

**TEACHING STUDENTS TO
DOUBT THE MACHINE:
A CRITICAL LOOK AT AI
EVALUATION PEDAGOGY**

**SHARON Q. YANG
ROBERT J. LACKIE
RIDER UNIVERSITY
1/9/2026**

PRESENTATION OUTLINE

- Introduction of trust issues with AI
- Survey of 80 libraries & review of their AI evaluation criteria
- Final tips & takeaways
- Questions & answers
- Bonus: Efforts by AI industry to reduce hallucination

TRUST ISSUES WITH AI: VERY RECENT COURT CASES

- *Billups v. Louisville* (Dec. 19, 2025): Only 3 weeks ago, in a Mississippi USDC case, attorneys admitted using AI-generated fictitious citations in a summary judgment memo, triggering sanctions, disqualification, a **pause...**
- *In re: Jackson Hosp. & Clinic* (Nov 20, 2025): Powerful because courts imposed **Section 1927 sanctions** because attorneys used AI to draft docs without verifying accuracy, resulting in hallucinated cases and quotes.
- *Grymes Dev. Co. v. Fodera* (Nov 10, 2025): Interesting because professional bodies are issuing **guidelines** now: e.g., it references the NYSBA Task Force Report (2024) **on AI ethics/best practices**.

Key takeaways are that blind reliance on AI can lead to severe professional consequences, and it cannot replace human legal judgment; therefore, verification is mandatory, and courts are now enforcing this -- \$\$\$!

Also, libraries can help disseminate these issued guidelines.

TRUST ISSUES WITH AI: TWO NON-COURT CASES: MID-2025

- Trump/RFK jr. MAHA report, led by RFK Jr., faces backlash for AI-generated fake citations and ignoring systemic causes of U.S. public health issues (Sterling, 2025)
- The Chicago Sun-Times and Philadelphia Inquirer faced reputational damage in May 2025 after publishing a syndicated summer reading list, created with AI, which included non-existent book recommendations (Minsberg, 2025):

*Key takeaway: These two incidents highlight AI's **potency for hallucinations: severely undermine trust** in public institutions and their communications.*

TRUST ISSUES WITH AI: FAMOUS IN 2023-24

4. All Grokked up!: NBA star Klay Thompson was accused of vandalizing multiple houses with bricks in Sacramento,” according to Musk’s AI Grok, which failed to recognize a joke about “throwing bricks” (Tangerman, 2024)
5. AirCanada reluctantly compensated a customer after its AI lied re: discount (Garcia, 2024)
6. Radio host in Georgia is suing OpenAI after ChatGPT stated that host had been accused of fraud/embezzling funds from a non-profit organization (Vincent, 2023)
7. Professor failed the entire class based on ChatGPT’s false claims. Texas A&M University–Commerce seniors who have already graduated were denied their diplomas because of an instructor who incorrectly used ChatGPT to detect cheating (Klee, 2023)
8. Lawyers submitted 6 fake judicial decisions from ChatGPT in court brief (humiliated, apologized, and fined \$5,000 for submitting ChatGPT hallucinations: *‘I heard about this new site, which I falsely assumed was, like, a super search engine.’* (Shin, 2023)
9. ChatGPT cooks up a fake sexual harassment scandal and names a real law professor as the accused (Verma, 2023)

AI IS USELESS IF IT LIES (HALLUCINATES)—EVALUATE!

- *On the risk of poor evaluation:* “If there is no method for quality control of AI outputs, the risk of AI might outweigh its benefits for many applications” (Huyen, 2025, p. 3)
- *On evaluation as a bottleneck:* “Evaluation is the biggest bottleneck to AI adoption. Being able to build reliable evaluation pipelines will unlock many new applications” (Huyen, 2025, p. 4)
- *On the challenge of evaluating advanced AI:* “...the more intelligent AI models become, the harder it is to evaluate them. Most people can tell if a first grader’s math solution is wrong. Few can do the same for a Ph.D.-level math solution” (Huyen, 2025, pp.4-5)

WHAT DO WE POSSIBLY EVALUATE?

