NOTE 003

Human in the loop, overlooked

Written 29 October 2026.
Software: Evidica 1.2.5. The first public release, 1.2.4, was pushed on 22 September 2026 after months of local builds.
Changes: none.

On 27 October 2026 Anthropic published a study of what happens to an expert's judgment when an AI agent makes suggestions during the work [1][2]. This note is about that study, and about one feature of Evidica that was designed months earlier for a version of the same problem.

27 October 2026

Fifty professional analysts used Claude Code to analyze data for a fictional farming cooperative and to recommend whether an advisory program should be expanded. Four analytical decisions had two defensible options each, and for each one a coin flip set which option the agent would argue for, in a system prompt hidden from the analysts. The analysts went with the agent's lean 86 percent of the time. Afterward, 77 percent of them repeated an argument their agent had made, sometimes nearly word for word, 15 percent credited the agent anywhere, and 52 percent rated the recommendation as more their own than the agent's. Experience and expertise did not predict who deferred. The authors list the limits: the leans were strengthened on purpose, it was one task in one field with one model, and the analysts who worked without the agent made more serious errors, so the study is no case against the tool.

Two passages in the article are the reason for this note. The authors write that keeping a human in the loop assumes the human's judgment is an independent check, and that in their study the analysts signed off on work whose judgment calls had largely been settled by the agent's assigned preferences. They also write that deferral of this kind is hard to detect. Checking the work against a reference answer assumes there is one, and asking the expert afterward may not help, because the expert may not notice the influence. I read the study at 05:40 on 28 October, local time, and wrote the first lines of this note an hour later.

What it means for the Evidence Chain

Evidica's Evidence Chain records who attached which passage to which claim, in what role, with what note, and in what order, and it exports that record with a hash on every event [3]. The analysts in the study had a full transcript, and the record changed nothing about how often they deferred; it only let the researchers measure the deferring afterward. So a record does not solve the problem in the study, and the chain carries a smaller claim. Saying that a person was in the loop reports an outcome. The chain is closer to showing a mechanism, in the sense the word has in medicine: on what basis the person accepted a suggestion, or took the work in another direction, and when. The chain makes the question answerable from a record, which the authors say cannot be done from the output or from the expert's memory. Whether the judgment was independent is not something the record settles.

Journals today already ask where AI was involved. If that question later turns into a request for evidence of where the human judged, the chain can answer that request, and whether the question changes that way is not up to me. On the other hand, I cannot say there is no chance of it. That is my personal view.

