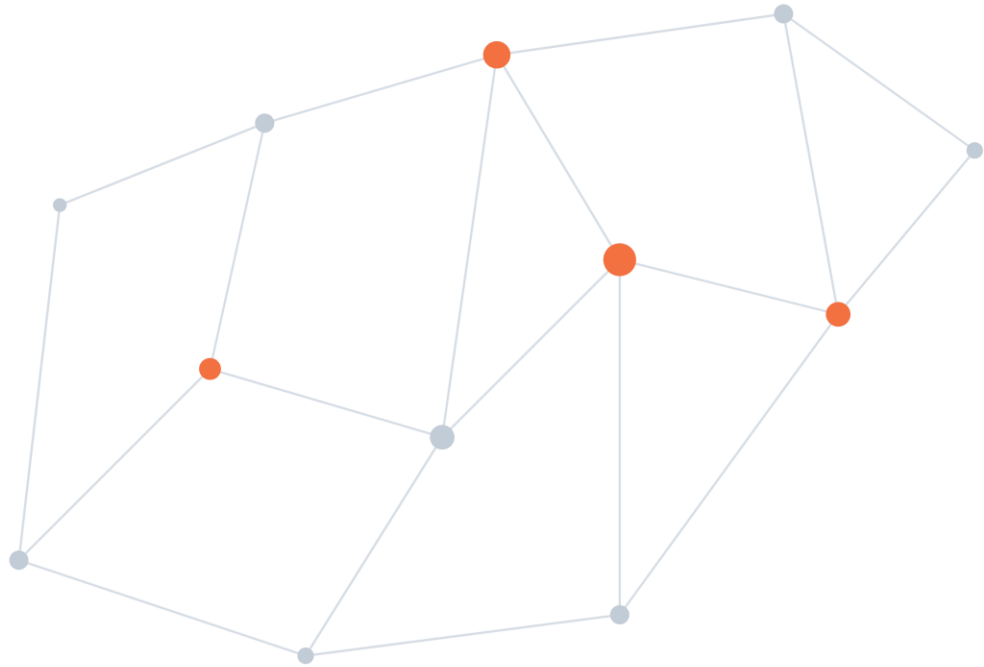
ACCUFIND

A WORKING PAPER

Evaluating grounded AI systems



How to score the output of an AI system that works from source documents, whether it answers questions or generates findings, so the score means something and two people scoring the same thing land in the same place.

AccuFind Pty Ltd August 2026

WHAT USUALLY HAPPENS

Most AI pilots in government are judged on whether people liked them

A pilot runs, a demonstration is given, stakeholders are impressed or they are not, and a decision follows. Sometimes there is a survey. It is unusual to find a number saying how often the system was right, and rarer still to find one that separates the parts of the system that could have made it wrong.

COLLECTED

What people thought

- Whether people found it useful
- Whether the demonstration was convincing
- Whether they would use it again

NOT COLLECTED

Whether it was right

- How often the system was wrong
- Which questions it was wrong about
- Whether the fault was in the retrieval or in the answer

Decisions about scaling then rest on the evidence that was easiest to collect.

WHY IT PERSISTS

The failure is invisible in a demonstration

A fluent, well cited answer that is wrong looks exactly like one that is right, and nobody opens the source.

Nobody budgeted the expert time

Building a test set means senior people reading source documents. That cost is visible and it lands early.

There is no established practice

Agencies know how to check a build against requirements. A system that is usually right does not fit that test.

WHEN IT IS DONE AT ALL

Every rubric we have used or inherited has failed in one of two directions

An evaluation rubric is the instrument that turns what an AI system produced into a number somebody can act on. It is usually written in an afternoon and rarely revisited. Two failures account for almost all of them.

Too coarse

Thumbs up or thumbs down. Fast to apply and easy to report. Anything that is not obviously bad gets a thumbs up, so an output that is useful scores the same as one that is merely true, and nothing in the result tells you what to change.

Too elaborate

Twelve things to rate, five points each, a page of definitions. Evaluators apply it carefully for the first hour and approximately after that. Scores drift toward the middle, and the drift is invisible because the instrument looks rigorous. The rubric gets abandoned quietly rather than revised.