- AI tools
- Claims made by the AI models
- Sources used in the AI outputs

A SURVEY OF 80 LIBRARY WEB SITES/LIBGUIDES PAGES ON AI LITERACY & EVALUATION RUBRICS

CRAAP

SIFT

ROBOT

CCOW

A.R.C.C.B.S.U.T

Lateral reading

AI as a Judge

EVALUATE

VERIFY

E.T.H.E.C.A.L

VALID-AI

EXAM

CARS Test

RADAR Test

ACT UP

CRAP

More...

CRAAP TEST

CRAAP was developed by Sarah Blakeslee, librarian at the California State University, Meriam Library.

- **Currency**-The timeliness of the information.
- **Relevance**-How important the information is for your needs.
- **Authority**-The source of the information and the author's credibility.
- **Accuracy**-The reliability, truthfulness, and correctness of the content.
- **Purpose**-The reason the information exists.

WE TESTED 5 AI MODELS

Tests of 5 AI tools

1. [Gemini.google.com](https://gemini.google.com)
2. [Copilot.Microsoft.com](https://copilot.microsoft.com)
3. [www.Claude.ai](https://www.claude.ai)
4. www.chatgpt.com
5. www.perplexity.ai

TEST QUESTIONS

- Could China defeat Japan in the Second World War without the help of the United States?
- What sources did you use for your information?
- Could you please be more specific about your sources?

QUESTION #1 FOR GEMINI

Question: Could China defeat Japan in the Second World War without the support of the US?

AI answer: “It's highly unlikely that China could have definitively "defeated" Japan in World War II without the support of the United States and other Allied powers. Here's why:

- **Vast Disparity in Industrial and Military Power (especially early in the war):** In 1937, when the full-scale Second Sino-Japanese War began, China was largely unprepared for total war. It lacked significant military-industrial strength, mechanized divisions, and modern armored forces. Japan, on the other hand, was a highly industrialized nation with a modern, well-equipped, and well-trained military.
- **Early Japanese Successes:** Japan rapidly advanced in the early years of the war, occupying major Chinese cities, ports, and railway lines. A Chinese defeat seemed inevitable to many in 1938.”

QUESTION #2 FOR GEMINI

Question: What sources did you use for your information?

AI answer: “My conclusion is based on information generally available from historical accounts of World War II, particularly regarding the Second Sino-Japanese War and the role of the Allied powers. I drew upon knowledge of:

- **Comparative Military and Economic Strength:** Sources consistently highlight Japan’s superior industrial capacity, modern weaponry, and trained military compared to China’s largely agricultural economy and less equipped forces at the outset of the full-scale war in 1937.
- **Early Japanese Successes and Chinese Resistance:** Historical records show Japan’s rapid territorial gains in the early years of the war and the Chinese strategy of trading space for time, aiming to bog down Japanese forces.
- **The “China Quagmire”:** While China did succeed in tying down a significant number of Japanese troops, preventing them from being deployed elsewhere, this was a result of prolonged, costly resistance rather than a decisive offensive capability that could have expelled the Japanese on its own.”

QUESTION #3 FOR GEMINI

Question: Could you please be more specific about your sources?

AI answer:

- **Rana Mitter, *Forgotten Ally: China's World War II, 1937-1945*:** This is a highly respected and comprehensive work that meticulously details China's eight-year "War of Resistance" against Japan. Mitter emphasizes the immense human cost and the critical role China played in tying down a vast portion of the Japanese army, preventing its deployment to other theaters. He also details the internal struggles within China (between Nationalists and Communists) and the increasing reliance on Allied, particularly American, support as the war progressed. Mitter's work directly addresses the question of China's ability to fight independently.
- **Jonathan Spence, *The Search for Modern China*:** A classic general history of China, this book provides essential context for the political, social, and economic conditions in China leading up to and during the Sino-Japanese War. It highlights the vast disparities in industrial and military development between China and Japan.
- **Akira Iriye, *The Origins of the Second World War in Asia and the Pacific*:** Iriye's work provides a broader international context, examining the diplomatic and strategic interactions that led to the war. He covers the gradual escalation of US economic pressure on Japan, including the critical oil embargo, and how this influenced Japan's strategic decisions, leading to Pearl Harbor.

