Level 1: Foundations | All Librarians

Module 05

Critical evaluation of AI output

Librarians have been teaching critical evaluation of sources for decades. AI does not require a new framework; it requires applying what you already know to a new type of source. This module bridges that connection.

By Yulia Brusova

25 min read · Reviewed July 2026

WHAT YOU'LL BE ABLE TO DO

1. Apply at least three verification strategies to AI generated text
2. Identify the claim types most likely to be hallucinated and explain why each is high risk
3. Connect AI critical evaluation to the six frames of the ACRL Information Literacy Framework
4. Explain AI evaluation to a student or patron using a live demonstration approach
5. Identify at least two categories of systematic bias in AI output and their implications for library practice
6. Develop a personal checklist for evaluating AI output tailored to your daily work context
7. Describe the difference between fact checking an AI generated claim and verifying a citation
8. Apply a calibrated trust framework to distinguish lower risk from higher risk AI outputs
9. Apply the RACBAC framework to an AI generated output, and explain why declining to use AI for a given task is itself an expression of AI literacy

> *Critical evaluation of information sources is foundational to library practice, and it is the professional competency that transfers most directly to working with AI. The questions librarians have always brought to sources apply here without modification: Who produced this, and by what process? What are the structural limitations of how it was created? What kinds of errors does this source type characteristically make? Can I verify the specific claims it makes? What might it systematically omit? The source type is new; the professional reasoning is not. In order to apply that reasoning effectively, however, a librarian needs a clear understanding of the specific failure modes of AI generated content, because the hallmarks of AI errors are distinct from the errors produced by other source types, and recognizing them requires deliberate familiarity.*

What hallucinations look like in practice

Hallucinated content has a characteristic signature that librarians, once they learn to recognize it, begin to see reliably: it is specific and confident. A general claim such as "AI has transformed academic research workflows" carries low hallucination risk. A specific claim with named details attached, such as a citation with journal name, volume, issue, page range, and DOI; a statistic with a precise percentage attributed to a named report; a quote attributed to a specific person with a publication date: all of these carry substantially higher risk. Such specificity is exactly what language models are trained to produce, because specificity is what confident, authoritative text looks like. The model predicts the tokens that follow a seemingly plausible claim, and those tokens often include seemingly plausible bibliographic details that do not correspond to real documents.

The practical implication is that the most dangerous AI outputs in a library context are the ones that look the most authoritative. For example, a hallucinated citation does not announce itself as fabricated. It includes a seemingly realistic journal name from the appropriate field, a publication year within the expected range, an author name with appropriate disciplinary credentials, and a volume and issue number that fall within the journal's publication history. The DOI may be syntactically correct but resolve to nothing, or may resolve to a different article entirely. Such a fabrication is indistinguishable from a real citation until it is verified, which is precisely why verification cannot be optional.

Several categories of claim warrant consistent verification regardless of how confident the AI response appears:

- **Specific statistics and percentages**: numbers are among the most commonly hallucinated details; the model has learned that quantitative claims carry authority and will produce seemingly plausible figures
- **Named citations with full bibliographic details**: the most frequent and consequential hallucination type in library and research contexts
- **Quotes attributed to specific people**: the model may correctly attribute a general position to a person while fabricating the specific words
- **Historical dates and specific events**: particularly vulnerable to subtle errors that are difficult to catch without direct verification
- **Policy and legal claims**: training data may be outdated, and the consequences of acting on incorrect policy information in a professional context are significant

By contrast, certain AI outputs carry lower hallucination risk: structural recommendations, tone and style guidance, brainstorming lists, and summaries of documents the librarian has provided are all lower risk categories because they are not factual claims about the world, or because the model is working with provided content rather than generating from training data. In order to build reliable AI evaluation instincts, it is useful to internalize this risk gradient and apply it automatically when reviewing AI outputs.

Verification strategies for AI generated claims

There is no shortcut to verification, but there are strategies that make it efficient enough to be sustainable in daily professional practice. In order to verify AI output systematically without treating every use of AI as a research project, librarians need a small set of reliable strategies that can be applied quickly to the specific claim types most likely to be wrong.

