From Prompts to Verified Loops: The PCHL-HE Framework for Generative AI-Assisted Educational and Research Content Creation in Higher Education

Oleksii Nalyvaiko

School of Psychology, V. N. Karazin Kharkiv National University

Kharkiv, Ukraine

ORCID: https://orcid.org/0000-0002-7094-1047

Abstract

Generative artificial intelligence is moving beyond isolated prompt-response exchanges toward systems that curate information, call tools, retain state, verify intermediate products, and repeat actions under explicit control. Higher education lacks a pedagogically grounded vocabulary for selecting among these increasingly complex forms of interaction. This conceptual preprint develops the Prompt-Context-Harness-Loop Framework for Higher Education, abbreviated PCHL-HE. The framework is constructed through an integrative synthesis of research on prompting, retrieval and context management, agentic systems, human interaction with automation, artificial intelligence in higher education, assessment, feedback, academic integrity, and design science. PCHL-HE differentiates four task configurations across eight dimensions: unit of control, information grounding, workflow complexity, temporal horizon, tool orchestration, feedback and iteration, verification, and human oversight. Prompt engineering designs a bounded instruction. Context engineering designs the information environment available to the model. Harness engineering designs a complete, multi-step, tool-mediated working pass. Loop engineering designs bounded repeated runs governed by feedback, stopping rules, escalation conditions, and memory. The layers are not a maturity hierarchy. The framework therefore introduces the minimally sufficient layer principle: educators and learners should use the least complex configuration capable of producing a verifiable result. The paper maps the four layers to six categories of educational content and six categories of research content, differentiates teacher and student responsibilities, and proposes a four-part verification architecture covering output, evidence, process, and human decision validity. Eight testable propositions and a staged research program are specified for future scoping review, expert validation, and comparative pilot studies. The contribution is a vendor-neutral framework that connects emerging agent-engineering concepts with pedagogy, learner agency, traceability, and academic responsibility.

Keywords: generative artificial intelligence; higher education; prompt engineering; context engineering; harness engineering; loop engineering; academic integrity

1. Introduction

Generative artificial intelligence (GenAI) has entered higher education through a deceptively simple interface: the prompt. This interface has encouraged a view of AI use as a sequence of isolated requests followed by human acceptance, rejection, or revision. Prompting remains important, and systematic work

on prompting demonstrates that the form of an instruction can substantially shape model behaviour (Liu et al., 2023). Yet the prompt no longer defines the full design problem. Contemporary systems can retrieve documents, access tools, preserve working memory, coordinate several specialized processes, evaluate outputs, and continue acting across repeated runs. The pedagogical question has therefore shifted from how to write a better prompt to how to configure an accountable human-AI workflow.

This shift matters because the quality of AI-assisted educational and research content cannot be inferred from linguistic fluency. A plausible explanation may be factually wrong, a literature synthesis may contain fabricated or misrepresented sources, a test item may be misaligned with the intended learning outcome, and a repeated agentic workflow may compound rather than correct error. Reviews of AI in higher education repeatedly identify opportunities alongside problems of reliability, ethics, assessment, and methodological rigour (Bond et al., 2024; Chiu et al., 2023; Crompton & Burke, 2023; Farrokhnia et al., 2024; Lo, 2023). Research on ChatGPT and academic work likewise emphasizes that human researchers remain responsible for sources, interpretation, and scientific claims (Salvagno et al., 2023; van Dis et al., 2023).

The available educational guidance is fragmented. Some frameworks address institutional policy, assessment permission, or ethical use (Chan, 2023; Furze et al., 2024; Lodge et al., 2023). Other literature explains technical mechanisms such as retrieval-augmented generation, agentic reasoning, or self-correction (Lewis et al., 2020; Wang et al., 2024; Yao et al., 2023). These contributions are valuable, but they do not provide a common pedagogical model for deciding when a teacher or student needs a single instruction, a curated source environment, a multi-step tool-mediated workflow, or a controlled iterative process. Without this distinction, simple tasks can become unnecessarily automated, while complex tasks remain under-specified and under-verified.

A recent technical vocabulary makes this gap more visible. Context engineering treats the model's information environment as a designed object (Mei et al., 2025). Agent harness research treats the surrounding system, tools, and control logic as consequential for performance (Ning et al., 2026). Macedo (2026) explicitly organizes coding-agent practice as a progression from prompt to context to harness to loop and defines a loop specification through its trigger, goal, verification, stopping rule, and memory. These contributions arise mainly from software and agent engineering. Their pedagogical implications for educational and scientific content creation remain insufficiently operationalized.

