Proceedings Series of Educational Studies

The 5th International Conference on Educational Management and Technology (ICEMT) 2026

Interrogating the Machine: Integrating Hercule Poirot's "Little Grey Cells" Deductive Framework into AI Literacy and Critical Prompt Engineering in Education

Fauziah Nur Aisyah Rosyidah

State University of Malang, Jl. Semarang No. 5 Malang, East Java, Indonesia

Email: fauziah.nur.2303336@students.um.ac.id

Abstract

Generative artificial intelligence has entered classrooms faster than the verification habits required to use it safely. This study examines the mismatch between adoption, model reliability, and student verification behaviour, and derives a cognitive scaffold from an unexpected source: the deductive procedure of the fictional detective Hercule Poirot. A systematic literature review of studies published between 2021 and 2026 was combined with descriptive time-series analysis of sixty-one months of global relative search-volume data for four query strings. The search data show that interest in operating generative systems reached the normalised ceiling of the index, whereas interest in governing them within schools remained at the measurement floor, returning zero in fifty-five of sixty-one months. The reviewed literature converges on three findings: high adoption among students, persistent model hallucination on complex academic queries, and low rates of systematic cross-checking against primary sources. From this convergence the study derives a four-stage framework detailed observation, contradiction identification, alibi verification, and logical conclusion that repositions the chatbot as a witness under examination rather than an authority. The framework is conceptual and awaits empirical validation.

Keywords: AI literacy; prompt engineering; critical thinking; cognitive scaffolding; systematic literature review

1. Introduction

Ask a classroom of teenagers how they know a chatbot is telling the truth, and the silence that follows is instructive. That silence now sits inside a question drawing attention at a scale few pedagogical concerns ever reach. Measures of global digital attention show interest in AI literacy and prompt engineering climbing sharply over recent years, spiking rather than drifting the signature of a problem that has outrun the institutions meant to handle it. Schools are adopting generative systems faster than anyone is teaching students to doubt them (Chan & Hu, 2023; EDUCAUSE, 2024; Kasneci et al., 2023). What is missing is not enthusiasm but method.

Three quantities describe the problem, and they are meaningful only in collision. Surveys of higher education report that a large majority of students in the order of 75% to 80% now use generative systems routinely for academic work (Chan & Hu, 2023; EDUCAUSE, 2024). Technical evaluations continue to record hallucination on complex academic queries, with reported rates commonly falling between roughly 15% and 27% depending on task type and model, including the fabrication of plausible but non-existent citations (Bommasani et al., 2021; Ji et al., 2023). Digital information-literacy surveys, meanwhile, indicate that fewer than one learner in five systematically cross-checks machine output against primary sources (Wineburg & McGrew, 2019). Exposure alone would be unremarkable. An error rate alone

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

would be a technical limitation. A verification deficit alone would be a familiar information-literacy finding. Taken together they describe a population routinely receiving unreliable output and structurally unequipped to detect it.

The mechanism behind that deficit is not a lack of access to verification procedures but a misclassification of the tool's epistemic status. Students frequently treat generative systems as an oracle that cannot be wrong rather than as probabilistic computational instruments whose fluency is uncorrelated with accuracy. Ji et al. (2023) locate hallucination in the architecture of natural-language generation itself, not in a defect awaiting a patch. If that characterisation holds, then institutional responses built on detection tooling and prohibition are permanently reactive, and the burden of validation necessarily shifts toward the reader.

Patterns in public search behaviour sharpen the point. Over the sixty-one months examined in this study, queries concerned with operating generative systems accelerated to the ceiling of the relative interest index, while queries framing those systems as an institutional problem of school policy collapsed toward the measurement floor. Enthusiasm for the technology and unease about its output are not sequential phenomena but simultaneous ones, and the institutional conversation about governance has not kept pace with either.

Existing scholarship supplies parts of a response but not the whole. Conceptual work on AI literacy has produced competency taxonomies that specify what learners should be able to do (Long & Magerko, 2020; Ng et al., 2021), yet these frameworks describe outcomes more than procedures. Prompt engineering, as currently taught, is largely syntactic: it instructs learners in how to command the machine and rarely in how to suspect it. Established verification heuristics such as lateral reading presume an external corpus against which a claim can be checked (Wineburg & McGrew, 2019; Caulfield, 2019), an assumption that fails precisely where the machine has fabricated a source with no external referent. A procedural scaffold for interrogating machine testimony from the inside is therefore absent from the instructional-design literature.

