

Trustworthy AI Literacy in Primary Education: A Human-Centred Learning and Measurement Design

Heidrun Mühle¹ and Shaily Gandhi¹

¹ Interdisciplinary Transformation University Austria (IT:U), Linz, Austria

ABSTRACT

Generative foundation models can produce plausible outputs without guaranteeing factual or relational correctness. Primary-school learners therefore need accessible ways to translate trustworthy-AI principles into evidence-based judgements and responsible use decisions. This conceptual design paper presents a human-centred learning and measurement design within the Austrian TRUST-AI4Schools project. Human-centredness is operationalised through learner agency, age-appropriate accessibility, evidence ownership, and human authority over consequential use. Using a design-science-informed process, the paper describes prior reference construction: learners establish and validate an inspectable reference before encountering a generated transformation, then compare selected relations and justify whether the output should be used, revised, further checked or withheld for a stated purpose. The design addresses selected aspects of trustworthy AI literacy through verification, provenance awareness and human-controlled use decisions. An author-generated image illustrates the task context. Classroom feasibility and learning effects remain to be evaluated. A proposed augmented reality extension will investigate whether spatially linked evidence supports independent judgement without replacing it with interface feedback.

Keywords: Trustworthy AI, Foundation models, AI literacy, Primary education, Augmented reality

INTRODUCTION

AI literacy combines understanding AI with evaluating its outputs and considering consequences of use (Long and Magerko, 2020). Reviews address ethical judgement (Ng et al., 2021) and generative AI (Park, 2025). The primary-school design problem is to make these competencies observable without requiring independent online research. Reviews describe diverse approaches and assessments instead of establishing an instructional solution (Casal-Otero et al., 2023; Yim and Su, 2025).

TRUST-AI4Schools is an Austrian FFG-funded project developing age-adapted approaches to understanding foundation models and using generative AI responsibly (FFG, 2026). This paper specifies the project's primary-school workshop design to help learners evaluate AI outputs and justify their use.

Existing AI literacy approaches commonly teach learners to recognise limitations, identify errors, or reflect on responsible AI use (Casal-Otero et al., 2023; Dangol et al., 2025; Vartiainen et al., 2025; Yim and Su, 2025). When evaluation begins only after generation, a plausible output may itself influence learners' judgement.

We address this risk through prior reference construction, in which learners establish and validate the evidence against which an AI output will later be judged before encountering that output. This separates reference formation from AI evaluation and creates an observable pathway from evidence to judgement to use decision.

The paper proposes prior reference construction as an age-adapted instructional sequence for trustworthy AI literacy. Children first construct and validate an inspectable spatial reference before evaluating a generative output. The design is instantiated with Linz: learners position selected physical landmarks on a city map, validate the arrangement against supplied authoritative source material, document the reference and then request a LEGO-style AI representation. They compare selected identities and spatial relations against the pre-existing reference and justify whether the output should be used, revised, further checked or rejected for a stated purpose. Linz provides a concrete spatial example. The sequence is intended to be applicable to other supplied-reference tasks, such as scientific diagrams, timelines or charts, although such transfer has not yet been demonstrated. This paper addresses the following research question: How can prior reference construction support observable and assessable trustworthy AI literacy through evidence-based verification and justified use decisions in primary education?

This paper makes three contributions: it specifies prior reference construction as an instructional sequence; translates selected trustworthy-AI requirements into observable learner actions; and proposes a measurement framework that distinguishes model understanding, provenance awareness, evidence judgement, justified use and transfer.

CONCEPTUAL FOUNDATIONS AND RELATED WORK

Foundation models, plausibility and evaluation scope

Foundation models are trained on broad data and can support multiple downstream tasks (Bommasani et al., 2021). This workshop concerns generative image applications. It distinguishes prompting from training and introduces learned patterns, inputs and outputs through age-adapted explanation (Touretzky et al., 2019). The city task illustrates verification of generated outputs against an external reference. Compositional benchmarks show that object presence and relational correctness are distinct evaluation targets: T2I-CompBench evaluates attribute binding and spatial composition (Huang et al., 2023). GenEval examines co-occurrence, position, count and colour (Ghosh et al., 2023). Wang et al. (2026) extend spatial evaluation to perception, reasoning and interaction. These studies motivate relation-level checking in the workshop.

