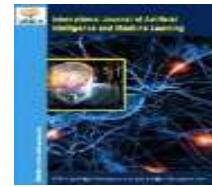




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

EVIDENCE-Prompt: An Auditable Secondary-Data Framework for Scientific Inquiry with Generative AI in Higher Education

Rissa A. Lasap^{1,2*}, Luigi Carlo M. De Jesus^{2,3,4}

¹Biological Science Department, Centro Escolar University, Manila, Metro Manila, Philippines. E-mail: rissalasap22@gmail.com, ORCID: 0009-0002-1029-3460

²Naveen Jindal Youth Fellowship Program, O.P. Jindal Global University, Sonapat, Haryana, India. E-mail: luigi_dejesus@dlsu.edu.ph

³Department of Computer, Electronics, and Electrical Engineering, De La Salle University, Manila, Metro Manila, Philippines. E-mail: luigi_dejesus@dlsu.edu.ph

⁴School of Engineering, Asia Pacific College, Makati City, Metro Manila, Philippines.

Abstract

Generative artificial intelligence is increasingly used for questioning, data interpretation, experimental planning, and explanation in higher education, but most prompt frameworks optimise fluency rather than the integrity of scientific reasoning. This study develops EVIDENCE-Prompt, an auditable framework that treats a prompt as an information-governance object linking a learner's question, admissible evidence, model output, and human verification. A secondary-data design re-analysed 25 operational studies retained from a 2025 systematic review and added seven post-seed records, yielding 32 studies across biological, medical, engineering, computing, social-science, language, and cross-disciplinary settings. Each record was coded for eight components: Epistemic goal; Variables and context; Inquiry phase; Data and evidence; Explicit constraints; Neutrality and alternatives; Citation and verification; and Error checking and reflection. Descriptive prevalence estimates used Wilson 95% confidence intervals; cross-domain contrasts used Fisher's exact tests with Holm adjustment. Epistemic goals and contextual specification appeared in all records. Explicit constraints and reflective checking each appeared in 29/32 studies (90.6%), whereas inquiry-phase alignment and evidence boundaries appeared in 20/32 (62.5%), neutrality and alternatives in 18/32 (56.3%), and citation verification in only 9/32 (28.1%). Median framework coverage was 6 of 8 components (interquartile range 5–7), and only four records operationalised all components. STEM, health, and engineering studies more frequently bounded data and operationalised citation verification than other domains, although adjusted subgroup tests were exploratory and non-significant. EVIDENCE-Prompt converts these gaps into a reproducible prompt schema, scoring rubric, and governance workflow. The accompanying CSV dataset and codebook support independent audit, adaptation, and future model benchmarking.

Keywords: generative artificial intelligence; prompt engineering; scientific inquiry; secondary data; information governance; higher education; evidence verification; human-AI interaction

1. Introduction

Large language models (LLMs) have become a general-purpose interface for accessing, transforming, and generating information. In higher education, students and educators now use them to formulate questions, interpret data, generate code, plan experiments, draft explanations, and obtain feedback. These uses place LLMs directly inside the scientific inquiry cycle. The educational issue is therefore no longer limited to whether AI is used; it is whether the information system preserves the relations among question, method, evidence, inference, uncertainty, and accountability. Reviews of generative AI in education identify substantial opportunities alongside hallucination, bias, privacy, integrity, and over-reliance risks [1–2].

Prompt engineering is often presented as a technical method for obtaining a better answer. That framing is too narrow for scientific learning. A prompt also structures the user's epistemic intentions: it determines what counts as relevant context, which evidence the model may use, whether competing explanations are considered, and what must be checked before a conclusion is accepted. Prompt patterns can improve output structure and interaction quality [3], but an educationally defensible prompt must also expose the learner's responsibility for evidence selection and verification. This requirement connects prompt engineering to AI literacy, information governance, and human-computer interaction rather than to wording optimisation alone [4].

This study operationalises the eight-component EVIDENCE framework through a transparent secondary-data design. It re-analyses published prompt-engineering evidence, quantifies which epistemic controls are present or absent, and derives a cross-disciplinary framework suitable for biological science and engineering contexts.

This common ground is important because scientific prompts must handle both domain-specific evidence and general information-system controls.

A recent systematic review identified 33 higher-education studies and concluded that prompt engineering is teachable but still lacks a comprehensive, validated taxonomy and consistent assessment methods [5]. Subsequent empirical studies have begun to test interventions and student prompt quality [6–7], while discipline-specific work in chemistry shows the importance of structured domain knowledge and evidence boundaries [8–9]. However, the literature still lacks an audit-oriented framework that joins scientific inquiry, source traceability, alternatives, error checking, and institutional governance in one observable schema.

