-5 (advanced)
2: MedTech/Biotech	5	MedTech (n=4), Biotech (n=1)	TRL 2-4 (intermediate)
3: EdTech / Mixed	3	EdTech (n=2), Hardware/Robotics (n=1)	TRL 1-3 (early stage)

C. Protocol

Weeks 1 and 2 covered orientation and framework walkthrough with an example project; weeks 3 to 10 involved framework application through weekly three-hour sessions (90 minutes lecture, 90 minutes

guided project work); week 11 was devoted to the framework evaluation survey; weeks 12 to 14 provided ProjectIA software access and proofpoint submissions; week 15 completed the software evaluation survey; and week 16 finalized the integration evaluation survey, completed by all 13 participants.

V. Results

A. Framework Evaluation: H1 Confirmed with a Usability Cost

TABLE II
Framework Evaluation (N=13)

Metric	Mean	SD	Finding
V1: Fidelity to roadmap	4.77	0.44	Excellent
V2: Completeness	4.62	0.51	Excellent
U1: Risk discovery	5.00	0.00	UNANIMOUS – H1 confirmed
U2: Decision impact	4.85	0.55	12/13 made concrete pivots
U3: Recommendation	5.00	0.00	UNANIMOUS
US2: Autonomous completion	3.15	0.69	CRITICAL – instructor dependency
US3: Time efficiency	2.54	0.88	CRITICAL – avg. 60h total
US4: Logical sequence	4.62	0.87	Coherent progression
Q1–Q2: Quality (avg.)	4.70	0.56	Excellent

H1 is confirmed (Table II). All 13 participants identified risks previously invisible to them (U1: 5.0/5.0) and all 13 would recommend the frameworks to peers (U3: 5.0/5.0); 12 of 13 made concrete decisions attributable to framework insights (U2: 4.85, SD=0.55). Three types of blind spot drove the unanimous utility scores. Technical-market dependency failures: seven participants found their technical solution assumed unvalidated customer infrastructure or behavior; a MedTech participant developing an AI diagnostic tool discovered through Framework 09 interviews that clinical adoption required HL7 EHR integration not included in the prototype specification, before hardware procurement that would have locked in an incompatible architecture. Problem misframing: six participants found through Frameworks 01 and 02 that they were solving the wrong problem; a CleanTech participant believed the challenge was that municipal waste had no commercial value, while futures analysis surfaced that the actual barrier was municipalities lacking incentives to separate waste streams. IP exposure: participants without prior IP strategy encountered university ownership provisions through Framework 07 requiring pre-incorporation negotiation, with several within weeks of decisions that would have created material post-incorporation disputes.

Average total time was approximately 60 hours (SD=12.3). The Miro canvas for Framework 02 instructs entrepreneurs to conduct 8 to 12 stakeholder interviews with defined severity and frequency metrics; that prescribed field work, not the canvas structure itself, accounts for the majority of the time investment. One participant described the distribution precisely: the canvas takes two hours; the work to fill it honestly takes twenty-five. That observation is theoretically accurate within CLT terms [2]: the twenty-five hours of field work is the germane load through which the system produces its outcomes. Reducing it would not improve the system; it would remove the layer from which all downstream value derives.

B. Software Evaluation: H2 Confirmed, H3 Confirmed and Refined

TABLE III
Software Evaluation (N=13)

Metric	Mean	SD	Finding
SU1: Feedback relevance	4.38	0.87	Good
SU2: Promotes reflection	4.31	1.03	Good – Socratic design works
SU4: Quality of pushback	3.92	0.95	Acceptable, inconsistent
C15: Metacognitive awareness	4.77	0.44	HIGHEST result – software eval.
C16: AI equivalence to mentor	2.85	0.90	LOWEST software score
C18: Complementarity	4.69	0.85	EXCELLENT – H2 confirmed
F2: Hallucinations (5=never)	3.23	1.36	CRITICAL – H3 confirmed
F3: Session memory	3.31	0.95	CRITICAL – H3 confirmed
F5: Trust in evaluation	4.08	1.26	Good – pre-incident baseline
I1: Framework–software transfer	3.23	1.36	Friction – manual overhead

H2 is confirmed (Table III): complementarity scored 4.69/5.0 (SD=0.85). Participant responses identified two consistent mechanisms. Role differentiation: across all three groups, participants described a clear

functional division between AI and human mentoring; one Group 1 participant stated that software feedback was very rich in specific details while the mentor usually gave a more broad overview, and that therefore they complement each other; another drew the boundary precisely, noting that software is ideal for ensuring technical and scientific rigor while humans are ideal for relationship strategy and market vision. A Group 3 participant identified what made AI interrogation distinctive: it exhaustively and repeatedly asks questions and pushes for changes in a way that a client would not tolerate from a human mentor, a friction cost in human relationships that becomes an asset in a structured validation tool.

