How we judge an analysis when the best method is not known in advance

18 Nov 2026

Work in progress. For evaluators and methodologists who want to know how we decide whether one way of answering a question from documents is better than another, and for the people and agents who run those tests for Rubicon. Part of Pilots, not passengers — reconstructible reporting in evaluation.

Rubicon does not fix a method. A ranking, a contribution question, a typology, a judgement against a rubric and "what is going on here" each call for a different way of working, and the model chooses one for each question. What stays fixed is the record: quotations checked against their documents, every number counted by fixed rules, and the steps set out so that somebody else can rebuild them (see the jury test).

That makes testing awkward. If we knew the best method for each question in advance, we could score each run against it. We do not, and for most evaluation questions nobody does. So we judge an analysis on several criteria at once, and we judge mostly by comparison: two or three approaches to the same question, side by side, rather than each one against a standard.

What we judge

Accuracy. A reviewer who has not seen which approach wrote which answer reads each answer against its documents and lists every statement the documents do not support. We report the errors per 100 sentences of the answer's body, with a 95% interval. One blind review finds about half to two thirds of an answer's errors, so a difference of a few errors between two approaches is within the noise; a round that decides a change gets two reviews per answer, merged error by error.

Coverage. Where a question has an answer key (the things a good answer must include), we count the items each answer misses, scoring by hand on blind copies. Where there is no key, judges compare coverage against the case's rubric.

The trail. How much of the answer a reader can walk back: the share of its sentences supported by a count the fixed rules made or a quotation they checked, as against the analyst's own reading. Reading is often the right tool. The question is whether a sentence was left to reading when a step could have counted it.

Workflow fit. Whether the workflow did by fixed rules what fixed rules could do, used steps the question needed and no others, and counted the right unit (people rather than documents, say, when one person was interviewed twice). This is the criterion closest to "the best method", and the one we can least score alone.

Writing, evidence and calibration. Whether the answer is clear to its reader, whether each claim cites what supports it, and whether its confidence matches its evidence. **Parsimony** (padding, restating the question, narrating the method) is reported on its own line and never rewards brevity as such.

Cost. What the run cost at list price, since a method that is as good and cheaper wins.

Why comparison rather than scores

We tried scoring each answer on its own against a rubric. Judges marked 84% of the scores "met", and the scores agreed with head-to-head verdicts on only 11 of 18 comparisons. A judge asked "is this good?" says yes; a judge asked "which of these two is better at X?" has to look.

Comparison has its own trap: the order in which two answers are shown changes the verdict. In 160 stored pairs of a judge and a case, 36% of verdicts reversed when the order was swapped. So each pair is judged in both orders, by two judge models from providers other than the one that wrote the answers, and a verdict counts only where the same judge gives it both ways round.

Each case then scores +1, −1 or 0 on each criterion, and the criteria are weighted into one composite: accuracy three times, coverage twice, writing, evidence and calibration once each. One approach is called better than another only if a sign test over the cases' composites says so at $p < 0.05$. Otherwise the result is "not separable", which means the round was too small to tell, never that the two are equal.

Comparing two or three approaches at once

Because the best method is not known in advance, the approaches compared have to come from somewhere. Three sources are useful:

- the workflow the model chooses for itself;
- the workflow it proposes when, after the run, it is asked whether the same steps would have served the question better, or what step it would add;
- a plain model answering the same question without Rubicon.

With three approaches we judge each pair in both orders, which makes six judgements per judge per case, rather than asking a judge to rank all three at once. A three-way ranking has six possible orders of presentation, and nothing in the order effect we measured in pairs suggests it shrinks with more answers on the page.

The trail and workflow fit are the newest criteria, and neither is yet in the composite. The trail is measured by fixed rules, so the judges do not score it. To judge workflow fit, the judges see the workflow as well as the answer: the steps each approach ran with their settings, and the draft with each sentence tagged by the count or quotation it depends on. They are told not to prefer a workflow for having more steps or fewer. Workflow fit compares two Rubicon workflows with each other; a plain model's answer has no workflow to judge.

What this cannot tell us

- The judges are models. Their agreement with each other, and with people, is something we measure as we go rather than assume.
- The workflow proposed in a debrief comes from the same model that ran the first one, so the comparison tests its second thoughts, not a method from outside its habits.
- Much of the testing runs on small made-up or public corpora, chosen so that results can be shown. A method that wins there may not win on a long, messy evaluation archive.

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