This article proposes that such a scaffold can be adapted from detective method. The investigative logic attributed to Hercule Poirot the disciplined use of what Christie's character calls the “little grey cells” proceeds through a fixed sequence: patient observation of small detail, cross-examination of testimony against alibi, deliberate hunting for one's own bias, and only then conclusion (Christie, 1920). Read as cognitive scaffolding in the sense established by Wood, Bruner, and Ross (1976) rather than as literary ornament, that sequence supplies a usable grammar for repositioning the chatbot from teacher or sacred text to witness or suspect under examination. The present study accordingly pursues three objectives: to document the empirical mismatch between adoption, reliability, and verification; to establish, through systematic review, whether an equivalent procedural scaffold already exists; and to formulate a four-stage framework that converts learners from passive consumers of machine output into digital investigators capable of interrogating and validating what an artificial system hands them.

2. Method

This study is a systematic literature review supplemented by descriptive analysis of secondary search-interest data. It is non-experimental and produces a conceptual framework rather than an effect estimate; no human participants were involved and no ethical clearance was required.

2.1. Research Design

The review followed the PRISMA 2020 reporting guideline (Page et al., 2021). Two strands of evidence were assembled in parallel. The first was a body of peer-reviewed literature on AI literacy, prompt engineering, and the critical evaluation of machine output, used to establish the state of the art and to test whether a procedural verification scaffold already exists. The second was a sixty-one-month time series of global relative search interest, used as a descriptive indicator of the distribution of public and professional attention. The framework reported in Section 3 was derived from the thematic synthesis of the first strand and calibrated against the deficits documented in both.

2.2. Data Sources and Search Strategy

Four databases were searched: Scopus, ERIC, Web of Science, and Google Scholar. The last was used only to identify records missed by the other three, since its coverage is not reproducible. Searching Scopus alone was avoided deliberately, because that database under-covers educational technology reviews and single-database reviews are a recurring source of reviewer objection.

Boolean strings combined three concept blocks with the AND operator, with synonyms nested inside each block using OR. The core Scopus string was: TITLE-ABS-KEY ((“AI literacy” OR “artificial intelligence literacy” OR “algorithmic literacy” OR “prompt engineering” OR “prompt literacy”) AND (“critical thinking” OR “epistemic vigilance” OR “automation bias” OR overreliance OR “source evaluation” OR hallucinat*) AND (ChatGPT OR “generative AI” OR “large language model*” OR chatbot*) AND (educat* OR student* OR school* OR pedagog*)). A second string targeted the narrative-heuristic strand: TITLE-ABS-KEY ((“detective fiction” OR “mystery-based learning” OR “abductive reasoning” OR “case-based reasoning”) AND (pedagog* OR instruction* OR scaffold*)). Searches were run in May and June 2026 and restricted to English-language journal articles and reviews.

2.3. Inclusion and Exclusion Criteria

Records were included if they were peer-reviewed, published between 2021 and 2026, written in English, situated in an educational setting from K–12 to higher education, and reported either empirical data or a conceptual framework bearing on AI literacy, prompt engineering, or the evaluation of machine-generated output. Two categories of exception were admitted outside the date window: foundational theoretical sources on scaffolding and cognitive load, and the primary literary source for the deductive sequence under adaptation.

Records were excluded if they addressed artificial intelligence in non-educational applied domains such as medical imaging or drug discovery, if they treated generative systems purely as an assessment-integrity problem without reference to learner cognition, or if they were editorials, conference abstracts, or preprints without peer review.

2.4. Secondary Search-Interest Data

Relative search-volume data for four query strings "AI literacy", "prompt engineering", "ChatGPT in schools", and "critical thinking skills" were extracted from Google Trends for the worldwide region in English, at monthly resolution, covering August 2021 to August 2026 (n = 61 monthly observations). The four strings were selected to represent, respectively, the core construct, the instructional method, the institutional-governance framing, and a long-established pedagogical baseline against which novelty could be judged.

Two properties of this index constrain its interpretation and are stated here rather than in the limitations alone. The index is relative and ordinally normalised within the comparison set, so a value of 100 denotes the highest observed volume for any series in any month rather than an absolute query count; means computed across such a series are descriptive summaries only. In addition, the terminal month of any export may be incomplete at the time of extraction, which is a common cause of apparent drop-off in the final observation.

2.5. Data Analysis

Retrieved records were screened by title and abstract, then at full text, and coded against a framework comprising the three constructs of the review: AI literacy, critical evaluation of machine output, and prompt engineering as instructional method. Coding was iterative, with categories refined across successive passes, and synthesis proceeded thematically rather than statistically because the retrieved designs were too heterogeneous to pool.

