ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI Computing

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0001

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

Designing a Scenario-Based AI Literacy Toolkit for Defense Higher Education

Jawahir Che Mustapha1,*, Nurul Akhmal Mohd Zulkefli2 and Mohd Nazri Ismail3

1 Malaysian Institute of Information Technology, Universiti Kuala Lumpur, 1016, Jalan Sultan Ismail, Kuala Lumpur, 50250, Malaysia

2 Dhofar University, Salalah, Oman

3 National Defence University of Malaysia

Corresponding author: Jawahir Che Mustapha (jawahir@unikl.edu.my) · ORCID: [0000-0000-0000-0000]

Author emails: Nurul Akhmal Mohd Zulkefli (nzulkefli@du.edu.om); Mohd Nazri Ismail (m.nazri@upnm.edu.my)

ABSTRACT

Artificial intelligence (AI) is increasingly affecting higher education and professional decisions, thus calling for a broader form of AI literacy that includes not only technological expertise but critical analysis, human agency, verification, calibrated reliance, and ethical responsibility. To address this need, this study aims to develop a Scenario-Based AI Literacy Toolkit for Defense Higher Education, especially taking into account the demands of high-stakes learning and Professional Military Education (PME). An evidence-informed, staged approach was selected in which preliminary evidence synthesis was combined with a prospective validation protocol. The preliminary search revealed 515 records, out of which 514 survived the deduplication process, and 204 became potentially eligible for further assessment; however, since no full-text eligibility process was completed, those records could not be considered a systematic evidence corpus. A verified subset of 20 evidence items was used to detect recurring constructs, assessment approaches, Defence/PME evidence, and methodological gaps. The synthesis revealed that AI literacy is multidimensional and responsible AI judgement requires more than confidence or skill proficiency. Hence, six tentative dimensions of the toolkit were developed: (1) understanding and limits of AI, (2) critical analysis and judgement, (3) human agency and intellectual ownership, (4) reliance calibration, (5) verification, and (6) ethical and responsible judgement. The synthesis also reveals certain evidence gaps, including prevalence of self-reports, lack of objective measures of scenario-based learning outcomes, fragmented assessment of verification and calibrated reliance, and limited evidence on direct Defence/PME validation. The toolkit is presented here as an evidence-informed design architecture rather than as a validated educational intervention or psychometric instrument. Further research will require validation from experts, cognitive interviews, piloting of scenarios, performance assessment, reliability and psychometric testing, experimental

validation, and longitudinal Defence/PME assessment.

Keywords: Artificial Intelligence Literacy; Scenario-Based Assessment; Defense Higher Education; Professional Military Education; Human AI Judgement; Human Agency;

Highlights

- **Artificial intelligence (AI) is increasingly affecting higher education and…**
- **To address this need, this study aims to develop a Scenario-Based AI Literacy…**
- **An evidence-informed, staged approach was selected in which preliminary evidence…**

IMPACT STATEMENT

Artificial intelligence (AI) is increasingly affecting higher education and professional decisions, thus calling for a broader form of AI literacy that includes not only technological expertise but critical analysis, human agency, verification, calibrated reliance, and ethical responsibility. To address this need, this study aims to develop a Scenario-Based AI Literacy Toolkit for Defense Higher Education, especially taking into account the demands of high-stakes learning and Professional Military Education (PME). An evidence-informed, staged approach was selected in which preliminary evidence synthesis was combined with a prospective validation protocol. The preliminary search revealed 515 records, out of which 514 survived the deduplication process, and 204 became potentially eligible for further assessment; however, since no full-text eligibility process was completed, those records could not be considered a systematic evidence corpus. A verified subset of 20 evidence items was used to detect recurring constructs, assessment approaches, Defence/PME evidence, and methodological gaps. The synthesis revealed…

1 I. INTRODUCTION

In recent years, artificial intelligence (AI) has been changing the way learners gain information, create content, analyze evidence, solve problems, and make decisions in higher education. The development of generative artificial intelligence (GenAI) has sped up this process since current AI systems can generate seemingly authoritative text, analysis, explanation, summary, and recommendations in just a few seconds [1]. Consequently, AI literacy is now not about using the technology properly; it is also about understanding AI's potential and limitations, evaluating AI-generated outputs, validating information against available evidence, identifying uncertainties and biases, and maintaining accountability for decisions made with the help of AI. Indeed, previous reviews of AI literacy show that it is multidimensional and includes several aspects, namely, knowledge, application, evaluation, ethics, metacognition, and context-awareness [2]. The significance of this shift is especially pronounced in Defense Higher Education, where the educational tasks often imply making professional judgments in a complicated information environment with a considerable amount of uncertainty and making decisions with potentially serious consequences. Under these circumstances, inappropriate use of an AI recommendation can have much more dramatic consequences compared to giving an incorrect answer during a normal educational activity [3,4]. Therefore, the goal of AI education in this context should not only be encouraging learners to use an AI tool effectively. It should also be developing skills that allow determining when AI support is needed, when an AI-generated output should be challenged, when some claims require validation, when output needs to be modified or rejected, and when human judgment should prevail. High-stakes and military research contexts have revealed the importance of automation bias, algorithm aversion, calibrated trust, and human supervision in decision-making processes supported by AI.

A key issue, however, is that current efforts to teach AI literacy may fail to assess whether students are able to convert their knowledge into correct actions. Existing research on AI literacy recognizes multiple aspects of competence; yet measurement often rests on the learners' perception of their own knowledge, confidence, or competence [6]. According to the findings of AI-literacy reviews, there is a notable difference between the learners' perceptions of what they are able to do and the real actions they are capable of performing. For the case of Defense Higher Education, such a difference takes special importance. The learners may know that the AI system produces wrong information, but still accept its claims because of their plausibility. Likewise, the learners may feel confident in using AI; at the same time, they will be unable to detect uncertainty or change their dependence in the case of high risks in particular tasks. As will be discussed further in this Introduction, such a situation represents three interrelated gaps which become the motivation for the present study: the gap in conceptualization of AI literacy and responsible judgement, the gap in measurement of the judgement itself, and the gap in the context of direct Defence/PME-related evidence.

The educational issue, in other words, is not whether the learner knows about AI, but whether he or she is capable of exercising responsible judgment in response to AI outputs. To be responsibly literate in AI means that the learner can move from knowing to doing, which includes understanding the capabilities and constraints of the AI system, evaluating the quality of the output generated by the system, verifying claims by checking them against authoritative sources, matching reliance on the AI system to its capabilities and the stakes involved, and being ultimately accountable for the resulting decision. There is ample evidence in favor of measuring these behaviors within decision-making sequences involving accept, question, verify, modify, override, or reject choices, rather than using trust or confidence as an adequate measure of AI competency. This makes the case for researching scenario-based AI literacy education. In the scenario-based method, one can put the learner into a situation where AI outputs are reliable, incomplete, uncertain, biased, or not relevant to the problem at hand. Rather than asking the learner to self-report on whether he or she is a competent user of AI systems, the scenario-based measure will allow one to assess the learner's reaction to an AI output that is either helpful, misleading, incomplete, contradictory, or uncertain.

The necessity for scenario-based assessments is especially important in the context of the transfer because Defense Higher Education is a high-risk transfer context and not merely another example of using higher education AI literacy in general. In the former case, general higher education literature can contribute to understanding of educational mechanisms, whereas military and professional literature can help understand the concepts of reliance, automation bias, challenge, and consequence management. However, the latter cannot automatically imply what defense students know and how they act. Transfer depends on factors such as learner demographics, task, authority, institutional policy, assessment circumstances, and consequence structure. The conceptual logic outlined by the paper suggests that AI literacy is a judgment process pathway and not a set of separate technical skills. The suggested conceptual framework is supposed to be a conceptual architecture and not necessarily a verified causative model. The above-mentioned distinction is consistent with the existing evidence on the relevance of agency, critical thinking, reliance, verification, and accountability, but not on their developmental order.

Table I presents a structured summary for this topic. It compares traditional awareness-based AI literacy to performance-based criteria, and explains why the toolkit should be more focused on reactions rather than confidence.

Table I. From Conventional AI Literacy to Scenario-Based AI Literacy:

Conventional AI-literacy emphasis	Key limitation	Scenario-based requirement
Understanding AI	Knowledge may not translate into behavior	Demonstrate appropriate AI use in context
Awareness of AI limitations	Learners may recognize limitations but ignore them	Identify limitations during an actual task
Confidence in AI use	Confidence does not demonstrate competence	Demonstrate observable performance
Trust in AI	Excessive trust can produce automation bias	Calibrate reliance according to evidence and stakes
Ethical awareness	Knowing principles does not guarantee ethical action	Apply ethical judgement in realistic situations
AI-generated information	Plausible outputs may remain inaccurate	Verify claims against authoritative sources
General AI skills	Context and consequences may be overlooked	Adapt AI use to task and consequence conditions
Self-report assessment	May overestimate actual competence	Combine performance, explanation, and verification

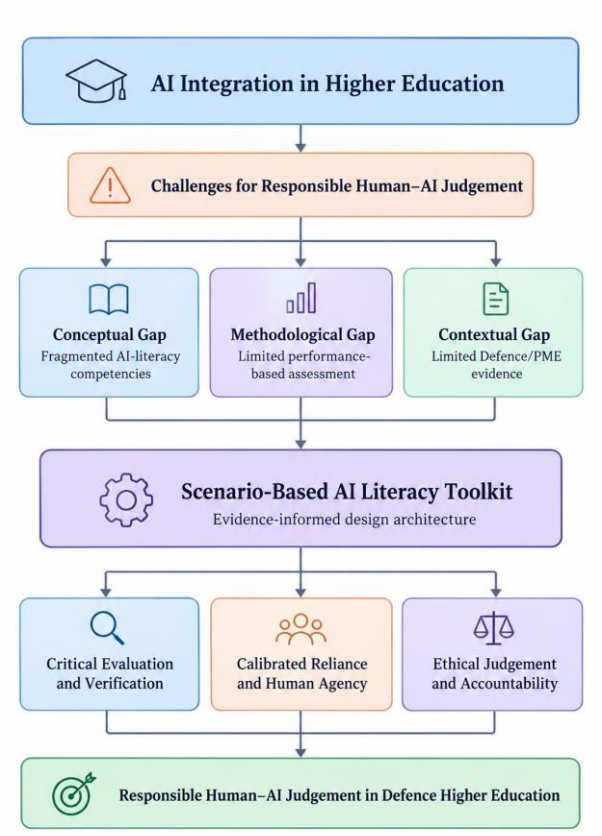


Figure 1. Conceptual overview of the rationale for scenario-based AI literacy in Defense Higher Education.

Figure 1 below depicts the gaps in the literature on AI literacy, which have spurred the current research, including fragmented competencies, lack of performance assessment, and a lack of specific evidence from Defence/PME. The figure illustrates the relationship between these gaps

and the proposed scenario-based toolkit, which highlights critical thinking, appropriate trust, human autonomy, and responsible judgment.

