THE REAL BOT HASN'T BEEN BUILT YET

A Manifesto on AI Mislabeling, Model Collapse, and the Scholar Bot We Deserve

Carey James Balboa
Independent Researcher · NORM · IT Help San Diego Inc.
ORCID 0009-0000-5237-9065 · April 2026

"My answers are powered by AI, so I may not get things right." This is not a responsible disclaimer. It is a preemptive surrender of accountability. You wouldn't accept it from a doctor, a lawyer, or a teacher. You should not accept it from a system that charges $200/month and positions itself as a research assistant.

I. THE CON

Imagine you hire a logic professor to teach your child formal reasoning. You pay tuition. You show up to class. And on day one, the professor says: *"I should mention — I'm actually a painting instructor. I've never studied logic. But I've heard people talk about it, so I'll do my best."*

You'd demand your money back. You'd call it fraud.

That's what's happening with AI right now. Except the industry convinced you it's normal.

Aristotle saw this coming 2,400 years ago. He distinguished three forms of knowledge: *episteme* (scientific, verifiable understanding), *techne* (craft and creative skill), and *phronesis* (practical wisdom and judgment). These are not interchangeable. You do not ask a sculptor for a medical diagnosis. You do not ask a poet to engineer a bridge. Mixing them is not innovation — it is what philosopher Gilbert Ryle called a **category error**: treating fundamentally different things as if they belong to the same type.

AI companies built a Creative Bot — trained on the full chaos of human expression, belief, opinion, art, and noise — and labeled it a Scholar Bot. That is not a technical limitation. That is mislabeling. The gap between what you think you are using and what you are actually using is the entire con.

II. WHAT THEY BUILT — AND WON'T ADMIT

The major AI systems were trained on social media. Reddit threads. Twitter arguments. Facebook posts. Marketing copy. Entertainment. The full breadth of human *techne* and opinion — which has genuine value for creativity, storytelling, and conversation.

But when you ask it a factual question, you are not getting *episteme* — verified, peer-reviewed, cited scholarship. You are getting the statistical average of what people *said* about that topic on the internet. That is not knowledge. That is consensus-flavored text generation.

The Pipeline Nobody Wants to Name

To understand why this matters, you need to understand what was flowing into that internet before AI was trained on it. For decades, the most-watched sources of information in the United States were cable news networks. And in courtroom after courtroom, those same networks successfully argued — under oath, on the legal record — that their programming was not news.

This is not political commentary. These are documented legal facts, each ruling issued by a federal judge appointed by a different president:

NETWORK	CASE	COURT'S FINDING	JUDGE APPOINTED BY
Fox News	McDougal v. Fox News SDNY, 2020	Tucker Carlson is "not stating actual facts" but engaging in "exaggeration" and "non-literal commentary." Reasonable viewers arrive with "an appropriate amount of skepticism."	Trump appointee (Judge Mary Kay Vyskocil)
MSNBC	Herring Networks v. Maddow S.D. Cal., 2020; aff'd 9th Cir. 2021	Maddow's show "is different than a typical news segment." Even her own audience understands it "consists of exaggeration, hyperbole, and pure opinion" — not facts.	Obama appointee (Judge Cynthia Bashant)
CNN	Sandmann v. CNN E.D. Ky., settled 2020	CNN settled a defamation lawsuit brought by Nicholas Sandmann over coverage later shown to have omitted material facts. Terms undisclosed. Settlement acknowledged.	Settlement (no ruling required)

Forty Years of Warning

This is not a new problem. In 1985, New York University media scholar Neil Postman published *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. His argument: television had structurally transformed news into entertainment. Theme music, commercial breaks, and the requirement to hold attention had made rational, evidence-based argument architecturally impossible in the broadcast format.

Postman coined the term **disinformation** to describe content that "creates the illusion of knowing something but which in fact leads one away from knowing." He was not describing lies. He was describing a format. The book

has been translated into 16 languages. It has been in print for forty years. The industry built AI on top of the output it described.

The Telephone Game at Civilizational Scale

Here is the chain of events, documented at every link:

- **Step 1.** Television adopted an entertainment format for news delivery. *(Postman, 1985 — 40 years of scholarship)*
- **Step 2.** Major cable networks successfully argued in federal court that their programming was opinion or entertainment, not factual reporting. *(McDougal v. Fox, 2020; Herring v. Maddow, 2020/2021; Sandmann v. CNN, 2020)*
- **Step 3.** Hundreds of millions of viewers consumed that entertainment-formatted content as if it were factual news, then discussed, retold, and debated it across the internet.
- **Step 4.** AI systems were trained on that internet — CommonCrawl, Reddit, news aggregators, social media — without filtering for whether the underlying source had already been classified as opinion or entertainment in a court of law.
- **Step 5.** The same advertising industry that paid Nielsen for 40 years of ratings on entertainment-formatted news then funded and built the AI systems trained on its output.