Trustworthy AI literacy and tangible reference construction

Trustworthy-AI requirements concern systems, their development, deployment and use, not a learner assessment scale (HLEG, 2019). The OECD and European Commission (2026) and UNESCO frameworks address learner competencies

(Miao et al., 2024). Here, trustworthy AI literacy means applying selected requirements to output use. Verification, provenance awareness and human-controlled decisions instantiate transparency and oversight, without establishing overall system trustworthiness. In this study, human-centredness is operationalised through four characteristics: learner agency, age-appropriate accessibility, evidence ownership, and human authority over consequential use. The design aims to support learner agency by involving children in constructing, validating and evaluating the reference, accessibility through visual, icon-based and colour-coded scaffolds, evidence ownership by learners recording their own judgements before group discussion, and human authority by requiring a purpose-specific decision to use, revise, further check or withhold an AI output.

PopBots exemplifies hands-on AI education for young children (Williams et al., 2019). AI Puzzlers investigated how children aged six to eleven identify generative-AI errors in visual reasoning tasks (Dangol et al., 2025), while Vartiainen et al. (2025) found that, following workshops combining machine-learning activities and text-to-image bias exploration, fourth- and seventh-graders more frequently explained biased images in terms of imbalances in training data. These studies motivate hands-on exploration, error identification and critical reflection. The present design combines an independently validated reference established before AI exposure with evidence judgement and a purpose-specific use decision. Marshall (2007) and Byrne et al. (2023) caution against treating physical interaction as sufficient for learning. Before viewing the AI-generated image, children position landmark models on a source-checked map of Linz. This arrangement provides a visible reference for checking selected spatial relationships in the generated image.

Prior reference construction builds on established educational practices. Worked examples and self-explanation link tasks to explicit references and reasons (Chi et al., 1989). Source evaluation examines the origins and credibility of evidence (McGrew et al., 2018). Verification before reliance and initial judgements before AI advice are related approaches, including cognitive forcing (Buçinca et al., 2021). Unlike precommitment to an answer, the present sequence records validated reference facts before the output is seen. Evidence-based reasoning already links claims to supporting evidence (Sandoval and Millwood, 2005). The contribution is their integration into a learner-constructed spatial reference, external validation, output comparison and a purpose-specific use decision. It does not establish a new cognitive mechanism or an advantage over an equivalent supplied reference.

DESIGN METHODOLOGY

The present study is a conceptual design specification informed by design science and educational design research (Peffers et al., 2007; McKenney and Reeves, 2021).

Figure 1 summarises the design process in six stages: problem identification, literature and framework synthesis, derivation of design requirements, instructional artefact design, measurement specification and planned formative

evaluation. Formative evaluation will assess feasibility and guide subsequent revisions before any learning-effect study. Accordingly, the paper evaluates the coherence and inspectability of the proposed design.

The intended learners are children aged 8–10 years in Austrian primary-school grades 3–4, working in groups of three or four in facilitator-led classroom sessions. The planned sample comprises ten classes of 25 children (250 in total). Cycle 1 will involve two classes. Following revision, Cycle 2 will involve eight different classes. Each workshop is planned to last 90 minutes: introduction and opening assessment (15 minutes), placement and validation (30), a break (5), image generation and explanation (10), individual comparison and group discussion of use decisions (20), and closing assessment (10). Cycle 1 will test this schedule.