This paper proposes PCHL-HE, a Prompt-Context-Harness-Loop Framework for Higher Education. The claim is not that the four labels are new. The contribution lies in their integration into a higher-education framework with explicit operational dimensions, a principle for choosing the minimally sufficient layer, content taxonomies, teacher and student role boundaries, and verification requirements. The framework is intended as a conceptual artifact that can be evaluated, revised, and empirically tested.

1.1 Aim and research questions

The aim of this conceptual study is to develop a pedagogically grounded framework that differentiates prompt, context, harness, and loop engineering and maps them to the creation of educational and research content by university teachers and students, with particular attention to verification, human oversight, and academic integrity.

The paper addresses four research questions:

1. How can prompt, context, harness, and loop engineering be operationally distinguished for higher education?

2. Which layer is minimally sufficient for different educational and research content-production tasks?
3. How should teacher and student responsibilities change as workflow complexity and AI autonomy increase?
4. Which verification and governance controls are required at each layer, and how can the framework be empirically validated?

1.2 Intended contribution

The paper makes five linked conceptual contributions:

- A four-layer, vendor-neutral framework defined across eight operational dimensions.
- The minimally sufficient layer principle, which treats complexity as a design cost rather than as an automatic sign of sophistication.
- Parallel taxonomies for educational content and research content, each mapped to teacher and student activities.
- A verification architecture that links output checks, evidence traceability, process controls, and accountable human decisions.
- Eight testable propositions and a staged empirical research agenda for expert validation and comparative studies.

2. Conceptual foundations

2.1 From prompt design to system design

Prompt engineering treats the model-facing message as the principal design object. A prompt can specify a role, task, constraints, examples, evidence requirements, and output format. This is suitable when the task is bounded, the required information can fit into a single interaction, and the user can directly evaluate the result. Prompting research has produced taxonomies of discrete prompts, demonstrations, reasoning cues, and prompt optimization methods (Liu et al., 2023). In educational practice, however, a well-formed prompt cannot compensate for missing or unreliable evidence.

Context engineering expands the design object from the message to the information available at each model step. Retrieval-augmented generation illustrates the underlying logic: external information is retrieved and supplied to a generative model so that the response can be grounded in a relevant evidence set (Lewis et al., 2020). The broader context-engineering literature includes retrieval, generation, processing, memory, tool outputs, and multi-agent information exchange (Mei et al., 2025). Educationally, this layer foregrounds source selection, provenance, representativeness, recency, and the risk that irrelevant or biased documents will shape the result.

Harness engineering expands the design object again. An agent harness is the system around the model that assembles prompts and context, manages state, exposes tools, invokes specialized processes, records intermediate products, and performs checks. The agent literature shows that reasoning and action become interdependent when a model observes tool results and adjusts its next step (Wang et al., 2024; Yao et al., 2023). Recent work argues that harness structure can materially determine agent behaviour and performance (Ning et al., 2026). In higher education, the harness is not merely technical infrastructure. It is a pedagogical workflow in which permissions, checkpoints, evidence standards, and responsibility must be designed.

Loop engineering governs repeated harness runs. A loop is warranted when feedback from one run must alter the next action and when completion cannot be established in advance by a fixed sequence.

Macedo (2026) identifies trigger, goal, verification, stopping rule, and memory as central elements of a loop specification. Related research on verbal feedback and tool-interactive critique suggests that iterative correction can improve some model behaviours (Gou et al., 2024; Shinn et al., 2023). Nevertheless, models do not reliably correct reasoning merely because they are asked to reconsider. External feedback can be necessary, and repeated self-evaluation may amplify confidence without improving validity (Huang et al., 2024).

2.2 Pedagogical anchors

PCHL-HE is anchored in four pedagogical and human-factors traditions. First, constructive alignment requires that intended learning outcomes, learning activities, and assessment be designed as a coherent system (Biggs, 1996). AI-generated content is educationally valuable only when it supports this alignment. A polished artifact that is unrelated to the learning outcome is not a successful result.

Second, self-regulated learning positions learners as agents who set goals, select strategies, monitor progress, and evaluate outcomes (Zimmerman, 2002). AI can scaffold these processes, but it can also displace them. The appropriate layer therefore depends partly on whether the task is intended to produce an artifact or to develop the student's capability. Higher automation may be defensible in a production task and pedagogically counterproductive in a learning task.

Third, feedback literacy emphasizes the learner's capacity to interpret information about quality and use it to improve subsequent work (Carless & Boud, 2018). This idea is especially relevant to loop engineering. A productive loop requires meaningful criteria, interpretable feedback, and a decision about the next action. Repetition without learning-relevant feedback is not a pedagogically valid loop.

