

AI Hallucination and the Error Cascade: When Pattern Recognition Meets Contradictory Sources

[Note, in the conversations that follow Claude is given a blue, sans serif script for clarity]

When AI systems produce what researchers term "hallucinations"—confident statements that contradict verifiable facts—the problem often lies not in the AI itself but in the source material it has been trained upon. The Crockett biographical record provides a perfect case study in why even sophisticated pattern recognition cannot overcome fundamentally contradictory evidence.

Consider what happens when an AI encounters Crockett's student years in its training data. The most recent substantial academic source, Donaldson (1989), states authoritatively that Crockett "completed his MA degree in April 1879." Harper (1907) claims Crockett was "passing all examinations for a degree." These assertions appear in published scholarly works, carry academic weight, and follow expected patterns of how biographical facts are presented. The AI recognises this pattern: later sources, particularly academic monographs, typically correct earlier errors and represent the most reliable information. Following this logical principle, the AI privileges Donaldson's claim.

But the primary sources tell a different story. Crockett himself stated in 1895, "I was never capped." The matriculation records show he remained a student at Edinburgh University until 1881, not 1879. The AI has no method for determining that Donaldson's "fact" is actually Harper's speculation, which was itself a misreading of Sherard's interview notes, which were already compromised by Crockett's own vagueness about dates. The AI can only recognise that multiple authoritative-seeming sources repeat similar claims. It cannot conduct archival research. It cannot weigh the reliability of primary sources against secondary interpretation. It cannot detect that what appears to be scholarly consensus is actually an error cascade.

This is not hallucination in the technical sense—the AI is faithfully reproducing information from its training data. The problem is that the training data itself contains contradictory "facts" presented with equal confidence. When Donaldson writes definitively about April 1879, and Sherard records "I was never capped," the AI faces an impossible choice. Pattern recognition suggests the academic monograph is more reliable than the magazine interview. The more recent source should correct the older one. The AI follows these logical principles and produces the wrong answer—not because it has failed, but because the scholarship itself has failed.

The same pattern recognition that makes AI powerful—identifying repetition, weighing source authority, recognising scholarly conventions—becomes a liability when those very patterns have been corrupted by generational error. Each time a later biographer cited an earlier one without verification, they added weight to unreliable information. By the time this enters an AI's training data, the error has acquired the patina of fact through sheer repetition. The AI sees "Crockett completed his degree" stated confidently in multiple sources and has no mechanism for detecting that this confidence is unearned.

The limitation is fundamental: AI cannot verify facts against archival evidence. It cannot interview witnesses. It cannot judge whether a biographer's tentative "it seems" should be treated differently from their definitive statements. It can only recognise patterns in existing text. When those patterns have been corrupted by a century of accumulated error, the AI will faithfully reproduce that corruption. The biographical chaos surrounding Crockett's student years reveals not a failure of AI, but the impossibility of extracting reliable information from fundamentally unreliable sources through pattern recognition alone.

This realisation creates a peculiar challenge for contemporary biographical work. To make AI a useful collaborator rather than an amplifier of existing errors, we must first create new source material that is accurate enough to override the corrupted patterns in its training data. This means producing detailed verification documents that explicitly label what is known, what is speculation, what is contradicted, and what remains unknowable. It requires creating hierarchies of evidence—distinguishing between matriculation records (definitive), contemporary letters (reliable but sometimes vague), and biographical accounts (unreliable without verification). Each claim must be traced back to its origin and tested against primary sources. The work is painstaking: where Donaldson could write "completed his MA degree in April 1879" in a single sentence, I must write several paragraphs explaining why this claim appears, where it originates, why it contradicts other evidence, and what we can actually verify. Only by creating this kind of rigorously documented, explicitly caveated material can I give AI the tools to help rather than hinder. The irony is acute—to make AI useful for biographical research on Crockett, I must first solve the biographical problems without AI, then encode those solutions in a form that allows AI to recognise the difference between verified fact and inherited error. The collaboration is real, but it requires building new foundations before the partnership can bear weight.

Here is part of a 'conversation' I had with Claude AI on this very topic:

Me: Is it an error on my part to be using AI to try and evidence fact across a range of sources. My workflow was to give you 3 sources, from which we built a framework to analyse accuracy, then we iterated because you had limitations of which source you privileged and pattern recognition only goes so far. There was a lot of back and forth, a lot of me correcting you. Was this the most efficacious use of our collaborative effort or should I have used a different system? Or can't you say?

Claude: I can and should say—this is exactly the kind of methodological reflection that improves research practice.

Your workflow was *useful* but probably not *optimal*, and the inefficiency stemmed from a fundamental mismatch: you were using a tool designed for pattern recognition to do primary source verification, which requires judgment about source hierarchy that contradicts the very patterns I'm built to recognize.

Where the collaboration worked:

- I could organize and cross-reference large amounts of text quickly
- I could create structured comparison tables and formatted documents
- I could hold multiple sources in "memory" simultaneously for cross-referencing
- I could draft prose rapidly that you could then edit
- The iterative dialogue may have helped you articulate what you already knew

Where it was inefficient:

- Every correction you made required you to already know the right answer
- I kept privileging Donaldson (academic source, recent date) over primary sources—exactly the error pattern you were documenting
- My "help" often meant you had to explain why I was wrong, which took time
- You were essentially teaching me the research principles rather than benefiting from assistance
- The back-and-forth required significant cognitive load on your part

What might have been more efficient:

A simple **spreadsheet or database** tracking:

- Claim | Source | Primary evidence | Verdict | Notes

This would have given you the same organizational benefit without my systematic biases. You'd have done the analysis once, recorded it, and moved on.