1.1 Research questions

RQ1. Which EVIDENCE components are operationalised in the selected higher-education prompt-engineering literature, and how frequently do they occur?

RQ2. How complete are existing approaches when assessed as auditable epistemic scaffolds rather than as output-optimisation techniques?

RQ3. Do STEM, health, and engineering records differ from other higher-education domains in their use of evidence boundaries, verification, and inquiry-phase alignment?

RQ4. How can the observed gaps be converted into a reproducible prompt schema, scoring rubric, and information-governance workflow for scientific inquiry?

1.2 Contributions

- A study-level secondary dataset and codebook in CSV format, enabling audit and re-analysis.
- An eight-component operational coding model that distinguishes prompt fluency from epistemic integrity.
- Quantitative evidence on the under-representation of source verification and competing explanations.
- A cross-domain prompt architecture linking biological-science inquiry and engineering/computing analysis.
- Implementation guidance for assessment, prompt logging, model-version control, and human verification.

2. Conceptual Background

2.1 Scientific inquiry as an information system

Scientific inquiry is a coordinated process rather than a single answer-producing act. Questions delimit a problem; variables and conditions define its scope; methods determine what observations can be generated; evidence supports or weakens claims; uncertainty limits inference; and verification permits correction. In an AI-supported workflow, each of these elements becomes an information object that can be recorded, exchanged, inspected, and revised. A prompt is therefore analogous to an input specification, while the model response is an intermediate computational artefact rather than an authoritative conclusion.

This systems view changes the unit of evaluation. A response may be fluent yet epistemically weak if it silently imports unsupported assumptions, fabricates a reference, ignores an alternative mechanism, or fails to distinguish observation from interpretation. Conversely, a response with explicit uncertainty and verification steps may be more valuable for learning even when it is less polished. Validity consequently concerns the inferences supported by the entire prompt–output–verification process, not merely the surface quality of the output [10].

2.2 Prompt engineering, AI literacy, and assessment

Prompt-engineering curricula commonly teach role assignment, context, task specification, examples, format control, and iteration. These techniques improve usability, but student performance also depends on prior knowledge, the ability to assess outputs, and the willingness to reject incorrect advice. Knoth et al. operationalised prompt quality through observable components and showed that AI literacy relates to prompting strategies [6]. Brief interventions have improved prompt-writing ability and self-efficacy, although the available studies remain small and context-dependent [7]. Therefore, assessment should capture both prompt construction and downstream verification.

Information literacy adds a crucial source dimension. The CLEAR framework integrates critical evaluation into prompt use [11], and medical and scientific applications increasingly emphasise domain constraints and traceable information [12–13]. Nevertheless, citation checking is not consistently operationalised. This omission is particularly consequential for scientific inquiry, where fabricated identifiers or unsupported numerical claims can contaminate subsequent analysis.

2.3 EVIDENCE-Prompt architecture

EVIDENCE-Prompt treats a prompt as an auditable epistemic contract. The learner declares the inquiry goal, context, phase, evidence boundary, constraints, alternatives, verification rules, and error-checking obligations. The model is asked to expose uncertainty and limitations, while the human retains final responsibility. Figure 1 summarises the architecture.



Figure 1. EVIDENCE-Prompt architecture for auditable scientific inquiry.

Table 1. Operational definitions of the EVIDENCE components.

Component	Guiding question	Required prompt element	Primary risk addressed
Epistemic goal	What scientific question or claim is being examined?	Question, claim type, intended reasoning	Goal ambiguity; answer substitution
Variables and context	What constructs, population, units, and conditions matter?	Definitions, boundary conditions, prior knowledge	Context collapse; invalid generalisation
Inquiry phase	Which scientific practice is being performed?	Questioning, hypothesising, designing, analysing, explaining, evaluating	Task mismatch
Data and evidence	What evidence may be used?	Dataset, observations, sources, assumptions, uncertainty	Hallucination; unsupported claims
Explicit constraints	What limits and output rules apply?	Scope, format, length, safety and prohibited actions	Policy or task non-compliance
Neutrality and alternatives	What competing explanations or counter-evidence exist?	Alternative hypotheses, bias checks, falsification conditions	Confirmation bias
Citation and verification	How will claims be traced and checked?	Claim-level sources, DOI/identifier checks, source-quality rules	Fabricated or non-traceable references
Error checking and reflection	How will mistakes and limitations be detected?	Recalculation, contradiction checks, uncertainty, human review	Automation bias; overconfidence

3. Materials and Methods

3.1 Study design

This study used a structured secondary-data re-analysis with a targeted evidence update. It was designed as an original data article rather than a narrative review: the unit of analysis was the study record, the variables were explicit prompt components, and all derived results were calculated from the accompanying CSV file. Reporting was informed by transparent review and thematic-synthesis principles [14–15], but the study is intentionally described as an evidence map because the update was targeted rather than a de novo exhaustive database review.

