Checking an AI answer against a published study – MECC train the trainer

22 Oct 2026

For evaluation commissioners and practitioners wondering whether an AI-assisted analysis of interviews can be trusted, and how they would know. Part of Pilots, not passengers — reconstructible reporting in evaluation.

We asked Rubicon, our plugin for Claude, one evaluation question about a set of published interviews, and then checked its answer in two ways: against what the original researchers concluded from the same interviews, and sentence by sentence against the transcripts themselves, by a reviewer that did not know which text it was reading.

The interviews and the question

Making Every Contact Count (MECC) trains health and care staff to use everyday contacts with the public for short conversations about health. In the North East and North Cumbria, a "train the trainer" course was meant to produce people who would then train colleagues in their own organisations, which is called cascading. Researchers at Northumbria University interviewed 13 people who attended that course, and deposited the anonymised transcripts in the UK Data Service's ReShare repository under a Creative Commons licence (10.5255/UKDA-SN-857461). The same team published two papers on them (Nichol et al. 2026, *British Journal of Health Psychology*; Nichol et al. 2026, *International Journal of Behavioral Medicine*).

The question we gave Rubicon, in full: "What stopped attenders of the MECC Train the Trainer session from going on to deliver (cascade) the training in their own organisations, and what helped the ones who did? Read across all 13 attender interviews, comparing those the document list records as having cascaded with those it records as not having cascaded, and including what interviewees report about other attenders they trained, managed or met."

What Rubicon found

The full report is public: click any number in it to see the passages behind it. The run took 57 minutes. At Anthropic's API list prices its model use would have cost about USD 24, but it ran on a Claude Max subscription (the USD 200-a-month tier), where it used about 1% of the plan's weekly allowance. Rubicon coded 809 passages, each with its quotation checked against the transcript, grouped them into 20 kinds of barrier and help, and counted each kind for the 7 attenders who cascaded and the 6 who did not.

- **Backing for delivery, not just for attending.** All 7 who cascaded described an organisation that wanted MECC delivered, against 2 of the 6 who did not.
- **Time.** All 6 who did not cascade described other work getting in the way of their own delivery, against 2 of the 7 who did.
- **Not decisive with this few interviews:** having delivery in the job (5 of 7 against 3 of 6), a line manager standing in the way (3 of 6 against 1 of 7), and a ready route to learners (6 of 7 against 3 of 6).
- **Confidence did not separate the groups.** Five of the 6 who did not cascade felt able to deliver, while several who did were nervous. Lack of confidence was, though, the commonest reason interviewees gave for *other* attenders not delivering.
- **A limitation it states itself:** all 7 who cascaded are women, against 3 of the 6 who did not, so the contrast cannot be told apart from other differences between the groups.

Rubicon called a difference clear only where it would survive one interview being read the other way. Every number in the report is counted by fixed rules from the coded passages, and every claim opens onto the quotations behind it. Before the answer was finished, a check listed the judgements that could go the other way and added a limitations line for each one that would change a conclusion.

Against the published study

The published papers drew on more than Rubicon did: all 21 interviews (adding six people who did not attend and the two course trainers) and comments from 373 post-course surveys. They also asked a wider question. So we compared only the findings that fall inside our question.

On those, the two agree on the main lines. Both find that backing for attending was not enough without backing for delivery, that time and capacity held people back, that one attender cascaded without training in the role because the support was strong enough, and that enthusiasm for MECC did not decide who delivered. The mixed-methods paper profiles each of the 13 attenders, and on whether delivery was part of their role it agrees with Rubicon on 11 of the 13. The differences are mostly about where a line is drawn: whether a supportive line manager counts as organisational backing, or whether knowing that delivery was expected is the same as knowing what it would involve. Our comparison sets out each difference with the transcript evidence for both readings.

We are grateful to Beth Nichol and her colleagues at Northumbria University, who did what few researchers do: they deposited their interview transcripts openly alongside their published study. That is excellent practice, and without it neither this example nor any check of it would be possible. We have every reason to think they designed and carried out their study with care and skill. The comparison above is not a verdict on it: an AI model read the papers' results against only the 13 attender transcripts, while the authors drew on 21 interviews and 373 survey responses and wrote for a journal's readers rather than for our one question.

A blind check against the transcripts

We then gave a reviewer, Claude Opus on its own, the 13 transcripts and two unlabelled texts: Rubicon's answer and the papers' results on attenders and cascade, in the authors' own words, each stripped of headings and references. The reviewer tested every sentence that says something about the evidence and listed the root errors, separating substantive ones, which would change what a reader concludes, from minor ones. We ran it twice with the order swapped.

Rubicon's answer had about 2.5 root errors per 100 sentences tested (4 errors in 160 sentences, and 4 in 153 with the order swapped), of which 1 or 2 were substantive. The main one matters: both reviewers counted at least two more of the 6 who did not cascade as having organisational backing, which would make that difference unclear rather than clear. Rubicon's own check had named that judgement as one that could go the other way, and its limitations say so, but the headline still reads more firmly than an independent reader would put it.

Tested the same way, Rubicon's answer was at least as accurate as the published results. That is not a ranking of the two. The papers were written for a wider aim, from more evidence than the reviewer was given, and for a journal rather than for one question.

We also gave Claude without the plugin the same transcripts and the first sentence of the question above, and put its answer through a blind check of the same kind: the answer and its review.

What this shows, and what it does not

It shows that one run of an AI analysis, on long, real interviews, can produce an answer that agrees with a peer-reviewed study on its main findings, holds up sentence by sentence against the transcripts, and lets a reader see exactly which judgements its conclusions rest on.

It does not show that every run will do as well, that the answer is complete, or that AI analysis can replace the kind of study the authors did, which designed the research, conducted the interviews and asked the questions in the first place. Without their decision to deposit the transcripts openly, neither the analysis nor the check would have been possible.

We sometimes use AI to help draft these pages, but the final text is ours.