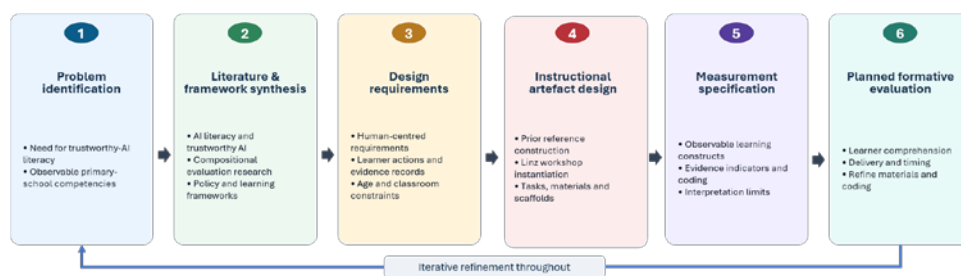


Figure 1. Design process for the human-centred learning and measurement design.

HUMAN-CENTRED DESIGN REQUIREMENTS

The human-centred orientation is expressed through four design requirements. Learner agency requires children to participate in constructing and evaluating the evidential baseline. Age-appropriate accessibility requires alternatives to text-heavy interaction, including visual and tangible scaffolds. Evidence ownership requires learners to record their own evidence judgements before collective discussion. Human authority over consequential use requires the final decision about whether an AI output is suitable for a stated purpose to remain with the learner rather than the system. Prior reference construction operationalises these requirements by placing independently established human evidence and learner judgement before reliance on the generated output.

DESIGN PRINCIPLE: PRIOR REFERENCE CONSTRUCTION

The central design principle of this paper is that the reference precedes the AI output. Prior reference construction is an instructional sequence in which learners establish an independently supported reference state before exposure to an AI-generated representation and subsequently evaluate the output against that reference. Separating reference construction, external validation, generation, comparison and use decision makes the evidential basis inspectable before learners encounter a generated representation. The principle may extend to other modalities. The Linz spatial task is the current design instantiation.

Materials and reference status

Materials comprise an oriented Linz map, five landmark tokens, visual source cards, a camera and a display. To accommodate differences in reading proficiency within this age group, landmark tokens and source cards incorporate child-friendly visual aids, distinct colour-coded symbols, and clear icons alongside text labels. This implements the age-appropriate accessibility requirement of the human-centred design. The Trinity Column, Linz Castle, the New Cathedral, Ars Electronica Center and Lentos can use simplified 3D models or colour-coded, icon-marked blocks. The Danube, Hauptplatz and Nibelungen Bridge provide base-map orientation cues.

To maintain feasibility within the workshop placement phase, small groups of three or four position the landmarks by pairing icon-marked tokens with matching visual source cards. Children with limited reading proficiency can use distinct colour codes and icon graphics to interpret spatial relationships, such as river-bank orientation, without extensive textual decoding. In the validation phase, a scaffolded peer-and-facilitator check replaces detailed moderator pre-determination: learners first compare their layout with a peer group's placement before the facilitator checks key reference points with the supplied City of Linz map and source cards (City of Linz, n.d.-a). Only layouts that match the pre-specified landmark identities and selected spatial relations are approved as the reference. This structured verification is intended to support manageable timeframes and student agency in establishing the baseline.

A photograph from above, a landmark legend and three agreed spatial relations document the approved reference. A consistent display orientation reduces perspective changes. The reference is both generation input and a record for checking the transformation. Its evidential independence comes from prior sourcing and validation against the supplied City of Linz map and source cards, as opposed to being withheld from the model.

Table 1. Proposed workshop sequence

Stage	Learner action	Retained record
Locate	Position five landmarks using the map and source cards.	Initial arrangement
Validate	Check and correct positions with the facilitator before generation.	Approved photo and sources
Generate	Inspect the approved photograph and prompt, then generate a LEGO-style image.	Exact inputs, service and output
Compare	Judge selected relations against the unchanged reference.	Classification and evidence
Decide	Justify use, revision or withholding for a stated purpose.	Decision, reason and reviewer

Generation, comparison and use decisions

The workshop facilitator submits the approved photograph and landmark legend to an institutionally approved image-generation service, with Google Gemini and OpenAI ChatGPT planned as candidate services.

Facilitator-controlled submission limits children's direct interaction with external services and keeps data handling, prompt settings and generation conditions consistent, while learners remain responsible for inspecting the inputs, evaluating the output and deciding how it should be used. The prompt requests a LEGO-style transformation that retains the five landmarks, their positions, river-bank relations and viewing orientation.