Several participants described value in accessing the system between mentor sessions, citing continuity between customer pain points and technical features. That perceived continuity, however, must be read against the session memory limitation confirmed in H3: the v1.0 architecture accessed only the immediately preceding session, meaning any longitudinal coherence participants experienced depended on their own documentation practices rather than on system architecture.

The highest-scoring metric in the entire software evaluation was metacognitive awareness (C15: 4.77/5.0, SD=0.44), meaning participants developed a reliable sense of where the AI's limits were, a result that matters for the scaffolding argument even if its persistence after program completion requires Cycle 2 testing. The boundary of H2's confirmation: two Group 2 participants noted that human mentors demonstrated superior contextual judgment; the AI complements human mentoring where methodological rigor and continuous availability are the binding constraints, but is not perceived as equivalent where contextual strategy, political judgment, or domain expertise are required.

H3 is confirmed with an important refinement. System logs recorded 8 hallucination incidents across 13 participants. The two most consequential each produced an explicit and persistent trust breakdown. In the first, the Co-Pilot cited a 2023 Nature Biotechnology paper that does not exist; the participant spent 45 minutes attempting to locate it before confirming the fabrication, and subsequently verified every AI-cited source independently, inverting the tool's intended function. In the second, the Validation Agent rejected an experimental protocol by citing a measurement standard (TIA-455-61B) in a context where it does not apply; the participant's 15 years of industry expertise detected the error. The refinement to H3 is that fidelity failures carry expertise-asymmetric consequences: domain experts detected the errors, though detection itself carried a cost; participants without equivalent domain expertise would have lacked the reference point to detect the fabrication at all, accepting the AI's constraints as authoritative and revising valid experimental work on that basis.

Memory failures confirmed H3 on the session-amnesia dimension. The v1.0 retrieval logic accessed only the immediately preceding proofpoint; when a participant's target segment shifted between sessions, the Validation Agent evaluated later submissions without referencing earlier definitions. One participant described what the system should have done: checking in real time whether the Value Proposition and Key Benefits being written actually address the most severe and frequent pain points identified in the Pattern Analysis section. Without persistent memory, the guidance-fading effect [2] is architecturally blocked.

Eight of 13 participants ended the evaluation cycle unable to close Technology-Solution Fit: their validated problem, designed solution, technology under development, and promised benefits did not hold together as a coherent whole when examined against each other. In most cases the gap was not caused by missing field work; participants had conducted the required interviews and experiments. What they had not resolved was whether their validated findings were mutually consistent. The lean startup methodology identifies market rejection as the primary trigger for pivoting [20]: the entrepreneur goes to market, customers do not respond, and the team revises direction. What this cycle observed is a different and earlier trigger that lean startup methodology does not name: the entrepreneur's own validated findings ceasing to be consistent with each other, before the market has had any opportunity to respond. A participant who validated a severe hospital problem, designed a solution around it, and then discovered that their current technology cannot deliver what that solution requires has not been rejected by the market. Their own evidence has become internally contradictory. Technology-Solution Fit is the proofpoint that surfaces that contradiction, precisely because the Roadmap requires all validation streams to be examined together rather than sequentially in isolation. The split follows consistent TRL-stage lines: all five Group 1 participants (TRL 4 to 5) reached full completion, while all eight participants in Groups 2 and 3 showed incomplete or partial fit. This pattern is surfaced here as Exploratory Hypothesis E1: the Technology-Solution Fit gap is an internally generated consistency failure distinct from market rejection pivots, more prevalent at early TRL stages, and traceable through artifact log analysis.

Confirming E1 is a primary investigative target for Cycle 2.

VI. Discussion

A. Framework Utility and the Paradox of Productive Difficulty

The framework data present an apparent contradiction. All 13 participants unanimously agreed that the frameworks helped them discover previously invisible risks and all 13 would recommend them to peers, yet those same participants rated the time required as the lowest-scoring dimension of the entire evaluation, averaging just 2.54 out of 5.0 for approximately 60 hours of total work. That combination is only a design failure if the time cost and the utility are treated as independent variables. They are not. The MedTech participant did not discover the HL7 EHR integration gap by reading a framework; he discovered it by interviewing clinicians. The CleanTech participant did not find that he was solving the wrong problem by filling out a canvas; he found it by talking to municipalities. A framework designed to surface technical-market dependencies must require the activities that make such discoveries possible; shortening the process would not make the system more efficient, it would make it less honest. Within CLT terms [2], the field work behind a single canvas is germane load: the effortful schema-building activity through which entrepreneurial judgment develops. The frameworks were not designed to minimize that load; they were designed to direct it toward the validation activities most likely to surface the risks that matter.