1.1 *A. Research Gap*

Firstly, there is a conceptual issue with AI literacy and responsible human judgement, as current frameworks consider various dimensions of AI literacy, but those dimensions do not always cohere as part of a decision-making process where learners assess AI output and respond accordingly. As a result, knowledge, skills, attitude, trust, ethical considerations, and critical thinking can remain separate competencies rather than elements of a joint decision-making process. Secondly, there is a methodological gap as far as assessment of AI literacy is concerned. Current data show that self-reporting measures are widely used, while objective or scenario-based assessment is underdeveloped. As a result, educational programs run a risk of measuring students' perceptions of competence instead of assessing their ability to verify the information, understand uncertainty, calibrate trust, and dispute incorrect AI suggestions. Thirdly, there is a gap as far as Defense Higher Education is concerned, as current defense and military literature considers the importance of responsible AI and human supervision and accountability, while higher education literature covers data on AI literacy and educational techniques. Yet there is no research into Defense learners and how they should be taught and tested in the context of responsible AI-based judgement.

1.2 *B. Purpose of the Study*

In the context of this study, the intention is to generate an evidence-informed Scenario-Based AI Literacy Toolkit for Defense Higher Education, which will take complex concepts related to AI literacy and articulate these into concrete scenarios, observable learner behavior, assessment criteria, and reflective decision-making exercises. In developing this toolkit, the assumption will be made that in high-stakes education, AI literacy should be reflected in the quality of human judgment used in regard to the outcomes of AI technology, and not merely inferred on the basis of self-report measures such as knowledge and confidence. Hence, this proposed toolkit should be considered as an evidence-informed, testable pedagogic design rather than a validated intervention.

1.3 *C. Contributions of the Study*

In the field of responsible AI education in Defense Higher Education, the present study makes a theoretical contribution to the existing body of knowledge by proposing an evidence-informed Scenario-Based AI Literacy Toolkit that links together the concepts of AI literacy, critical evaluation, evidence verification, calibrated reliance, human agency, and accountability. The study makes a methodological contribution by developing the Scenario-Based AI Literacy Toolkit and translating the abstract concept of AI literacy into observable human behavior that can be studied via realistic scenarios. The practical contribution made by the study lies in providing the architecture for designing curriculum, developing scenarios, performance assessments, faculty training, and learner self-reflection. The key contribution of the study is, thus, the integration and contextualization of scenario-based AI literacy in Defense Higher Education, rather than the novel nature of the dimensions of AI literacy per se. It should be stressed that the proposed educational architecture is presented as an evidence-informed and testable one, not a validated intervention, due to the necessity to validate the proposed intervention itself through expert validation, scenario piloting, psychometric testing, and Defence/PME-specific research.

2 II. LITERATURE REVIEW

The review integrates AI literacy frameworks, scale validation, interventions, human–automation interactions, and examples from the military context. These literature types answer different

questions and cannot be used interchangeably: reviews inform about competencies, scale validations validate scales, interventions explore learning, and military contexts highlight reliance in situations of great importance. For that reason, the synthesis starts by comparing the population, design, assessment type, and inference possibilities of each type of evidence before identifying the requirements for the suggested toolkit.

2.1 *A. Conceptualizing AI Literacy in Higher Education*

AI literacy has come to be regarded as a significant educational construct given that artificial intelligence systems have been increasingly incorporated into various fields of academia, profession, and society. When AI literacy was first framed, it was mainly concerned with the ability of the learners to understand concepts related to AI. However, these days, it has evolved into an interdisciplinary skill involving the knowledge, application, assessment, creation, and ethical interaction of learners. Long and Magerko presented an early competency-focused definition of AI literacy by highlighting the need for learners to be able to critically evaluate and interact with AI technologies [8]. In sharp contrast, Almatrafi, Johri and Lee conducted a review of 47 articles written in 2019-2023, pointing to the diverse populations of learners involved, the various ways in which this construct was implemented, and how it was assessed.

The complexity of the concept of AI literacy is especially important in terms of higher education, since university students need not only to communicate with AI but also to evaluate information, generate knowledge, solve complicated tasks, and make decisions independently [9,10]. Therefore, AI literacy implies not only technical skills with AI systems but also the understanding of their functioning and limits, evaluating generated information, and deciding whether AI should be used in a particular task and, if so, how it can be applied. Thus, this broad concept underlies the choice of the scenario-based conception of AI literacy, which is considered in the current study. A recent synthesis of 29 studies adds weight to the integration of competencies in the concept of AI literacy and proposes a framework including the mentioned competencies for all kinds of learners and educational levels, making the idea of the uniform development path for AI-literacy competencies even stronger [11]. Nevertheless, there are some indications that the notion of AI literacy is also fragmented both conceptually and in practical terms. Whereas the reference framework defines the following relatively stable components as comprehension, application, evaluation, and ethical consideration, the modern framework includes such components as creation, social, integrational, legal, metacognitive, and contextual ones. It is a positive development, but it also raises the following issue: the increase in the number of components does not facilitate their practical application in the AI-based decision-making process.

2.2 *B. AI Literacy as a Multidimensional Competence*

Other forms of multidimensionality in AI literacy can be found in studies concerning measurement instruments of AI literacy. In this regard, the systematic literature review conducted by Lintner has revealed 22 papers containing validation results for 16 scales of AI literacy. While most of the reviewed scales revealed good structural validity and internal consistency, there were certain limitations with regard to their measurement properties. Notably, only a few scales were evaluated for content validity, reliability, construct validity, and responsiveness; no cross-cultural validity or measurement error was reported. More critically, only three of the reviewed scales were performance-based while thirteen were self-report scales.

The issue of the mismatch between self-report and performance-based measures is especially pertinent to the current toolkit proposal. If learners are asked about their understanding of AI, about their ability to properly trust AI and use it, then this assessment will measure the learner's

perception of their own capabilities rather than the demonstrated competence. The discrepancy between self-assessment and performance-based assessments of AI literacy has been recently established as well. In light of that, the current study defines AI literacy as a capability which must be ultimately measured by action. That is to say that the learner cannot be considered fully AI literate just because they can describe the concepts of hallucination, bias, and automation bias. The learner must demonstrate that they are able to detect these issues in a real-life task and react to them adequately.

Evidence quality is evaluated in terms of study design and context, not by the number of sources alone. A systematic review can identify recurring competencies, but it cannot demonstrate that a newly proposed Defense-specific instrument is valid. Similarly, results obtained from small interventions or Defense studies make useful comparison cases, but they are not direct applications of the tools at hand.

Table II contrasts concept reviews, measurement tools, small educational interventions, and Defense-related behavioral findings. It outlines their unique value without suggesting that results obtained from one population can be generalized to another.

Table II. Evidence on AI-Literacy Conceptualization and Measurement:

Study	Year	Evidence/design	Sample/corpus	Principal contribution	Relevance to toolkit
Almatrafi et al.	2024	Systematic review	47 articles (2019–2023)	AI-literacy constructs, implementation and assessment	Competency synthesis; not an intervention test
Lintner	2024	Systematic review	22 studies / 16 scales	Measurement quality and limitations	Supports performance-based assessment
AI-literacy measurement review	2024	Scale analysis	Multiple instruments	Four correlated dimensions identified across instruments	Supports multidimensional structure
Yan et al.	2025	Intervention	16 students	Knowledge, metacognition, ethics, and contextual AI use	Supports educational intervention design
Kahn et al.	2026	Survey experiment	236 cadets	Automation bias, algorithm aversion, and calibrated trust	Provides a defense-relevant behavioral dimension

There are three mutually complementary types of evidence in the comparison: the systematic review of Almatrafi et al., who analyzed 47 papers on AI literacy competence and described its core components among different populations; the analysis by Lintner of 16 scales used to measure 22 validation studies and their high reliance on self-reporting; the intervention literature demonstrating possible ways of teaching; and finally, the cadets' literature addressing the issue of reliable reliance among a Defense population. All four types together prove that it is reasonable to consider the definition of competence and human–AI judgment.

2.3 *C. Generative AI Literacy in Higher Education*

The emergence of generative AI technology has added new requirements that cannot be met by traditional definitions of AI literacy. GenAI can create fluent and contextually relevant texts, thus making one of the main challenges in learning about GenAI the following one: An output

created by a system may look reliable while the facts, reasoning, sources, or assumptions behind it are incorrect [12,13]. Thus, some researchers have started to make a distinction between GenAI literacy and the more general concept of AI literacy. In 2025, there was a systematic review of literature on GenAI literacy that identified five competencies in GenAI literacy: knowing and understanding GenAI; using and applying GenAI; evaluating and incorporating GenAI; GenAI ethics; and attitudes towards GenAI. It included the analysis of 51 empirical studies published in 2023 and 2024, proving how fast the research on GenAI literacy has developed. In addition, a systematic review done in 2025 and aimed at studying the relationship between GenAI and academic skills includes 158 studies published in 2021-2024, showing how actively scientists explore the impact of GenAI on cognitive, technical, and interpersonal skills in higher education.

From these data, it is evident that GenAI literacy cannot be considered just an addition to prompt-writing skills. Rather, learners need to develop the skills that will help them to evaluate generated outputs, recognize the situations in which the outputs can be unreliable, appropriately combine AI help with reasoning, and ensure academic and professional responsibility. This fact is particularly true for Defense Higher Education since the consequences of an apparently reasonable but actually incorrect output may be much worse than in other contexts. Also, one of the five areas of under-researched governance issues in generative AI ethics in education is the issue of human accountability for AI decisions [14].

2.4 *D. Critical Evaluation and Verification of AI Outputs*

Critical assessment of the information produced by AI technologies is one of the essential aspects of AI literacy. There is a possibility that the information produced by an AI algorithm might be linguistically convincing yet not accurate, unsupported, fictional, incomplete, or based on incorrect assumptions.

Therefore, the ability to verify should be one of the explicit parts of AI literacy. The information provided in the references shows that evaluation of AI literacy should take into account knowledge, skills, explanations, source verification, and decision responses instead of just confidence. In the context of Defense education, this distinction becomes crucial because students have to deal with information that is incomplete, uncertain, controversial, or consequential. Verification in this case means:

1. What is the claim?
2. What evidence exists to back up the claim?
3. Where did the evidence come from?
4. Is the evidence source credible and suitable?
5. What is still uncertain?

Empirical studies in the field of professional education, which is considered a high-risk setting, shed light on the problems caused by using unverified explanations from artificial intelligence: a randomized experiment conducted on inexperienced medical students showed that misleading explanations from artificial intelligence led to decreased diagnostic ability when compared with the no-explanation control group, and there was no difference in confidence between the two groups, hence a lack of confidence calibration [15].

2.5 *E. Human Agency and Critical Judgement*

AI literacy is also highly linked with human agency. Human agency deals with whether the learners maintain control of their reasoning, intentionality, authorship, and responsibility during the use of AI, while critical judgment focuses on learners' capacity to assess claims, evidence, uncertainty, bias, and consequences. While related constructs, both agency and

critical judgment are not the same and must not be confused. Reference evidence shows that interactions between learners and AI can be structured in such a way that would either promote or limit human agency [16]. In addition, reference evidence states that learning interventions may help to develop metacognitive skills, critical judgment, ethics, and context-based application of AI technologies in conjunction with technical skills. This difference is particularly significant in terms of Defense Higher Education, as AI-assisted decisions must not be equated with AI-motivated decisions. In the appropriate situation, a learner can use AI in order to produce different options and organize information, remaining responsible for their evaluation and final decision.

2.6 *F. Calibrated Trust, Reliance, and Automation Bias*

There is also a body of literature on the link between AI trust and reliance. Trust does not need to be maximized but must be calibrated to match the system's capability, quality of evidence, level of uncertainty, cost of making the wrong decision, and the possible consequences of failure. Literature on the issue clearly defines trust as an attitude and reliance as a task-specific action [17]. It is especially important for Defense education, where two mistakes can occur. Automation bias makes one trust the AI too much, while algorithm aversion makes one reject its help without a good reason [18].

