

WHICH RELIABILITY TERM EACH VENDOR DOCUMENT FILLS

Vendor-Evidence Scorecard

Put one vendor's documents on the table. Score how well each one fills the four terms of the reliability equation, and note the source backing the score. The blank cells are your procurement to-do list.

BRING One vendor proposal, with its model card and supporting documents

How well does the vendor's evidence fill each reliability term?

Score 0 if no evidence, 1 if asserted only, 2 if evidence is given but transfer to your deployment is unargued, 3 if evidence is dated, sourced and tied to your environment. In Evidence, name the document: model card, system card, eval report, security review or RFP response.

Criterion	Score (0-3)	Evidence
P(error): is the accuracy claim backed by a named benchmark, version, date, scaffold, contamination protocol and confidence interval?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
P(error) transfer: is there an argument the score holds on your data, users and tools, not just the benchmark's?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
P(undetected error): does the evidence name an independent verifier that catches wrong outputs before they become an action?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
P(uncontained error, undetected): does the evidence name the hard action boundaries that stop a missed error reaching live systems?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
E(severity escaped): is a harm unit and a harm tier attached to an escaped error, with a named escalation owner?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
Helpfulness claims: is there evidence of what catches sycophantic or over-compliant responses at runtime?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
Safety claims: do safety tests connect to runtime monitors and to the permissions that bound a safety miss?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	
Document coverage: are model card, system card, eval report, security review and RFP response all present and consistent?	<input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3	

The blank cells - your procurement to-do list

List the cells the vendor left empty. For each, write what operational evidence must be true, and who owns producing it, before believing the purchase is safe.

Companion worksheet to **Essay 06 · The Model Card Won't Save You**, in the series **Architecting the AI Coworker**. · Dr Peter McCann Strain · Fill this in against one real agent, action class or vendor. © 2026 Peter McCann Strain.

Series Companion + all 22 worksheets: **Release_v12/Series_Companion.pdf**