The most important strategy is citation verification before claim verification. If AI produces a citation, the first question is not whether the claim the citation supposedly supports is true; it is whether the cited document exists and says what the AI says it says. This two step sequence is more efficient than attempting to verify the underlying claim through other means, because if the citation is fabricated, the claim itself may also be unreliable and should be sourced independently regardless. For example, a librarian who receives an AI generated paragraph asserting that "Smith and Jones (2021) found that 73% of academic libraries report..." should immediately run that citation through a library database before engaging with the 73% figure at all. If the Smith and Jones article does not exist, the figure requires independent sourcing. If it does exist, the article itself should be consulted to confirm that it actually contains that finding.

The second strategy is asking AI to characterize its own uncertainty, which is more reliable than it might initially appear. Asking "How confident are you in this specific claim? What are you uncertain about? Where might I find a primary source for this?" will often produce a useful acknowledgment of uncertainty, not because the model has access to a confidence metric it would otherwise conceal, but because such questions are answered through the same prediction process as everything else, and the model has learned that acknowledging uncertainty is the appropriate response in certain conversational contexts. Such responses should be treated as leads for further investigation rather than as definitive assessments of reliability, but they are genuinely useful for identifying which specific elements of a complex response deserve the most verification effort.

The third strategy is triangulation with authoritative sources. For any factual claim that will be used in professional work (advice to a patron, content in a LibGuide, information in a

committee report), verification through at least one authoritative source that the librarian can cite independently is a professional standard, not an optional precaution. For example, if AI reports that a particular database covers a specific range of years or disciplines, the library's own database documentation or the vendor's platform page is the authoritative source, and that is where the claim should be confirmed. Such triangulation takes additional time and is the reason that AI does not actually accelerate research workflows in which accuracy is non negotiable; it accelerates drafting and thinking workflows, which is a different and genuine value.

Fact checking AI output against authoritative sources

Fact checking AI generated content is a distinct activity from verifying a citation, and the distinction matters practically. Verifying a citation establishes whether a document exists and whether it says what AI claims. Fact checking a claim establishes whether the underlying assertion is accurate, independently of what AI said and independently of whether AI cited anything in support of it. Both activities are necessary in different circumstances, and confusing them leads to verification gaps.

For factual claims that AI makes without citation support (figures, dates, characterizations of policy, descriptions of professional standards), the appropriate fact checking process begins with identifying what type of authoritative source would be definitive for that category of claim. For example, a claim about ALA policy should be verified against ALA's own published documents. A claim about ACRL subcompetencies should be verified against the ACRL AI Competencies document itself. A claim about federal law should be verified against the actual statutory text or a recognized legal resource. A claim about database coverage should be verified against the database vendor's documentation. Such source type matching is precisely the professional skill that librarians teach students under the heading of source authority, and it applies without modification to AI fact checking.

The most common fact checking error among librarians beginning to use AI is using AI to fact check AI. Asking a different AI tool whether a claim produced by the first AI tool is accurate does not constitute verification. For example, asking Claude whether a statistic produced by ChatGPT is correct may produce a confident confirmation, a confident contradiction, or an acknowledgment of uncertainty, none of which constitutes verification, because neither model has access to the authoritative source. Verification requires consulting the primary or authoritative source directly. Additionally, web search is not automatically sufficient for verification; search results may themselves draw on AI generated content or on sources that repeat an inaccurate claim widely enough to make it appear credible. In order to verify AI

generated factual claims reliably, the authoritative source must be consulted, not merely a source that agrees with the AI.

When to trust and when to verify: calibrating professional skepticism

Treating every AI output as requiring full verification is not professionally sustainable. A librarian who must independently verify every sentence of every AI drafted email will abandon the workflow within a week, not out of carelessness but because the overhead eliminates the value. In order to use AI effectively and responsibly, practitioners need a calibrated approach that applies verification effort proportionately to actual risk, not uniformly to all outputs.