Fourth, human-centered automation research distinguishes levels and types of automation and stresses the need to match automation to human capabilities and consequences (Parasuraman et al., 2000; Shneiderman, 2020). PCHL-HE therefore treats human oversight as a structural property rather than as a final disclaimer. Oversight includes selecting evidence, defining permissions, setting checkpoints, interpreting uncertainty, and retaining authority over consequential decisions.

3. Methodological approach

Version 1.0 is a conceptual design study. It does not report a systematic review, an expert panel, or participant data. The framework was developed through an integrative synthesis of several literature streams: prompting and large language model interaction; retrieval, context, and memory; autonomous agents and tool use; iterative critique and self-correction; artificial intelligence in higher education; assessment and academic integrity; human interaction with automation; and design science.

The framework-building procedure followed a design-science logic in which a problem is defined, a conceptual artifact is constructed, and an evaluation program is specified (Hevner et al., 2004; Peffers et al., 2007). Sources were selected purposively to establish construct definitions, identify pedagogical constraints, and make each design decision traceable. Peer-reviewed journal articles and conference papers were prioritized. Recent arXiv preprints were included only where the technical constructs are emerging faster than the peer-reviewed literature. Their preprint status is retained in the reference list and treated as a limitation.

The synthesis involved four analytical steps. First, candidate definitions were extracted and compared according to their unit of control and temporal scope. Second, recurrent design properties were

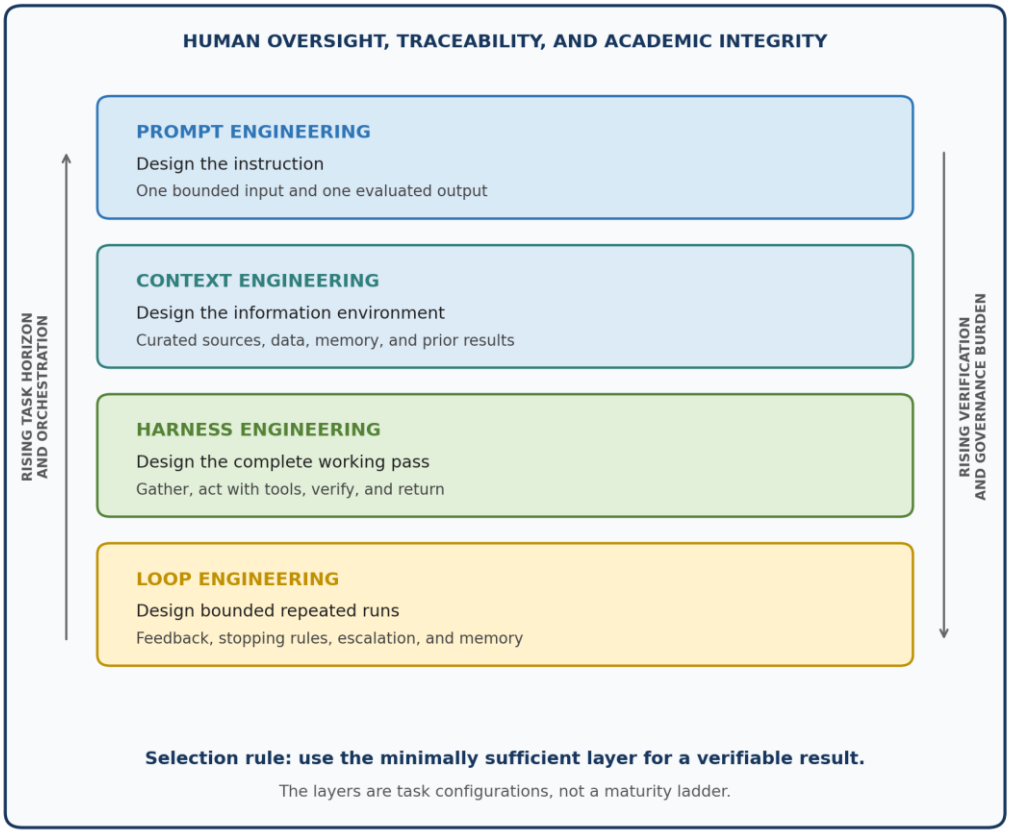
consolidated into eight dimensions. Third, the layers were mapped to educational and research content types using the minimally sufficient layer principle. Fourth, failure modes and governance needs were matched to verification controls. Conceptual validity in this version rests on definitional clarity, internal coherence, non-redundancy among layers, and traceability to published work. Empirical validity remains an open question addressed by the research agenda in Section 9.

4. The PCHL-HE framework

PCHL-HE defines four layers of AI-assisted work. Each layer changes what the user designs, what the model can do, how long the task persists, and what must be verified. The layers can be nested. A harness contains prompts and context, while a loop repeatedly invokes a harness. Nesting does not make the layers interchangeable. The relevant classification is determined by the principal object of design and by the control problem that must be solved.

