

Last quarter, we ran the same behavioral risk intake note through four large language models. Three rated it "elevated." One rated it "low." The note described a current employee who had asked about access logs for a colleague's account, had changed a password outside business hours, and had recently been placed on a performance plan. Nothing in the note was alarming on its own. Together, it was ambiguous.

The question on the table was not "which model is right?" It was "what do we do now, and who owns the answer?" This post covers how we handle moments like that, why we refuse to treat model agreement as proof, and how to document the disagreement so that a reviewer two years from now can understand what happened.

Why disagreement is the useful part

When models disagree, the instinct is to resolve the conflict quickly. Pick the majority, average the scores, or choose the answer that sounds most confident. All three moves throw away the most valuable information in the room.

Disagreement usually means one of a few things is true. The input is missing context that a human would need. The categories being used are ambiguous. The models were trained on different signals and weight the same evidence differently. Or the case genuinely sits near a boundary where reasonable reviewers would split. Each of these tells you something different about what to do next.

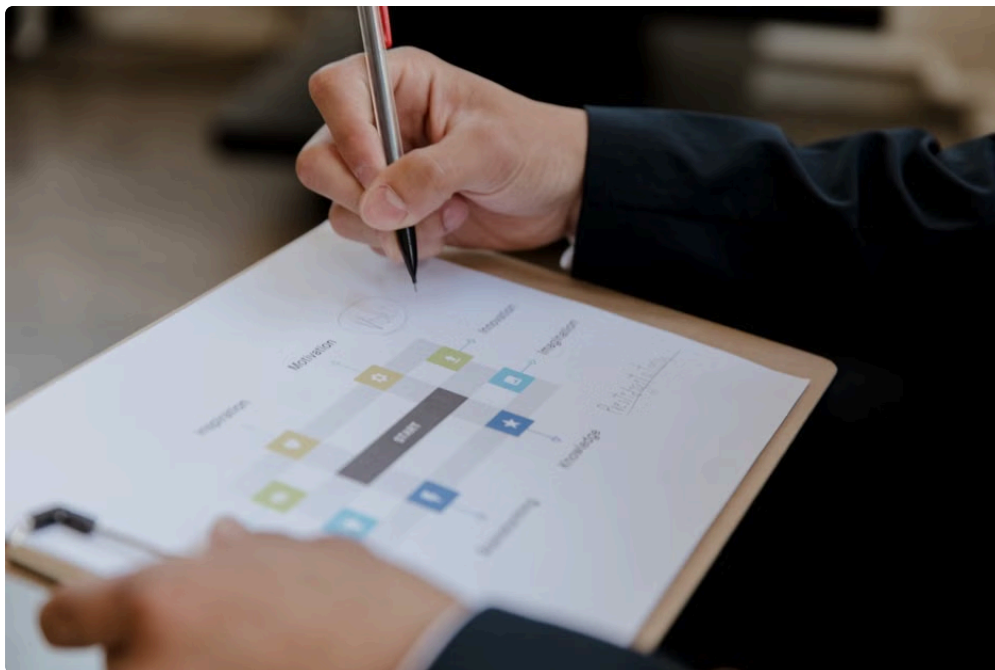
We treat disagreement as a signal about missing information. Our working rule is simple: if the models split, the case is not ready for a decision. It is ready for a question. The question is usually "what would we need to know to tell these apart?"

Do not average, and do not count votes

Two habits cause more damage than any others in multi-model review.

The first is averaging. If one model says 8 out of 10 and another says 3 out of 10, the number 5.5 is not a risk score. It is a number that describes nothing. Averaging assumes the scales mean the same thing across models, that each model's error is independent, and that the midpoint is a meaningful place to land. None of those assumptions usually hold.

The second is voting. Three models agreeing is not three independent confirmations. They may share training data, share the same blind spots, or respond to the same wording in the prompt. Agreement among correlated systems can feel like evidence while telling you very little. We have seen models converge on the wrong answer because the intake form used a loaded phrase, and all of them picked up the phrase.



Side-by-side model outputs are useful precisely because they let a human see where the reasoning diverges. A good multi-LLM orchestration tool should make that divergence visible rather than hiding it behind a single blended result. Suprmind.ai, for example, is built around presenting multiple model responses together so reviewers can compare the reasoning rather than just the verdict. That is the right direction. The comparison is the work product, not a step on the way to a number.