Such calibration rests on two dimensions: the stakes of the output and the claim type. Stakes refer to the consequences of an error. A hallucinated figure in a draft email to a colleague for internal planning purposes carries lower stakes than the same figure in a LibGuide accessible to hundreds of students or in a committee report that will inform a budget decision. The same AI error, in different contexts, requires different verification responses. Claim type, as discussed in the prior section, refers to the structural risk profile of specific kinds of AI generated content; citations, statistics, and quotes carry higher risk than structural recommendations and editing.

The practical result of applying both dimensions is a tiered approach to verification. High stakes outputs with high risk claim types (any AI generated content that will be published, presented to administration, used to advise a patron on an important decision, or incorporated into instruction materials) require verification of all factual claims through authoritative sources before use. Medium stakes outputs with mixed claim types such as internal planning documents, draft materials that will be reviewed before use, and brainstorming artifacts require verification of specific high risk elements while lower risk elements can be accepted provisionally. Low stakes outputs composed of lower risk claim types such as structural advice for a draft document, tone and style suggestions, and reorganization recommendations can generally be evaluated on their professional merits without independent source verification, though the librarian's own professional judgment remains the final filter.

In order to develop genuine calibration rather than a mechanical checklist, it is useful to develop the habit of identifying, before using AI for any task, what category of output is being produced and what the consequences of an undetected error would be. Such explicit risk

assessment takes approximately ten seconds and substantially changes the verification behavior that follows. Additionally, the calibration should be communicated explicitly when working with colleagues who are newer to AI, because the instinct to either fully trust or fully distrust AI outputs is common among beginners, and neither instinct produces responsible practice.

Three layers of discernment

Dakan and Feller introduce a three-layer model of evaluation they call discernment, one of the most practically useful frameworks in the AI Fluency course for library professionals (Dakan & Feller, AI Fluency: Framework & Foundations, Anthropic Academy, 2025). The first layer is **product discernment**: is the output factually accurate, appropriate for the audience, and actually responsive to what was asked? This is the layer most librarians begin with, and it is necessary - but not sufficient on its own. The second layer is **process discernment**: how did the AI arrive at this output? This layer matters most in longer or more complex exchanges, where subtle problems accumulate. For example, the AI may quietly introduce an assumption partway through a response, build subsequent reasoning on that assumption, and produce a plausible-looking conclusion that is wrong from the second paragraph forward. Such errors do not surface in a quick read; they require following the logic rather than scanning the result. The third layer is **performance discernment**: is the interaction itself working? There is a recognizable point in an exchange where refining the prompt is no longer improving the output - where the conversation has gone off track in a way that makes starting over more efficient than continuing to correct. Such a judgment applies to the interaction as a system, not just to any individual response. Together, these three layers move critical evaluation from a one-time check at the end of an interaction to a practice applied throughout - which is where critical thinking actually belongs.

Connecting to information literacy frameworks you already teach

The ACRL Framework for Information Literacy provides six frames that structure how librarians conceptualize and teach critical engagement with information sources. Each frame connects directly to the evaluation of AI generated content, and for librarians who already teach these frames, that connection provides a ready made professional vocabulary for integrating AI evaluation into existing instruction without constructing an entirely new pedagogical apparatus.

Authority is Constructed and Contextual is the frame most directly applicable to AI evaluation. AI has no authority in the framework's sense; it does not derive credibility from institutional position, expertise, peer review, or community recognition. It synthesizes patterns from sources that may or may not carry authority, without preserving or communicating the authority of those sources. For example, when a student asks AI about the history of a research method, the response synthesizes from whatever sources were in the training data, weighted by statistical patterns rather than by scholarly standing. The productive question for instruction is not "Is AI authoritative?" but "What would the actual authority be for this claim, and how do I find it?"

Information Creation as a Process frames AI particularly well because it foregrounds the production mechanism. AI text is generated through statistical prediction, not through research, expert judgment, or evidence evaluation. Students who understand the generation process, even at the conceptual level covered in Module 01, evaluate AI output more accurately than students who treat it as a search result.