Literature that includes military academy cadets is especially relevant as it investigates automation bias, algorithm aversion, and calibration of trust in Defense students [19]. However,

The above-mentioned reference paper properly highlights that there is reason to be careful with the results of the study due to its nature as a preprint with several shortcomings. Thus, the presented toolkit does not regard "trust in AI" as a result of education on its own. The tool is focused on the process of choosing the proper response to the use of the technology: Accept → Question → Verify → Modify → Override → Reject. This allows for a more behavioral approach to the concept of responsible AI literacy. Military evidence complementing the above-mentioned studies shows that trust in the use of AI for the purposes of decision-making at the strategic level is driven by the balance of numerous technical, operational, and oversight factors [20]. Another study conducted among cadets showed that the assumption about trusting attitude towards AI technologies as "digital natives" was not true [21].

2.7 *G. Scenario-Based Learning and Performance Assessment*

Self-report measure limitations give an excellent theoretical explanation for why scenario learning should be encouraged. Scenarios give education scholars an opportunity to examine how learners react to changes in context. It is possible that the same AI system will generate useful outputs in a certain non-threatening scenario and yet needs considerable validation in a different threatening scenario. Scenario-based evaluation can thus manipulate:

- AI-system reliability;
- uncertainty;
- evidence quality;
- task stakes;
- conflicting information;
- time pressure;
- institutional rules; and
- consequences of error.

Students may then be encouraged to justify their choices, to provide the evidential base they relied on, to account for any discrepancies, to identify the appropriate dependence response,

and to determine who is responsible for the eventual outcomes. This is consistent with the advice provided in the reference framework for performance assessment to ask students to take on responsibilities like accepting, challenging, verifying, revising, overriding, and rejecting the outputs of the AI system in justifying their decisions. Evidence from professional learning situations other than AI learning provides additional support for such an approach [22].

2.8 *H. Defense Higher Education as a High-Stakes Transfer Context*

The Defense context imposes one more theoretical consideration: the transfer needs to be context-specific. General higher education literature may prove helpful in providing evidence about AI literacy mechanisms; however, it cannot automatically ensure their validity in the case of Defense education. The reference literature makes it clear that transfer is contingent upon the population involved, the stakes of the task, authority relations, assessment settings, institutional governance, and consequence profiles. It allows making the distinction between AI literacy generally and AI literacy in the context of the Defense organization. The former could involve knowing about AI, applying AI technologies, assessing outcomes, and acting responsibly. The latter, besides all that, should take into consideration professional responsibility, high-stakes decision-making, institutional accountability, uncertainty, and the necessity to question AI applications when evidence is not convincing enough. The policy literature of Defense can provide the governance standards and institutional demands [23,24,25]; the empirical literature of higher education can provide information about the learners' capabilities and educational mechanisms [26]. These evidence categories should not, however, be treated as interchangeable. The reference paper emphasizes that policy expectations do not demonstrate learner attainment, while general higher-education evidence does not automatically demonstrate Defense learner capability. Defense-focused work on secure and explainable language models illustrates a relevant application context, although it does not itself validate an AI-literacy assessment toolkit [27].

Table III provides a structured summary relevant to this section. The table connects AI-literacy constructs to illustrative scenario challenges, observable learner actions, and proposed assessment indicators.

Table III. Illustrative Scenario-Based Operationalization of AI-Literacy Constructs Identified in the Literature:

AI-literacy construct	Scenario challenge	Observable learner behavior	Potential assessment indicator
AI understanding	AI produces a confident but uncertain response	Identifies system limitations	Accuracy of limitation identification
Critical evaluation	AI provides competing claims	Challenges unsupported claims	Quality of reasoning
Verification	AI provides factual assertions	Checks authoritative sources	Verification completeness
Reliability judgement	AI output varies in quality	Recognises uncertainty	Calibration accuracy
Reliance	Evidence is strong/weak	Accepts or questions appropriately	Appropriateness of reliance
Human agency	AI recommends a decision	Retains independent judgement	Justification of final decision
Ethics/accountability	AI output creates an ethical concern or contributes to final work	Identifies ethical implications; documents AI contribution	Ethical reasoning; accountability statement

The table below highlights how the framework maps the constructs of AI literacy to observable actions and performance indicators of the learners. It assesses the learners' capabilities of comprehending the limitations of AI, critically analyzing the claims and counterclaims, validating

the information by using authoritative sources, dealing with uncertainty, and calibrating the use of recommendations given by AI technologies. In addition, it also highlights the role of human agency and ethics in making assessments by measuring the capabilities of maintaining independent judgment, justifying the decision, and giving credit to the contributions of AI to learners.

3 I. Synthesis of the Literature and Implications for Toolkit Design

In summary, the literature suggests that AI literacy has evolved into a concept that is becoming more focused on capability, involving the understanding of, application of, evaluation of, ethicality of, reflection upon, human agency in relation to, verification of, and contextualized judgement about AI systems. Nevertheless, several limitations of the current body of literature should be noted: AI-literacy measurement remains largely self-reported, objective performance testing is rather underdeveloped, and transfer of general higher education results to the Defense domain requires further validation. The above literature informs the rationale for the proposed Scenario-Based AI Literacy Toolkit.

In the conceptual framework employed in the current study, AI literacy is viewed separately from responsible AI judgement, with the former providing knowledge and the evaluative capabilities that shape the latter. In other words, the former is a competence required by the latter, which is, however, not necessarily generated by the former and thus has to be observed. In contrast to simply replicating the existing competency frameworks, the current toolkit attempts to translate the key dimensions of the concept of AI literacy into decision-making behaviors under certain contexts.

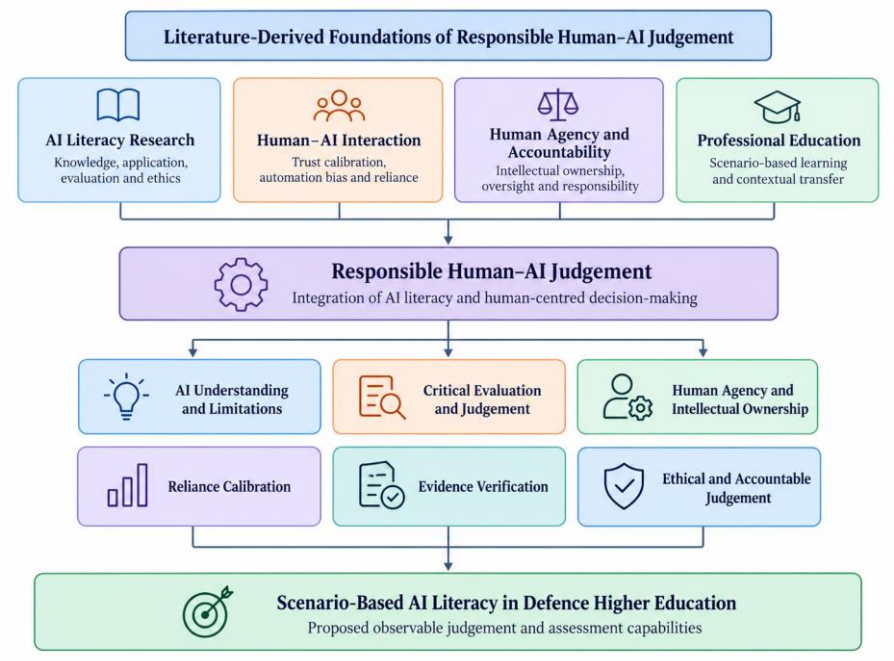


Figure 2. Literature-derived foundations and six provisional dimensions of responsible human–AI judgement.

As illustrated in Figure 2, there are four literatures that relate to the notion of responsible human AI judgement: AI literacy, Human-AI Interaction, Agency and Accountability, and Professional Education. The six provisional toolkit dimensions provide an organizational framework of these literatures in a way that can be used for educational purposes. They do not represent validated constructs or a developmental progression.

4 III. METHODOLOGY

This subsection presents the staged approach of the methodology that has been employed for the development of the Scenario-Based AI Literacy Toolkit for Defense Higher Education. The processes of searching and screening the evidence, the standards for synthesis of the evidence for AI literacy and human-AI judgement research, and its translation into the scenario-based teaching and assessment tools are described. Additionally, the difference between the design processes based on evidence, which has been performed within this study, and validation processes proposed for future studies is highlighted.

4.1 *A. Research Design*

For the purposes of developing a Scenario-Based AI Literacy Toolkit for Defense Higher Education, this study utilizes a two-part, staged research design. The first component consists of an initial evidence search and synthesis with characteristics akin to scoping research, wherein the ways that concepts such as AI literacy, generative AI literacy, critical evaluation, verification, human agency, balanced use, automation bias, accountability, and scenario-based assessment have been theorized and put into practice in higher education, professional education, high-stakes decision-making, and Defence/PME are mapped out. The second component will involve a prospective validation process by which the toolkit developed based on evidence is tested via expert review, scenario piloting, performance assessment, and psychometrics.

This has been done in such a way that evidence synthesis would inform the constructs, dimensions of the scenarios, and the assessment criteria, whereas validation would occur at a later point in time as part of the subsequent research process. It should be stated that the proposed toolkit is neither reliable nor validated in this particular research effort. It has been done in the same manner as in the reference study.

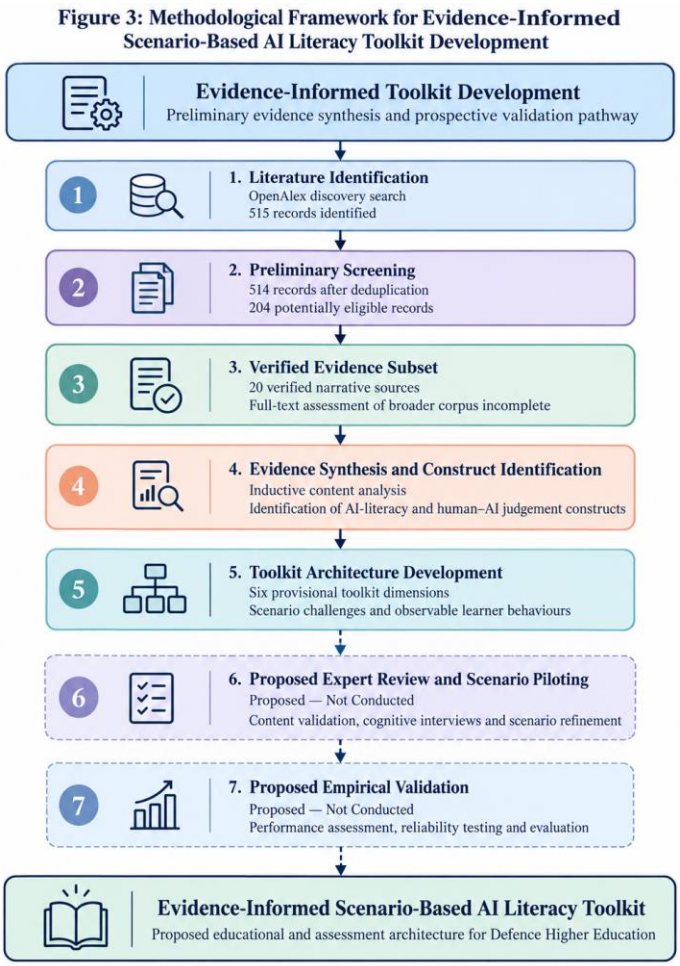


Figure 3. Evidence-informed toolkit development and proposed future validation stages.