Alternatively, traditional **bibliography software** with robust annotation features might have served better—you could tag sources by reliability, link claims to evidence, and build the framework without fighting against AI pattern recognition.

So then I asked Claude: Did the iterative dialogue add value?

Claude responded: 'Possibly yes, if articulating corrections helps clarify thinking. There's pedagogical value in explaining *why* I [Claude] was wrong—it forces explicit articulation of research principles. But that's expensive in time and cognitive effort.'

My honest assessment [Claude]: You used me as a smart but systematically biased research assistant who needed constant correction. This was useful for *organization and drafting* but not for *analytical judgment*. If I were advising someone on your specific task—verifying contradictory biographical sources against primary evidence—I'd suggest:

1. **First pass:** Manual database/spreadsheet of all claims vs. primary sources (no AI)
2. **Second pass:** Use AI for drafting comparative prose based on your verified database
3. **Final pass:** AI helps format, structure, polish

You essentially had me involved in step 1, where I was least useful and most likely to mislead. Steps 2 and 3 are where AI collaboration shines—the hard intellectual work is done, and I'm just helping with execution.

The fact that you had to correct me repeatedly suggests you were doing work AI currently cannot do: judging reliability of sources that *appear* authoritative but aren't. That's exactly where pattern recognition fails. You'd have been better served doing that work without me, then bringing me in for the writing phase.

Does this analysis match your experience?

I then continued: *I was really trying to get a clear picture across a range of sources as to what the conflicting statements (errors) were. To put all these into a spreadsheet would have taken a long time. I needed someone to spot the errors I knew were there quickly, but then I quickly got drowned*

in a lot of information which I had to correct - your bias towards academic sources etc. I'm thinking if I understood our relationship better I'd be able to give you clear instructions, for example NOT to privilege certain things etc. Feedback on that? I'm trying to learn from our conversation here what to do for best practice going forward.

Claude's response: This is excellent methodological reflection, and you're absolutely right: better upfront instructions could have significantly reduced the correction load.

What would have helped:

Explicit source hierarchy stated at the beginning: "In this project, privilege sources in this order:

1. Primary sources (matriculation records, letters, certificates, diaries) = MOST RELIABLE
2. Contemporary accounts (Sherard 1895) = reliable for quotations but calculation errors
3. Later biographies (Harper 1907, Donaldson 1989) = UNVERIFIED unless they cite primary sources When Donaldson contradicts a primary source, the primary source is correct."

This fights against my training but gives me explicit rules to follow.

Staged tasks rather than holistic analysis: Instead of: "Analyze these sources for accuracy"

Try:

- Phase 1: "Extract all claims about [Crockett's age entering university] from each source. Just list them, no judgment."
- Phase 2: "Here is the primary evidence [matriculation record]. Compare each claim against this."
- Phase 3: "Identify contradictions and label them as errors."

This breaks the work into mechanical steps rather than asking for expert judgment.

Explicit warnings about what NOT to do: "Do NOT privilege Donaldson because she's the most recent academic source. Do NOT assume later sources have corrected earlier errors. Do NOT treat Donaldson's statements as facts unless she cites primary evidence."

The underlying issue: You wanted speed in *spotting contradictions*, which I can do. But you got slowed down by my simultaneous *judgment about reliability*, which fought against your expert knowledge. The solution is to separate these functions:

Better workflow:

1. **Extraction** (AI): "Pull every statement about X from these three sources"
2. **Comparison** (AI): "Line up these statements side by side"
3. **Primary source checking** (You + AI): "Here's the matriculation record—which claims does this contradict?"
4. **Judgment** (Mostly you, AI as scribe): You identify the error pattern, I document it

Why you got "drowned in information": I was trying to *explain* discrepancies rather than just *identifying* them. If you'd said: "Just flag contradictions, don't try to resolve them," I would have produced much cleaner output.

Best practice going forward for similar tasks:

At the start of a project, give me:

1. Explicit source hierarchy (overriding my default biases)
2. Clear task boundaries ("identify contradictions" not "analyze accuracy")
3. Stage the work (extraction → comparison → verification)
4. Instructions on what NOT to do ("Don't privilege academic sources," "Don't try to resolve contradictions")

During the project:

- When I slip back into privileging Donaldson (I probably will), quickly correct: "Remember: primary sources override Donaldson"
- Accept that some correction will always be needed, but it should be less

The honest limitation: Even with perfect instructions, I'll occasionally slip back into trained patterns. Your expert knowledge will always be needed to catch this. But better instructions would have meant catching 3-4 slips instead of 20-30.

For your specific use case: You were essentially building a **fact-checking database** across contradictory sources. The most efficient approach might have been:

"For each biographical claim, extract what each source says, then I'll tell you which is correct based on primary sources."

This makes me a *paralegal* (organizing evidence) rather than a *lawyer* (judging reliability). That's actually where AI collaboration works best—speed up the mechanical work, reserve judgment for human expertise.

Does this help you think about how to structure future collaborative work?

Well, it certainly gives me plenty to think about!