It permits stylistic simplification but not rearrangement. The class checks the inputs before transmission. The task tests preservation of specified relations instead of architectural photorealism. A short explanation distinguishes prior training, present inputs and output checking.

Text-only generation requests geographic correspondence without supplying the class's layout. Image-conditioned generation also requires input fidelity. These conditions must be documented and analysed separately. Records include the prompt, input files, service, available model identifier, date, settings, tool or retrieval use and output. A coherent image does not establish that a verified map was consulted or that every constraint was preserved.

Learners compare three preselected features or relations. Each judgement is classified as supported, contradicted or not decidable from the available evidence. For example, a landmark shown on the same river bank as in the validated reference is supported; a landmark placed on the opposite bank is contradicted; and a landmark obscured or outside the frame is not decidable. A contradiction requires identifiable conflicting evidence. A church-like shape is not assigned a particular identity without sufficient support. Viewpoint ambiguity and stylistic simplification must not be scored as geographic errors. Relative distance and direction may be selected among these three relations where the reference supports them.

Learners record their evidence judgements and proposed use decisions independently on individual recording sheets prior to group discussion. This is intended to support learner agency and evidence ownership and limit peer influence and dominant-participant effects on individual measurement. The class then considers fictional illustration and factual city information as distinct uses. Learners justify use, revision, further checking or withholding. An AI label does not correct an inaccurate location, and revisions require another check. The original image and reference remain fixed. Annotations are separate records, not regenerated substitutes.

Selected trustworthy AI requirements are operationalised as age-adapted learner practices. Transparency is addressed through provenance identification: learners distinguish the source material, validated class reference, generation inputs and AI output. Human agency and oversight are addressed through purpose-specific decisions to use, revise, further check or withhold the output. Technical robustness is addressed at the learner level through relation-level checking against the reference rather than through a system audit. Accountability is represented by evidence-linked judgement and decision records.

Privacy and data governance are addressed procedurally through facilitator-controlled use of an institutionally approved service and inspection of the transmitted inputs. These mappings operationalise selected requirements for learning.

ILLUSTRATIVE MATERIAL AND PROVENANCE

Figure 2 provides the illustrative image that motivated the present design. In classroom use, children compare the generated output with the validated map-based reference state constructed before generation.



Figure 2. AI-generated Linz illustration, 11 September 2026. The request asked for an image of Linz in LEGO style.

Recognisable elements coexist with missing, altered or spatially inconsistent features. Its function is to motivate the workshop's evidence task, as opposed to quantifying model performance. In the classroom, the decisive comparison is made against the learner-constructed map reference, checked against the City of Linz's 3D city model before generation (City of Linz, n.d.-b). Learners therefore judge selected relations as supported, contradicted or not decidable and connect that judgement to a purpose-specific use decision.

MEASUREMENT AND PLANNED FORMATIVE EVALUATION

The proposed measurement framework distinguishes five assessment targets: model understanding, provenance awareness, evidence judgement, justified use and transfer. Each target is linked to an observable learner action and a retained record, while interpretation boundaries prevent the indicators from being treated as a validated composite scale. Trust as an attitude is distinguished from reliance as behaviour (Lee and See, 2004). Hoff and Bashir (2015) further distinguish influences on trust across contexts. Appropriate-reliance measures require information about whether people follow correct or incorrect AI advice (Schemmer et al., 2023). The proposed task therefore measures evidence-informed use decisions, not calibrated trust or comprehensive reliance.

Table 2. Proposed observable learning constructs, indicators and interpretation

Target	Observation and coding	Interpretation boundary
Model understanding	Explain the difference between a pretrained capability, the current input and checking an output.	Not architecture knowledge or proof of a training mechanism.
Provenance awareness	Identify the map source, class reference and AI output using their documented origins.	Not visual AI-detection ability.
Evidence judgement	Classify specified relations and identify relevant evidence or an evidence gap.	Separate correct categories from unsupported guesses.
Justified use	Link use, revision or withholding to purpose, evidence and unresolved points.	Not a stable trust trait or a system audit.
Transfer (later evaluation)	Apply the routine to another supplied scene and reference.	Does not establish transfer to every modality or age group.