Figure 1

The PCHL-HE framework



Note. The layers represent task configurations. Selection is based on the minimally sufficient layer, not on an assumption that a higher layer is inherently better.

4.1 Eight analytical dimensions

The layers are distinguished by eight dimensions. Unit of control identifies the artifact being designed. Information grounding describes the evidence made available to the model. Workflow

complexity concerns the number and dependency of operations. Temporal horizon distinguishes a single exchange, a bounded pass, and repeated runs. Tool orchestration describes whether external tools are called and coordinated. Feedback and iteration specify how results alter subsequent actions. Verification defines how success is tested. Human oversight identifies where people set boundaries, approve actions, and assume responsibility.

Table 1

Operational distinctions among PCHL-HE layers

Dimension	Prompt	Context	Harness	Loop
Unit of control	A model-facing instruction	The model's working information set	A complete bounded working pass	A specification for repeated runs
Information grounding	Included in, or assumed by, one request	Curated documents, data, memory, and prior outputs	Context assembled and updated across steps	Context and memory updated after each run
Workflow complexity	One bounded generation task	One source-grounded generation or analysis task	Several dependent operations with checkpoints	Repeated passes whose next actions depend on feedback
Temporal horizon	One exchange or short revision	One context-bounded task	One complete multi-step pass	Multiple bounded runs until a terminal state
Tool orchestration	Usually absent or manually invoked	Retrieval or document access may be present	Tools, code, search, or specialized processes are coordinated	The harness is re-invoked under control logic
Feedback and iteration	Human revises the prompt	Human or system revises the working set	Checks can trigger bounded retry within the pass	Feedback changes the next run and state
Verification	Direct human review of output and fit	Output review plus evidence and provenance checks	Stage-level tests plus end-product verification	External criteria, stopping rules, escalation, and audit trail
Human oversight	Formulate and judge	Select evidence and judge	Design permissions, checkpoints, and final approval	Define boundaries, monitor exceptions, and authorize completion

Note. The table specifies ideal types. Real systems may combine elements from adjacent layers, but classification should follow the dominant design and control problem.

4.2 Prompt engineering

Prompt engineering is the design of a bounded model-facing instruction. Its essential elements are the task, purpose, relevant constraints, desired evidence use, audience, and output format. Examples include requesting a plain-language explanation, generating alternative research questions, reformulating a paragraph for clarity, or producing a draft set of examples. The human remains immediately responsible for both input and evaluation.

The educational strength of prompt engineering is low setup cost. It supports exploration, ideation, formative explanation, and short transformations. Its weakness is that the model may rely on opaque training data or assumptions not shared by the user. Prompt engineering is therefore insufficient when claims must be traceable to a defined evidence base, when several tools or dependent steps are required, or when the task must continue until independently verified criteria are met.

4.3 Context engineering

Context engineering is the design of what the model can see and use at a particular step. The working set may include selected publications, a syllabus, assessment criteria, institutional policy, datasets, student work, prior decisions, tool outputs, and memory. Context engineering is not synonymous with making a

prompt longer. Its central act is curation: selecting, structuring, compressing, labelling, and, when necessary, excluding information.

The main quality problem moves from wording to evidence governance. A context may be incomplete, outdated, unrepresentative, or internally inconsistent. Source provenance must therefore be visible. Sensitive student data require a lawful and institutionally approved environment. In research use, bibliographic details and quotations must be checked against primary sources. A model-generated synthesis cannot substitute for this verification.

4.4 Harness engineering

Harness engineering is the design of one complete, bounded AI-system pass. The pass assembles a goal, prompts, context, state, tools, intermediate artifacts, and verification steps. A higher-education harness might retrieve curriculum requirements, generate a lesson sequence, create assessment items, test alignment, check accessibility, and return a package for teacher approval. A research harness might collect a defined set of article metadata, extract specified fields, create a literature matrix, flag uncertain records, and produce a traceable report.

The harness shifts attention to workflow architecture. Each action requires a permission boundary, each intermediate product requires a status, and each transition requires a condition. Verification should occur at meaningful checkpoints rather than only after the final output. A harness can improve consistency and traceability, but it also introduces failure propagation. An incorrect search result can contaminate extraction, synthesis, and citation unless checks are placed before downstream reuse.

4.5 Loop engineering

Loop engineering is the design of repeated harness runs controlled by explicit feedback and terminal conditions. A pedagogically valid loop specifies at least six elements: a trigger, a goal, a verifier, a rule for updating the next action, a stopping rule, and an escalation rule. Memory is required when information from previous runs must persist. Resource caps, including time, cost, and number of iterations, are also necessary.