B. Complementarity Without Equivalence: What the Gap Reveals

The complementarity data tell a more precise story than the headline score suggests. Participants rated the AI as strongly complementary to human mentoring (C18: 4.69/5.0), the platform's second-highest result. Yet those same participants rated the AI as equivalent in usefulness to a human mentor at only 2.85/5.0 (C16), the lowest score in the entire software evaluation outside of fidelity failures. That nearly two-point gap is not a contradiction; it is the most analytically useful finding the study produced about the AI-mentor relationship. The system is valued where methodological rigor and continuous availability are the binding constraints. It is not considered a substitute where contextual strategy, political judgment, and domain expertise are required. Role differentiation emerged consistently across all three groups: one participant described the AI as ideal for ensuring technical and scientific rigor while humans remain ideal for relationship strategy and market vision. That division maps precisely onto what the Roadmap requires at each stage: structured evidence generation, where the AI adds value, and judgment about strategic positioning, where human expertise remains irreplaceable.

The platform's highest-scoring result overall was metacognitive awareness (C15: 4.77/5.0): participants nearly unanimously reported the ability to recognize when to seek human help rather than relying on the AI. That boundary-recognition capacity is a functionally important scaffolding property, representing the difference between a tool that builds dependency and one that develops judgment about its own limits. Whether that metacognitive awareness persists after program completion, when the scaffolding is removed, remains the primary unresolved question for Cycle 2.

C. Fidelity Failures and Expertise-Asymmetric Consequences

H3 is confirmed with an architecturally significant refinement. The two most critical software failures, hallucinations and session amnesia, were not independent bugs. They shared a common structural cause: version 1.0 assigned too many distinct functions to a single AI context without the structural separation needed to keep each function reliable. The hallucination problem traced to a specific enforcement gap: the citation constraint was applied in the formal validation component but not consistently in the conversational co-pilot. The session amnesia problem traced to a retrieval architecture limited to the immediately preceding session, blocking the longitudinal consistency checks that the guidance-fading mechanism requires [2].

The design-relevant refinement to H3 is that fidelity failures carry expertise-asymmetric consequences. Domain experts detected the errors, though detection itself carried a cost. The participant who spent 45 minutes chasing a fabricated citation subsequently verified every AI-cited source independently, inverting the tool's intended function. The participant with 15 years of industry expertise caught the misapplied measurement standard. Participants without equivalent expertise lacked the reference point to identify either fabrication, and would have revised valid experimental work on the basis of the AI's fabricated authority. That asymmetry is architectural, not incidental: the system's failure mode is most damaging precisely for the users it is designed to serve, consistent with the automation bias mechanism documented by Goddard et al. [28]. For any program considering AI tools for deep tech validation work, scores above 4.0 on both hallucination control and session memory should be treated as minimum thresholds before deployment.

Seven of 13 participants independently identified the same priority change: the ability to interact with the system through a conversational interface, asking questions and contesting outputs in real time rather than submitting structured text and receiving static feedback. Nine of 13 participants reported that data generated in the Miro frameworks was not used by the software, primarily the outputs of the first four frameworks. The framework rated as most critical for market validation had the lowest traceability score of all six frameworks evaluated. These are not usability complaints; they describe a structural gap between what the frameworks produce and what the software can receive. The architecture proposed for Cycle 2 addresses both failures by separating what version 1.0 conflated: a dedicated citation layer that makes unsourced claims architecturally impossible, and a persistent memory layer providing full project history for the guidance-fading mechanism [2].

D. Exploratory Patterns: E1 and E2

Two exploratory patterns emerged from the evaluation that the study was not designed to confirm. Their epistemic status must be distinguished carefully from the three confirmed hypotheses, and both are surfaced here as primary investigative targets for Cycle 2.