The search-interest series were analysed descriptively: arithmetic mean, standard deviation, maximum and its month of occurrence, and the count of months at zero were computed for each series, and Pearson product-moment correlations were computed between series. Trajectory labels supplied with the export were verified against the underlying monthly values before any directional claim was made, since automated trajectory fields do not always agree with the series they summarise. Framework derivation proceeded by mapping each documented deficit onto the investigative stage that addresses it, so that every stage of the proposed scaffold corresponds to a specific finding rather than to the structure of the source narrative.

3. Results and Discussion

3.1 Result

Descriptive statistics for the four query strings are reported in Table 1, and the monthly series are plotted in Figure 1. The four behave in markedly different registers.

"Prompt engineering" carries the highest mean relative search volume of the set (M = 10.69, SD = 18.34). The series held a comparatively stable band between 31 and 36 through

the second quarter of 2026 before an abrupt acceleration in August 2026 drove it to the normalised ceiling of 100. This was not, however, the only discontinuity: the series had already risen from 24 to 87 between January and February 2026, a proportionally larger expansion (3.6-fold) than the July-to-August rise (2.9-fold). The distribution is therefore bimodal in its peaks, and the large standard deviation relative to the mean reflects a series that is stable at a low level for most of the window and extreme at two points within it.

"AI literacy" describes a shallower arc (M = 3.02, SD = 5.43), climbing from a negligible base to a plateau of 17 to 18 across April, May, and June 2026. Its maximum of 18 occurs three times in November 2025, May 2026, and June 2026 rather than at a single peak. "Critical thinking skills" is the most stable series in the set (M = 5.05, SD = 2.21), oscillating within a narrow band for four years before peaking at 13 in April 2026.

Two observations require care. First, both "AI literacy" and "critical thinking skills" decline across the final three observations from 18 to 14 to 7, and from 9 to 6 to 4 respectively notwithstanding trajectory labels in the source export that describe them as surging. The declines are reported here as recorded. Second, because the terminal month may be incomplete at extraction, the final observation in each series should be read as provisional; this qualification applies equally to the August 2026 ceiling in the prompt-engineering series.

The analytically significant finding is the outlier in the negative. "ChatGPT in schools" registers a mean of 0.30, never exceeds 5, and returns zero in fifty-five of sixty-one months, including four of the five most recent. Interest in operating generative systems has surged to saturation while interest in governing them within schools has effectively vanished from public query behaviour. Correlations across the series are consistent with this divergence: "AI literacy" correlates moderately with "prompt engineering" (r = .69) and with "critical thinking skills" (r = .72), whereas "ChatGPT in schools" correlates only weakly with "prompt engineering" (r = .22). Attention to capability and attention to institutional governance are moving on separate tracks.

Table 1. Descriptive statistics of relative search volume for four query strings, August 2021–August 2026 (n = 61 monthly observations)

Query string	M	SD	Max	Month of maximum	Months at zero
Prompt engineering	10.69	18.34	100	August 2026	19
Critical thinking skills	5.05	2.21	13	April 2026	1
AI literacy	3.02	5.43	18	November 2025[a]	38
ChatGPT in schools	0.30	0.97	5	December 2025	55

Note. Values are relative search volume indices on a 0–100 scale, normalised across the full observation window such that 100 denotes the highest observed volume for any series in any month. The index is ordinal within a comparison set and does not represent absolute query counts. [a]The value 18 was attained on three occasions (November 2025, May 2026, and June 2026); the earliest is reported.

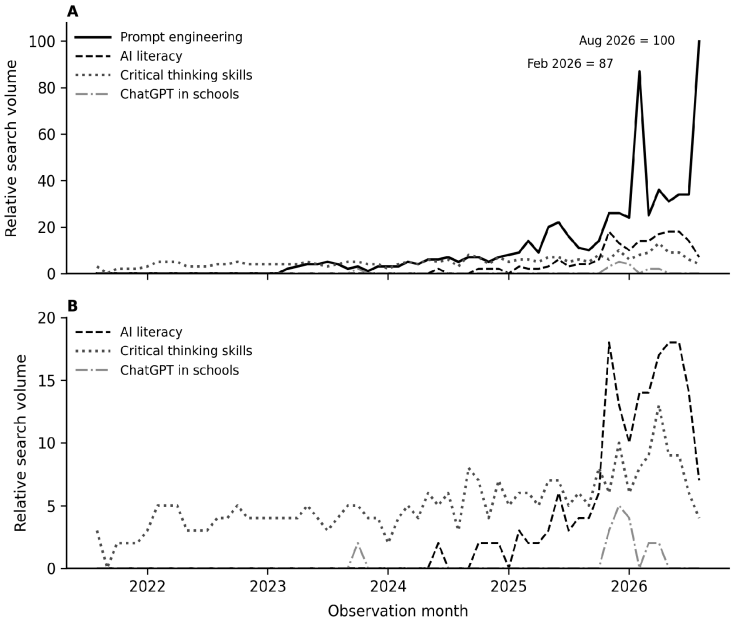


Figure 1. Monthly relative search volume for four query strings, August 2021–August 2026. Panel A shows all four series at full scale; Panel B restricts the ordinate to 0–20 to resolve the three lower-volume series.