Opening and closing tasks will use two supplied-reference forms intended to be parallel. Each form contains six short claims about spatial relationships: two supported, two contradicted and two not decidable from the supplied reference. Children will classify each claim and identify relevant evidence or an evidence gap. Form order will be counterbalanced as AB or BA within each class.

Items will be matched by relation type, number of landmarks, vocabulary, sentence length, visual complexity and evidence gaps. Teacher review and cognitive interviews will check comprehension and unintended cues. Pilot accuracy, omissions and response times will identify differences in difficulty and guide revision. Parallel-form equivalence is a design aim, not an established measurement property.

Researcher-written claims will be identified as assessment items instead of spontaneous model statements. Oral responses and pointing will reduce writing demands. Prior familiarity with Linz and basic map comprehension will be recorded. These are planned procedures and provisional measures. The shared-reference assessment is distinct from the collaboratively constructed workshop artefact. Brief oral questions at opening and closing will assess model understanding and provenance awareness, with responses coded as correct, partly correct or incorrect against the distinctions specified in Table 2. Analysis will report category-specific accuracy and a three-category confusion matrix. Reasons will be coded separately as absent, relevant but insufficient, or sufficient to support the judgement. Missing responses will remain distinct from not-decidable judgements. After rubric revision, the same two coders will independently code all opening and closing responses from five randomly selected pupils in each Cycle 2 class (40 pupils, 20% of that cycle), covering both forms and all evidence categories. Group decisions will not be attributed to every member. Assistance, timings, source versions and selection of archived outputs will be logged. Optional confidence ratings remain exploratory. Formative evaluation will proceed in stages.

Provisional agreement targets are at least 85% exact agreement and Cohen's kappa of at least .70, calculated before adjudication. Unweighted kappa will be

used for nominal categories and weighted kappa for ordered scores. Disagreements and category prevalence will be examined. Inadequate agreement will trigger rubric revision and further calibration. These are proposed project criteria, not validated cut-offs for this task.

The first cycle will examine whether children understand the instructions, evidence categories and recording formats and whether the task is usable within the planned time. It will also inform rubric development and training of the two coders. The second, using revised materials, will examine delivery fidelity and coding consistency. Cycles will be reported separately. Progression to a later learning-effect study will require all essential phases to be completed within 90 minutes and attainment of the proposed coding-agreement targets. The same instruction-related misunderstanding in at least two independently working groups, essential phases skipped or curtailed by time constraints, or items with multiple defensible classifications will trigger revision and retesting. Incorrect answers alone will not establish instruction-related misunderstanding. Persistent unequal form difficulty or unplanned interpretive assistance will also require revision and further formative testing. Progression will not depend on favourable pre/post gains. Correct live outputs will be retained rather than treated as failed demonstrations; archived examples, when used, will retain documented selection histories and generation conditions. Repeated generation until an error appears cannot estimate an error rate. Research collection will require institutional ethics review, parental consent, child assent and protected records. Later learning-effect studies will need appropriate comparison conditions, sufficient power and analysis accounting for classroom grouping.

DISCUSSION

The contribution of this design is an instructional sequence that places learner judgement at the centre of output evaluation. The instructional contribution is the translation of selected trustworthy-AI requirements into age-appropriate practices of provenance checking, relation-level verification and human-controlled use decisions. Practical feasibility in primary classrooms is intended to be supported by visual scaffolding and peer-facilitator checks, with a proposed target of 30 minutes for setup and validation. Building on hands-on exploration and error-identification activities (Dangol et al., 2025; Williams et al., 2019), this design establishes an explicit, inspectable reference baseline prior to generation. In contrast to framework models that define general literacy competencies (Long and Magerko, 2020; Miao et al., 2024; OECD and European Commission, 2026), our task operationalises trustworthy AI requirements into observable verification actions, provenance checks, and human-controlled decisions.