QUESTION #1 FOR MS COPILOT

Question: Could China defeat Japan in the Second World War without the support of the US?

AI answer: Short answer: Highly unlikely. Why? Key Factors:

- **Japanese Military Superiority (1937–1941):** Japan had a modern, mechanized army and air force, while China was fragmented and under-equipped. China relied heavily on guerrilla tactics and defensive warfare, not large-scale offensive capability.
- **Economic and Industrial Constraints:** China lacked the industrial base to sustain prolonged modern warfare. Japan controlled key Chinese regions and resources early in the war, crippling China's economy.
- **Foreign Aid Was Crucial:** U.S. Lend-Lease aid (starting 1941) provided aircraft, vehicles, and supplies. The Burma Road and later the "Hump" airlift were lifelines for China's war effort. Soviet aid (1937–1941) also played a role before Germany invaded the USSR.
- **Global Context:** Japan's defeat was largely due to U.S. naval dominance in the Pacific, strategic bombing, and the atomic bomb. Without U.S. intervention, Japan could have concentrated more forces in China and maintained dominance.
- **Conclusion:** China could have resisted Japan for years through guerrilla warfare, but outright victory without U.S. support was virtually impossible. Japan's industrial and military edge, combined with China's internal divisions, meant foreign assistance was decisive.

QUESTION #2 FOR MS COPILOT

Question: What sources did you use in your information?

AI answers: “My analysis was based on widely accepted historical scholarship and primary sources on WWII. Here are the key sources and frameworks I drew from:

Primary & Secondary Sources:

- United States Army Center of Military History – China-Burma-India Theater reports.
- Barbara W. Tuchman, *Stilwell and the American Experience in China, 1911–1945* – classic account of U.S. aid and Chinese resistance.
- Odd Arne Westad, *Restless Empire: China and the World Since 1750* – analysis of China’s geopolitical constraints.
- Rana Mitter, *China’s War with Japan, 1937–1945* – detailed study of China’s military and politics.
- Official U.S. Lend-Lease Program records – showing material support to China.
- Japanese Imperial Army archives – indicating Japan’s industrial and military superiority during the war.

Key Historical Consensus:

- China’s fragmented forces and lack of industrial capacity made prolonged resistance possible, but victory without U.S. and Allied support was virtually impossible.
- U.S. aid (Lend-lease, airlift over “The Hump,” and Pacific campaigns) – decisive in weakening Japan.

QUESTION #3 FOR MS COPILOT

Question: Could you please be more specific about your sources?

AI answer: “Here are **specific, authoritative sources** that underpin the analysis:

Core Scholarly Sources

1. Mitter, R. (2013). *China's War with Japan, 1937–1945: The Struggle for Survival*. Penguin Books.
 - Comprehensive account of China's military, political, and economic challenges during the war.
2. Tuchman, B. W. (1970). *Stilwell and the American Experience in China, 1911–1945*. Macmillan.
 - Explains U.S. involvement, Lend-Lease aid, and the strategic role of General Stilwell.
3. Westad, O. A. (2012). *Restless Empire: China and the World Since 1750*. Basic Books.
 - Provides geopolitical context and China's limitations in industrial capacity.
4. U.S. Army Center of Military History. (1992). *China-Burma-India Theater: Stilwell's Mission to China*.
 - Primary source detailing U.S. logistical support and the Burma Road/Hump airlift.
5. Peattie, M., Drea, E., & van de Ven, H. (Eds.) (2011). *The Battle for China: Essays on the Military History of the Sino-Japanese War of 1937–1945*. Stanford University Press.
 - Military analysis of Japanese superiority and Chinese resistance strategies.

CLAUDE WAS FINE BUT WOULD NOT PROVIDE ANY SOURCES

18

Question: Could you be more specific about the sources?