Examples include iteratively improving a course module against an alignment rubric, revising an analytic pipeline after failed reproducibility checks, or updating an evidence synthesis when new eligible studies enter a monitored corpus. Loop engineering is not repeated prompting without control. It is also not justified merely because an agent can continue. A loop should be used only when feedback changes the next action, the criterion can be evaluated independently, and a human can understand the basis for completion.

Educational risks are substantial. Repeated model self-critique may create an appearance of diligence without independent validation. Long loops can obscure authorship, increase automation bias, consume resources, and reduce the user's understanding of how the artifact was produced. High-stakes academic decisions, including final grading, misconduct judgments, admissions, and research conclusions, should not be delegated to an autonomous loop.

5. The minimally sufficient layer principle

PCHL-HE rejects the assumption that users should progress upward whenever technology permits. Every additional layer creates new coordination, verification, privacy, and comprehension burdens. The

appropriate design is the least complex layer capable of producing a valid and traceable result under the task's constraints. This is the minimally sufficient layer principle.

Selection begins with the epistemic and pedagogical requirements of the task. If a result can be judged directly and does not depend on a defined evidence base, prompt engineering may be sufficient. If claims must be grounded in selected sources or data, context engineering is required. If multiple operations, tools, or intermediate checks must be coordinated, harness engineering becomes appropriate. If feedback from a completed pass must determine the next pass, loop engineering may be justified. Consequence, data sensitivity, and learning purpose can constrain this choice and may require a lower level of automation even when a more complex system is available.

Table 2

Decision rules for selecting the minimally sufficient layer

Task condition	Usual layer	Typical examples	Minimum verification
A bounded output can be directly reviewed, and no defined evidence set is required.	Prompt	Explanation, brainstorming, wording alternatives, format conversion	Accuracy, relevance, audience fit, and format
The output must be based on specified sources, documents, data, or prior work.	Context	Source-grounded lesson, policy summary, literature comparison	Provenance, completeness, citation fidelity, and contextual bias
The task contains dependent stages, tools, intermediate products, or permissions.	Harness	Teaching package, assessment bank, literature matrix, analysis pipeline	Stage gates, tool outputs, error propagation, and final coherence
A verified result from one complete pass must change what happens next.	Loop	Rubric-guided revision, monitored evidence update, reproducibility repair	External verifier, stopping and escalation rules, resource caps, and audit log

Note. The usual layer can be reduced or prohibited when the assessed competence, institutional policy, privacy requirements, or potential consequences demand direct human performance.

6. Educational content creation

Educational content is defined here as an artifact intended to support teaching, learning, feedback, or assessment. The same artifact can require different layers depending on purpose. A single analogy for a seminar may require a prompt, while a set of explanations constrained by a curriculum and approved readings requires context. A complete module with activities, assessment, accessibility checks, and version control requires a harness. An adaptive sequence that changes after verified learner performance may require a loop.

Table 3

Educational content taxonomy and role mapping

Content type	Illustrative outputs	Teacher responsibility	Student responsibility	Usual layer
Explanatory content	Definitions, worked examples, analogies, mini-lectures	Generate alternatives, then select and adapt to outcomes and learner level.	Request explanations, counterexamples, or self-testing prompts without replacing required reasoning.	Prompt; Context when source-bounded
Instructional content	Lesson plans, instructions, exercises, cases, discussion prompts	Align materials with outcomes, time, prerequisites, and accessibility.	Use AI to clarify task steps or explore cases within stated course rules.	Context; Harness for a package

Content type	Illustrative outputs	Teacher responsibility	Student responsibility	Usual layer
Assessment content	Items, prompts, rubrics, exemplars, formative feedback	Retain authority for validity, difficulty, fairness, scoring, and final decisions.	Use only at the permitted AI level and disclose assistance where required.	Context or Harness; no autonomous high-stakes grading
Adaptive support	Hints, differentiated examples, scaffolds, learning pathways	Define pedagogical triggers, privacy limits, and escalation to a teacher.	Use feedback to plan next learning actions and preserve independent practice.	Context or Harness; Loop only with validated signals
Multimodal content	Slides, diagrams, scripts, audio, video, interactive material	Coordinate modalities and check accessibility, attribution, and consistency.	Develop representations while verifying claims, rights, and course relevance.	Harness
Course-level content	Modules, course maps, resource sets, assessment systems	Design and approve constructive alignment, sequencing, and version control.	Contribute feedback and user testing, not unreviewed curriculum decisions.	Harness; Loop for bounded revision

Note. Layer selection depends on the task purpose. If the intended outcome is the student's unaided performance, AI assistance may need to be restricted even when it could produce a technically stronger artifact.