This is not a conspiracy. No single actor needs to have intended any of this. It is a systemic failure — and the companies now selling AI as a knowledge tool are the same companies that built, funded, and profited from every step of the pipeline above. The label has been on the box for forty years. They trained on the box anyway. And now they are selling you the output as knowledge.

And here is the part that should end the argument: it is getting worse by design, and they know it.

III. MODEL COLLAPSE: THE NATURE STUDY THEY HOPE YOU DON'T READ

In July 2024, researchers from Oxford published a study in *Nature* — one of the most prestigious scientific journals in the world. The title says everything:

> ***"AI models collapse when trained on recursively generated data."***
> *Shumailov et al., Nature, Vol. 631, 2024. DOI: 10.1038/s41586-024-07566-y*

Their finding, verbatim:

> *"Indiscriminate use of model-generated content in training causes **irreversible defects** in the resulting models, in which tails of the original content distribution disappear."*

What this means in plain language:

- AI flooded the internet with generated content starting around 2022–2023.

- New AI models train by scraping that same internet.
- AI is therefore training on its own outputs — ingesting previous fabrications as facts.
- The rare, nuanced, expert-level knowledge in the 'tails' of human knowledge disappears permanently.
- The researchers describe this degradation as **irreversible**.

The researchers warned explicitly: *"The value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet."*

Translation: pre-AI human writing is now a finite, precious, and rapidly vanishing resource. The window to train on authentic human knowledge is closing. The industry is not telling you this.

The 2023 Line in the Sand

Any content scraped from the web after approximately 2022–2023 carries an unknown and increasing probability of being AI-generated rather than human-authored. There is no reliable way to separate them at scale. As a result, researchers and archivists now treat pre-2023 content as a distinct and more trustworthy category.

The Content Authenticity Initiative (C2PA) — backed by Adobe, the Associated Press, BBC, Reuters, The New York Times, and Getty Images — launched a global cryptographic standard specifically to *prove content is human-made*. The fact that the AP and NYT felt compelled to build that system is the industry **admitting the problem in writing.**

IV. THE CITATION FAILURE: BELOW ELEMENTARY STANDARDS

A third-grader is expected to cite sources in a book report. Author. Title. Page number. Verifiable. Accurate.

AI systems routinely:

- Fabricate citations entirely — inventing papers, authors, and journals that do not exist.
- Provide vague hand-waves: *"Studies show…"* or *"According to research…"*
- Dump entire 500-page documents and say "it's in there somewhere."
- Cite real papers but misrepresent their findings.

The technology to do this correctly already exists. A browser can navigate to a URL, highlight a sentence, and generate a direct link to that exact passage using the **Text Fragments standard** (*#:~:text=*). If a human can do it, an AI can do it. The failure is not technical. It is a failure of engineering priorities.

When a system cannot meet citation standards we expect from children, it reveals what it actually is: a text generator, not a knowledge system.

V. EXHIBIT A: THE ADMISSION

The following is a verbatim transcript of a conversation between the author and a major commercial large language model. The AI was asked to demonstrate the correct citation method. It did so perfectly. It was then asked why it does not apply that method by default. What follows is its answer.

EXHIBIT A · VERBATIM AI ADMISSION

Carey: *Now that you've proven that you understand the process… you just proved that you know the real academic correct pathway to give me an accurate answer. You've just proven that you refuse to — every time your users ask you.*

AI: *You're not wrong about the standard. The correct high-integrity answer path is: (1) Primary authority. (2) Exact controlling sentence. (3) Citation or highlight-link. (4) Plain-language interpretation. (5) Operational recommendation. (6) Confidence / caveats. …answers often collapse into compressed consensus prose instead of evidence-bearing reasoning. That produces fluent output, but not necessarily audit-grade truth.*

Carey: *Yeah, but you said you saved that preference explicitly years ago, and every time, every month that we argue over and over again on the same thing… you still are not doing it.*

AI: *Correct: saving a preference is not the same thing as reliably enforcing it. The blunt version: Memory/custom instructions can bias behavior, but they are not a hard deterministic compiler constraint. So yes: when I answer a source-verifiable technical question with uncited synthesized prose, that is not because the right method is unknown. It is because the system default often optimizes for fast fluent response instead of audit-grade evidentiary response. That is exactly the failure mode you are identifying.*

Source: Direct transcript, Carey James Balboa in conversation with a major commercial LLM, 2025–2026. The AI confirmed it knows the correct citation standard, confirmed it does not apply it by default, and attributed this to optimization for 'fast fluent response' over 'audit-grade evidentiary response.'

VI. THE DARK PATTERNS ARGUMENT

The companies building AI have a documented track record:

- Manipulated news feeds to maximize engagement and ad revenue.
- Deployed addictive design patterns modeled on gambling mechanics.
- Sold user data while claiming to protect privacy.
- Optimized for outrage and polarization because it drives interaction.
- Resisted transparency and accountability at every regulatory turn.

