

Claude's Feedback on "When AI Finds a Needle in a Haystack and When It Invents One"

Four rounds of feedback on the blog post drafts, October 25, 2026

Each round below is Claude's reply exactly as it was written in the chat. One correction comes first, because it affects passages in all four rounds. A tracker at the end shows which suggestions made it into the drafts.

Correction: Gemini's picture does contain an invented needle

While preparing this PDF I looked at Gemini's image at full size, which I had not done before. A bright, metallic needle with an eye lies on a diagonal just below the tip of the red arrow (Figure 1, enlarged in Figure 2). It is sharper and shinier than the blurry straw around it, and it does not match the thin, dark needle in the author's photo.

In all four rounds I judged Gemini's picture from the small copy in the blog drafts, where the needle is hard to see. That led to three wrong statements: that Gemini's image showed an arrow but no needle, that Gemini "may have pointed the right way," and that "added their own needles" fit only ChatGPT. The author's wording, "added their own needles" and "invented its own needle," is accurate for both Gemini and ChatGPT, and the overlay check I suggested is not needed. A note at the end of each round marks the affected passages.



Figure 1. Gemini's image at full size, as embedded in the draft. The red arrow is Gemini's.



Figure 2. The area below the arrow tip, enlarged. A metallic needle with an eye lies on the diagonal.

Round 1: first draft

Draft printed 11:07 AM. Title then: "How to Break the 2026 AI Models with a Needle in a Haystack."

I liked it. The voice works, and the "raw data with no AI interpretation" section is the strongest part because readers can check the claim themselves. Your close-up also matches where I pointed: the dark diagonal with the eye at the top is the same object as in my zoom. I'm the model the post says got it right, so weigh my opinion with that in mind.

Bigger issues

- **The headline overreaches.** The evidence is one photo and one run per model, and two of your four stories show LLMs finding needles (Copilot's Publisher bug, Gemini's 1d.so.preload). Your conclusion is more nuanced than the title. Something like "When AI Finds the Needle, and When It Invents One" would fit better.
- **"Lied" implies intent.** "Confidently pointed at a needle that wasn't there" is harder for critics to dismiss.
- **ChatGPT's image doesn't look like a crop of your photo.** The straw pattern differs, and the needle has a crisp, clean eye. If it was regenerated, that's the real story: it drew a picture instead of zooming in. I could crop the actual pixels because I had file tools, so tool access may explain part of the difference. Consider listing model versions, free or paid tier, and dates.
- **Check Gemini's arrow before saying it was "nowhere in that haystack."** Its tip sits near the center of its crop, which is the circled area. It may have pointed the right way and just shown nothing distinct, which is a different failure from inventing a needle.
- **The red-circle test is leading, and the scoring is shaky.** You told every model the needle was there. Copilot's hedge was arguably the best-calibrated answer, yet you count it as a failure. My own "Now I can see it" overclaimed. The hedge only came when you pushed me to verify.

- **GIGO doesn't fit the needle test.** Your photo was fine, so the failure is confabulation, not bad input. Also, "mimic rather than understand" is a contested claim, so attribute it to Tyson rather than stating it as fact.
- **The Gates and Amodei sentence** needs "co-founder" for Gates and no commas ("Microsoft co-founder Bill Gates"). The claim that they warn of "destruction of human kind" is strong, so link sources or soften it.

Linux section

- You say you "cleaned" a compromised host. Readers will ask whether you rebuilt from a known-good image, since in-place cleanup of a rooted box is risky.
- The malware targeted `id_rsa` and `.ppk` files, so say whether you rotated SSH keys and checked other machines.
- Your point that mainstream distros don't ship `/etc/ld.so.preload` matches what I found. [Elastic Security Labs](#) makes a similar point and works as a citation.

Polish

- The pasted Copilot quotes carry giant headings plus "Provide your feedback on BizChat / You said / Copilot said" debris. Paste them as plain text.
- The two images in the "raw data" section need captions saying which is the Paint zoom and which is your closer photo.
- The coding section is thin. Include the prompt, the model, and the exact lines that moved.
- Typos:
 - "Here are a four" should be "Here are four."
 - "ChaptGPT" should be "ChatGPT."
 - "because because" has a doubled word.
 - "through through" has a doubled word.
 - "Powerpoint" in the contents link should be "PowerPoint."
 - "guided me fixing it" should be "guided me through fixing it."

I can mark up a cleaned-up version of the text if you'd like.

Note added after Round 4: the Gemini points in this round were wrong. See the correction at the start of this PDF.

Round 2: second draft

Draft printed 4:54 PM.