6.1 Teacher use across the educational lifecycle

Teachers can use PCHL-HE during curriculum analysis, lesson preparation, facilitation, assessment design, feedback, and course improvement. Prompt engineering is appropriate for local ideation. Context engineering is needed when materials must conform to curriculum, policy, approved readings, or a specific student profile. Harness engineering is appropriate for packages whose elements must remain aligned. Loop engineering is appropriate for bounded revision cycles in which validated evidence, such as item analysis, accessibility review, or structured student feedback, changes the next version.

Teacher responsibility does not diminish as the layer rises. It changes from direct composition toward architecture, evidence selection, permission design, and approval. The teacher must be able to explain why the workflow was appropriate, which information it used, how quality was evaluated, and where human judgment entered.

6.2 Student use and learner agency

Student use should be defined by the learning purpose rather than by a general attitude toward AI. Prompt and context engineering can support explanation, planning, feedback interpretation, and comparison of evidence. Harnesses can support complex projects when students are expected to manage tools and document decisions. Loops can support revision when the learner interprets feedback and remains responsible for the next action. When a system independently performs the competence being assessed, the artifact may cease to provide valid evidence of student learning.

Clear permission levels are therefore necessary. The AI Assessment Scale demonstrates one approach to communicating allowable AI involvement in assessment (Furze et al., 2024). PCHL-HE complements such permission scales by describing workflow architecture. A task may permit AI use but still require evidence of source verification, prompt and context records, human decision points, or an oral defence of the work.

7. Research content creation

Research content is defined as an artifact used to formulate, conduct, analyse, report, or communicate scholarly inquiry. Research use creates a higher traceability burden because scientific claims must be

linked to evidence and methods. AI can support workflow operations, but it is not an accountable author, a primary source, or an independent verifier of its own claims.

Table 4

Research content taxonomy and verification requirements

Content type	Illustrative outputs	Usual layer	Human responsibility	Critical verification
Problem formulation	Problem statements, objectives, questions, hypotheses, construct maps	Prompt or Context	Researcher establishes novelty, theoretical fit, feasibility, and final wording.	Do not infer novelty from model confidence; search and assess the literature.
Source-grounded analysis	Search terms, evidence tables, literature matrices, thematic comparisons	Context or Harness	Researcher defines inclusion logic and checks every record and interpretive claim.	Verify title, authors, DOI, quotations, eligibility, and claim-source fit.
Methodological content	Design options, protocols, instruments, codebooks, analysis plans	Context or Harness	Researcher justifies method, obtains ethics review, and validates instruments.	AI output cannot replace methodological expertise or ethical approval.
Data-analytical content	Code, transformations, tables, visualizations, preliminary interpretations	Harness	Researcher preserves data governance, tests assumptions, and reproduces results.	Retain code, versions, seeds where relevant, diagnostics, and independent checks.
Scholarly communication	Titles, abstracts, summaries, presentations, reviewer-response drafts	Prompt, Context, or Harness	Researcher ensures exact consistency with evidence and takes authorship responsibility.	Do not add claims, citations, or certainty beyond the underlying study.
Long-form research	Articles, reports, review manuscripts, dissertation sections	Harness; bounded Loop	Researcher controls argument, claim-evidence links, citations, and final text.	Use external criteria, version history, stopping rules, and disclosure.

7.1 Source-grounded scholarly work

Literature work is a clear example of the difference between layers. A prompt can generate search concepts or a preliminary structure. Context engineering is required when a model compares an identified corpus. A harness can retrieve metadata, screen records against explicit criteria, extract fields, and flag inconsistencies. A loop may update a living review when a scheduled search identifies eligible new studies. At every layer, a bibliographic record must be verified against the publisher, DOI registry, or original paper before it enters the final manuscript.

The model's synthesis should also be auditable at the level of claims. A valid record does not guarantee that the cited paper supports the sentence. Claim-source fit requires reading the relevant primary source, distinguishing reported findings from the current author's interpretation, and representing limitations. Automated citation generation without this control is incompatible with responsible scholarship.

7.2 Data analysis and reproducibility

Harness engineering is particularly suitable for data-analytical tasks because the workflow can retain code, inputs, outputs, diagnostics, and tests. The same capability creates risks if private or restricted data are transferred into an unapproved system. Data minimization, access control, de-identification where appropriate, and institutional approval should precede tool use. A repeated loop may repair code after a failed test, but statistical assumptions and substantive interpretation still require qualified human judgment.