Information Has Value connects to AI through questions of attribution and training data. AI does not cite its sources, which means the value chain of the information it synthesizes (who produced it, under what conditions, with what rights) is entirely obscured. Copyright, attribution, and scholarly credit questions are all embedded in this frame's application to AI.

Searching as Strategic Exploration is the frame most often violated by patrons who use AI as a search tool. AI does not search databases; it generates text. A patron who asks AI for sources on a topic is not retrieving indexed records from a controlled collection. The model is producing seemingly credible citations from statistical patterns. The key distinction for patron instruction is between AI as a thinking tool (appropriate) and AI as a search tool (inappropriate for anything requiring authoritative retrieval).

Scholarship as Conversation highlights a limitation of AI that students often do not consider: AI cannot represent current scholarly debate. Its training has a cutoff date, and even within that date, the representation of any given scholarly conversation is shaped by what was most statistically prevalent in the training data, not by what is most significant in the field. Use AI for synthesis; go to databases for understanding current debate.

Research as Inquiry most clearly articulates why AI does not replace library research. Inquiry involves forming questions, evaluating evidence, reaching conclusions, and revising them, an iterative process of intellectual engagement with sources. AI can support inquiry at specific points, but it cannot conduct it. Instruction that helps students understand this distinction provides durable professional value regardless of how AI tools evolve.

Bias, gaps, and what AI systematically gets wrong

Hallucination refers to AI generating false specific claims. Bias is a different and in some ways more consequential problem: AI systematically overrepresenting certain perspectives, populations, languages, and epistemological frameworks while underrepresenting others, in ways that reflect the composition of the training data and the choices made during fine tuning. In order to evaluate AI output as a professional, a librarian needs to understand not only whether a specific claim is accurate but whether the overall response reflects a complete and representative view of the relevant knowledge landscape.

The most significant bias pattern in large language models is the overrepresentation of English language, Western, and American sources and perspectives. Models trained predominantly on English language web content and books produce outputs that reflect English language norms, examples, and frameworks as default. For example, a librarian asking AI to describe research library practices will receive a response shaped primarily by practices at American and British research universities, not because other models of library practice do not exist, but because they are underrepresented in the training data relative to the volume of English language library literature. Such a response may be accurate for its context while being significantly misleading about the global landscape of the profession.

A second systematic gap is recency. Models have a training cutoff date, and the cutoff affects reliability differently depending on the subject area. For a domain such as AI itself, where developments occur monthly, a model trained through late 2024 is substantially outdated on its own field. For a domain such as the history of medieval manuscripts, the same cutoff matters much less. In order to apply this awareness practically, librarians should identify the training cutoff of any AI tool being used, note the rate of change in the subject area being addressed, and treat AI outputs on fast moving topics with proportionally higher skepticism.

A third pattern is the overrepresentation of majority views and consensus positions at the expense of legitimate minority scholarly perspectives. Models trained on large corpora tend to converge on the most statistically common framing of a topic, which may exclude emerging research, contested findings, or perspectives associated with less published communities. For example, a patron researching a topic where community based participatory research has produced findings that conflict with dominant quantitative studies may receive AI output that represents only the dominant position, because that position has more representation in the training data. In order to counteract these patterns, librarians should treat AI's representation of any contested intellectual landscape as a starting point requiring bibliographic expansion rather than a complete account.

From spotting bias to auditing for it: systematic bias evaluation

Recognizing bias in a single AI response, the skill the previous section builds, is necessary but not sufficient for a library that has embedded AI into its systems. When AI shapes cataloging, discovery ranking, recommendations, and patron interactions at scale, bias stops being a property of individual outputs a librarian happens to read and becomes a pattern operating across thousands of interactions no one is reading. ALA's AI guidance responds to this by asking libraries to move from spotting bias to auditing for it, calling for evaluation of "AI systems for potential bias in datasets and algorithms before adoption" and for "regular audits of tools used in cataloging, reference, recommendations, and patron interactions." The distinction is between an evaluative habit applied by an individual and an institutional practice applied to a system, and the second is what protects patrons the individual never sees.