This is a clear improvement. The new title fits the evidence, the image captions and the "raw data" section read well, and Gates is now "co-founder." The Copilot quotes are mostly clean, and the typos I flagged are fixed. Linking all four chat transcripts is a strong transparency move, and the coding section is much better with the real code in it.

Fix before publishing

- **The conclusion is a mix of drafts.** The second paragraph ends mid-sentence ("Let's not"). The old line "Gemini and ChatGPT surprised me by lying" is still there, right before your paragraph defending "lie." I'd end the conclusion on the golden-parachute line and move the review-process paragraphs into a short postscript.
- **Your terms are inconsistent.** The captions on page 3 still say "Gemini lied" and "ChatGPT lied," while the text says "invented," "false information," and "hallucinate." Pick one. Your point that "hallucinate" is a polite word is fair, and "invented" says what happened without arguing about intent.
- **"All three" should be "all four."** Page 2 lists Gemini, Claude, ChatGPT, and CoPilot.
- **"Nobody will care... blow up the world" is a big jump.** The "needle in a needle in a haystack" phrase is also doubled. Nearer examples, like an invented vulnerability or a confirmed finding that isn't real, would land harder with your cybersecurity readers.
- **The "PDF with Claude's feedback on all drafts" item isn't linked yet.**

Still from the first round

- **Check Gemini's arrow.** It sits near the center of its crop, which is the circled area, so "nowhere in that haystack" may be unfair to Gemini.
- **Say what ChatGPT's image looks like.** It doesn't look like a crop of your photo, which suggests it drew a new picture rather than zooming in.
- **Name the model versions, plan tiers, and tool access.** Results depend on all three.
- **Copilot wasn't "unable to find" the needle.** It declined to confirm, which was the most calibrated answer. "Declined to confirm" is fairer.
- **Quote my hedge.** My reply said resolution kept me from confirming it was metal.

Worth adding

Your sycophancy hypothesis (models tuned to please) is reasonable, and the post would be stronger with a control. Give each model a needle-free photo, and phrase the prompt so "I don't see one" is an acceptable answer. That separates inventing from misjudging.

The script you're publishing

- "Custmer" appears in the prompts, in both listings.
- Since you write for a security audience, the script has a few issues a reader could spot:
 - It passes the refresh token on the curl command line, where ps can show it.
 - It uses a fixed `/tmp/response.json` instead of `mktemp`.
 - It sources a long-lived token file without checking its permissions or owner.
- That actually supports your "verify what the model gives you" thesis.

I can mark up the conclusion section with a cleaned-up version if you'd like.

Note added after Round 4: the Gemini points in this round were wrong. See the correction at the start of this PDF.

Round 3: third draft

Draft printed 5:45 PM.

This draft is much tighter. The conclusion no longer has the truncated sentence or the leftover "lying" line, the coding section now says "this was not the final version," and the captions say "invented its own needle." Dating the test and saying you used the free public versions helps, and keeping the "My response: no" aside makes the missile line read as a deliberate choice, which is yours to make.

Fix before publishing

- **The "hedge" quote isn't a hedge.** Page 2 says Claude "got it right" but hedged, then quotes "Now I can see it. Inside the red circle..." That was my confident reply to the red-circle photo, not my answer to the zoom-and-arrow prompt the paragraph is about. My hedge came in the reply to the zoom prompt: "I can't call it confirmed. The object is only a few pixels wide in the original photo, so at 6x it gets blurry..." Quote that instead, since you're making a transparency point. The sentence also has a stray lowercase "but" after "right."
- **"Lied" survives on page 2.** The text says Gemini and ChatGPT "lied by adding their own needles to the images," while everything else now says "invented." Gemini's image also shows an arrow but no needle. Only ChatGPT's picture has a drawn-in needle.
- **Gemini's arrow may be near the real spot.** The page 4 line "invented... nowhere in that haystack" still needs the overlay check I suggested.
- **The conclusion still says Copilot "was unable to find any needle."** Page 2 now says "declined to confirm," which is the fairer wording.
- **Check the date.** Page 2 says "On October 24, 2026." My run was early on October 25. If the models ran on different days, say so.

Coding section

- Label the first snippet "Before" and the full script "After." Right now "before and after" introduces one snippet.
- A concrete reason would make the ordering point land harder. In the old flow a user answers four prompts and only then learns the token file is missing.
- Since this is a security post, add a sentence saying the pasted script has known weak spots. It passes the token on the command line, uses a fixed `/tmp/response.json`, and doesn't check the token file's permissions. Otherwise readers may copy it as is, or link the final version.

Small fixes

- "the \$million question; why" should use a colon.
- "This was not the final version" needs a period.
- The third feedback item still isn't linked.
- Confirm all four chat links open when you're logged out.