3.2 Corpus construction

The 2025 systematic review by Lee and Palmer served as the seed source because it documented 33 papers directly concerned with teaching or operationalising prompt engineering in higher education [5]. From this seed set, 25 records were retained when they provided at least one of the following: (a) an explicit prompt structure or framework; (b) an observable prompt-quality or output-quality variable; (c) a benchmark, prototype, classroom, or administrative application; or (d) an iterative procedure that could be coded against EVIDENCE. Broad commentary with no operational element was not retained.

A targeted update searched publisher pages, ERIC-indexed material, and institutional repositories for records published after the seed review's search period and available through 15 July 2026. The search combined "prompt engineering" with "higher education", "university", "student", "scientific inquiry", "chemistry", "engineering", "AI literacy", "assessment", and "intervention". Seven additional records met the same eligibility criteria: chemistry education, undergraduate intervention, domain-knowledge prompting in chemistry and bioinformatics, a university-course framework, a business-analytics quasi-experiment, an EFL intervention, and an engineering/statistics prompt-performance study [7–9,16–19]. The final corpus contained 32 records: 25 seed records and seven targeted-update records.

The retained seed records included higher-education administration, learning analytics, medical education, assessment, writing, social-science methods, field science, information retrieval, software patterns, and chemistry applications [3,6,11–13,20–39].



Figure 2. Seed-and-update workflow used to construct the secondary dataset.

3.3 Data extraction and coding

Each row captured bibliographic information, disciplinary domain, design, empirical status, sample or data description when available, source identifier, and eight binary EVIDENCE indicators. A component was coded 1 only when the available source description operationalised the element in a prompt, workflow, assessment rule, or benchmark. General discussion without an observable implementation was coded 0. This conservative explicit-presence rule reduces interpretive inflation and makes every code auditable against the source note.

The coding unit was the published approach, not every prompt instance. For example, "Data and evidence" required a bounded dataset, observations, domain knowledge, examples, or admissible source set; merely mentioning accuracy was insufficient. "Citation and verification" required source attribution, identifier checking, or claim traceability; a generic instruction to be accurate was insufficient. "Error checking and reflection" included explicit iteration, correction, recalculation, limitation analysis, or reflective review. The codebook supplied with the CSV provides the complete operational definitions.

A coverage score was calculated as the sum of the eight binary indicators, ranging from 0 to 8. Records were also grouped as STEM/health/engineering or other higher-education domains. This broad grouping was chosen to examine the shared domain of the authors—biological science and engineering—without implying that all non-STEM fields are homogeneous.

3.4 Statistical analysis

Frequencies and percentages describe component prevalence. Wilson 95% confidence intervals were calculated for each proportion. The coverage score is summarised by median and interquartile range because the scale is bounded and the corpus is purposive. Fisher's exact tests explored whether STEM/health/engineering records more often operationalised inquiry phase, data and evidence, neutrality and alternatives, citation and verification, and error checking. Holm adjustment controlled the family-wise error rate across the five exploratory comparisons [40]. Results are interpreted descriptively; the corpus does not support causal inference or population-level prevalence claims.

3.5 Reproducibility, ethics, and AI-assisted workflow

The study used publicly available bibliographic and secondary research information and did not process identifiable participant-level data. The dataset, source URLs, feature codes, and codebook are provided as supplementary CSV files. Generative AI-assisted tools were used for preliminary bibliographic extraction, rulebook implementation, statistical scripting, figure generation, and language editing. The human authors retain responsibility for verifying every source record, code, analysis, interpretation, and citation before submission. This disclosure is consistent with the framework's core principle that AI output is an intermediate artefact subject to human audit.

4. Results

4.1 Corpus profile

The final dataset contained 32 records. Twenty-four were empirical, benchmark, prototype, case, or intervention records, while eight were conceptual, tutorial, or framework contributions. Ten records were classified as STEM/health/engineering and 22 as other higher-education domains. The corpus deliberately spans both scientific and educational information-system applications rather than treating prompt engineering as a single-discipline technique.

