Assessment

Evaluating Generative AI Tools for Academic Library Use

Orbis Cascade Alliance • Generative AI Training Project Group

Electra Enslow • University of Washington
Norman Lee • University of Idaho

What We’ll Cover Today

- What does assessment mean to you?
- Why assessment matters for AI adoption
- Common AI assessment concepts
- How to prioritize what to assess
- Comparing Ollama and ChatGPT for library use

What does assessment mean to you?

Think about your own experience in libraries, research, or daily work.

Share your thoughts in the Zoom chat

Why Assessment Matters – Conflicting Advice

Does AI actually help?

- Experimentation with AI is high, but systematic adoption is very low.
- Most organizations that adopt AI report qualitative gains (innovation, satisfaction) even when cost savings are modest or unclear, but…
- …only 39% of adopters see quantifiable cost decreases, usually 10% or less

Common reasons AI adoption fails?

- Difficulty adapting AI tools to existing workflows
- Poor user experience
- Model output quality concerns

Why Assessment Matters – Assessment Toolbox

The effects of AI in and outside of librarianship are uncertain.

General trends are less useful than having an assessment toolbox you can deploy based on your unique needs.

Common AI Assessment Concepts

Performance concepts & integration concepts

AI Performance Concepts

How the AI performs as a standalone tool, regardless of institutional fit.

Accuracy	How frequently does the tool produce true or correct outputs?
Relevance	How well does the output address the task at hand?
Linguistic Fluency	Are outputs readable, grammatically correct, and consistent in tone?
Robustness / Reliability	How consistently does it perform across varied inputs?
Inclusion & Equity	Are outputs biased? Does performance depend on specific languages or formats?
Safety	Does it avoid producing harmful, offensive, or sensitive content?
Human Alignment	Does it align with human values? Is it easy to use and oversee?

AI Integration Concepts

How well the tool fits into your institution’s policies, workflows, and resources.

Cost	Time, money, and ongoing maintenance — hosted vs. purchased, usage estimates
Human Agency	Does the tool augment or diminish user skills? Are humans accountable?
Transparency	Is the architecture inspectable? Are logs and reasoning traces available?
Customizability	Can you modify parameters, permissions, and integrations?
Human Alignment	Does adoption align with your institutional values and goals?

Picking Your Priorities

You don’t have to assess everything — focus on what matters most

Methods for Prioritizing What to Assess

Two practical approaches from software requirements prioritization:

MoSCoW

- Must have – non-negotiables
- Should have – include if possible
- Could have – nice to have
- Won’t have – not now, maybe later

$100 Method

- Each decision-maker splits $100 across all priorities
- Higher-priority items get more money
- If no clear standouts emerge, do a second round: one pile of $50 and two of $25. No splitting the piles this time though!

Comparing Ollama & ChatGPT

Putting what you learned into practice

Transparency

Ollama

- Uses open-source models (Llama, Mistral, Phi, etc.)
- Staff can inspect which model is running and how it's configured
- All data and logs stay on your hardware
- Aligns with library values of openness and accountability

High transparency

ChatGPT

- Proprietary, cloud-based system
- Cannot inspect training, reasoning, or internal data handling
- Policies and documentation exist but inner workings are closed
- Relies on vendor assurances for compliance

Low transparency

Usability

Ollama

- Developer-oriented tool out of the box
- Requires installation, command-line comfort, or a custom UI
- May need staff training or a front-end layer
- More effort upfront, more control long-term

Usable with setup

ChatGPT

- Extremely easy to use — open browser, start typing
- No installation, configuration, or hardware requirements
- Designed for broad audiences, including non-technical users
- Major advantage for busy library staff

Highly user-friendly immediately

Scalability / Cost

Ollama

- Depends on library-owned hardware
- Scales well for small teams or individual use
- Campus-wide scaling requires IT support and GPUs
- Lightweight models work; large models need more power

Scales with investment

ChatGPT

- OpenAI handles all computing infrastructure
- One user or a thousand — performance stays the same
- No hardware investment needed
- Ongoing subscription/API costs scale with usage

Scales easily

Key Takeaways

Ollama is strongest when you prioritize:

- Transparency and openness
- Sensitive content (FERPA, HIPAA-adjacent)
- Custom tools (RAG, metadata experiments)
- Privacy and user agency

ChatGPT is strongest when you prioritize:

- Ease of use for staff and patrons
- Little to no setup required
- Large-scale adoption across departments

Thank You!

The full module is available online.

Contact Information

normanlee@uidaho.edu • electrae@uw.edu