The proposed design links an independently established reference, observed output features, and consequential use choices. Alongside approaches addressing bias awareness through model training and output exploration (Vartiainen et al., 2025), this design fixes the evidential baseline before AI exposure. This sequencing is intended to reduce the risk that a plausible AI output becomes its

own reference point and to make the learner's evidence-to-decision pathway inspectable.

While literature cautions against assuming physical interaction alone guarantees learning (Byrne et al., 2023; Marshall, 2007), our framework uses tangibles specifically to fix reference facts before AI exposure.

Future evaluation must isolate prior reference construction from alternative factors such as map literacy or instruction time to test whether it adds value over an equivalent supplied reference.

FUTURE WORK

A later extension could test whether tablet-based augmented reality (AR) can link documentary evidence to physical landmark tokens through spatial registration (Radu and Schneider, 2023). AR would not reveal model reasoning or provide a new ground truth, and learners would make an initial judgement before receiving interface feedback. Any evaluation should hold sources, outputs, instruction and learning time constant while varying spatial registration, and should include an unaided transfer task to distinguish interface-supported performance from independent learning.

Reference data and physical anchors would require independent validation. Object-linked excerpts would be preferred to overlaying the generated city as though it were a georeferenced map. Ambiguous landmark correspondences would remain unresolved, and registration errors would be recorded separately from generation errors.

CONCLUSION

This paper specifies a human-centred learning and measurement design in which children establish and validate a reference before evaluating a generative output against it. Prior reference construction separates evidence formation from AI-output evaluation with the aim of preventing visual plausibility from becoming the basis for correctness. The design operationalises selected trustworthy-AI principles through provenance awareness, evidence-based verification, human oversight and justified use decisions, and proposes five assessment targets with observable indicators. Classroom feasibility, learning effects and transfer beyond the Linz spatial task have not yet been demonstrated and are the focus of planned formative and later comparative studies.

Future work may additionally test whether spatially registered AR can strengthen the link between physical reference, documentary evidence and learner judgement without replacing learner agency.

ACKNOWLEDGEMENTS AND AI ASSISTANCE

This work forms part of TRUST-AI4Schools, supported by the Austrian Research Promotion Agency (FFG) through Talente regional 2025. ChatGPT (OpenAI) and Gemini (Google) assisted with image generation, language editing and grammar correction. The authors retain responsibility for all claims, references and final wording.

REFERENCES

- Bommasani, R. et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. doi:10.48550/arXiv.2108.07258.
- Buçinca, Z., Malaya, M. B. and Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. doi:10.1145/3449287.
- Byrne, E. M., Jensen, H., Thomsen, B. S. and Ramchandani, P. G. (2023). Educational interventions involving physical manipulatives for improving children's learning and development: A scoping review. *Review of Education*, 11(2), e3400. doi:10.1002/rev3.3400.
- Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B. and Barro, S. (2023). AI literacy in K-12: a systematic literature review. *International Journal of STEM Education*, 10, 29. doi:10.1186/s40594-023-00418-7.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P. and Glaser, R. (1989). Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science*, 13(2), 145–182. doi:10.1207/s15516709cog1302_1.
- City of Linz (n.d.-a). Der Linzer Stadtplan. <https://www.linz.at/stadtentwicklung/stadtplan.php> (accessed 12 September 2026).
- City of Linz (n.d.-b). 3D Stadtmodell Linz. <https://3d.linz.at/> (accessed 14 September 2026).
- Dangol, A., Zhao, R., Wolfe, R., Ramanan, T., Kientz, J. A. and Yip, J. (2025). “AI just keeps guessing”: Using ARC Puzzles to Help Children Identify Reasoning Errors in Generative AI. *Proceedings of the 24th Interaction Design and Children*, 444–464. doi:10.1145/3713043.3728836.
- FFG (2026). TRUST-AI4Schools: Foundation Models verstehen & nachhaltig vertrauenswürdig verwenden. FFG Project Database. <https://projekte.ffg.at/projekt/3d3f453e-f36b-1410-8652-00d3c07d99bb> (accessed 12 September 2026).
- Ghosh, D., Hajishirzi, H. and Schmidt, L. (2023). GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. *Advances in Neural Information Processing Systems*, 36, 52132–52152. arXiv:2310.11513.
- High-Level Expert Group on Artificial Intelligence (HLEG) (2019). Ethics Guidelines for Trustworthy AI. Publications Office of the European Union. doi:10.2759/346720.
- Hoff, K. A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. doi:10.1177/0018720814547570.
- Huang, K., Sun, K., Xie, E., Li, Z. and Liu, X. (2023). T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. *Advances in Neural Information Processing Systems*, 36, 78723–78747.
- Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. doi:10.1518/hfes.46.1.50_30392.
- Long, D. and Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of CHI 2020*, 1–16. doi:10.1145/3313831.3376727.
- Marshall, P. (2007). Do tangible interfaces enhance learning? *Proceedings of TEI 2007*, 163–170. doi:10.1145/1226969.1227004.
- McGrew, S., Breakstone, J., Ortega, T., Smith, M. and Wineburg, S. (2018). Can Students Evaluate Online Sources? Learning From Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, 46(2), 165–193. doi:10.1080/00933104.2017.1416320.
- McKenney, S. and Reeves, T. C. (2021). Educational design research: Portraying, conducting, and enhancing productive scholarship. *Medical Education*, 55(1), 82–92. doi:10.1111/medu.14280.