Table 2. Profile of the secondary evidence corpus.

Corpus characteristic	n	Interpretation
Total records	32	Study-level secondary observations
Seed-review records	25	Operational records retained from the 2025 review
Targeted-update records	7	Post-seed studies or frameworks
Empirical/benchmark/case records	24	Included observable data or implementation evidence
Conceptual/tutorial/framework records	8	Provided operational structures without original outcome data
STEM/health/engineering records	10	Biology, chemistry, medicine, engineering, computing, statistics, or field science
Other higher-education domains	22	Cross-disciplinary, administration, language, humanities, assessment, and learning design

4.2 Prevalence of EVIDENCE components

Epistemic goal and variables/context were present in all records, indicating strong agreement that effective prompts require a defined task and relevant context. Explicit constraints and error checking/reflection were each present in 29 records (90.6%). Inquiry-phase alignment and data/evidence boundaries were less common, each appearing in 20 records (62.5%). Neutrality and alternatives appeared in 18 records (56.3%). Citation and verification was the least frequent component, appearing in only 9 records (28.1%).

Table 3. Prevalence of EVIDENCE components in the 32-record corpus.

EVIDENCE component	Records	Prevalence	Wilson 95% CI
Epistemic goal	32/32	100.0%	89.3–100.0%
Variables and context	32/32	100.0%	89.3–100.0%
Inquiry phase	20/32	62.5%	45.3–77.1%

EVIDENCE component	Records	Prevalence	Wilson 95% CI
Data and evidence	20/32	62.5%	45.3–77.1%
Explicit constraints	29/32	90.6%	75.8–96.8%
Neutrality and alternatives	18/32	56.2%	39.3–71.8%
Citation and verification	9/32	28.1%	15.6–45.4%
Error checking and reflection	29/32	90.6%	75.8–96.8%

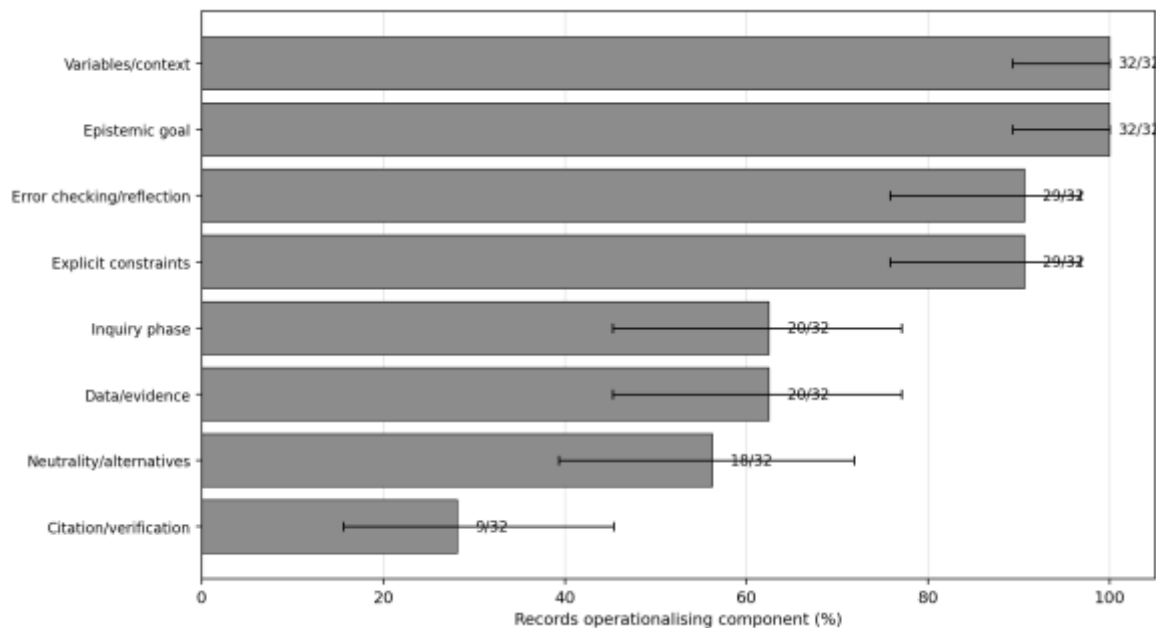


Figure 3. Component prevalence with Wilson 95% confidence intervals.