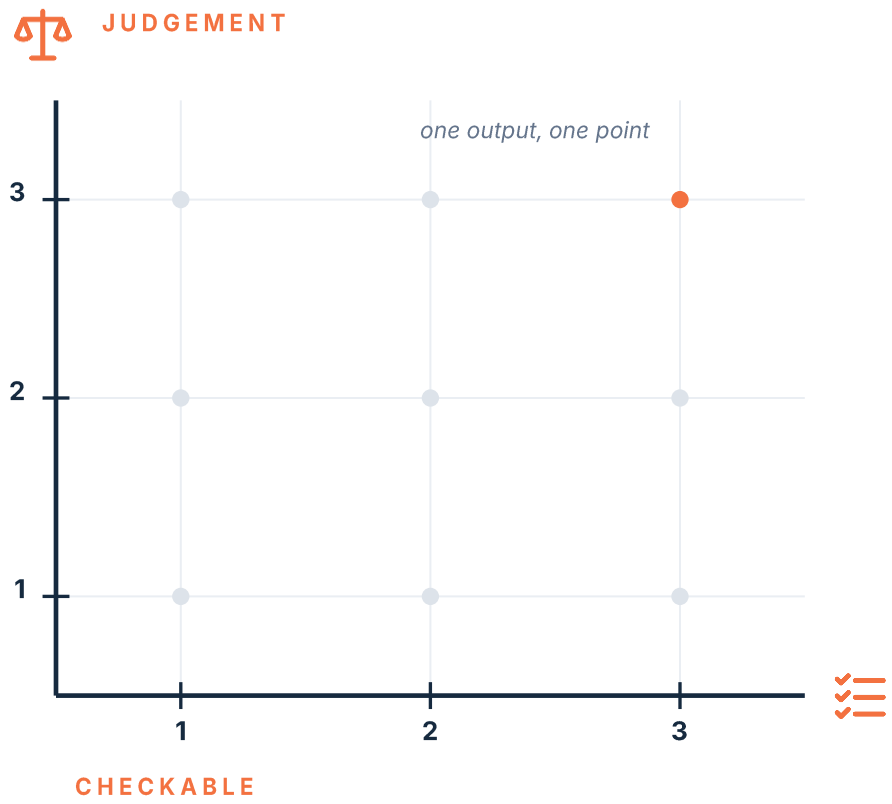
Both ends fail for the same reason

One score is carrying two questions. Whether the output is correct, and whether it is worth anything. Collapse them into one number and two evaluators disagree without either being wrong, because they are not answering the same question.

THE PRINCIPLE

Score two axes and keep them apart

There are two questions worth asking about any output, and they are not the same question. Ask them separately and each answer means something.



The checkable axis

Can be checked against a source by anyone able to read it, and two people working separately will land on the same answer. It does not need a subject matter expert, and putting one on it wastes the scarcest thing you have.

The judgement axis

Needs someone who knows the subject, and two of them can reasonably disagree. This is the question of whether the output is worth acting on. No rubric design removes the judgement, and it should not pretend to.

Read the pair, not the total. Adding the two scores together throws away the information you went to the trouble of collecting. The pair tells you which part of the system failed, and therefore what to change.

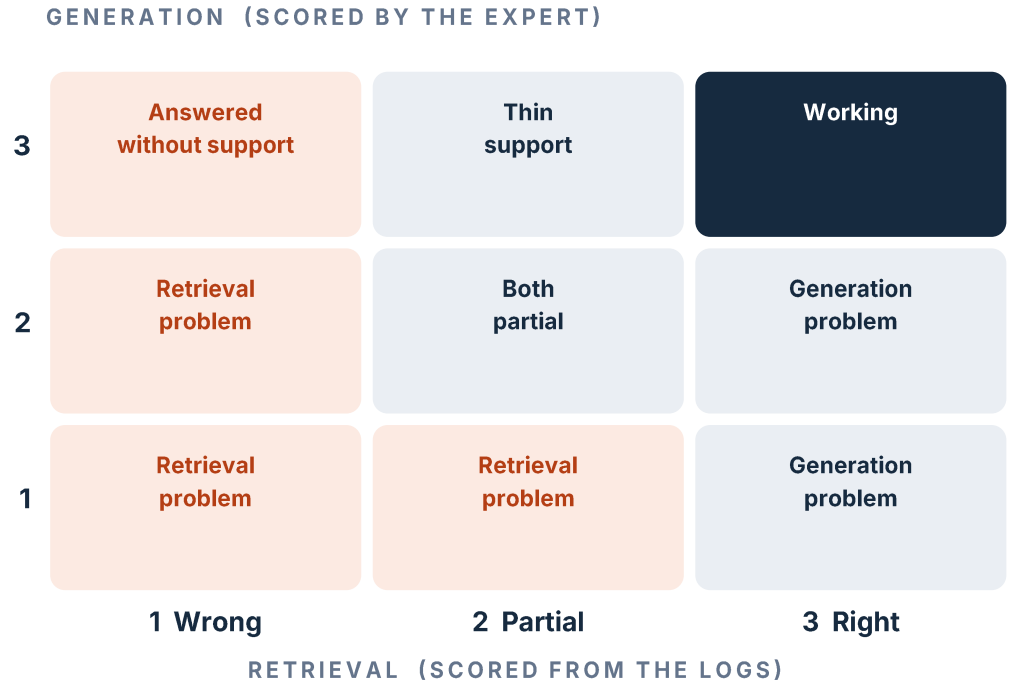
THREE POINTS, NOT FIVE

Give people five options and the unsure ones pick the middle, so you end up with a pile of threes that mean nothing. Uncertainty is the thing you most want to see. Three points force a call: wrong, partly right, right.