Auditing has two moments, and both matter. The first is before adoption: evaluating a tool for bias as a condition of acquiring it, rather than discovering the bias after it is already shaping patron experience. For example, a library considering an AI recommendation feature for its discovery layer can ask the vendor what the system was trained on, test it against queries in non-English languages and about marginalized communities, and check whether the results systematically surface the same narrow range of materials, all before signing the contract. The second moment is recurring: because a tool that was acceptable at adoption can drift, and because bias often becomes visible only at scale, the guidance asks for regular audits rather than a one-time check. Such an audit examines whether the discovery system's ranking is quietly reducing the visibility of certain materials, whether AI-suggested subject headings reproduce outdated or offensive terminology for materials about particular communities, and whether the recommendations a patron receives narrow rather than widen the range of what the collection offers.

The specific harm these audits are designed to catch is representational. The ALA guidance names it directly, acknowledging that AI systems "underrepresent non-English languages, dialects, and marginalized communities," and this is exactly the harm that is invisible in any single interaction and unmistakable in the aggregate. For example, an AI cataloging assistant that performs well on mainstream English-language monographs may systematically mishandle materials in other languages or misdescribe materials documenting Indigenous or immigrant communities, and only an audit that deliberately tests those categories will surface the pattern. In order to conduct such audits meaningfully, the guidance also calls on libraries to "train staff to recognize algorithmic bias and its impact on marginalized communities," because an audit is only as good as the person interpreting its results, and recognizing that a ranking pattern disadvantages a particular community is a professional judgment that requires both technical awareness and knowledge of who the collection serves.

Two further commitments from the guidance turn a bias audit from a diagnosis into a practice. The first is that AI-enabled services must always preserve "a clear path to human assistance," so that a patron disadvantaged by an algorithmic pattern is never trapped inside it; the human path is both a service guarantee and a safety valve for bias the audit has not yet caught. The second is that libraries should "use or contribute to open-source AI tools explicitly aiming to reduce bias," which points beyond the individual institution toward the collective work of building tools that are more equitable at the source. For example, a library that documents a bias pattern it found in a vendor tool, and reports it both to the vendor and to a professional community, is contributing to a body of shared evidence that no single audit produces. What none of this delegates to the audit is the underlying professional responsibility: the judgment about whether a tool's biases are tolerable for a specific community, and about what to do when they are not, remains with the librarians who know that community and are accountable to it.

Teaching AI evaluation to students and patrons

In order to teach AI evaluation effectively, the approach that consistently produces the most durable learning is demonstration rather than instruction. A librarian who spends twenty minutes explaining that AI fabricates citations produces less behavioral change in students than a librarian who spends ten minutes demonstrating the fabrication live and allowing students to encounter it directly. Such a demonstration makes the abstract concrete in a way that explanation cannot, and the resulting skepticism is calibrated and experiential rather than theoretical.

A reliable demonstration sequence proceeds as follows. First, ask the AI tool to identify three peer reviewed sources on a topic relevant to the course for which instruction is being delivered, a topic the students know something about, so they can evaluate plausibility. Second, attempt to locate each cited article in an appropriate library database while students observe. Third, report the results transparently: typically at least one citation does not exist, at least one has bibliographic details that do not match any real document, and at least one may exist but may not say what the AI claims. Fourth, ask students directly: given what you just saw, how does this change how you will use AI in your research process? Such a question invites students to draw their own conclusions rather than receiving a rule, which produces more reliable behavior change. Additionally, this demonstration takes approximately ten minutes and can be integrated into an existing single shot instruction session without displacing other content.

For patron interactions at the reference desk, the most effective approach is normalization rather than warning. When a patron mentions that they used AI in their research process, the productive response is not an expression of concern but a practical question: "Did you verify those sources? Let's take a look together." Such framing positions verification as the expected professional next step, which it is, rather than as a corrective response to a mistake. For example, a patron who arrives with an AI generated list of sources can be guided through the verification process as a reference interaction, which serves both the immediate research need and the patron's long term information literacy.