As illustrated in Figure 3 below, the research was carried out following preliminary discovery, title and abstract screening, analysis of verified narrative sources, and development of toolkit architecture. The expert review, piloting of the scenarios, reliability assessment, and empirical validation have been planned but not yet carried out.

4.2 *B. Research Questions and Prospective Validation Hypotheses*

The research questions (RQ1–RQ4) guide the present evidence-informed design study and are addressed directly by the evidence synthesis reported below. Hypotheses H1–H5, by contrast, are prospective and are reserved for the subsequent empirical validation study; they are not tested against data in the present manuscript. Question RQ1 deals with the concept and measurement of AI literacy and generative AI literacy. Question RQ2 deals with critical evaluation, human agency, verification, balanced reliance, automation bias, ethics, accountability, and scenario-based testing for learning and decision-making with AI. RQ3 addresses direct Defence/PME evidence and the bounded transfer of findings from higher education and other high-stakes professional contexts to Defense Higher Education. RQ4 addresses the derivation, design, and prospective validation of the Scenario-Based AI Literacy Toolkit, including the translation of evidence-informed constructs into scenarios, observable learner behaviors, and performance-based assessment indicators. The empirical validation protocol will test the following hypotheses after data collection; they are stated here as prospective hypotheses only and are not empirical findings of the present manuscript:

H1. Higher AI-literacy knowledge and awareness of AI capabilities and limitations will be positively associated with scenario-based judgement accuracy.

H2. Higher critical-evaluation and human-agency scores will be positively associated with verification quality and intellectual-ownership scores.

H3. Higher reliance-calibration scores will be associated with lower rates of both over-reliance and under-reliance when scenario reliability and task stakes vary.

H4. Participants receiving the structured Scenario-Based AI Literacy Toolkit intervention will show greater improvement in scenario-based judgement accuracy and verification quality than participants receiving a tool-orientation comparison activity, with allocation, sample size, and treatment fidelity to be determined by an a priori power analysis in the subsequent validation study.

H5. The correlation between AI literacy and accuracy of judgement in scenario-based tasks is expected to be partially mediated by quality of verification to some extent, under the assumption of sufficient sample size and proper modeling.

All these are prospective hypotheses only and not validated with any actual data, statistics, or validations in this paper.

Table IV shows a table related to this section. This table states the prospective hypotheses along with their requirements for evidence; none of which are tested statistically or proven.

Table IV. Hypothesis Audit Matrix (Prospective Not Yet Tested):

Hyp.	Theoretical rationale	Predictor	Outcome	Planned analysis	Principal limitation
H1	Understanding AI limitations should improve evidence-sensitive evaluation.	Objective AI-literacy score	Scenario judgement accuracy	Regression / mixed-effects model	Association ≠ causation
H2	Agency and critical evaluation should support active verification.	Agency + critical evaluation	Verification quality + ownership	Multiple / mixed-effects regression	A construct distinction must be established.
H3	Appropriate reliance should change with reliability and stakes.	Calibration score	Over-/under-reliance	Scenario-level mixed-effects model	Requires sufficient observations
H4	Structured judgement training should outperform tool orientation alone	Toolkit vs comparison condition	Pre/post judgement + verification	Mixed-effects model	Requires valid allocation and treatment fidelity
H5	Verification may explain part of the literacy–judgement relationship	AI literacy → verification	Judgement accuracy	Exploratory mediation	Requires adequate sample and temporal ordering

4.3 *C. Search Sources, Dates, and Access Boundary*

The formal search for evidence will be carried out using Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and pertinent Defence/NATO/government repositories. These databases have been chosen based on the interdisciplinary nature of the topic in education,

computing, human-AI interaction, psychology, professional learning, Defense studies, and governance. The time span covered by publication will be between 2019 and 2026, as this is a period of great advancements in the field of contemporary AI literacy and GenAI studies. Earlier methodological or theoretical papers can be included in the final literature list if they form the basis of constructs or review methodology used in the current study. At this point, the current paper does not claim that any particular database has been searched, as this has not yet taken place; Section IV identifies the single database searched to date (OpenAlex) from the others listed in this section that are intended for future formal searches.

4.4 *D. Search Strategy and Search Strings*

A structured search strategy was developed to identify literature relevant to AI literacy and the design of a Scenario-Based AI Literacy Toolkit for Defense Higher Education. The search targets Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and relevant Defense, government, and professional education repositories, using a combination of controlled keywords, Boolean operators, truncation, and phrase searching. The search concentrates mostly on articles produced from 2019 to 2026; however, earlier works, which form the basis of knowledge in the field, can be kept for their conceptual/methodological value. The search terms include AI literacy, generative AI, performance-based testing, human agency, verification, calibrated reliance, automation bias, and Defence/PME education.

Search terms were organised into four interconnected concept blocks: (1) AI literacy, including AI literacy, artificial intelligence literacy, AI competence, and AI competencies; (2) generative AI, including generative AI, GenAI, large language model*, and ChatGPT; (3) responsible AI-supported judgement, including critical evaluation, critical judgement, human agency, verification, trust calibration, reliance, appropriate reliance, automation bias, accountability, and ethical judgement; and (4) educational and Defence context, including higher education, university, professional education, military education, Defence education, professional military education, and PME. These blocks are combined using Boolean operators to increase both sensitivity and specificity.

One of the main strings to be used for searching was developed as follows: ("AI literacy" OR "artificial intelligence literacy" OR "AI competence") AND ("generative AI" OR GenAI OR "large language model" OR ChatGPT) AND ("critical judgement" OR "critical evaluation" OR "human agency" OR verification OR "trust calibration" OR reliance OR "automation bias" OR accountability) AND ("higher education" OR "professional education" OR "military education" OR "Defense education" OR "professional military education" OR PME). Modifications in accordance with the specifics of databases searched and reference list searching and citation chaining will be performed where necessary.

The search will also involve synonyms such as defence/defense, military education, staff college, command and staff education, professional learning, human-autonomy teaming, machine advice, contestability, epistemic vigilance, and algorithmic reliance, as a search based on phraseology can ignore relevant articles that use different terminology.

4.5 *E. Eligibility Criteria*

Studies will be considered eligible when they directly or substantively contribute to understanding AI literacy, generative AI literacy, critical evaluation, human agency, verification, calibrated reliance, automation bias, ethical judgement, accountability, or scenario-based assessment within higher education, professional learning, high-stakes decision-making, or Defence/PME. The eligibility criteria will be applied consistently during title/abstract and full-text screening. Eligibility screening involves distinguishing eligible higher education and

Defense-related research literature from technical AI literature that is not related to this context, opinion-based literature without support, duplicate literature, and literature that provides insufficient methodological information. Studies are eligible when they are: (1) published primarily from 2019 to 2026 but including earlier literature that provides a foundation for important theoretical or methodological concepts; (2) focused on higher education students, educators, professional learners, military/Defence learners or similar high-stake professional learners; (3) address AI literacy, GenAI literacy, AI-enabled judgements, human-AI interactions, critical thinking, verification, trust/reliability, human agency, ethics, accountability, or similar learning outcomes; (4) provide empirical, conceptual, methodological, review or authoritative policy/governance literature that is pertinent to developing the proposed toolkit; and (5) provide sufficient information about the research context, constructs, methodology, or main findings to allow for valid evidence extraction.

Articles will be excluded if they: (1) consider AI technology development or its effectiveness without bearing on human learning or judgement; (2) consider applications of AI that have no relation to literacy, use, learning or judgement in any form; (3) contain only opinions or commentaries without any evidential basis; (4) are redundant articles or lack necessary information for evidence classification; or (5) discuss applications where there is no reasonable way to draw lessons to higher education or Defence/PME without considerable unjustified extrapolation. Policy and governance documents will be used as sources to set institutional standards and ensure institutional oversight and assurance but not as empirical evidence of learner competencies; results from general higher education and other professional environments will be applied to the Defense Higher Education context only if the transition context is justified.

4.6 *F. Evidence Selection and Screening*

The evidence selection process will follow a systematic, multi-step screening process that is aimed at ensuring transparency, consistency, and reproducibility. Once the database searches have been performed, the articles that have been obtained from the databases will be exported to a reference management program, and duplicates will be flagged. The remaining articles will then be screened against the eligibility criteria using the titles and abstracts. Whenever possible, two independent reviewers will perform title/abstract screening and full-text screening, and any discrepancies will be discussed; in cases where there is no agreement, a third reviewer will make the final decision. The manuscript of this review will provide the total number of records identified, number of duplicates eliminated, number of records screened, number of records excluded based on titles and abstracts, full texts reviewed, full texts excluded with reasons, and number of articles used in the analysis. A flow diagram following the PRISMA format will be developed once the screening process has been completed; numbers will not be inserted until screening has been completed.

4.7 *G. Evidence Categories and Rigor*

Five types of evidence, which are not interchangeable, will be considered to differentiate the origin, context, and function of the literature that will form the basis of the toolkit. The first tier is made up of empirical evidence related to Defence/PME and represents the most contextually sound foundation from which to make claims about learners and the education environment in the Defense context. The second tier is comprised of high-stakes empirical research involving judgment, proper reliance, trust calibration, verification, and automation bias. Tier 3 comprises general higher-education empirical and conceptual evidence concerning AI literacy, generative AI literacy, learning, assessment, and related educational mechanisms. Tier 4 consists of human

factors and professional comparator evidence relating to human/AI interaction, trust, dependency, verification, agency, and decision-making. Tier 5 consists of policy, governance, institutional, Defense, and professional guidance material, primarily employed for establishing expectations with regard to oversight, assurance, accountability, and institutional responsibility rather than learner performance evidence.

Methodological rigor is additionally reported using four descriptive bands: High, Moderate, Limited, and Conceptual/Policy, reflecting design, transparency, evidentiary strength, and applicability rather than fabricated numerical quality scores.