Exploratory Hypothesis E1 addresses the Technology-Solution Fit gap described in Section V. Eight of 13 participants could not achieve internal consistency across their validated findings, following consistent TRL-stage lines. The lean startup methodology names market rejection as the primary pivot trigger [20]; what this cycle observed is a different and earlier trigger that lean startup methodology does not name. Whether that pattern reflects a design gap in earlier frameworks, a TRL-stage-specific challenge, or a characteristic of the multi-domain cohort cannot be determined from survey data alone; artifact log analysis is required to trace the mechanism.

Exploratory Hypothesis E2 addresses the group-differentiated pattern in H2 data. More advanced participants (Group 1, TRL 4 to 5) consistently rated the AI as not equivalent to a human mentor while simultaneously rating it as fully complementary, a stable and deliberate division of labor. Less advanced participants (Groups 2 and 3) rated the AI closer to equivalent in usefulness. Given the small group sizes involved (n=5 and n=3), this difference must be treated as a pattern consistent with the hypothesis rather than a confirmation of it. St-Jean and Tremblay's [24] finding that mentees reporting the greatest relief from support showed the most fragile post-support outcomes provides the theoretical basis for this concern. Confirming whether it applies here requires the unassisted performance comparison that Cycle 2 is designed to deliver.

E. Limitations and Cycle 2 Requirements

Five constraints bound this cycle. First, qualitative evidence derives exclusively from open-ended survey items; response depth is bounded by the survey format, and behavioral nuances that would surface in extended interviews are not captured. Second, the professional expert cohort likely underestimates usability barriers and overestimates error-detection capacity for less experienced users. Third, the sub-group patterns in E1 and E2 rest on small n values and must not be generalized beyond the sample without Cycle 2 replication. Fourth, the redesigned architecture proposed for Cycle 2 has not yet been built or tested; it is a reasoned response to the failures this cycle identified, not a demonstrated solution. Fifth, and structurally most important: this cycle was not designed to measure unassisted performance at program completion, and it did not. Thirteen participants reported the system as valuable and complementary; what the evaluation cannot say is whether those 16 weeks left them more capable of conducting comparable validation work without the scaffold. Measuring whether the system develops or substitutes for judgment requires a within-subject comparison of supported and unassisted performance on tasks of equivalent difficulty, administered after the intervention ends. That design was deferred to Cycle 2 by intent and remains the primary methodological requirement for the next research cycle.

VII. Conclusion

Structured validation frameworks operationalizing the Technology Startup Roadmap [1] surface risks that professional experience alone does not, unanimously, across domains, and specifically because the field work they require generates evidence that no AI platform can substitute. That result is not a usability problem to be engineered away; it is the mechanism through which the system produces value, consistent with CLT's account of germane load [2].

The AI platform complements human mentoring where methodological rigor and continuous availability are the binding constraints, and does not replace it where contextual strategy and domain judgment are required. The nearly two-point gap between perceived complementarity (C18: 4.69/5.0) and perceived equivalence (C16: 2.85/5.0) is a more precise instrument for evaluating AI mentoring tools than overall satisfaction scores. Metacognitive awareness (C15: 4.77/5.0) emerged as the platform's highest result, indicating that participants developed

reliable judgment about the system's limits, an important but not sufficient scaffolding property.

Fidelity failures in AI validation tools are not uniformly distributed. They are most damaging for the users who most need the tool and who are least able to detect fabricated authority. That asymmetry is architectural, not incidental, and it establishes a design requirement for any system deployed as a primary validation guide in contexts where users lack independent verification capacity.

This cycle produces four contributions with distinct standing. Two are validated: eight operationalized canvases with auditable sufficiency criteria achieving unanimous risk-discovery scores with a professional doctoral cohort across five technical domains; and the expertise-asymmetric fidelity failure pattern establishing a design requirement for AI validation systems. Two are exploratory and constitute the primary hypotheses for Cycle 2: E1, that the Technology-Solution Fit gap is an internally generated consistency failure distinct from market-rejection pivots; and E2, that less experienced participants are at higher substitution risk and may develop less independent judgment during supported intervention. The scaffolding-substitution boundary, introduced here as a theoretically grounded and empirically testable design constraint, defines the primary outcome measure for Cycle 2: a within-subject comparison of supported and unassisted performance on validation tasks of equivalent difficulty, administered after the intervention ends. That comparison is what any AI mentoring system in this domain should ultimately be evaluated against.

References

[1] [Author details withheld for blind review], Proofpoint Roadmap for Technology Startups, Draft v0.3, working paper, 2025.

[2] J. Sweller, J. J. G. van Merrienboer, and F. Paas, "Cognitive architecture and instructional design: 20 years later," Educational Psychology Review, vol. 31, no. 2, pp. 261–292, 2019.