Furthermore, the language used in patron instruction matters considerably. Framing AI as "unreliable" or "dangerous" tends to produce binary responses: patrons either dismiss the concern and continue using AI uncritically, or overcorrect and avoid it entirely. Neither response reflects the calibrated professional judgment the librarian is trying to model. The more accurate framing is that AI has specific strengths and specific failure modes, verification is the professional practice that allows one to benefit from the strengths while managing the failure modes, and the librarian is the professional most qualified to teach that practice. Such framing positions the library's expertise as essential to AI use rather than opposed to it.

Building a personal AI evaluation checklist

There is no doubt that the most durable professional practice is one that has been made explicit, documented, and tested rather than simply internalized as a vague disposition toward skepticism. In order to evaluate AI output consistently across the full range of tasks and contexts a librarian encounters, it is useful to develop a personal checklist, a specific, written set of questions applied deliberately to AI outputs before they are used in professional work. Such a checklist is not bureaucratic overhead; it is the professional equivalent of a pilot's preflight check, converting an evaluative disposition into a reliable procedure.

An effective AI evaluation checklist for library practice should address four categories of question. The first is output type: What kind of content did AI produce? Does it contain specific factual claims, citations, statistics, quotes, or legal and policy information? Identifying the claim types present determines which verification steps apply and how much verification effort is warranted.

The second category is stakes: Where will this output be used? Is it an internal draft for my own reference, a patron facing document, a published resource, or a committee report? Higher stakes require more thorough verification. The calibration framework described in the prior section applies here as the decision rule.

The third category is claim specific verification: For each high risk claim identified in the first step, what is the authoritative source, and have I consulted it? For citations: does the document exist, and does it say what AI claims? For statistics: what is the primary source, and does the figure appear there? For policy and legal claims: what is the governing document, and does it support the AI's characterization?

The fourth category is gap assessment: What does this response not address that would be relevant to a complete professional answer? AI outputs tend to present confident syntheses that omit contested perspectives, recent developments, and minority scholarly positions. Asking explicitly what is missing is as important as verifying what is present.

In order to make such a checklist actionable rather than aspirational, writing it out in the specific language that works for one's own professional context, keeping it brief enough to complete in under two minutes, and applying it consistently to all AI outputs intended for professional use rather than selectively to outputs that seem suspicious is the approach that produces durable habit. Additionally, sharing the checklist with library colleagues creates a common professional vocabulary for AI evaluation that strengthens the team's collective practice, and contributes to the kind of institutional AI literacy that ACRL subcompetencies 3.1 through 3.3 describe as the professional standard.

The right to refuse, and RACBAC as a starting framework

ACRL's October 2025 AI Competencies for Academic Library Workers includes a framing that practitioners sometimes overlook amid the document's emphasis on building AI fluency: a right to refuse. The framework states plainly that adoption of AI technologies is neither necessary nor beneficial in all cases, and it positions that statement as part of AI literacy itself, not as an exception to it. For a librarian who has just worked through this module's calibrated skepticism, this framing should feel familiar rather than surprising: the same professional judgment that determines how much verification a given AI output requires is the judgment that, in some cases, concludes AI should not be used for the task at all. AI literacy, in other words, includes the literacy to decline. A librarian who chooses not to use AI for a specific task, having considered it and found it unsuitable, is exercising the same competency as a librarian who uses AI well; declining to adopt is not a failure of AI fluency but an expression of it.

The Critical AI Literacy Framework, published in the International Journal of Librarianship in 2025, offers a concrete tool for the kind of evaluation this right to refuse depends on: RACBAC, an acronym for Relevance, Accuracy, Coverage, Bias, Authority, and Currency. Librarians who have taught the CRAAP test for source evaluation will recognize the family

resemblance immediately, and that resemblance is the point: RACBAC takes the evaluative habits already built around print and database sources and applies them, term by term, to AI generated output. Relevance asks whether the AI's response actually addresses the question asked, rather than a related but different question, which AI does with some frequency. Accuracy asks whether the specific factual claims in the response can be verified, the question this module has emphasized throughout. Coverage asks what the response leaves out, namely which perspectives, time periods, or subtopics are absent, a question that connects directly to the bias and gaps discussion earlier in this module. Bias asks whose framing, language, and assumptions the response reflects by default. Authority asks what gives this response any claim to credibility at all, given that AI itself has none in the traditional sense. Currency asks how the model's training cutoff and the pace of change in the subject area affect the reliability of what it produced.