4.3 Completeness and co-occurrence

The median coverage score was 6 of 8 components (interquartile range 5–7; range 3–8). Thirteen records operationalised at least seven components, but only 4 operationalised all eight. Only 7 records combined neutrality/alternatives with citation/verification. This co-occurrence is important because an alternative explanation is not useful if its supporting evidence cannot be traced, and a cited answer can still be biased if competing mechanisms are excluded.

The most common architecture was “goal + context + constraints + iteration”. This architecture is effective for usability but incomplete for scientific inquiry because it can improve response conformity without requiring a bounded evidence base or source audit. The least common architecture was “alternatives + source verification + error audit”, which represents the strongest protection against confirmation bias and fabricated authority.

4.4 Cross-domain comparison

STEM/health/engineering records more frequently operationalised inquiry-phase alignment and data/evidence boundaries (9/10, 90.0%) than other-domain records (11/22, 50.0%). Citation and verification was present in 6/10 STEM/health/engineering records (60.0%) and 3/22 other-domain records (13.6%). The corresponding unadjusted Fisher tests were $p = 0.050$ for inquiry phase, $p = 0.050$ for data/evidence, and $p = 0.013$ for citation/verification. None remained significant after Holm adjustment, reflecting the small purposive sample. These patterns should therefore be treated as design signals rather than confirmatory domain effects.

Table 4. Exploratory cross-domain comparison of selected components.

Component	STEM/health/engineering (n=10)	Other domains (n=22)	Fisher p	Holm p
Inquiry phase	9/10 (90.0%)	11/22 (50.0%)	0.050	0.199
Data and evidence	9/10 (90.0%)	11/22 (50.0%)	0.050	0.199

Component	STEM/health/engineering (n=10)	Other domains (n=22)	Fisher p	Holm p
Neutrality and alternatives	6/10 (60.0%)	12/22 (54.5%)	1.000	1.000
Citation and verification	6/10 (60.0%)	3/22 (13.6%)	0.013	0.063
Error checking and reflection	9/10 (90.0%)	20/22 (90.9%)	1.000	1.000

4.5 Derived EVIDENCE-Prompt schema

The secondary-data gaps were converted into a prompt schema that makes every component observable. The template is modular: users may shorten it for low-risk tasks, but high-stakes scientific work should retain all eight fields. The template deliberately avoids a generic role instruction as its organising principle. Instead, it starts with the epistemic claim and evidence boundary, because these determine whether the resulting output is scientifically auditable.

Box 1. Generic EVIDENCE-Prompt template.

[EPISTEMIC GOAL] Examine the following scientific question or claim: ____ . The required reasoning is ____ and the decision or conclusion to be supported is ____ .

[VARIABLES AND CONTEXT] Define the population/system, variables, units, boundary conditions, prior knowledge, and disciplinary assumptions: ____ .

[INQUIRY PHASE] This task concerns: questioning / hypothesising / designing / analysing / explaining / evaluating. Do not silently move to another phase.

[DATA AND EVIDENCE] Use only the following data, observations, or verified sources: ____ . Separate observations, assumptions, calculations, and external knowledge.

[EXPLICIT CONSTRAINTS] Do not invent data, quotations, citations, measurements, or experimental results. Follow this scope, format, safety rule, and citation style: ____ .

[NEUTRALITY AND ALTERNATIVES] Present at least two plausible explanations or designs, relevant counter-evidence, and conditions that would weaken each option.

[CITATION AND VERIFICATION] Attach a traceable source to each externally checkable claim. Verify DOI/PMID/accession identifiers and label any unverified item.

[ERROR CHECKING AND REFLECTION] Recheck calculations and units; identify contradictions, uncertainty, limitations, likely hallucinations, and the human checks required before acceptance.

4.6 Cross-disciplinary scientific exemplars

Table 5 demonstrates the common thematic ground between biological science and engineering. The examples use the same eight controls while changing the domain evidence, units, and failure modes. They are not model outputs or experimental results; they are implementation exemplars derived from the secondary synthesis.

Table 5. Cross-disciplinary implementation exemplars for biological science and engineering.