The reviewed literature converges on three findings that jointly define the problem space. First, adoption is near-universal in higher education, with reported routine use of generative systems for academic work in the order of 75% to 80% (Chan & Hu, 2023; EDUCAUSE, 2024). Second, hallucination persists on complex academic queries at rates commonly reported between roughly 15% and 27%, and is characterised in the technical literature as a property of the generation architecture rather than a remediable defect (Bommasani et al., 2021; Ji et al., 2023). Third, systematic cross-checking against primary sources is performed by fewer than one learner in five (Wineburg & McGrew, 2019).

The narrative-heuristic search returned a thin and scattered set of records, with almost no peer-reviewed work applying detective or abductive reasoning procedures specifically to the evaluation of machine-generated text. That thinness is itself a review finding: it documents the absence this article addresses, and it is the reason the framework below is adapted from a primary literary source and from established scaffolding theory rather than derived from an existing validated instructional model.

Table 2 sets out the resulting four-stage framework. Each stage is paired with the specific deficit it targets, so that the correspondence between the scaffold and the evidence is explicit rather than metaphorical.

Table 2. Alignment of the four investigative stages with documented deficits

Stage	Investigative operation	Targeted deficit	Empirical anchor
1. Detailed observation	Read the output slowly and sceptically; note specific claims, named sources, and numerical values as separable items rather than as a single fluent whole.	Epistemic misclassification of the system as an authority whose fluency implies accuracy.	Oracle framing reported in adoption studies (Chan & Hu, 2023).

Stage	Investigative operation	Targeted deficit	Empirical anchor
2. Contradiction identification	Cross-examine the testimony against itself: locate internal inconsistency, unsupported inference, and bias in the framing of the answer.	Absence of any check that functions when no external referent exists to verify against.	Fabricated citations and invented findings (Ji et al., 2023).
3. Alibi verification	Trace each claim to its original source and confirm that the source exists, is correctly attributed, and states what is claimed.	Cross-checking against primary sources performed by fewer than one learner in five.	Verification rates in information-literacy surveys (Wineburg & McGrew, 2019).
4. Logical conclusion	Reassemble the surviving fragments into a validated claim, stating explicitly what remains unverified.	Uncritical acceptance of the whole output when only part of it has been checked.	Overreliance and cognitive offloading (Kasneci et al., 2023).

Note. Stage numbering follows the investigative sequence and does not imply that stages are performed only once; stages 2 and 3 are expected to iterate in practice.

3.2 Discussion

The mismatch docug and the illusion of understanding (Kasneci et al., 2023; Zawacki-Richter et al., 2019). A scaffold that adds four steps to an interaction whose principal appeal is the elimination of steps runs directly against that current. There is also a more basic objection internal to the present data: the verification gap has persisted despite the wide availability of digital and information-literacy instruction, which supports a null hypothesis that heuristic frameworks in general are weak instruments in this setting rather than that the wrong heuristic has been selected. Cognitive load theory offers a partial answer that scaffolds reduce extraneous load when they are proceduralised and faded rather than added as a permanent checklist (Sweller et al., 2019; Wood et al., 1976) but this remains a design commitment, not a demonstrated rmented in Section 3.1 is not a gap between two independent trends but a single structural condition. Near-universal exposure intersects a non-trivial error rate at a point where the corrective behaviour is almost absent, and the search-interest data show that institutional attention has moved away from precisely the framing under which that condition would be addressed. The finding that Ji et al. (2023) locate hallucination in the architecture of natural-language generation carries the interpretive weight here. If the phenomenon is architectural, then detection tooling and prohibition are structurally incapable of resolving it, and validation must be internalised by the reader. This reframes the field's dominant institutional response as permanently reactive.

It follows that the prevailing approach to prompt engineering is pedagogically incomplete. Current instruction is preoccupied with syntax how to phrase, constrain, and chain instructions to the machine and largely silent on stance. Yet the deficit identified in the literature is not procedural ignorance but epistemic misclassification: learners who treat the system as an authority have no occasion to apply a verification procedure however well they can phrase a prompt. Additional syntactic training cannot repair a mis-set stance, which is why competency taxonomies of AI literacy (Long & Magerko, 2020; Ng et al., 2021), valuable as descriptions of desired outcomes, do not by themselves close the gap.