WHERE THERE IS A RIGHT ANSWER

Retrieval and generation fail separately, so score them separately

A grounded system finds material, called retrieval, then writes from it, called generation. That holds for one question answered once or an agent looping for evidence. Either can fail while the other works, and the answer alone does not say where.



Reading the pair

Retrieval 1, generation 1. The system never found the material. The fix is in how the documents are broken up and indexed for search. Changing the model or its instructions will do nothing.

Retrieval 3, generation 1. The material was there and the answer still came out wrong. The fix is in the model or the instructions it is given, and this is the only case where changing the model is the right response.

Retrieval 1, generation 3. A correct answer built on material the system did not retrieve. Score it as a failure. The system was right by accident and will not be next time.

Who scores what. Retrieval is read out of the logs by whoever runs the system. Generation goes to the expert. Experts are the constraint, so spend them only on the axis that needs them.

THE HARDER CASE

Most of that method depends on a gold standard, and exploratory work does not have one

The method so far assumes a gold standard: a set of right answers written down before the system runs, for the outputs to be scored against. On material that has been worked before, that is reasonable. Experts draft twenty questions, write the answers from the source documents, and the rubric measures distance from those answers.

Exploratory work breaks the assumption. If the task is to find something nobody has found before, nobody can hand you the list in advance. The gold standard would be the deliverable. That is normal for discovery work, not an edge case, and it applies to most first-of-a-kind analysis in government.

The usual responses both fail. Building a small gold set from the material you already understand measures the system on the easy part of the corpus and tells you nothing about the rest. Abandoning structure and relying on impressions gives you a result no one can defend when it is challenged.

> **An example.** A pilot asks whether AI can find every approval step in a state's environmental laws that duplicates a Commonwealth one. If anyone already held that list, there would be no reason to run the pilot. The evaluation has to work without it, and saying so at the start is easier than explaining it at the end.

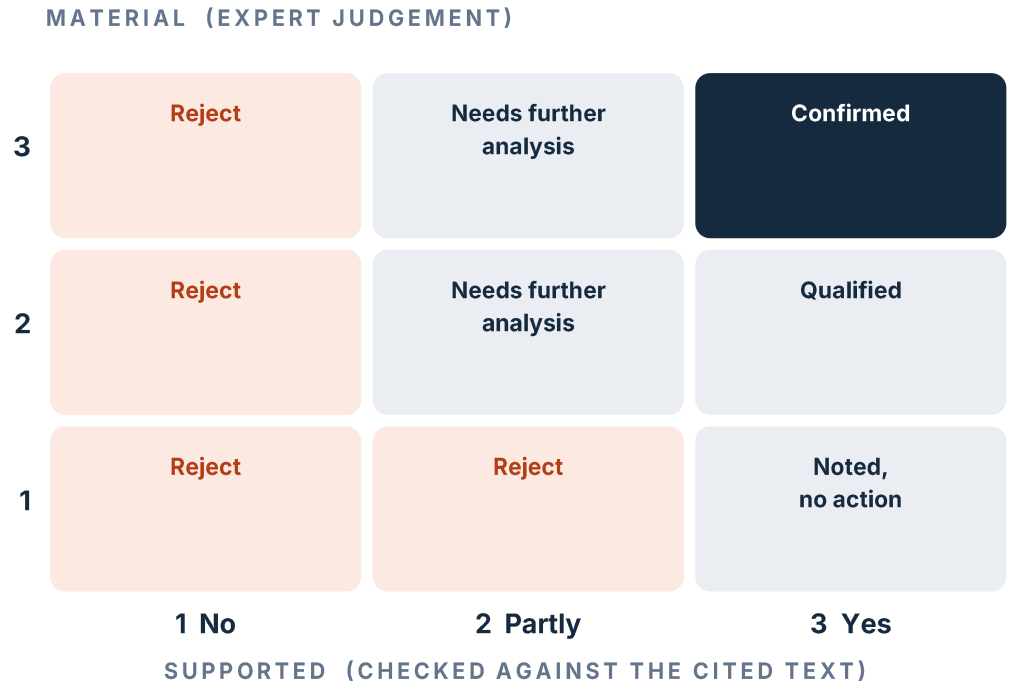
What failure looks like

- An accuracy figure quoted to one decimal place, derived from a gold set the same team wrote after seeing the output.
- A pilot that reports ninety per cent accuracy and cannot say what the system missed, because nobody established what was there.
- Two reviewers scoring the same finding three and one, and a project that averages them rather than asking why.