For example, a librarian evaluating an AI generated literature overview for a LibGuide can run through RACBAC in the time it takes to read the response once: does it answer the actual question (Relevance), are the cited studies real and correctly characterized (Accuracy), does it represent the full range of scholarly positions on the topic or only the dominant one (Coverage), does it default to particular regional or disciplinary framings (Bias), what would actually establish this as a trustworthy overview (Authority), and is this a fast moving area where the response may already be outdated (Currency). Such a pass takes a few minutes and produces a specific list of what to verify, rather than a vague sense that the response seems fine. In order to build toward the personalized checklist described above, RACBAC is a useful starting structure for a librarian who has not yet developed one, and a librarian who already has a checklist will recognize most of RACBAC's questions as already present in it.

What neither RACBAC nor any other framework can do is make the underlying decision: whether a given AI output, however carefully evaluated, is good enough to use for a specific purpose, or whether the task would be better served without AI at all. That determination, informed by the framework but not made by it, remains the librarian's professional judgment, exercised anew for every task, every time.

PRACTITIONER NOTE

Keeping a printed AI evaluation checklist visible at the reference desk serves as a consistent reminder to apply it, not only when a response seems suspicious. The most dangerous AI outputs are often the ones that seem most authoritative, and the checklist is most valuable precisely in those moments when nothing seems wrong. When introducing colleagues new to AI to this practice, sharing the checklist as a practical

tool is more effective than explaining the reasoning first; the reasoning becomes clearer through use. The single question that proves most useful to add to any checklist is: 'What would the primary source be for this specific claim, and have I looked at it?' For library professionals, that question is already second nature for evaluating student research; applying it to AI output is a professional skill transfer, not a new skill.

KEY TAKEAWAYS

- Specific, seemingly confident claims, including citations, statistics, and attributed quotes, carry the highest hallucination risk; verify before use.
- Verify citations before claims; never use one AI tool to fact check another.
- Apply calibrated skepticism: match verification effort to the stakes of the output and the risk profile of the claim type.
- All six ACRL Information Literacy frames, namely Authority, Creation, Value, Searching, Scholarship, and Inquiry, connect directly to AI evaluation.
- AI systematically overrepresents English language, Western, and majority view perspectives; treat any contested topic as requiring bibliographic expansion.
- Beyond spotting bias in one output, ALA's guidance asks libraries to audit AI systems for bias before adoption and on a recurring basis, since underrepresentation of non-English languages and marginalized communities is invisible in a single interaction and unmistakable at scale; always preserve a clear path to human assistance.
- A written personal evaluation checklist converts a professional disposition into a reliable, repeatable procedure.
- ACRL's October 2025 competencies affirm a right to refuse: adoption of AI is neither necessary nor beneficial in all cases, and declining to use AI for a task is itself an expression of AI literacy, not a failure of it.
- RACBAC (Relevance, Accuracy, Coverage, Bias, Authority, Currency), from the Critical AI Literacy Framework (International Journal of Librarianship, 2025), adapts the CRAAP test for evaluating AI generated output and is a useful starting point for a personal checklist.

References

APA 7TH EDITION

1. American Library Association. (2026). *Guidance on the use of artificial intelligence in libraries.* https://www.ala.org/tools/standards-and-guidelines/guidance-use-artificial-intelligence-libraries
2. Association of College and Research Libraries. (2016). *Framework for information literacy for higher education.* American Library Association. https://www.ala.org/acrl/standards/ilframework
3. Association of College and Research Libraries. (2025, October). *AI competencies for academic library workers.* American Library Association. https://www.ala.org/acrl/standards/ai

ACRL AI Competencies covered

Analysis & Evaluation

Ethical Considerations

Sub-competencies: 3.1, 3.2, 3.3, 2.4 · ACRL AI Competencies (2025)