[3] J. S. Bruner, The Process of Education, 2nd ed. Harvard Univ. Press, 1977.

[4] L. S. Vygotsky and M. Cole, Mind in Society: Development of Higher Psychological Processes. Harvard Univ. Press, 1978.

[5] R. A. Baron and M. D. Ensley, "Opportunity recognition as the detection of meaningful patterns," Management Science, vol. 52, no. 9, pp. 1331–1344, 2006.

[6] T. R. Holcomb et al., "Architecture of entrepreneurial learning," Entrepreneurship Theory and Practice, vol. 33, no. 1, pp. 167–192, 2009.

[7] D. S. Siegel and M. Wright, "Academic entrepreneurship: Time for a rethink?" British Journal of Management, vol. 26, no. 4, pp. 582–595, 2015.

[8] M. Obschonka et al., "Artificial intelligence and entrepreneurship: A call for research," Entrepreneurship Theory and Practice, vol. 49, no. 3, pp. 620–641, 2025.

[9] T. L. Ahlgren et al., "Assisting early-stage software startups with LLMs," Information and Software Technology, vol. 187, 107832, 2025.

[10] J. M. Haynie et al., "A situated metacognitive model of the entrepreneurial mindset," Journal of Business Venturing, vol. 25, no. 2, pp. 217–229, 2010.

[11] D. M. Townsend and R. A. Hunt, "Entrepreneurial action, creativity, and judgment in the age of artificial intelligence," Journal of Business Venturing Insights, vol. 11, e00126, 2019.

[12] J. H. Park, S. J. Kim, and S. T. Lee, "AI and creativity in entrepreneurship education," AI, vol. 6, no. 5, Article 100, 2025.

[13] Y. Ji et al., "Human-machine co-creation: The effects of ChatGPT on students' learning performance," IEEE Transactions on Learning Technologies, vol. 18, pp. 412–425, 2025.

[14] H. Thus, P. Bachmann, and A. Meier, "Chatbot-based tutoring," Journal of Computer Assisted Learning, vol. 40, no. 2, pp. 486–501, 2024.

[15] E. F. Risko and S. J. Gilbert, "Cognitive offloading," Trends in Cognitive Sciences, vol. 20, no. 9, pp. 676–688, 2016.

[16] A. R. Hevner et al., "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[17] K. Peffers et al., "A design science research methodology for information systems research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, 2007.

[18] J. C. Mankins, Technology readiness levels: A white paper. NASA, Advanced Concepts Office, 1995.

[19] M. Heder, "From NASA to EU: The evolution of the TRL scale in public sector innovation," The Innovation Journal, vol. 22, no. 2, pp. 1–23, 2017.

[20] T. Eisenmann, E. Ries, and S. Dillard, Hypothesis-Driven Entrepreneurship: The Lean Startup. Harvard Business School Background Note 812-095, 2021.

[21] E. Ries, The Lean Startup. Crown Business, 2011.

[22] E. L. Bjork and R. A. Bjork, "Making things hard on yourself, but in a good way," in Psychology and the Real World, 2nd ed. Worth Publishers, 2011, pp. 56–64.

[23] E. St-Jean and J. Audet, "The effect of mentor intervention style in novice entrepreneur mentoring relationships," Mentoring and Tutoring, vol. 21, no. 1, pp. 96–119, 2013.

[24] E. St-Jean and M. Tremblay, "Mentoring for entrepreneurs: A boost or a crutch?" International Small Business Journal, vol. 38, no. 5, pp. 424–448, 2020.

[25] O. N. Oyelade, H. Wang, and K. Rafferty, "SMAR+NIE IdeaGen," Expert Systems with Applications, vol. 279, Article 127455, 2025.

[26] M. MacArthur, "Large language models and the problem of rhetorical debt," AI and Society, advance online publication, 2025.

[27] A. Camuffo et al., "A scientific approach to entrepreneurial decision making," Management Science, vol. 66, no. 2, pp. 564–586, 2020.

[28] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review," Journal of the American Medical Informatics Association, vol. 19, no. 1, pp. 121–127, 2012.

[29] E. J. Huang, M. Easterday, and E. Gerber, "AI that helps us help each other," Proceedings of the ACM on Human-Computer Interaction, vol. 9, no. 7, pp. 1–30, 2025.

[30] M. Catta-Preta et al., "Innovation flow: A human-AI collaborative framework," Applied Sciences, vol. 15, no. 22, 11951, 2025.