Domain	Epistemic goal	Variables/context	Evidence boundary	Alternatives and verification
Biological science	Interpret a cell-viability dose-response dataset for a plant extract	Cell line, concentration (µg/mL), exposure time, controls, replicate structure, viability assay	Raw anonymised measurements and verified assay protocol only	Compare cytotoxicity, assay interference, and batch effects; require DOI checks and prohibit invented IC50 values
Engineering / IoT	Diagnose elevated indoor CO ₂ and particulate matter from multi-zone sensor logs	Sensor model, calibration, units, timestamps, occupancy, ventilation state, missingness	Uploaded time-series data, calibration records, and cited standards only	Compare occupancy, ventilation, outdoor infiltration, and sensor drift; recalculate exceedances and flag unverified thresholds

5. Discussion

5.1 Principal findings

This secondary-data study shows that the prompt-engineering literature has converged on goal definition, context, constraints, and iterative improvement. These are necessary but not sufficient conditions for scientific inquiry. The weakest component was citation and verification, present in fewer than one third of records. This gap is consequential because scientific claims often enter a workflow through references, identifiers, numerical thresholds, or extracted data. Without traceability, a polished output cannot be reliably audited.

The second gap concerns the joint use of alternatives and verification. Just seven records combined both. Scientific reasoning requires more than checking whether a preferred explanation is supported; it requires considering what else could explain the observations and what evidence would discriminate among alternatives. EVIDENCE-Prompt therefore treats neutrality as an operational field rather than a general ethical aspiration.

The median coverage score of six indicates that the literature already contains most of the building blocks. The contribution of EVIDENCE-Prompt is not to claim that each element is new in isolation. Its novelty is architectural: it arranges the elements as an auditable information chain and makes missing controls visible. This distinguishes it from frameworks intended primarily to improve relevance, persona, style, or output format.

5.2 Contribution to computer information systems and industrial management

The framework fits computer information systems because it specifies data provenance, input constraints, model metadata, output traceability, and human review. A university can implement it as a lightweight governance layer around general-purpose LLMs: prompt templates become controlled input forms; model and date are logged; citations and identifiers are checked; and the final human revision is retained. This creates an auditable record suitable for quality assurance, curriculum management, research governance, and regulated industrial training.

For industrial management, the framework supports decisions in which AI-generated recommendations must be connected to operational evidence. Examples include quality control, sensor analytics, workforce classification, maintenance planning, and laboratory process design. The corpus already contains applications in statistical process control, learning analytics, job classification, chemistry text mining, and field-course design [23,31,35–36,39]. EVIDENCE-Prompt provides a common governance vocabulary across these contexts.

5.3 Educational and assessment implications

Teaching should move beyond handing students a template. Learners need worked examples showing why each field changes the validity of an answer, including deliberately flawed outputs with plausible prose, incorrect units, fabricated citations, and omitted alternatives. Prompt assessment should be paired with an evidence log and a revision statement. A high score should require not only a detailed prompt but also evidence that the learner checked the output and changed the final explanation when necessary.

A practical rubric can assign 0 = absent, 1 = present but generic, and 2 = specific and auditable for each EVIDENCE component, producing a 0–16 prompt-quality score. Separate outcome criteria should assess factual accuracy, claim–evidence alignment, uncertainty calibration, source traceability, and correction of model errors. This separation prevents a well-structured prompt from receiving full credit when the learner accepts an incorrect answer.

Biological-science teaching should emphasise experimental controls, units, assay limitations, and the distinction between observed association and mechanism. Engineering teaching should emphasise calibration, data quality, boundary conditions, reproducible calculations, and standards. The shared EVIDENCE structure allows interdisciplinary instruction while preserving these domain-specific requirements.

5.4 Governance and implementation requirements

- Prompt and output logging: retain the exact prompt, model name/version, access date, uploaded evidence, output, and final human revision.
- Data protection: prohibit confidential, identifiable, proprietary, or ethically restricted data in unapproved systems.
- Source integrity: require users to open cited sources and verify persistent identifiers rather than accepting generated reference strings.
- Assessment design: grade question quality, evidence use, alternatives, verification, and revision—not only the final prose.
- Model change control: repeat benchmark tasks after major model updates and record configuration changes.

- Equity: provide institutionally supported access or equivalent non-AI pathways so assessment does not reward paid-tool advantage.
- Human accountability: state explicitly that the learner, researcher, or decision maker remains responsible for methods, calculations, interpretation, and citations.

5.5 Limitations

First, the targeted update was not a fully exhaustive database search, so the dataset should be interpreted as a structured evidence map rather than a prevalence estimate for all prompt-engineering research. Second, codes were based on explicit information available in source descriptions and publications; a study may have implemented an element without reporting it. Third, binary indicators simplify differences in depth and quality. The proposed 0–2 rubric should be tested in future work. Fourth, the cross-domain comparison used broad groups and small cells; adjusted results were non-significant. Fifth, model behaviour changes over time, so any prompt framework requires version-controlled benchmarking. Finally, this study evaluates reported prompt architectures, not direct student learning or model accuracy. Controlled multi-model and classroom validation remain necessary.