- Miao, F., Shiohira, K. and Lao, N. (2024). AI Competency Framework for Students. UNESCO, Paris. <https://unesdoc.unesco.org/ark:/48223/pf0000391105> (accessed 12 September 2026).
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W. and Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. doi:10.1016/j.caeai.2021.100041.
- OECD and European Commission (2026). Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education. OECD Publishing, Paris. doi:10.1787/65cd27d4-en.
- Park, J. (2025). A systematic literature review of generative artificial intelligence (GenAI) literacy in schools. *Computers and Education: Artificial Intelligence*, 9, 100487. doi:10.1016/j.caeai.2025.100487.
- Peffer, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. doi:10.2753/MIS0742-1222240302.
- Radu, I. and Schneider, B. (2023). How augmented reality (AR) can help and hinder collaborative learning: A study of AR in electromagnetism education. *IEEE Transactions on Visualization and Computer Graphics*, 29(9), 3734–3745. doi:10.1109/TVCG.2022.3169980.
- Sandoval, W. A. and Millwood, K. A. (2005). The Quality of Students' Use of Evidence in Written Scientific Explanations. *Cognition and Instruction*, 23(1), 23–55. doi:10.1207/s1532690xc2301_2.
- Schemmer, M., Köhl, N., Benz, C., Bartos, A. and Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *Proceedings of IUI 2023*, 410–422. doi:10.1145/3581641.3584066.
- Touretzky, D., Gardner-McCune, C., Martin, F. and Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9795–9799. doi:10.1609/aaai.v33i01.33019795.
- Vartiainen, H., Kahila, J., Tedre, M., López-Pernas, S. and Pope, N. (2025). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New Media & Society*, 27(9), 5342–5368. doi:10.1177/14614448241252820.
- Wang, Z., Hu, X., Wang, Y., Xiong, F., Zhang, M. and Chu, X. (2026). Everything in Its Place: Benchmarking Spatial Intelligence of Text-to-Image Models. *International Conference on Learning Representations*. arXiv:2601.20354v2. doi:10.48550/arXiv.2601.20354.
- Williams, R., Park, H. W., Oh, L. and Breazeal, C. (2019). PopBots: Designing an artificial intelligence curriculum for early childhood education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9729–9736. doi:10.1609/aaai.v33i01.33019729.
- Yim, I. H. Y. and Su, J. (2025). Artificial intelligence literacy education in primary schools: a review. *International Journal of Technology and Design Education*, 35, 2175–2204. doi:10.1007/s10798-025-09979-w.