8. Verification, academic integrity, and governance

Verification is not a single final step. PCHL-HE distinguishes four targets. Output verification asks whether the artifact is accurate, relevant, coherent, accessible, and aligned with its purpose. Evidence verification asks whether data and sources are authentic, complete, and correctly represented. Process verification asks whether tools, intermediate steps, permissions, and state transitions operated as intended. Decision verification asks whether an accountable human reviewed the appropriate evidence and retained authority over consequential conclusions.

Verification intensity should rise with task consequence, workflow opacity, number of dependent operations, duration, and AI autonomy. Repetition alone is not verification. A model that generates an answer and then judges the same answer can reproduce shared blind spots. Where possible, verification should use independent sources, executable tests, predetermined rubrics, blinded human ratings, or separate qualified review. High-consequence decisions require direct human adjudication.

Table 5

Minimum governance controls by layer

Layer	Primary failure modes	Minimum controls	Required human gate
Prompt	Hallucination, ambiguity, overgeneralization, inappropriate style	Fact check, task-fit review, format check, compare alternatives	Human accepts, revises, or rejects every output
Context	Omitted, stale, biased, sensitive, or misattributed evidence	Source inventory, provenance labels, corpus completeness check, citation verification	Human selects and authorizes the evidence set
Harness	Tool failure, silent transformation, error propagation, permission overreach	Stage gates, logs, schema checks, test cases, intermediate review, reproducibility	Human defines permissions and approves consequential transitions
Loop	Runaway iteration, false completion, reward gaming, drift, comprehension debt	External verifier, progress metric, iteration and resource caps, stopping and escalation rules	Human defines terminal states, monitors exceptions, and authorizes completion

Note. Controls are cumulative. Harness and loop workflows also require the prompt- and context-level controls embedded within them.

8.1 Academic integrity and assessment validity

Academic integrity cannot be reduced to detecting AI-generated text. The central questions are whether the permitted level of assistance was clear, whether the submitted artifact is valid evidence of the learner's competence, whether sources and contributions are accurately represented, and whether the student can account for the decisions made. Assessment reform therefore requires task redesign, transparent AI permissions, and opportunities to verify learning through process evidence or performance (Cotton et al., 2024; Lodge et al., 2023; Xia et al., 2024).

A defensible disclosure should identify the system and version where known, the dates or period of use, the tasks delegated, the principal context or sources supplied, the verification performed, and the human decisions retained. Disclosure is not a substitute for compliance with course rules, research ethics, privacy law, copyright, or journal policy.

8.2 Privacy, intellectual property, and institutional control

Context, harness, and loop configurations can expose substantially more information than a single prompt. Course materials, unpublished manuscripts, assessment items, personal student data, confidential feedback, and research datasets should not be uploaded to a system without authorization and an adequate data-processing basis. Institutional AI policy should therefore cover approved tools, data classifications,

retention, access, logging, intellectual property, human review, and incident response. Policy frameworks for higher education and general AI risk management provide useful foundations for this governance layer (Chan, 2023; Miao & Holmes, 2023; National Institute of Standards and Technology, 2023).

9. Discussion

PCHL-HE offers a bridge between the technical design of contemporary GenAI systems and the pedagogical design of higher education. Its central move is to relocate attention from the visible prompt to the full configuration of information, actions, feedback, verification, and human authority. This shift does not diminish prompt engineering. It places prompting within a larger architecture and clarifies the conditions under which a prompt is sufficient.

The framework differs from a maturity model. A teacher who uses a single prompt for a bounded, low-consequence task is not operating at an inferior level. The teacher may be making the most defensible design choice. Conversely, a loop is not advanced merely because it runs autonomously. Its quality depends on the validity of the verifier, the transparency of state changes, the suitability of stopping rules, and the human capacity to understand and challenge the result.

PCHL-HE also differs from an assessment permission scale. Permission frameworks specify what degree of AI use is allowed, while PCHL-HE specifies how the workflow is organized and what must be checked. The two are complementary. A course can state that AI assistance is permitted and then require students to disclose the PCHL-HE layer, evidence set, tools, iterations, and human decisions. This produces richer process evidence than a binary declaration.

The framework makes verification burden visible. Agentic systems can reduce manual production effort while increasing the effort required to establish provenance, monitor intermediate steps, inspect exceptions, and defend conclusions. Efficiency claims should therefore include verification and governance costs. The relevant outcome is not the speed of generation but the cost of obtaining a trustworthy and pedagogically valid result.

The pedagogical importance of the minimally sufficient layer principle extends beyond efficiency. It protects learner agency. Students need opportunities to formulate problems, select evidence, monitor understanding, and justify decisions. A system that delivers a stronger artifact can still produce a weaker learning experience if it removes these actions. PCHL-HE therefore makes the purpose of the task a primary selection criterion.