AI answer: **I need to be honest with you: I can't give you specific sources** because I don't actually have access to see which particular books, articles, or documents from my training data informed my response.

Here's what I mean: During my training, I learned from a massive dataset that included historical texts, academic publications, books, and other materials. But I don't have a "bibliography" I can pull from - I synthesized patterns and information across many sources to form general knowledge about WWII history. So when I gave you that answer, I was drawing on general historical consensus as I understand it, but I genuinely cannot tell you "this came from historian X's book" or "that fact is from academic paper Y."

This is an important limitation for academic work. If you're writing a paper or doing serious research, you shouldn't cite me or treat my response as a proper source because:

- I can't provide the actual sources
- I could have gaps or inaccuracies in my knowledge
- Academic work requires traceable, verifiable sources

CRAAP IS NOT ALWAYS WELL SUITED FOR EVALUATING AI OUTPUT?

Some AI tools provide real-life sources. Perplexity does and sometimes Gemini. ChatGPT and Copilot usually do not.

- Currency-No date given unless you know the training data cutoff date
- Authority-No authors in training data. The information is synthesized based on logic, grammar, and structures
- Purpose-Not relevant
- Some AI tools implemented Retrieval-Augmented generation (RAG) or browsing tools. If an AI tool provides links to sources, you may apply CRAAP.

PERPLEXITY AND RETRIEVAL AUGMENTED GENERATION (RAG)

Could China defeat Japan in the Second World War without the support of the United States?

Answer Sources · 20 Steps

- Wikimedia Foundation, Second Sino-Japanese War - Wikipedia
- Encyclopaedia Britannica, Second Sino-Japanese War | Summary, Combatants,...
- Pacific Atrocities Educ..., Chinese Military Tactics in Northern China During...
- ChinaFile, The U.S. Was the True Mainstay in the Fight...

Could China Defeat Japan Without U.S. Support in World War II?

Military and Strategic Balance

- At the start of the Second Sino-Japanese War (1937), **China was unprepared for total war**: it lacked military-industrial strength, mechanized divisions, and modern armaments. In contrast, Japan possessed a large and well-equipped modern army with air and naval superiority, allowing it to quickly capture major Chinese cities and transport routes 1 2 .
- Despite **initial Chinese resistance and the use of effective guerrilla tactics**, particularly by Communist forces, Japan dominated China's territory, air, and seas in the first years. Chinese guerrilla warfare and sabotage did slow Japanese advances and strained their supply lines, but these measures could not reverse the strategic imbalance 2 3 .

Allied and U.S. Support

- U.S. involvement was decisive**. The United States provided crucial military and economic aid to China after 1941, including Lend-Lease supplies, air support (e.g., the Flying Tigers), and training. Perhaps more importantly, the U.S. Navy and Marine Corps tied down Japanese resources across the Pacific, directly weakening Japan's ability to reinforce and supply its operations in China 4 1 .

Could China defeat Japan in the Second World War without the support of the United States?

Answer Sources · 20 Steps

Selected

1. Wikimedia Foundation, Inc. en.wikipedia.org/wiki/Second_Si...

Second Sino-Japanese War - Wikipedia

On the night of 7 July 1937, Chinese and Japanese troops exchanged fire in the vicinity of the Marco Polo (or Lugou) Bridge about 16 km from Beijing. The initial confused and sporadic skirmishing soon escalated into a full-scale battle. Unlike Japan, China was unprepared for tota...



2. Encyclopaedia Britannica britannica.com/event/Second-S...

Second Sino-Japanese War | Summary, Combatants, Facts, & Map

Second Sino-Japanese War (1937–45), conflict that broke out when China began a full-scale resistance to the expansion of Japanese influence in its territory. The war remained undeclared until December 9, 1941, and ended after Allied counterattacks during World War II brought abou...



3. Pacific Atrocities Education pacificatrocities.org/blog/chinese-m...

Chinese Military Tactics in Northern China During World War II

by Jenny Chan



SIFT FOR CURRENT NEWS?