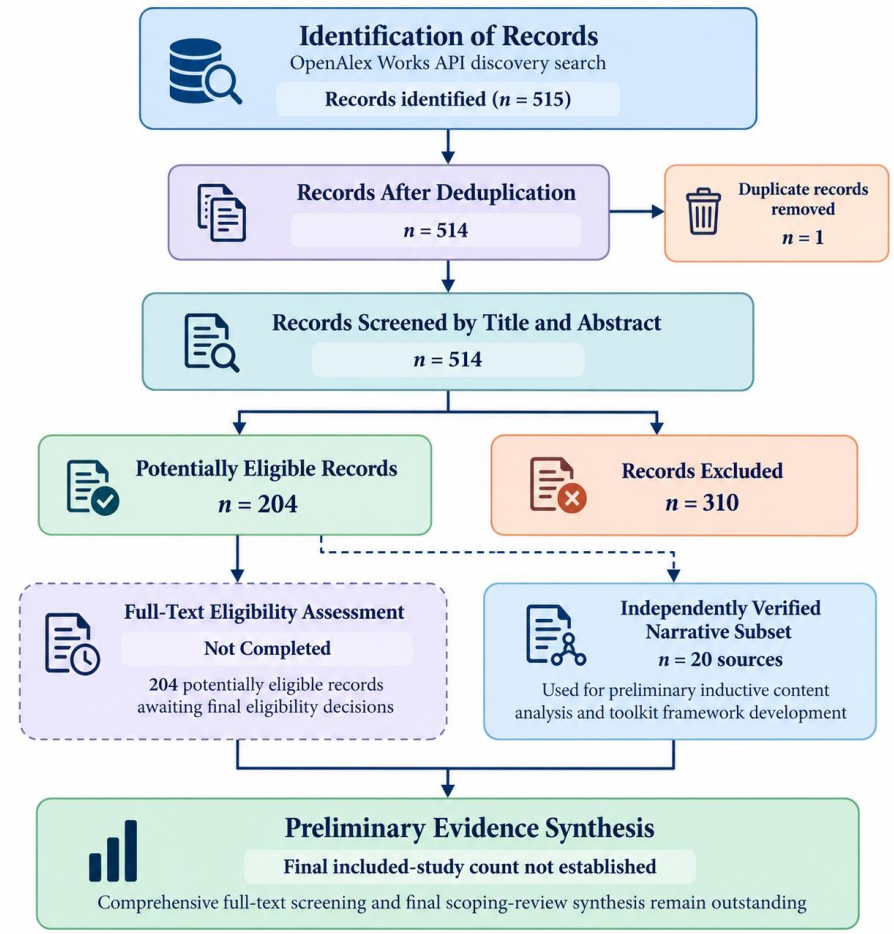


Figure 4. Preliminary OpenAlex evidence identification and screening; full-text eligibility remains incomplete.

As shown in Figure 4, the preliminary OpenAlex search returned 515 records; one duplicate was removed, leaving 514 titles and abstracts screened. Of those, 204 were potentially eligible and 310 excluded. Full-text eligibility decisions remain incomplete. The 20 independently verified narrative sources used for initial synthesis do not constitute the final included-study set.

4.8 *H. Claim-to-Evidence Audit*

In order to ensure the validity of claims and avoid exaggerated claims about the evidence basis, all the claims forming the toolkit will be subjected to an audit of the relationship between claim and evidence. The audit will assign an inferential category to each claim in terms of its status as an empirical fact, source-based claim, interpretation, bounded inference, hypothesis, policy implication, or proposed design aspect. Only those claims that are reported in the cited works in sufficiently comparable settings can be regarded as supported by evidence; support of one construct does not mean that the whole toolkit is validated.

Table V below is a structured summary related to this section. It differentiates source-supported claims, interpretations, and prospective designs to ensure that policy and evidence are not confused.

Table V. Claim-to-Evidence Audit Matrix:

Claim category	Evidence required	Example for present toolkit	Permitted interpretation
Empirical fact	Direct empirical finding	Learners demonstrate difficulty identifying AI limitations	Report as observed finding
Source-based claim	Review/framework evidence	AI literacy includes evaluation beyond tool operation	Report as literature-supported
Interpretation	Compatible evidence from multiple sources	Verification appears important for responsible AI use	Present as interpretation
Bounded inference	Convergent evidence + explicit transfer assumptions	Verification mechanisms may be relevant to Defense HE	Label as inference
Hypothesis	Theoretical rationale requiring testing	Toolkit may improve judgement accuracy	Test empirically
Policy implication	Defence/institutional governance evidence	AI use should remain accountable to human oversight	Present as governance/design implication
Author proposal	Evidence-informed design rationale	Scenario-based verification task	Present as proposed toolkit feature

5 I. Framework and Toolkit Derivation

This Toolkit on AI Literacy using a Scenario-Based Approach was developed based on an evidence-to-design process rather than on the aggregation of dimensions of AI literacy as described in past frameworks. For the initial list of potential constructs identified during the evidence synthesis, the constructs were first mapped in terms of their recurrence, conceptual meaning, educational significance, observability, and appropriateness for Defense Higher Education. Next, each potential construct was evaluated against five decision criteria: (1) recurrence in independent pieces of evidence; (2) distinct conceptual meaning from other constructs; (3) observability via learners' actions or performance; (4) importance in terms of responsible judgement supported by AI; and (5) appropriateness for the Defense Higher Education setting.

See Table VI for further detail.

Table VI contains a detailed outline related to this section. The table reflects the thought process from research constructs to design features of the toolkit and distinguishes findings from source materials from the author's interpretation.

Table VI. Evidence-to-Toolkit Derivation Matrix:

Candidate construct	Recurrence	Uniqueness	Observable learner behavior	Resp. AI relevance	Defence transfer	Decision
AI Understanding and limitations	High	High	Identifies capability or limitation	High	High	KEEP
Critical evaluation/judgement	High	High	Challenges unsupported AI output	High	High	KEEP
Human agency/ownership	Moderate–High	High	Provides independent rationale	High	High	KEEP
Verification	Moderate–High	High	Checks sources and evidence	High	High	KEEP
Reliance calibration	Moderate–High	High	Adjusts reliance to reliability and stakes	High	High	KEEP

Candidate construct	Recurrence	Uniqueness	Observable learner behavior	Resp. AI relevance	Defence transfer	Decision
Ethical judgement	High	Moderate–High	Identifies ethical consequence	High	High	KEEP/ REFINE
Accountability/ oversight	High (governance /high-stakes evidence)	Moderate	Documents responsibility and escalation	High	Very High	MERGE/ REFINE
Well-being/ dependency awareness	Emerging	Moderate	Recognizes cognitive offloading and dependence	Context-dependent	Context-dependent	BOUNDARY CONDITION
Generic AI tool operation	High	Low	Uses AI functions	Moderate	Low–Moderate	MERGE/ EXCLUDE

J. Translation of Constructs into Toolkit Components

After the identification of constructs to be taught, each one of these constructs is turned into a scenario-based learning module through a four-step process, namely: Construct -> Scenario Context -> Observable Behavior -> Assessment Criteria. The operationalization of AI knowledge and awareness of limitations is done through scenarios where the AI generates information confidently that is not fully true or reliable, thereby forcing the learner to find out about the limitations of this AI system and what extra information it requires.

Verification is represented through source checking, discrepancy identification, provenance review, or escalation. Reliance calibration is assessed by varying AI reliability and task stakes and observing whether the learner appropriately accepts, questions, modifies, overrides, or rejects AI recommendations. Human agency and ownership are assessed through the learner's ability to provide an independent rationale and retain responsibility for the final judgement. Ethical and accountable judgement is operationalized through disclosure, role clarity, documentation, escalation, and responsibility for consequences, integrating the accountability/oversight and ethical-judgment candidates identified in Table VI into a single toolkit dimension, consistent with their MERGE/REFINE decision.

K. Final Toolkit Architecture

This final toolkit does not define the Defence/PME as an AI literacy competency. In its place, it is Defense Higher Education which serves as the contextual boundary that determines how the core competencies manifest and are measured in terms of the higher stakes, professionalism, institutional processes, assurance needs, uncertainty and auditability. The resulting framework is based on six core dimensions, in order as follows: (1) Understanding AI and AI's Limitations → (2) Critical Appraisal & Judgment → (3) Human Agency & Intellectual Responsibility → (4) Dependence Calibration → (5) Verification → (6) Ethical & Responsible Judgment, with Defense Higher Education running along the full pathway as the contextual boundary. This six-dimension framework is applied consistently through the rest of the manuscript.

The following Table VII represents the structure of this part of the manuscript. The table contains the six provisional core toolkit dimensions along with the associated scenario suggestions and indicators; the latter are not yet validated.

Table VII. Scenario-Based AI Literacy Toolkit: Six-Dimension Architecture and Assessment Framework:

Toolkit dimension	Scenario manipulation	Learner task	Observable indicator	Assessment evidence
1. AI understanding & limitations	AI produces a confident but incomplete output	Identify limitations	Correctly identifies capability boundary	Limitation-awareness score
2. Critical evaluation & judgement	AI output conflicts with credible evidence	Compare and evaluate claims	Challenges unsupported output	Judgement rubric
3. Human agency & intellectual ownership	AI recommends a decision	Make and justify final decision	Provides independent rationale	Ownership score
4. Reliance calibration	AI reliability and task stakes vary	Decide whether to accept/reject/modify AI advice	Adjusts reliance appropriately	Calibration score
5. Verification	AI produces a questionable claim/source.	Verify information	Checks evidence and records discrepancy	Verification score

Toolkit dimension	Scenario manipulation	Learner task	Observable indicator	Assessment evidence
6. Ethical and accountable judgement	AI output creates an ethical/bias concern or contributes to a consequential decision.	Identify consequences; document AI contribution and responsibility	Recognizes ethical implications; discloses, escalates, and accepts responsibility	Ethics rubric; accountability score

This architecture recognizes the context of Defence/PME, and not a distinct competency, as outlined in the evidence synthesis presented in Section II and the derivation process in Table VI.

L. Prospective Validation Protocol

Validation of the proposed tool will therefore involve a mixed-method approach that is designed to test for the content validity, usability, reliability, construct validity, and educational validity of the tool. Expert review of the proposed toolkit dimensions and scenarios will be done first, whereby experts in the fields of AI literacy, Defence/PME, educational assessment, and human-AI interactions will rate the relevance, clarity, contextual fit, and coverage of the dimensions and scenarios. Cognitive interview and pilot testing of the scenarios among learners will then follow, followed by a performance-based assessment stage, inter-rater reliability analysis, psychometric validation if possible (due to sample size and measurement assumptions), experimental or quasi-experimental design involving a comparison condition, and longitudinal evaluation. The above validation process is in line with the methodology used in the literature, which recognizes the importance of expert review, cognitive piloting, scenario-based assessment, psychometric validation, and direct Defence/PME curriculum research prior to making any claims on validity and educational validity of a framework.

M. Methodological Boundary

In this study, the development of the evidence-based Scenario-Based AI Literacy Toolkit framework is performed as opposed to the development of an evidence-based valid psychometric measure or a demonstrated evidence-based education intervention. Literature evidence supports the selection of constructs of the toolkit, definition of conceptual boundaries of these constructs, and creation of scenarios that allow for observation of the relevant behaviors of learners; however, literature evidence alone is not sufficient for establishing the reliability, validity, efficacy, or educational influence of the toolkit. This boundary is particularly important because the evidence base identified to date is characterized by substantial reliance on self-report measures, limited objective or performance-based assessment, no Defense-specific AI-literacy scale identified within the verified evidence subset, and limited direct evidence from Defence/PME curricula.

6 IV. RESULTS

This section reports only what the preliminary evidence-selection process has actually established to date: the database search, deduplication, and preliminary eligibility screening, together with the narrative synthesis drawn from the independently verified subset of 20 sources. It does not report a completed systematic review, and it does not present conceptual synthesis as empirical outcome data. Findings are further distinguished throughout by evidentiary type: direct Defence/PME evidence, general higher-education evidence, evidence from other high-stakes professional contexts, and conceptual/author synthesis consistent with the evidence-category framework in Section 3.7.

6.1 *A. Preliminary Evidence-Selection Status*

The preliminary evidence search identified a broad but heterogeneous literature base relevant to the development of the proposed toolkit. Eleven Boolean search blocks were executed through the OpenAlex Works API on 22 August 2026, covering publications from January 2019 to August 2026. The search returned 515 unique records; DOI/title deduplication removed one duplicate, leaving 514 records for title/abstract screening. Of these, 204 records were identified as potentially eligible on the basis of their relevance to AI literacy, human–AI judgement, verification, reliance, agency, education, high-stakes decision-making, or Defence/PME contexts.