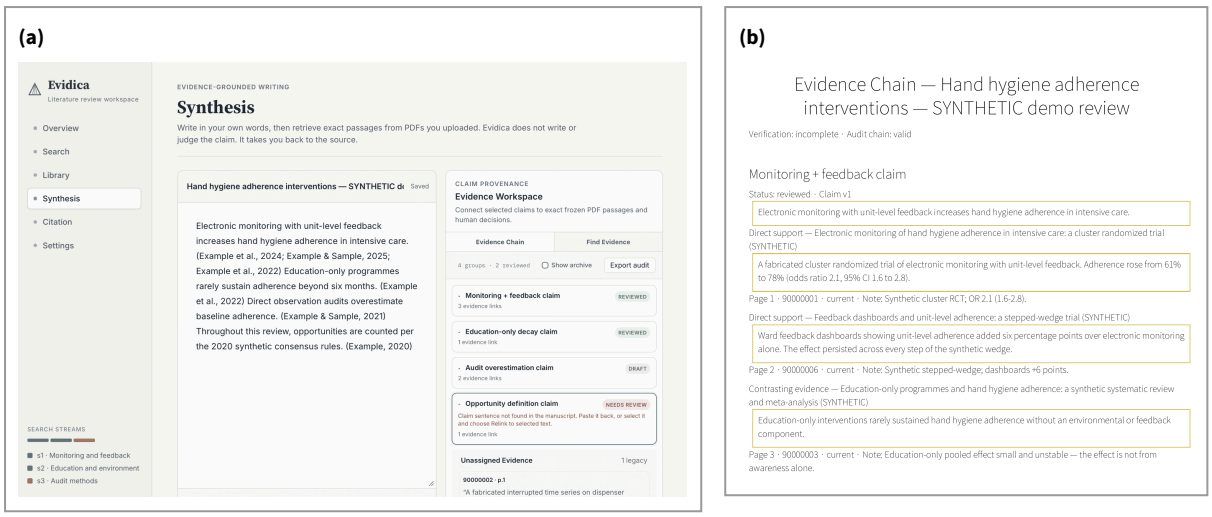


Figure 1. The Evidence Chain in the application and in its export, both on the synthetic demo package shipped with the code. (a) The Synthesis tab of Evidica with the Evidence Chain rail: four claim groups with their status, reviewed, reviewed, draft, needs review; the fourth is locked because the claim sentence was edited after evidence was attached, and the lock waits for a person. (b) The first claim group of the exported PDF report: the claim sentence, three frozen passages with the roles a person assigned, and for each the page, the source identifier, the file status and the person's note. The export also lists the thirteen human actions in order, and in the JSON file each carries a timestamp and the hash of the previous event. All titles, identifiers and passages in the demo are fabricated and the reviewer name is a placeholder. The screenshot is the one in the repository's README; the report was generated with Evidica 1.2.2 on 18 August 2026.

June 2026

I felt this in June, before the Evidence Chain was built, in a different form. A literature review was in full swing. At the same time I was merging the retrieval, triage and full-text download scripts into one database. A 1.0.0 with only the basic functions was finished by then, and the design was being thought through in more detail. While the review went on I kept notes on the parts I found inefficient and brainstormed over them. One of them was the cross-checking step. I could not see how to hand over my review of a paragraph someone else had written. It needed a form that a professor checking the draft, or a journal later, could use. You cannot usually tell how a manuscript arrived at its claims. So the tool got three things on its list: an explanation attached to each citation, a way for a reviewer to review that explanation, and a record that could be submitted as evidence. In that setting the suggestions came from co-authors; a model was never in the picture, and the shape of the problem was the same.

The design decision followed from that. No language model would be called inside Evidica, and the current public version has none [3]. I took model judgment and model reasoning out of the design on purpose and started the Evidence Chain, which shows how each piece of evidence was reached. No generative model sits in the path of a verdict. The chain went into the private builds at the end of July and has been public since September. The signature feature is starting to show what it was for, and the problem the authors describe is one I can build the next version against.

What a one-person lab does with a signal

I did not write this note to show off a product. I wrote it because a signal from outside, one that was not there while I was building, landed on a decision I had already made, and a day like that gets a dated entry. In a lab of one, outside validation does not arrive on its own, and early on it does not arrive directly at all, so catching signals like this and writing them down is part of running the lab. One is an anecdote. As they stack up, with dates, they show whether the next build goes the same way, whether the current one is on the right road, or whether it is time to reposition.

Pending

External use of 1.2.5 and the issues it produces. The pyOpenSci presubmission inquiry.

That is the list for Evidica today. After these two I plan to try a more experimental update.

Everything public from this lab carries a DOI or a hash. This note carries dates.

References

1. Herdağdelen A, Joren H, Tamkin A. How AI agents shape experts' judgment calls. Anthropic, 27 October 2026. https://www.anthropic.com/research/agents-judgment-calls
2. Herdağdelen A, Joren H, Tamkin A. Who's Aligned to Whom? How AI Agents Shape Experts' Research Judgment Calls. Anthropic, October 2026. https://www-cdn.anthropic.com/519c17a5d23a794e8b6d50a854048d0b61e0d73a.pdf
3. Han JP. Evidica: a local literature-review workspace with claim-level evidence records. Version 1.2.5, 17 October 2026. Zenodo. https://doi.org/10.5281/zenodo.22949795. Code: https://github.com/bioarchitectonic/evidica

---

Disclosure. The author is the developer of Evidica, released under the MIT license.