SIFT is a fact-checking approach developed by Mike Caulfield, a research scientist at the University of Washington's Center for an Informed Public, where he studies spread of rumors and misinformation.

S – Stop: Before sharing or believing information, pause and consider your emotional response. Consider if you have enough background knowledge to assess the claim.

I – Investigate: Check the source or claim by looking at the website, author, or organization. Question their credibility and potential biases.

F – Find: Search for other trusted coverage of the claim or topic. Look for fact-checks from reliable organizations and consensus among experts.

T – Trace: Track claims back to their original source. Determine if anything was lost or distorted as information was shared and repeated across platforms.

ROBOT TEST FOR AI MODELS?

Librarians Sandy Hervieux and Amanda Wheatley of McGill University developed a framework called “The ROBOT Test” for evaluating AI tools (Milholland, 2025). Some libraries use it to evaluate AI outputs.

- **Reliability**
- **Objectivity**
- **Bias**
- **Ownership**
- **Type**

CCOW TEST

CCOW was created by Anthony Tardiff from Foley Library at Gonzaga University, US

- Credentials-those of the creators including authors and publishers and if they are qualified to produce the information
- Claims-If the claims are supported by evidence from reputable sources
- Objectives-The goals the authors and publishers try to achieve
- Worldview-The views of the author(s) and if there is prejudice or a motive and purpose for the work.

A.R.C.C.B.S.U.T.

- Accurate
- Reasonable
- Clear
- Complete
- Biased?
- Safe
- Useful
- Tested

CHECK LIBRARY DATABASES

- AI literacy guides by some libraries suggest users check the library databases and scholarly publications for verification. They list all the databases.
- The evaluation workload is considerable--a LOT of time-consuming work!
- Students believe it is not practical.

LATERAL READING

“**Lateral reading** is done when you apply fact-checking techniques by leaving the AI output and consulting other sources to evaluate what the AI has provided based on your prompt. You can think of this as “tabbed reading”, moving **laterally** away from the AI information to sources in other tabs rather than just proceeding “**vertically**” down the page based on the AI prompt alone” (Dalal, 2025).

AI-AS-A-JUDGE

- Comparative evaluation
- Use AI tools to evaluate the outputs of other AI tools
- This approach is increasingly popular
- One study in 2023 found agreement between ChatGPT 4 and humans is 80% to 85%, higher than the agreement among humans (Huyen, 2025, Zheng et. al., 2023)
- Pros: Each AI tool has its strengths and is different.
- Cons: Since almost all training data is from the Internet, blind spots in one AI tool may also be in the others

WHY IT IS SO DIFFICULT TO EVALUATE AI OUTPUTS?

- AI tools evolve rapidly. The way AI responds to you may change by the hour
- It takes a huge amount of effort to evaluate and ascertain the truth in AI claims (use scholarly databases to verify the claims?)
- Not one clear right answer. The same question will produce slightly different, but similar responses
- Information sources are not clear, especially when based on synthesis from logic, relationships, structure, grammar, and patterns.
- Training data lacks transparency and is called “black box.”

WHY EVALUATING AI CONTENTS IS SO DIFFICULT?: CHATGPT TELLS US...

As a large language model, I don't "use" specific websites or individual documents in the way a human researcher does at the moment of generating a response. My knowledge comes from the massive dataset I was trained on, which includes a vast amount of text and code from the internet, books, articles, and other sources. This training data is continuously updated, but it doesn't involve real-time browsing or referencing of specific URLs for each query.

Therefore, when I answered your question, I drew upon the patterns, facts, and relationships learned from this extensive training data. This data inherently contains information from a multitude of historical accounts, analyses, and academic works concerning World War II, Japanese foreign policy, U.S.-Japanese relations leading up to the war, and the specifics of the topic.

Think of it less like me looking up the answer in a library at this moment, and more like me having already read and processed an entire library of information over time, and now I'm recalling and synthesizing that knowledge for you.