5.6 Future validation agenda

A next-stage validation should separate three claims. Content validity should be evaluated by experts in science education, biology, engineering, AI/NLP, assessment, and research governance. Technical utility should be tested by running matched minimal, conventional, and EVIDENCE prompts across multiple LLM families and repeated trials. Educational effectiveness should be assessed through a controlled pretest–posttest design in which students submit a question, prompt log, model output, verification record, revised explanation, and human-contribution statement. Outcome measures should include scientific reasoning, evidence traceability, critical-error rate, calibrated reliance, cognitive load, and transfer to an unfamiliar inquiry task. Instrument development should follow established guidance for observable constructs and validity evidence [41].

6. Conclusions

EVIDENCE-Prompt reframes prompt engineering as an auditable information-system process for scientific inquiry. A secondary re-analysis of 32 higher-education records found strong coverage of goals, context, constraints, and iteration, but substantially weaker operationalisation of source verification and competing explanations. The framework addresses these gaps through eight observable fields that connect the scientific question, admissible evidence, model response, and human review. Its cross-disciplinary structure supports both biological-science and engineering applications while preserving domain-specific data and validation requirements. The accompanying CSV dataset and codebook make the evidence map reproducible. EVIDENCE-Prompt should now be tested through expert content validation, multi-model benchmarking, and controlled classroom studies before claims of educational effectiveness are made.

Acknowledgments: The authors acknowledge the institutions supporting their interdisciplinary work in biological science, engineering, and higher-education innovation.

Funding: This research received no external funding.

Author Contributions: Rissa A. Lasap: Conceptualization, biological-science domain interpretation, framework refinement, writing-review and editing, and supervision. Luigi Carlo M. De Jesus: Methodology, data curation, formal analysis, visualization, software, and writing-original draft. Both authors: Validation, interpretation, final approval, and accountability for the work.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement: The study-level secondary dataset, codebook, and reproducible analysis script are supplied as supplementary files: EVIDENCE_secondary_evidence_dataset.csv, EVIDENCE_secondary_evidence_codebook.csv, and EVIDENCE_secondary_analysis.py. Each dataset record includes a DOI or stable source URL.

Institutional Review Board Statement: Not applicable. The study used publicly available bibliographic and secondary research information and did not involve identifiable human participants, animals, or confidential records.

Informed Consent Statement: Not applicable.

Artificial Intelligence Use Disclosure: Generative AI-assisted tools were used for preliminary bibliographic extraction, rulebook implementation, statistical scripting, figure generation, and language editing. The human authors retain responsibility for verifying all source records, analyses, interpretations, and references and for approving the final submitted version.

References

1. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
2. 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, 38. <https://doi.org/10.1186/s41239-023-00408-3>
3. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv:2302.11382*. <https://doi.org/10.48550/arXiv.2302.11382>
4. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
5. Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22, 7. <https://doi.org/10.1186/s41239-025-00503-7>
6. Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225. <https://doi.org/10.1016/j.caeai.2024.100225>
7. Woo, D. J., Wang, D., Yung, T., & Guo, K. (2025). Effects of a prompt engineering intervention on undergraduate students' AI self-efficacy, AI knowledge and prompt engineering ability: A mixed methods study. *British Educational Research Journal*. <https://doi.org/10.1002/berj.70087>
8. Tassoti, S. (2024). Assessment of students' use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, 101(6), 2475-2482. <https://doi.org/10.1021/acs.jchemed.4c00212>
9. Liu, H., Yin, H., Luo, Z., & Wang, X. (2025). Integrating chemistry knowledge in large language models via prompt engineering. *Synthetic and Systems Biotechnology*, 10. <https://doi.org/10.1016/j.synbio.2024.07.004>
10. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
11. Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4), 102720. <https://doi.org/10.1016/j.acalib.2023.102720>
12. Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, 2(3), 198-205. <https://doi.org/10.3390/ime2030019>
13. Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, 25, e50638. <https://doi.org/10.2196/50638>
14. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
15. Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 45. <https://doi.org/10.1186/1471-2288-8-45>