The tradeoff you cannot engineer away

In threat assessment, every threshold you set produces two kinds of error. A false positive flags someone who did not warrant it. A false negative misses someone who did. You cannot minimize both with the same setting, and the cost of each error is not symmetric.

The right balance depends on the context, the available safeguards, and the harm on each side. A false positive in an HR setting can damage a career, erode trust, and create legal exposure. A false negative can mean a missed warning sign. Neither outcome is acceptable, which is exactly why the decision cannot be delegated to a model.

Error type	What it looks like	Typical harm	Questions to ask before accepting it
False positive	Escalation of a case that does not warrant it	Unfair treatment, loss of trust, legal and reputational exposure	What review step follows? Who can close the case? How quickly?
False negative	A real concern is scored low or not surfaced	Missed intervention, preventable harm, failure of duty of care	What other signals exist? Is there a second channel for reporting?
Model disagreement	Models split across the threshold	Inconsistent handling if not documented	What information would resolve the split? Who decides?

Set the tradeoff explicitly, in writing, before the first case arrives. Changing it mid-stream because one case felt urgent is how organizations end up with inconsistent and indefensible records.

A working procedure for split decisions

Here is the checklist we use when models disagree on a risk level. It is deliberately short.

1. Record every model's output, rating, and reasoning in full, unedited.
2. Identify the specific factual claim or inference where the models diverge.
3. Write one sentence stating what additional information would separate the readings.

4. Gather that information, or record that it is unavailable and why.
5. Route the case to the named human decision-maker with the full comparison attached.
6. Record the decision, the rationale, and the reviewer's name and date.
7. Set a time-boxed follow-up, typically 14 days, to revisit the case if new facts arrive.

Notice what is missing. There is no step that says "take the highest score." There is no step that says "wait for consensus." The model outputs are inputs to a human judgment, not the judgment itself.

Name the human who owns the call

Every case note we write names the decision-maker. This is not a bureaucratic flourish. It is the single most important line in the record. When someone asks six months later why a case was closed, the answer should never be "the system flagged it" or "the AI said so." The answer should be "Dana Whitfield, Head of Trust and Safety, reviewed the comparison on the 3rd, concluded the access pattern had a documented business explanation, and closed the case with a note to revisit if the pattern recurs."

That sentence does three things. It assigns accountability to a person who can be questioned. It records the reasoning in human terms. And it makes the model's role visible without letting it stand in for the decision.

Radomir Basta, CEO and founder of Four Dots, has made a similar point in discussions of AI-assisted work: tools can accelerate analysis, but the organization remains responsible for what it does with the output. We agree. A governance program that cannot name a human owner for each high-stakes decision is not a governance program yet.



What defensible documentation looks like

A defensible record does not need to be long. It needs to be complete and honest. Our standard template includes the following.

- The date, the case identifier, and the named decision-maker.
- The full outputs from each model, preserved as received.
- A plain-language summary of where and why the models disagreed.

- The information gathered in response, with sources.
- The decision, the rationale in the reviewer's own words, and any dissent.
- A running list headed "What we still do not know."
- The time-boxed follow-up date and its owner.

That last list matters more than people expect. Every case has gaps. Writing them down prevents the quiet assumption that the matter is settled when it is only closed for now.

What we still do not know

We are still learning how to weight disagreement across different risk categories. We do not yet have a good measure of how often correlated models share the same <https://suprmind.ai/hub/insights/threat-assessment-meets-ai-governance/> blind spot in our domain. We are not sure whether showing reviewers the model reasoning changes their decisions in a systematic way, or whether it simply makes them more confident in whatever they were already inclined to do. We also have limited data on how often time-boxed follow-ups surface new information that changes the outcome.

If you are building a similar process, we would suggest tracking those questions in your own records. Revisit them quarterly. A governance practice that stops asking what it does not know is about to make an expensive mistake.

Trying it yourself

If you want to compare model outputs on your own intake material before committing to any workflow, Suprmind.ai offers a Start Free Trial with no credit card needed. The source material we reviewed did not list pricing, so check the current plans directly with the vendor before budgeting. Test the trial with cases you have already resolved, so you can judge whether the side-by-side view surfaces the disagreements your reviewers actually cared about.

The tool will not make the decision. Your named human will. Design the process so that is never in doubt.