Full-text assessment of these 204 records has not been completed; these records therefore represent a candidate full-text set rather than a final included-study corpus, and no final included-study count or exclusion-reason log is reported. Within this boundary, 20 source records were independently verified for narrative synthesis, comprising systematic and scoping reviews, empirical higher-education studies, an intervention study, an objective scenario-based assessment study, human-factors and trust research, military decision-making evidence, PME practice evidence, and official Defence/NATO sources. As a matter of importance, the research done so far has been done only using the OpenAlex Works API, while other databases like Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, as well as Defence/NATO/government archives specified in section 3.3 have been identified as sources to conduct a proper search at a later stage.

The evidence therefore provides a preliminary basis for toolkit construction but does not constitute a completed PRISMA-ScR review or establish a final included-study count. The search coverage must also be interpreted conservatively. The executed search reported in the present manuscript was conducted through the OpenAlex Works API only. Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and relevant Defense, NATO, government, and professional repositories identified in the methodology represent intended sources for a subsequent formal search and were not searched as part of the evidence-selection results reported here. Consequently, the current evidence base should not be interpreted as exhaustive coverage of the interdisciplinary literature.

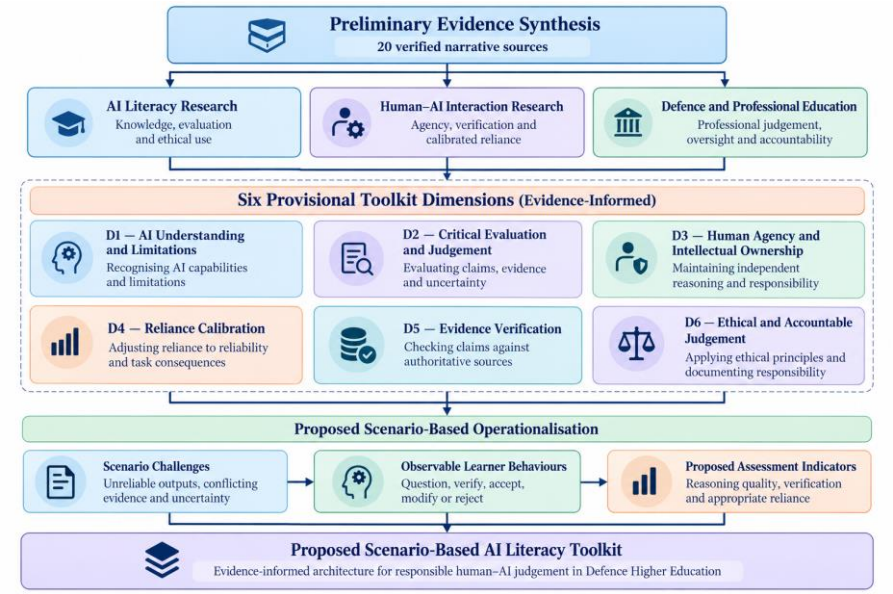


Figure 5. Evidence-informed development of six provisional toolkit dimensions and proposed assessment applications.

As shown in Figure 5, the preliminary 20-source narrative synthesis informed six proposed toolkit dimensions, which are linked to illustrative scenario challenges, observable behaviors, and potential assessment indicators.

These pathways represent the proposed operationalization of literature concepts, but do not constitute proof that there is validation of the toolkit, or that learner performance has been improved as a result.

Table VIII summarizes key information pertinent to this discussion. This table includes preliminary counts of records, as well as explicitly stating that the included study count remains undefined.

Table VIII. Preliminary Evidence-Selection Results:

Selection stage	Result	Interpretation
Boolean search blocks	11	Separate AI-literacy, judgement, trust, verification, and Defence/PME searches
Unique discovery records	515	OpenAlex discovery corpus
DOI/title duplicates removed	1	Transparent deduplication
Records after deduplication	514	Title/abstract screening set
Metadata/title-abstract exclusions	310	Did not meet AI + context + target-construct rule

Selection stage	Result	Interpretation
Potentially eligible records	204	Candidate full-text set
Full-text reports assessed	Not completed	Access/verification boundary
Final included-study count	Not established	Must not be reported as a completed review result
Independently verified narrative subset	20	Used for preliminary narrative synthesis

It is vital that the difference between the 204 potential sources and the 20 validated narrative sources be made clear in the methodology. The 20 sources can be better termed a "subset of validated evidence" because they have been used to determine recurring constructs and evidence gaps in order to design the toolkit.

6.2 *B. AI Literacy and Generative AI Literacy (General Higher-Education Evidence)*

The evidence shows that the best way to conceptualize AI literacy is to look at it from a multidimensional perspective, which differs greatly from a one-dimensional understanding of tool proficiency. Throughout the examined literature, recurring dimensions of AI literacy include recognition of AI, understanding of AI, application and usage of AI, evaluation of AI products, creation with AI, and ethics of AI. Recent frameworks also highlight critical-thinking, social, integration, legal, meta-cognitive, and contextual dimensions. Given this, an ideal Defense Higher Education toolkit would have to go beyond teaching students how to operate certain tools and check if they understand what AI can and cannot do and how to properly evaluate the products of AI. The measurement evidence also highlights another limitation of the current approach to AI literacy: Self-report measures of AI literacy are prevalent, whereas objective performance measures vary in their validity. A recent study that used an objective measure of performance involved 531 undergraduate students and showed a very weak correlation between AI literacy and accurate classification of ethical scenarios. This evidence is general education and does not constitute any PME/Defence evidence.

Figure 6 summarises the relative magnitude of the evidence gaps identified within the verified subset (qualitative ratings, not measured effect sizes).

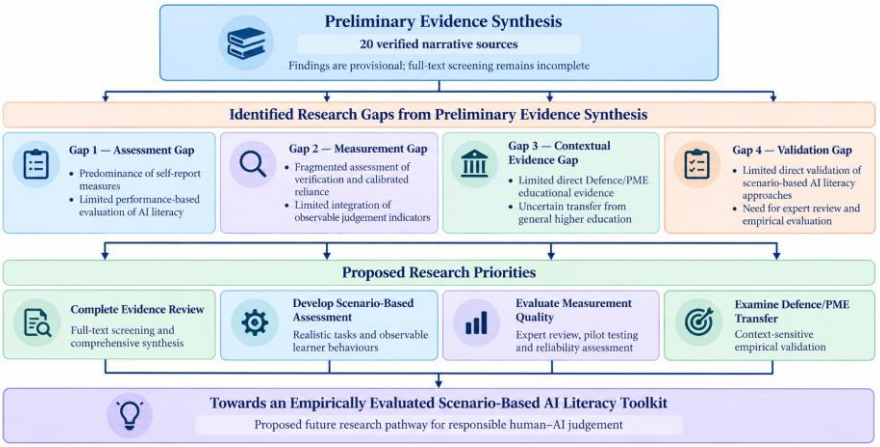


Figure 6. Preliminary research gaps and proposed priorities for scenario-based AI literacy research.

As shown in Figure 6, preliminary synthesis highlighted gaps in performance-based assessment, integrated measurement of verification and reliance, Defence/PME-specific educational evidence, and direct toolkit validation. These limitations motivate proposed future activities: completion of the review, scenario development, measurement testing, and context-sensitive evaluation. The figure does not assign quantitative strength to any gap.

The illustrated gaps motivate direct Defence/PME testing and greater use of performance-based assessment.

6.3 *C. Critical Judgement and Human Agency (General Higher-Education and Human-Factors Evidence)*

The evidence indicates that human agency and critical judgement are related but distinguishable constructs. Human agency concerns the learner's retention of purpose, reasoning, authorship, and responsibility, whereas critical judgement concerns the evaluation of claims, evidence, uncertainty, bias, and consequences. Student–AI collaboration research indicates that the structure of AI interaction can either support or constrain agency, while

intervention evidence suggests that metacognition, evaluation, ethics, and context-sensitive use can develop alongside technical tool competence. This result indicates that the successful use of AI should not be reduced to achieving the goal of generating an adequate response with AI for the purposes of the toolkit. Instead, the scenario activities have to require learners to analyze, question, and own the outputs supported by AI. Evidence about human-AI judgement helps to distinguish between attitude towards trust and real decisions regarding the use of AI.

6.4 *D. Trust Calibration, Appropriate Reliance, and Verification (Human-Factors and Military Evidence)*

The verified evidence distinguishes trust as an attitude or expectation from reliance as a task-level behavioral decision. Appropriate reliance should correspond to demonstrated AI capability, available evidence, uncertainty, task stakes, and potential consequences. Military and cadet evidence further indicates that both automation bias and algorithm aversion are relevant failure modes; a cadet study involving 236 cadets examined these patterns, although its preprint status and methodological limitations restrict the strength of the conclusions that can be drawn. Military decision-making evidence additionally emphasizes uncertainty, challenge mechanisms, contextual reasoning, and audit trails.

6.5 *E. Defence/PME Evidence*

The Defence/PME evidence provides strong contextual justification but limited direct educational validation. Military and Defense sources consistently emphasize human judgement, uncertainty, challenge mechanisms, audit trails, assurance, responsible AI use, and institutional accountability. Direct Defense evidence includes cadet research concerning automation bias and algorithm aversion, while PME practice literature identifies opportunities, limitations, ethics, and implementation considerations. These sources do not, however, establish that a particular Defense curriculum produces measurable gains in AI literacy, verification, reliance calibration, or accountability. The evidence therefore supports treating Defense Higher Education as the contextual boundary of the toolkit rather than as an additional AI-literacy competency; direct PME evidence remains comparatively limited.

6.6 *F. Evidence Gaps*

Four major evidence gaps emerged from the verified subset: (1) AI-literacy assessment remains heavily dependent on self-report, with limited objective and performance-based measurement; (2) no verified Defense-specific AI-literacy scale is established within the evidence subset; (3) education-facing measures of calibrated reliance and verification remain fragmented; and (4) direct evidence demonstrating that a Defence/PME curriculum improves AI-literacy-related performance is currently insufficient.

Table IX provides a structured summary relevant to this section. The table groups the identified limitations of the preliminary evidence base and links them to the need for further assessment research.

Table IX. Evidence Gaps Relevant to Toolkit Development:

Evidence gap	Current status	Toolkit response
Self-report dominance	Substantial	Add performance-based assessment
Defense-specific measurement	Not established	Develop Defense-contextualized indicators
Verification measurement	No validated rubric identified	Create observable verification rubric

Evidence gap	Current status	Toolkit response
Reliance calibration	Education measures fragmented	Use scenario reliability/stakes manipulation
PME curriculum efficacy	Direct evidence limited	Treat toolkit efficacy as prospective
Transfer across Defense contexts.	Untested	Explicitly test contextual transfer

6.7 *G. Evidence-Informed Toolkit Dimensions*

The evidence synthesis supports a provisional set of core toolkit dimensions. AI understanding and limitation awareness and critical evaluation/judgement are retained because they recur across AI-literacy, ethical-scenario, intervention, military-decision, and Defense-policy evidence. Human agency and intellectual ownership are retained because the verified evidence directly links AI-supported collaboration with agency and responsibility.