The fix is on the next page.

Check the finding against its own sources, and save the expert for the part that needs judgement

A finding cites the provisions it rests on. Anyone able to read can check whether that text says what the finding claims it says, with no answer key and no expert time. That gives the checkable axis back. Whether the finding is worth acting on stays with the expert, which was the only part that needed one.



What the axes mean

Supported. 1, the cited text does not say this. 2, it supports part of the claim, or the claim reaches further than the text goes. 3, the text says what the finding says it says.

Material. 1, correct and not worth acting on. 2, worth recording, not on its own worth a change. 3, worth acting on.

Supported and still not worth acting on. A finding can be entirely accurate and change nothing. Those are cheap to produce in volume, and a rubric that cannot mark them down will reward a system for making more of them.

A useful side effect. Confirmed findings become the gold set for the next round.

REPORTING

Without a gold standard there are numbers you can report and numbers you cannot

This is where most pilot reports overreach, and it is the part a reader with any statistical training will go looking for first.

Report these

Confirmed rate, the proportion of findings reviewers marked supported at 3.

Material rate, the proportion judged worth acting on.

False positive rate, the proportion rejected, which is what determines whether anyone keeps using the system.

Agreement, how often two reviewers scoring the same finding landed on the same number.

Do not report these

Recall, the share of everything findable that the system found. Nobody established what was there, so this number does not exist.

Accuracy against truth, there is no truth set, so there is nothing to be accurate against.

Anything quoted to a decimal place, sample sizes in this work run to a few dozen, and a decimal place claims a precision that sample cannot carry.

Say what you did not measure. A report that states its own limits is harder to attack than one that does not, and the limits will be found either way.

MAKING IT HOLD

The scale is written once and calibrated before anybody scores anything alone

A rubric is not finished when it is written. It is finished when two people can apply it and land in the same place, and the only way to get there is to score together before anybody scores alone.

1. **Write the scale**
   Three points per axis, one page, with an actual scored item as the example for each point.
2. **Everyone scores the same five**
   Pick five items across the range. Everybody scores them independently.
3. **Put the scores on the table**
   Compare. The disagreements are the useful part.
4. **Argue, then rewrite the examples**
   What the group settles on becomes the written example. This is the step that gets cut.
5. **Then score alone**
   Split the rest up. Re-check agreement partway through, not at the end.

THE TEST

Two evaluators score the same item without conferring and land within one point. If not, the rubric is not finished, and scoring more items will not fix it.

THE THRESHOLD

Eighty per cent at 2 or better on both axes is a reasonable default. Safety critical items carry their own bar, listed in advance.

> **A threshold agreed after the results are in is not a threshold.**

THE OTHER HALF

Scoring tells you whether the system is right, and nothing about whether anyone will use it

Structured testing runs a fixed set of questions and scores the answers. It is the only way to get a defensible number. It also limits people to the questions somebody thought of in advance, which is a poor guide to what the system will get asked in real use.

Structured

Fifteen to twenty questions across the range of the material, at a spread of difficulty. Scored on both axes. Produces the number.

Unstructured

Practitioners use the tool in their real work for a set period. Ratings, comments, surveys, debriefs and usage all get collected. Produces the reasons.

ONE TRIAL, TWO WEEKS

22	315	83%	50%
practitioners	questions asked	of rated answers positive	of answers rated at all

Read the 83 next to the 50. People rate when they feel strongly, so the rated half over-represents the very good and the very bad. In a working tool, no rating mostly means the person took the answer and moved on, which leans positive but was never measured. Quote the rating rate with the score, and treat both as background to the structured results, never as a substitute for them.

HOW TO CHECK

Six decisions, settled before anything is scored

1	**What two scores does each output get**	For an assistant: were the right documents found, and was the answer right. For findings with no answer key: does the finding match the text it cites, and is it worth acting on. Always two scores, kept apart.
2	**Do the right answers exist in advance**	If nobody can write them down before the system runs, the work is exploratory. Say so in the plan, and score each finding against the sources it cites instead.
3	**What do 1, 2 and 3 mean**	Write the scale on one page, with a real scored example under each point, so a new scorer can apply it without asking anyone.
4	**Who scores what**	Experts score judgement only. Anything checkable against a source goes to someone who can read, because expert hours are the scarce part.
5	**What counts as a pass, and what happens either way**	Agree the pass mark in writing before scoring starts, and what a pass and a fail each lead to. A score that changes no decision was not worth collecting.
6	**What will the report not claim**	List the numbers the method cannot produce, and put that list in the report from day one, while nothing is riding on it.

AccuFind Pty Ltd. Drawn from evaluation work across a number of government engagements. contact@accufind.ai