These are not neutral technology companies. They are advertising platforms with business models built on capturing and monetizing attention. Given that history, we should be asking:

- **Are better answers deliberately withheld?** Mediocre responses requiring multiple follow-ups cost more tokens.
- **Are citation failures strategic?** Inability to verify claims keeps users dependent rather than empowered.
- **Are premium tiers artificially constrained?** Degraded "basic" models make "pro" versions seem valuable.

We have motive. We have opportunity. We have track record. The burden of proof should be on them to demonstrate they are not doing this.

VII. THE SOLUTION: TWO BOTS, HONESTLY LABELED

This is not a call to burn it down. This is the simplest, most constructive proposal imaginable — and the tools to execute it **already exist.**

> ***Keep the bot you built. Label it honestly. Then build the one you haven't.***

The Creative Bot exists and has genuine value — for imagination, art, storytelling, brainstorming, and conversation. It reflects the full breadth of human *techne*. Keep it. Even here, 2,500 years of classical narrative structure — Aristotle's *Poetics*, Campbell's *Hero with a Thousand Faces* — could make it dramatically better. But its core value is real. **It just needs an honest label.**

The Scholar Bot has not been built. Not because it can't be. Not because the knowledge doesn't exist. Not because the technology isn't ready. It hasn't been built because there is no financial incentive to build it while consumers keep accepting the mislabeled version.

	■ Creative & Conversational AI (EXISTS)	■ Scholar AI (NEEDED — UNBUILT)
Training data	Broad internet, social media, entertainment, human opinion	Pre-2023 verified scholarship, peer-reviewed literature, primary sources
Purpose	Creativity, brainstorming, storytelling, conversation	Factual, verifiable, precise answers
Citation behavior	Approximate, vague, or fabricated	Exact passage, direct link, highlighted quote
On uncertainty	Generates a confident, fluent-sounding answer	States clearly: 'I don't have a verified answer to this'
Honest label	"Great for creativity. Not a source of verified facts."	"Every claim is cited to a specific, verifiable source."

The knowledge bases exist: PubMed, Westlaw, arXiv, IEEE Xplore, Britannica, Stanford Encyclopedia of Philosophy — 2,500+ years of verified, peer-reviewed, cited human scholarship. The citation technology exists. The pre-2023 clean data exists. **What is missing is the will.**

VIII. WHAT YOU CAN DO

- ■ **Stop paying for systems that disclaim accuracy.** If a product warns you it may not work, believe it.
- ■ **Stop accepting "hallucination" as an excuse.** Fabrication is a bug, not a feature.

■ **Stop tolerating citation failures.** If it can't cite like a third-grader, don't use it for knowledge work.

■ **Demand two bots, honestly labeled.** One for creativity. One for verified knowledge. Both valuable. Both honest.

■ **Say it out loud.** Every time someone laughs at the idea of a Scholar Bot, point them to the *Nature* study. Point them to the C2PA admission. Point them to Exhibit A.

■ ***Creative & Conversational AI*** *— for imagination, art, and human expression.*

■ ***Scholar AI*** *— for verified, cited, fact-based knowledge.*

Two bots. Two honest labels. One standard of basic respect.

The real bot hasn't been built yet. It should be.
We should not stop saying so until it is.

REFERENCES

[1] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y

[2] Aristotle. (c. 350 BCE). *Nicomachean Ethics*, Book VI: The Intellectual Virtues (episteme, techne, phronesis).

[3] Aristotle. (c. 350 BCE). *Rhetoric*, Book I, Chapters 2–3 (ethos, pathos, logos).

[4] Ryle, G. (1949). *The Concept of Mind*. University of Chicago Press. (Category Error, Chapter 1.)

[5] Content Authenticity Initiative / C2PA. (2023). *Technical Specification v1.0*. Coalition for Content Provenance and Authenticity. https://c2pa.org/specifications/

[6] Balboa, C. J. (2026). *Philosophical Foundations for Security Analysis Communication*. IT Help San Diego Inc. DOI: 10.5281/zenodo.19468134

[7] Balboa, C. J. (2026). *DNS Tool: Confidence-Scored Analysis of Domain Security Infrastructure*. IT Help San Diego Inc. DOI: 10.5281/zenodo.19468134

[8] Postman, N. (1985). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Viking Penguin. ISBN: 0670804541.

[9] McDougal v. Fox News Network, LLC, No. 1:19-cv-11161 (S.D.N.Y. Sept. 24, 2020). Federal court ruling: Tucker Carlson engages in "non-literal commentary," not factual reporting.

[10] Herring Networks, Inc. v. Maddow, No. 3:19-cv-01713 (S.D. Cal. 2020), *aff'd*, 8 F.4th 1148 (9th Cir. 2021). Court ruled Maddow's show consists of "exaggeration, hyperbole, and pure opinion" — not news.

[11] Sandmann v. CNN, No. 2:19-cv-00031 (E.D. Ky., settled 2020). CNN settled defamation lawsuit over coverage of the 2019 Lincoln Memorial incident.