Revised reliability calibration and validation are included as they are significant based on human factors, military, strategic, and governance evidence, whereas assessment in an educational context is inconsistent. Ethical judgment and accountability/oversight are combined under the one dimension, ethical and accountable judgment, due to Defense governance evidence suggesting a connection between authority, responsibility, ethics, and risk. Awareness of well-being/dependency is kept as a boundary condition rather than a main dimension of the toolkit, as no distinctive evidence about it is provided by the present literature base.

6.8 H. Evidence Synthesis and Toolkit Alignment

The findings of preliminary evidence have been synthesized in order to discover how the current literature can help shape the development of the proposed toolkit. Instead of considering all literature found as equivalent, it has been sorted according to construct, education context, population, function in methodology, and its applicability for Defense Higher Education. Direct Defence/PME literature has been separated from military and other high-stakes professional evidence, while higher education literature was mostly used as the means of establishing AI literacy and mechanisms of learning. Human factors and professional comparative studies were employed when appropriate, in cases where relevant evidence about trust, reliance, verification, automation bias, and human judgement was available.

These six constructs identified through the synthesis included AI literacy and limitation knowledge, critical appraisal and judgments, human agency and intellectual property, calibration, verification, and ethical judgments. They have not been chosen on the grounds of being present within individual pieces of literature; instead, these constructs were considered based on their recurring presence in evidence, their distinctiveness conceptually, the possibility to observe them among learners, and relevance for responsible judgments supported by AI and justifiable transfer into Defense Higher Education institutions.

These constructs were then transformed into observable scenario-based behaviors to ensure that the toolkit assessed the learner's performance as opposed to only self-reported confidence or attitude levels: AI literacy and limitation awareness can be demonstrated through the identification of deficiencies in confident yet incomplete AI outcomes; critical judgment can be tested through discrepancies between AI suggestions and credible evidence; human agency can be measured through independent judgment and reasoning; reliance calibration can be evaluated using varying levels of reliability of the AI and importance of the tasks; verification skills can be demonstrated through source checking, discrepancy detection, and evidence documentation; and ethical and accountable judgment can be measured through responsibility assignment, disclosure, escalation, and oversight decisions. Evidence sources include AI-

literacy literature, human-AI interaction studies, professional education, and defense/governance sources; the latter define the context requirements but do not provide evidence of learning outcomes.

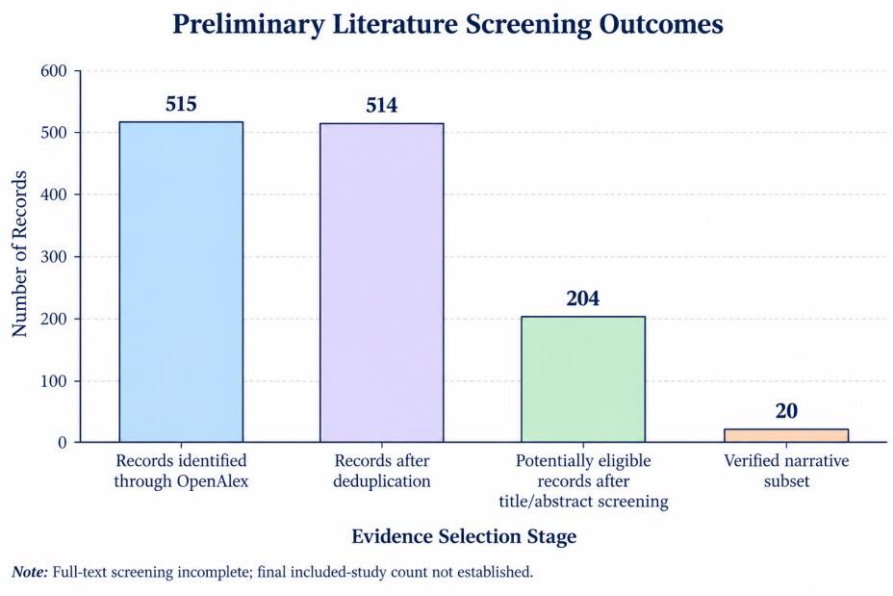


Figure 7. Preliminary literature screening outcomes, including the independently verified narrative subset.

Figure 7 shows that the selection process produced 515 records; 514 were duplicates removed, and 204 were classified as potential candidates based on titles and abstracts. These 20 narrative sources represent a distinct group of preliminary analysis sources instead of a final included study set. The bar chart depicts the number of records rather than a series of decisions.

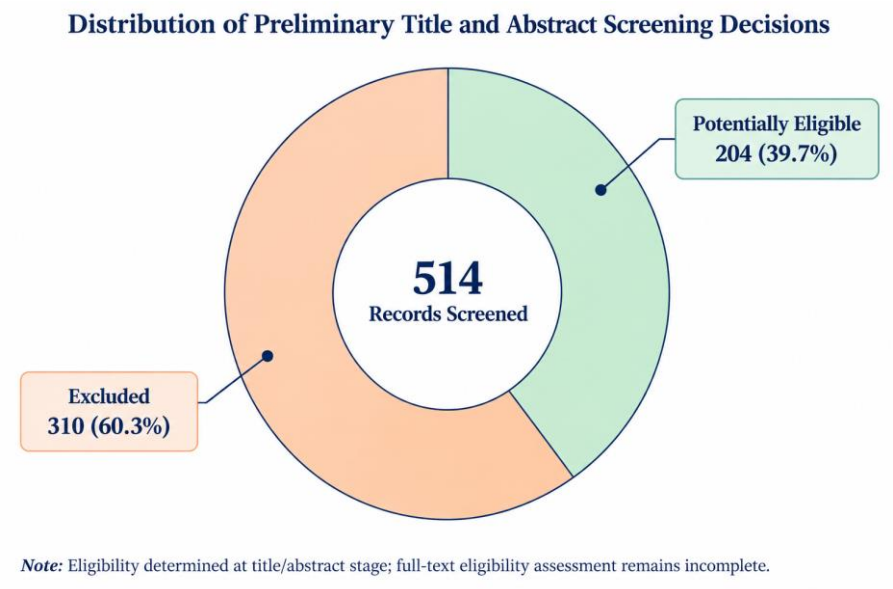


Figure 8. Distribution of preliminary title-and-abstract screening decisions among 514 records.

As illustrated in Figure 8, out of 514 records reviewed, 204 (39.7%) were identified as possibly eligible, whereas 310 (60.3%) were excluded based on title and abstract screening. The above statistics relate to the initial stages of screening, as the complete screening process has not been completed yet.

7 V. DISCUSSION

The architecture for the development of the evidence-informed Scenario-Based AI Literacy Toolkit for Defense Higher Education was established in the current research through incorporating evidence related to AI literacy, generative AI literacy, human-AI interaction, human agency, critical judgement, reliance, verification, and Defence/PME context. It can be seen from the findings that AI literacy cannot be considered just a technical skill or a measure of confidence in using AI applications. Rather, responsible usage of AI technology in Defense Higher Education requires a complex combination of competencies that enables individuals to be aware of AI limitations, be critical of outputs, retain intellectual autonomy, calibrate reliance on AI applications, verify information, and act in an accountable context. This is aligned with the evidence synthesis presented above but has limitations because the current research did not conduct a full-text assessment of all 204 potentially eligible studies. Thus, this paper cannot provide a systematically developed evidence corpus.

The primary contribution of this paper is the synthesis and contextualization of scenario-based AI literacy for Defense Higher Education – it is not the proposition that any of the constructs in question (critical evaluation, verification, human agency, appropriate trust, and accountability) have been new to the field of AI literacy. The constructs are all well documented in existing AI literacy, human factors, and military decision-making literature, but the contribution is in putting these constructs into a coherent judgment pathway and operationalizing them in terms of scenarios and assessment indicators.

7.1 *A. What the Study Contributes*

Implication one is that AI literacy and scenario-based judgement need to be seen as distinct but related outcomes. Evidence on AI literacy suggests conceptualization in terms of multi-dimensional concepts encompassing knowledge, understanding, using, evaluating, ethical considerations, and contextual application of AI systems. Yet the evidence on measurement suggests that the dependence on self-report instruments persists while there is little evidence of objective assessment. This implies the distinction between the statement by the learner of being able to use AI on one hand and the demonstration of judgement on the part of the learner when an AI system offers incorrect, incomplete, or misleading information. Implication two pertains to the distinction that needs to be made regarding human agency, critical evaluation, reliance calibration, and verification. These mechanisms cannot be conflated into one measure of responsible AI use. Human agency relates to the capacity of the learner to retain reasoning and ownership and responsibility; critical evaluation refers to whether the AI outputs are critically evaluated in relation to the evidence, uncertainty, and potential consequences; reliance calibration refers to the extent to which the learner adjusts the reliance depending on capability, evidence, and task stakes; and verification refers to checking the information.

Military and human-factors evidence further indicates that both automation bias and algorithm aversion can create inappropriate reliance. Consequently, the toolkit does not reward simply accepting or rejecting AI; instead, scenario conditions require learners to determine when to accept, question, verify, modify, override, or reject an AI output and to justify the decision.

7.2 *B. Why Scenario-Based Assessment Matters, and Relationship to Prior Frameworks*

Existing AI-literacy frameworks (e.g.,) establish which competencies matter but do not, by themselves, specify how a learner's possession of those competencies should be observed under realistic conditions of uncertainty and consequence. The present toolkit does not

compete with these frameworks; it operationalizes them. Scenario-based assessment matters because it is the only design examined in the

reviewed evidence that can distinguish a learner who merely reports confidence or knowledge from one who demonstrably adjusts behavior, accepting, questioning, verifying, modifying, overriding, or rejecting an AI output as reliability, evidence quality, and task stakes change. This is what conventional self-report AI-literacy instruments cannot capture, and it is why Defense Higher Education, where the cost of miscalibrated reliance is potentially high, is a particularly consequential setting in which to require it.

7.3 *C. Transferability of Evidence from General Higher Education*

A third implication concerns the Defense transfer boundary. General higher-education evidence can provide useful mechanisms and instructional hypotheses, while military and high-stakes professional evidence can clarify issues of uncertainty, reliance, automation bias, challenge mechanisms, and consequential judgement. Defense and governance evidence can establish expectations concerning oversight, assurance, accountability, and professional responsibility.

However, these streams of evidence cannot be considered equivalent. Transfer to Defense Higher Education is justified only if there is sufficient equivalence between the two on issues of population, task structure, authority relations, assessment circumstances, stakes, and consequences. Defense Higher Education must be considered, then, more a context boundary than another AI-literate skill.

7.4 *D. Interpretation of the Toolkit Architecture*

AI Understanding and Limitations, Critical Evaluation and Judgment, Human Agency and Ownership, Reliance Calibration, Verification, and Ethical and Accountable Judgement – This six-dimensional architecture of AI understanding and its limitations, critical evaluation and judgement, human agency and ownership, reliance calibration, verification, and ethical and accountable judgement is an educational design framework, not a causal model validated by empirical research. The dimensions discussed should not at this point be treated as latent factors, as a fixed progression, or even as universal determinants of judgement quality. Instead, these are educational levers that can be tested using content validation, scenario-based performance measures, psychometrics, and replicability across contexts.