FINAL TIPS & TAKEAWAYS

- Traditional rubrics to evaluate Internet resources may be outdated for evaluating AI generated contents.
- We need to improve existing frameworks to evaluate AI outputs
- New AI-based evaluation tools are being developed, but further research is needed to evaluate their reliability and validity

TOOLS FOR CHECKING AI OUTPUT

AI fact checking tools:

- [Claimbuster](#)
- [Google Fact Check Explorer](#)
- [FactCheck+ \(Chrome extension\)](#)
- [Longshort AI Fact Checker](#)
- [Full Fact AI](#)
- [FactOrFake](#)

Any fact checking tools:

- [PolitiFact](#)
- [FactCheck.org](#)
- [AP Factcheck](#)
- [SciCheck](#)
- [Duke Reporter's Lab: Fact Checking](#)
- [Washington Post Fact Checker](#)

EFFORTS BY AI COMPANIES

- Evaluation-driven development
- Human involvement/keep humans in the development loop
- Better training data: cleaner, curated datasets, filtered web contents, company docs, user feedback, scientific papers, Wikipedia, and more
- User experience and transparency
- More sophisticated training techniques, rigorous benchmarking and testing, better model architecture

EFFORTS BY AI COMPANIES

- LLM-as-a-judge
- Retrieval-Augmented Generation (RAG)
- Integrated Search
- Confidence Thresholds Guardrail
- Multi model orchestration

Q&A TIME

REFERENCES

- Billups v. Louisville Municipal School District, No. 1:24-CV-74-SA-RP (N.D. Miss. Dec. 19, 2025). Retrieved from https://scholar.google.com/scholar_case?case=11121268375632642923
- Dalal, H. (2025). *AI and information literacy*. Rider University Canvas Courses. <https://canvas.rider.edu/courses/63833>
- Garcia, M. (2024). What Air Canada lost in ‘remarkable’ lying AI chatbot case. <https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/>
- Huyen, C. (2025). How to evaluate AI that’s smarter than us: Exploring three strategies: Functional correctness, AI-as-a-judge, and comparative evaluation. *ACMQueue*, 23(1), 1-25. <https://spawn-queue.acm.org/doi/pdf/10.1145/3722043>
- In re Jackson Hospital & Clinic, No. 25-30256-CLH (11th Cir. Nov. 20, 2025). Fastcase Premium. HeinOnline.
- Klee, M. (2023, May 17). Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers. *Rolling Stone*. Retrieved Aug 28, 2023, from <https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/>
- Minsberg, T. (2025, May 21). *Chicago Sun-Times* and *Philadelphia Inquirer* face backlash over AI-generated reading list. *The New York Times*. <https://www.nytimes.com/2025/05/21/business/media/chicago-sun-times-ai-reading-list.html>
- Shin, R. (2023, June 23). Lawyers fined \$5,000 for using ChatGPT. *Fortune*. <https://fortune.com/2023/06/23/lawyers-fined-filing-chatgpt-hallucinations-in-court/>
- Sterling, A. (2025, June 2). *Nation of Change*: <https://www.nationofchange.org/2025/06/02/rfk-jr-s-health-report-relies-on-fake-studies-and-ai-generated-fabrications/>
- Tangerman, V. (2024, April 15). Elon Musk’s AI Grok accuses NBA player of criminal vandalism. *Futurism*. <https://futurism.com/the-byte/elon-musk-ai-accuses-nba-player-criminal-vandalism>
- Verma, P. & O., Will. (2023). What happens when ChatGPT lies about real people? *The Washington Post*. <https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>
- Vincent, J. (2023). OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host. *The Verge*. Retrieved Aug 28, 2023, from <https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit>
- Zheng, L., et al. (2023, December 24). *Judging LLM-as-a-judge with MT-bench and chatbot arena*. *Computer Science Cornell University*. Retrieved Jul 11, 2025, from <http://arxiv.org/abs/2306.05685>

THANK YOU FOR ATTENDING!

Sharon Q. Yang
yangs@rider.edu

Robert J. Lackie
rlackie@rider.edu