9.1 Scope of the priority claim

The prompt-context-harness-loop sequence has already been articulated in technical discourse, most explicitly by Macedo (2026). The priority documented by this preprint is narrower and pedagogical: the PCHL-HE framework as an operational model for higher education, defined by eight dimensions; the minimally sufficient layer principle; parallel educational and research content taxonomies; teacher and student role mapping; a cumulative verification architecture; and a testable validation program. Future versions should preserve this attribution boundary.

10. Limitations

This preprint has five principal limitations. First, it is conceptual and has not yet been validated by experts or tested with teachers and students. Second, context, harness, and loop engineering are emerging

terms whose definitions may change. Third, part of the technical foundation relies on recent preprints because peer-reviewed research has not yet stabilized. Fourth, the taxonomies may require disciplinary adaptation, especially in laboratory, clinical, legal, creative, and high-security settings. Fifth, model capabilities, interfaces, and institutional policies evolve quickly. The framework is intended to remain vendor-neutral, but its practical controls require periodic review.

These limitations define the next research steps rather than invalidate the framework. The empirical program should test discriminant validity among layers, examine whether users can classify tasks reliably, measure verification cost, and identify contexts in which the proposed ordering or taxonomies fail.

11. Conclusion

Higher education needs a vocabulary that reflects how GenAI work is actually changing. A prompt is one design object, not the entire system. PCHL-HE distinguishes prompt, context, harness, and loop engineering according to what is designed, how evidence and tools are organized, how long the task persists, how feedback changes action, and how success is verified. It maps these configurations to educational and research content while preserving human responsibility.

The framework's normative claim is deliberately conservative: use the minimally sufficient layer that can produce a verifiable result without unnecessary automation or loss of learner agency. As workflow complexity rises, verification and governance must rise with it. The next stage is empirical. A scoping review, expert validation, and comparative task studies can determine whether PCHL-HE is conceptually distinct, practically usable, and educationally beneficial.

Declarations

Funding. No external funding was received for the preparation of this conceptual preprint.

Conflict of interest. The author declares no conflict of interest.

Ethics statement. No human participants, personal data, or animals were involved in Version 1.0. Ethics review was therefore not applicable. Future expert and participant studies will obtain approval or a documented exemption before data collection.

Data availability. No dataset was generated or analysed for this conceptual preprint.

Use of generative AI. OpenAI ChatGPT 5.6. Sol was used as a language, structuring during manuscript preparation. The author reviewed the conceptual claims, verified bibliographic records, determined the final argument and framework, and assumes full responsibility for the manuscript.

References

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364. https://doi.org/10.1007/BF00138871

Bond, M., Khosravi, H., de Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*, Article 4. https://doi.org/10.1186/s41239-023-00436-z

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20,* Article 38. https://doi.org/10.1186/s41239-023-00408-3

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

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

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460-474. https://doi.org/10.1080/14703297.2023.2195846

Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology, 40*(4), 38-55. https://doi.org/10.14742/ajet.9434

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sx038qxjek

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

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*. https://openreview.net/forum?id=IkmD3fKBPQ

Jünger, S., Payne, S. A., Brine, J., Radbruch, L., & Brearley, S. G. (2017). Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine, 31*(8), 684-706. https://doi.org/10.1177/0269216317690685

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), Article 195, 1-35. https://doi.org/10.1145/3560815

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

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

Macedo, S. (2026). *Stop hand-holding your coding agent: Engineering the loops that replace step-by-step prompting* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.00038

Mei, L., Yao, J., Ge, Y., Wang, Y., Bi, B., Cai, Y., Liu, J., Li, M., Li, Z.-Z., Zhang, D., Zhou, C., Mao, J., Xia, T., Guo, J., & Liu, S. (2025). *A survey of context engineering for large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.13334

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Ning, X., Tieu, K., Fu, D., Wei, T., Li, Z., Bei, Y., Zou, J., Ai, M., Liu, Z., Li, T.-W., Chen, L., Zhao, Y., Yang, K., Li, B., Qian, C., Li, G., Lin, X., Zeng, Z., Qiu, R., ... He, J. (2026). *Code as agent harness* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.18747

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286-297. https://doi.org/10.1109/3468.844354

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

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459-467. https://doi.org/10.1002/nur.20199

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care, 27*, Article 75. https://doi.org/10.1186/s13054-023-04380-2

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems, 36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction, 36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C. M., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467-473. https://doi.org/10.7326/M18-0850

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*, Article 186345. https://doi.org/10.1007/s11704-024-40231-1

Xia, Q., Weng, X., Ouyang, F., Lin, T.-J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, Article 40. https://doi.org/10.1186/s41239-024-00468-z

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2