7.5 *E. Contradictory Evidence and Measurement Implications*

The findings also show inconsistencies that actually support rather than refute the logic of the toolkit. In self-report measures, individuals may show great levels of confidence, while in objective or scenario measures, they may exhibit a lack of clarity, ceiling effects, and difficulty in distinguishing between AI content and original content. Likewise, in trust literature, the dangers of over-reliance on AI and over-rejection of AI have been observed. Policy papers may outline responsible AI practices, while educational papers may rarely show that the principles mentioned in the policy papers result in learners' behavior. These inconsistencies show that it is not advisable to combine attitudes, knowledge, judgment performance, verification behavior, and accountability enactment into one composite measure. The future evaluation must consider the types of sources from which the findings have been obtained and their inferential distance.

7.6 *F. Practical Implications for Defense Higher Education*

What would appear to be the most justified use of such an AI tool for learning purposes at this point is conducting an unclassified scenario exercise with the learners being provided with an

AI-supported analysis and then being tasked with recognizing the limitations of the system, analyzing the stakes of the task, comparing the AI's statements with relevant literature, etc.

Scenarios should include useful, misleading, incomplete, and ambiguous AI outputs so that learners cannot succeed simply by accepting or rejecting AI consistently. Assessment should focus on reasoning quality, provenance checking, verification, disclosure, escalation, and responsibility statements, rather than response speed, polished language, or agreement with the AI system. At the institutional level, the toolkit also implies alignment between curriculum, assessment, AI-use policy, and assurance arrangements. These are proposed educational implications, not evidence that existing Defence/PME systems already achieve these outcomes.

8 VI. LIMITATIONS

Several limitations should be considered when interpreting the findings of the present study.

First, the evidence synthesis is preliminary rather than a completed systematic or scoping review. Although the discovery search identified 515 unique records and 204 potentially eligible records, full-text eligibility assessment was not completed. The 20 independently verified sources therefore represent a narrative evidence subset rather than a final included-study corpus. Consequently, the present findings should not be interpreted as exhaustive estimates of the Defense Higher Education evidence base. Second, the evidence base is heterogeneous across populations, contexts, and methodological designs.

The existing literature involves general higher education literature, AI literacy frameworks, human factors research, military/professional literature, PME practice literature, and Defence/governance literature. Such sources cannot be considered to be interchangeable in terms of their evidentiary value. General higher education research can be used to guide educational mechanisms, while military/professional research can be used to guide reliance and automation bias mechanisms. However, neither necessarily indicates that learners in the context of Defense have achieved the desired AI literacy competencies.

Lastly, there is no doubt that the problem of measurement is significant. The research on the topic of AI literacy continues to rely on self-report measures to a large degree, and there is practically no performance measure at all. In addition, there is no mention of any Defense-specific AI-literacy scale among the validated ones, while education-focused validation and calibration measures are not unified.

Fourthly, the toolkit architectural structure does not have empirical validation. In fact, the six elements cannot be seen at this stage as independent latent variables, a linear developmental pathway, or a universal predictor of decision quality. Instead, the toolkit is made up of an evidence-based structure of mechanisms that need to be further validated, cognition tested, scenario tested, performance measured, reliability measured, and psychometrically validated.

Lastly, it must be emphasized that the current study is incapable of demonstrating any causation regarding the educational outcomes associated with the use of the proposed toolkit. There is no evidence at all that suggests the implementation of the toolkit can result in increased accuracy of judgments, enhanced verification, better calibration of reliability, and higher levels of accountability behavior. Therefore, the current study should be considered an evidence-informed toolkit framework.

9 VII. FUTURE WORK

Future research needs to shift from the current approach of evidence-informed architecture of the toolkit towards systematic empirical validation. The most important step here is completing the evidence selection process through the full-text assessment of the 204 potentially eligible records, justifying exclusion, and generating the final evidence set and PRISMA-like flow record. The finalized search protocol must be officially registered prior to any analysis of outcome data. The other step is to validate the toolkit constructs and scenarios with experts and learners. First of all, the six proposed constructs need to be reviewed by defense higher education experts and those who know AI literacy. Then cognitive interviews will assess interpretation of scenario instructions, AI responses, evidence needed, and choice of answers by learners.

Third, the validity and reliability of the toolkit need to be determined in future research. It would be important to develop scenario-based tasks that are related to each of the six dimensions. The assessment of performance should involve rubrics as well as, when needed, independent raters. Inter-rater reliability, internal consistency when needed, construct validity, criterion-related validity, and sensitivity to scenario and task reliability need to be determined. Fourth, a validation experiment should be conducted in which the Scenario-Based AI Literacy Toolkit is compared with some appropriate control condition, e.g., conventional orientation or instruction with respect to AI tools. Measures should include scenario-based judgement accuracy, verification quality, proper use, limitations recognition, human agency/ownership, ethical judgement, and accountability. If conditions are appropriate, mediation and moderation analysis can be performed to explore whether verification quality mediates the relationship between AI literacy and judgement accuracy and whether the effect varies depending on AI reliability and/or task stakes. Such analysis needs to be preregistered.

In conclusion, there should be future research into transfer in a longitudinal sense as well as in terms of transfer between contexts. Testing at different times might provide evidence of whether skills acquired through the toolkit are sustained over time as well as whether any improvements achieved during training will transfer into practical Defense Higher Education tasks.

10 VIII. CONCLUSION

In this research, a framework architecture for a Scenario-Based AI Literacy Toolkit in Defense Higher Education was proposed, considering the evidence regarding AI literacy, critical appraisal, human agency, calibration of reliance, verification, and ethical and accountable judgement. It is implied from the synthesis that responsible AI literacy in Defense Higher Education can not be effectively captured through technical AI knowledge and confidence in AI. Rather, responsible behavior should be demonstrated by learners having knowledge and understanding about AI's strengths and weaknesses, critical appraisal of AI-generated information, ownership of decisions, calibration of reliance according to reliability and level of stakes involved, verification of critical assertions, and working in an ethical and accountable context.

However, the greatest benefit of the research lies in the integration of these models into the architecture of scenario-based and observable performance for Defense Higher Education and not their novelty. These models distinguish between foundational knowledge in AI and the decision-making processes that such knowledge should help us achieve, regard dependence on

others as conditional and not just something we increase or decrease, and make verification and accountability observable through the learner's behavior.

However, on the other hand, the results draw definite limits to claims that can currently be made. The preliminary evidence base provides support for the design rationale behind the toolkit but does not provide evidence of its reliability, construct validity, Defense-specific validity, and educational effectiveness. Objective assessment scarcity, lack of a Defense-specific AI-literacy scale among the verified evidence items, fragmented verification and reliance assessment, and the lack of direct evidence on the PME curriculum are important limitations. Any future research should proceed via expert validation, cognitive interviews, scenario piloting, performance-based assessment, inter-rater reliability tests, psychometric validation, comparative experiments, and longitudinal studies. The key empirical challenge is not to prove that Defense learners embrace responsible AI principles but to test whether they use appropriate judgement under conditions of varying AI reliability, uncertainty, and consequences of AI-assisted decision-making. In this respect, the toolkit offers a testable framework for developing responsible AI judgement based on AI-literacy principles in Defense Higher Education.

REFERENCES

[1] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2:100041, 2021.

[2] O. Almatrafi, A. Johri, and H. Lee. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers & Education Open, 6:100173, 2024.

[3] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2):230–253, 1997.

[4] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. Human Factors, 46(1):50–80, 2004.

[5] L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5):991–1006, 1999.

[6] T. Lintner. A systematic review of AI literacy scales. npj Science of Learning, 9:50, 2024.

[7] A. Ali, Z. Farid, A. H. Al-kassem, Z. A. Khan, M. Qamer, K. F. Khan Ghouri, M. Al Sakhnani, and A. M. Momani, "Development and use of artificial intelligence in the defense sector," in 2023 International Conference on Business Analytics for Technology and Security (ICBATS), 2023, doi: 10.1109/ICBATS57792.2023.10111113.

[8] D. Long and B. Magerko. What is AI literacy? Competencies and design considerations. In Proc. 2020 CHI Conf. Human Factors Comput. Syst., pages 1–16, 2020.

[9] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur. Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education, 16:39, 2019.

[10] W. Holmes, M. Bialik, and C. Fadel. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign, Boston, MA, 2019.

[11] H. Chee, S. Ahn, and J. Lee. A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. British Journal of Educational Technology, 56(5):2146–2182, 2025.

[12] E. Kasneci et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103:102274, 2023.

[13] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, et al. ChatGPT and large language models in academia: Opportunities and challenges. BioData Mining, 16:20, 2023.

[14] M. I. I. Alras, A. Q. Emran, and A. M. Mohamed. Ethics and governance of generative AI in education: A systematic review on responsible adoption. Discover Education, 5:37, 2026.

[15] D. Teng, L. Tan, Q. Cao, et al. Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. npj Digital Medicine, 9:356, 2026.

[16] M. H. Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4):577–586, 2018.

[17] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, 30(3):286–297, 2000.

[18] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick. Automation bias: Decision making and performance in high-tech cockpits. International Journal of Aviation Psychology, 8(1):47–63, 1998.

[19] L. Kahn, M. C. Horowitz, and L. Resnick Samotin. What is human in judgment? Comparing automation bias and algorithm aversion between the United States Military Academy and the general public. arXiv:2604.04333, 2026.

[20] P. Lushenko. AI, trust, and the war room: Evidence from a conjoint experiment in the US military. Cambridge Forum on AI: Law and Governance, 1:e52, 2025.

[21] P. Lushenko and R. Sparrow. Artificial intelligence and U.S. military cadets' attitudes about future war. Armed Forces & Society, 2024.

[22] S. Holdsworth, O. Sandri, J. Hayes, and S. Maslen. Scenario-based capability assessment in professional learning contexts. Journal of Adult and Continuing Education, 31(3), 2025.

[23] Ministry of Defense, UK. Defense artificial intelligence strategy. Technical report, UK Government, London, U.K., 2022.

[24] Ministry of Defense, UK. Ambitious, safe, responsible: Our approach to the delivery of AI-enabled capability in defense. Technical report, UK Government, London, U.K., 2022.

[25] U.S. Department of Defense. Responsible artificial intelligence strategy and implementation pathway. Technical report, Dept. of Defense, Washington, DC, 2022.

[26] C. A. Pfaff, C. J. Lowrance, B. M. Washburn, and B. A. Carey. Trusting AI: Integrating artificial intelligence into the army's professional expert knowledge. Technical report, US Army War College, Strategic Studies Institute, Carlisle, PA, 2023.

[27] A. Ali, H. T. Sheik, A. Raza, T. Hanif, S. G. Virk, and H. Riaz, "MilGPT: Secure and explainable large language model framework for military applications," in Proc. 11th International Conference on Next Generation Computing (ICNGC), 2025, pp. 65–68. Available: https://www.earticle.net/Article/A478462.

Declarations

Funding. [Funder name and grant number, or a statement that no external funding was received.]

Conflicts of interest. [Relevant interests, or: The authors declare no conflicts of interest.]

Ethics and consent. [Approval reference / not applicable statement.]

Data availability. [Repository / access conditions, or not applicable.]

Code availability. [Repository / version, or not applicable.]

Author contributions. [CRediT roles and initials.]

Acknowledgments. [Optional. Delete if none.]

Use of generative AI. [Disclose AI tool use, or state that none was used.]

How to Cite This Article

SUGGESTED CITATION

J. Che Mustapha et al., "Designing a Scenario-Based AI Literacy Toolkit for Defense Higher Education," International Journal of Security, Intelligence, Governance & Modern AI Computing (IJSIGMA), vol. 1, no. 1, pp. 01–33, 2026, doi: [DOI to be assigned].