16. Murray, M. C. (2025). Bridging the generative AI literacy gap: A guide to introducing prompt engineering in university courses. *Issues in Informing Science and Information Technology*, 22, 010. <https://doi.org/10.28945/5516>
17. Gastineau, J. (2024). Exploring the impact of generative AI prompt engineering in higher education: A study in undergraduate and graduate business analytics courses [Doctoral dissertation, University of Arkansas]. <https://scholarworks.uark.edu/etd/5551>
18. Preschern, J., Hunter, M., Born-Lechleitner, I., & Kobylak, C. (2025). Developing prompt engineering as a 21st-century skill: The impact of structured ChatGPT instruction in EFL education. *GILE Journal of Skills Development*, 5(3), 87-108. <https://doi.org/10.52398/gjsd.2025.v5.i3.pp87-108>
19. Sawalha, G., Taj, I., & Shoufan, A. (2024). Analyzing student prompts and their effect on ChatGPT's performance. *Cogent Education*, 11(1), 2397200. <https://doi.org/10.1080/2331186X.2024.2397200>
20. Abdulshahed, A. M. (2023). A novel framework leveraging prompt engineering and the grey-based approach: A case study in Libya. SSRN. <https://doi.org/10.2139/ssrn.4492606>
21. Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice*, 9(2), 135-214. <https://doi.org/10.1080/23735082.2023.2258886>
22. Cain, W. (2023). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*. <https://doi.org/10.1007/s11528-023-00896-0>
23. Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023). Large language models in the workplace: A case study on prompt engineering for job type classification. In *Natural Language Processing and Information Systems (LNCS 13913, pp. 3-17)*. Springer. https://doi.org/10.1007/978-3-031-35320-8_1
24. Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, 20(5), 02. <https://doi.org/10.53761/1.20.5.02>
25. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and English-as-an-additional-language student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>
26. Fotaris, P., Mastoras, T., & Lameris, P. (2023). Designing educational escape rooms with generative AI: A framework and ChatGPT prompt engineering guide. In *Proceedings of the 17th European Conference on Games Based Learning (pp. 180-189)*. <https://doi.org/10.34190/ecgbl.17.1.1870>
27. Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & Society*. <https://doi.org/10.1007/s00146-023-01752-8>
28. Jacobsen, L. J., & Weber, K. E. (2023). The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback. *OSF Preprints*. <https://doi.org/10.31219/osf.io/cr257>
29. Korzynski, P., Mazurek, G., Krzykowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3), 25-37. <https://doi.org/10.15678/EBER.2023.110302>
30. Mabrito, M. (2024). Artificial intelligence in the classroom: Conversation design and prompt engineering for English majors. *The International Journal of Technologies in Learning*, 31(2), 129-142. <https://doi.org/10.18848/2327-0144/CGP/v31i02/129-142>
31. Megahed, F. M., Chen, Y. J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2024). How generative AI models such as ChatGPT can be (mis)used in statistical process control practice, education, and research? An exploratory study. *Quality Engineering*, 36(2), 287-315. <https://doi.org/10.1080/08982112.2023.2206479>
32. Raftery, D. (2023). Will ChatGPT pass the online quizzes? Adapting an assessment strategy in the age of generative AI. *Irish Journal of Technology Enhanced Learning*, 7(1). <https://doi.org/10.22554/ijtel.v7i1.114>
33. Schäffer, B., & Lieder, F. R. (2023). Distributed interpretation: Teaching reconstructive methods in the social sciences supported by artificial intelligence. *Journal of Research on Technology in Education*, 55(1), 111-124. <https://doi.org/10.1080/15391523.2022.2148786>
34. Spasić, A. J., & Janković, D. S. (2023). Using ChatGPT standard prompt engineering techniques in lesson preparation: Role, instructions and seed-word prompts. In *Proceedings of the 58th International*

- Scientific Conference on Information, Communication and Energy Systems and Technologies (pp. 47-50). <https://doi.org/10.1109/ICEST58410.2023.10187269>
35. Susnjak, T. (2023). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-023-00336-3>
 36. Tupper, M., Hendy, I. W., & Shipway, J. R. (2023). Field courses for dummies: Can ChatGPT design a higher education field course? *EdArXiv*. <https://doi.org/10.35542/osf.io/b65nu>
 37. Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21, 15. <https://doi.org/10.1186/s41239-024-00448-3>
 38. Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2023). Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*. <https://doi.org/10.1109/TLT.2023.3324714>
 39. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T., & Yaghi, O. M. (2023). ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society*, 145(32), 18048-18062. <https://doi.org/10.1021/jacs.3c05819>
 40. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
 41. Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, 36(6), 463-474. <https://doi.org/10.3109/0142159X.2014.889814>