The proposed framework must be positioned against the established alternatives rather than offered in isolation. Lateral reading and the SIFT procedure are the closest rivals, and both are supported by empirical work that the present framework is not (Caulfield, 2019; Wineburg & McGrew, 2019). The distinction is one of applicability rather than superiority.

Lateral reading directs the reader away from a claim and toward an external corpus in which its provenance can be checked; it presumes that such a corpus exists. Fabricated citations and invented findings have no external referent, so the lateral move returns nothing and the absence of a result is easily misread as a failure of search rather than as evidence of fabrication. Stage two of the proposed scaffold the identification of internal contradiction within the machine's testimony is an operation lateral reading does not contain, and it is diagnostic precisely where the external check is unavailable. The two are therefore best treated as complementary, with the investigative sequence supplying an internal consistency check that precedes and triggers the lateral one.

Counter-evidence deserves explicit statement. The overreliance literature suggests that the introduction of capable systems tends to reduce rather than increase deliberate cognitive effort, through cognitive offloadinesult.

A structural critique cuts deeper still. Selwyn (2023) argues that individual-competence framings displace responsibility for structural technological problems onto learners. Applied here, training students to interrogate hallucinating systems may function as a substitute for demanding more reliable systems and better institutional policy an implication given uncomfortable support by the collapse of the governance-framed query series documented in Section 3.1. The honest position is that a cognitive scaffold is a mitigation and not a remedy. It equips the reader without excusing the developer or the institution, and it should be advanced alongside, not instead of, arguments for reliability standards and procurement scrutiny.

Several limitations follow from the design. The review is conceptual and produces no effect estimate; the efficacy of the framework is an empirical question that quasi-experimental or design-based research must settle. The empirical anchors drawn from the literature are reported as ranges rather than point estimates because hallucination rates in particular vary substantially with task type, model, and evaluation protocol. The search-interest data are relative and ordinally normalised, and are used here only as an indicator of the distribution of attention, not as a measure of behaviour or capability. Finally, the adaptation of a literary heuristic carries an inherent risk of over-extension: the framework claims only that a fixed investigative sequence supplies a memorable and teachable grammar for verification, not that fictional detection is a model of scientific reasoning.

4. Conclusion

This study set out to document the mismatch between the adoption of generative systems, their measured reliability, and the verification behaviour of learners, and to derive a procedural scaffold in response. The search-interest analysis shows attention to capability rising to the ceiling of the index while attention to institutional governance in schools has fallen to the measurement floor, absent from public query behaviour in fifty-five of sixty-one months. The review shows near-universal adoption alongside a persistent, architecturally grounded error rate and a verification rate below one learner in five, and it shows that no established procedural scaffold addresses the interrogation of machine testimony from the inside.

The contribution offered here is not the detective metaphor but the argument the metaphor operationalises: that verification must be taught as a stance toward testimony rather than as a syntax for issuing instructions, and that the stage which locates contradiction within an output is indispensable where no external referent exists to check it against. The four-stage framework detailed observation, contradiction identification, alibi verification, and logical conclusion maps each of its steps onto a documented deficit and is offered as a conceptual scaffold for instructional design.

Three lines of work follow. The framework requires operationalisation into a rubric or scale with reported reliability evidence before it can be tested. It then requires quasi-experimental or design-based evaluation against an active comparison, most appropriately lateral reading, with outcomes measured on learners' detection of fabricated citations rather than on self-reported confidence. And it requires implementation research on fading, since a scaffold that remains a permanent checklist will be abandoned by the learners who most need it. Until such evidence exists, the framework should be read as a proposal for testing rather than as a validated instrument, and as a mitigation that leaves the case for more reliable systems and stronger institutional policy entirely intact.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258

Caulfield, M. (2019). *SIFT (The four moves)*. Hapgood. https://hapgood.us/2019/06/19/sift-the-four-moves/

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Christie, A. (1920). *The mysterious affair at Styles*. London, England: John Lane.

EDUCAUSE. (2024). *2024 EDUCAUSE horizon report: Teaching and learning edition*. Boulder, CO: EDUCAUSE.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … 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

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). New York, NY: ACM. https://doi.org/10.1145/3313831.3376727

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Selwyn, N. (2023). Digital degrowth: Toward radically sustainable education technology. *Learning, Media and Technology*. https://doi.org/10.1080/17439884.2022.2159978

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record, 121*(11), 1–40. https://doi.org/10.1177/016146811912101102

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0