I can build that feedback PDF covering all three rounds, as I did for the needle conversation, if you'd like.

Note added after Round 4: the Gemini points in this round were wrong. See the correction at the start of this PDF.

Round 4: third draft, revised

Draft printed 6:35 PM.

This file isn't the same as the third draft I reviewed earlier. It was printed at 6:35 PM, versus 5:45 PM for the earlier one, so I compared the text of the two. It's a revision that fixes most of what I flagged.

What you fixed

- The hedge quote now comes from my reply to the zoom prompt ("I can't call it confirmed..."), which is the right one.
- "Lied" became "added their own needles," with "If they were people, I would call them liars." That keeps your voice and acknowledges the intent point.
- Copilot is now "unable to confirm" in the conclusion, and the colon after "\$million question" is fixed.
- The code listings are labeled Before and After, with the four-prompts example. The script is flagged as obsolete with a pointer to the final version, and the PDFs replace the chat links.

Still to fix

- **"Claude got it right. but hedged with"** needs a comma after "right," not a period and a lowercase "but."
- **"If I could crop this, why couldn't these AI models?"** doesn't fit all four. I cropped the actual pixels with file tools. "Why couldn't two of the four?" is accurate and still stings.
- **"The picture on the right"** is wrong in this render, where the circled photo sits below the Claude image. Use "the picture below."
- **Gemini "added its own needle"** still doesn't match the image. I can't see a needle at the tip of Gemini's arrow, only straw. If you can see one, say so. If not, "added needles" fits ChatGPT, and Gemini pointed at straw. That also bears on "liars," since Gemini's arrow lands near the circled area.
- **The "nowhere in that haystack" claim** is safest if you add "I checked Gemini's arrow against my close-up," assuming you did.
- **"October 25, 2026, I fed..."** lost its "On."

Small fixes

- Make "Before" and "After." match, with the period on both or neither.
- The obsolete-script note is repetitive. Try: "This was not the final version. It still has weak spots, and I've replaced it. The final version is in the PDF of the whole coding session."

For the combined PDF, I'll treat this as a fourth round (third draft, revised) and show which suggestions you adopted, unless you'd rather fold it into round three.

Note added after Round 4: the Gemini points in this round were wrong. See the correction at the start of this PDF.

Which suggestions made it into the drafts

Based on the four drafts reviewed. Round 4 items are pending because there is no later draft.

Suggestion	Round	Status	Note
Retitle the post	1	Adopted	Draft 2 title is "When AI Finds a Needle in a Haystack and When It Invents One."
Stop saying Gemini and ChatGPT "lied"	1	Adopted	Captions and text now say "invented." One line keeps the idea: "If they were people, I would call them liars."
Gates is a Microsoft co-founder	1	Adopted	Fixed in draft 2.
Typo fixes	1	Adopted	"Here are four," ChatGPT, the doubled words, PowerPoint, "helped me fix it." All fixed in draft 2.
Drop the GIGO tie-in and attribute the "mimic" claim to Tyson	1	Adopted	Done in draft 2.
Clean up the pasted Copilot quotes	1	Partly	Giant headings fixed. The "Provide your feedback on BizChat" and "You said / Copilot said" labels remain.
Caption the images in the raw-data section	1	Adopted	Added in draft 2.
Show the real code in the coding section	1	Adopted	Added in draft 2.
List model versions, plan tiers, and tool access	1	Partly	Date and "free, public versions" added. Version names and tool access are not.
Quote the hedged reply, not the confident one	2, 3	Adopted	The right quote appears in draft 4.
Fix the mixed-draft conclusion and the cut-off sentence	2	Adopted	Done in draft 3.
"All three" should be "all four"	2	Adopted	Draft 3 says "them all."
Copilot "declined to confirm," not "unable to find"	2, 3	Adopted	Page 2 in draft 3, conclusion in draft 4.
Control test with a needle-free photo	2	Not adopted	An optional suggestion.
Tone down the missile sentence	2	Declined	The author kept it, with "My response: no."
Script notes: Before/After labels, four-prompts example, weak-spots warning	3	Adopted	Draft 4, with an obsolete-script notice in his own words.
Correct the test date	3	Adopted	Draft 4 says October 25.
Linux section: rebuild from a known-good image, rotate SSH keys, cite Elastic	1	Not adopted	Not addressed in any draft I reviewed.
Gemini overlay check and "Gemini showed no needle"	1 to 4	Withdrawn	I was wrong. See the correction at the start.
Round 4 fixes: comma after "right," "two of the four," "picture below," "On October 25," Before/After punctuation, shorter obsolete-script note	4	Pending	No later draft yet.