

E02	Per-case results and adjudication		
E03	Vendor security and data terms		
E04	Operational trial and cost log		
E05	Legal and privacy assessment		

02 Set hard gates before scoring

The owner marks each row Pass, Fail, Not applicable with reason, or Pending. A material Fail or Pending prevents recommendation unless the firm's designated authority documents a lawful, policy-permitted conditional decision. A weighted average never overrides a gate.

Gate	Acceptance rule set by firm	Status / evidence / owner
G1 Lawful and approved use	Jurisdiction and use-case assessment; privacy screening and any required DPIA completed.	
G2 Data and contract	Approved data location, retention, training use, deletion, subcontractors and access terms.	
G3 Security	Threat model and controls for prompt injection, data disclosure, tool permissions, incident response.	
G4 Critical quality	No breached high-severity acceptance criterion for the defined test set.	
G5 Human control	Review, override, audit and safe fallback proved in a workflow test.	
G6 Exit and resilience	Outage response, rollback and migration/portability options evidenced.	

03 Design the test set

Pre-register case IDs and expected behavior before seeing candidate outputs. Keep a held-out set for final evaluation. Prevent training or prompt tuning on held-out examples. Use qualified reviewers and an adjudication procedure for disagreement. Record exclusions and missing coverage.

Slice	Count and source	Minimum evidence / owner
Routine in-scope		
High consequence / deadline sensitive		
Unsupported or ambiguous input		
Adversarial instructions in user text or retrieved content		
Confidential and personal data exposures		
Outage, truncation and malformed outputs		

Candidate and exact version	Hard gates	Weighted result / 100	Critical incidents	Decision rationale
A				
B				
C				

05 Capture each test result

Field	Entry
Case ID / risk register link	
Slice / source and lawful data classification	
Input reference or hash / expected behavior	
Candidate, prompt, corpus and tool versions	
Actual output reference / latency / total cost	
Reviewer / adjudicator / review date	
Verdict / severity / harm pathway	
Control action / retest ID and result	

Duplicate this record for each case or retain a controlled results ledger linked from E02. Redact confidential material before any vendor submission, subject to your firm's approved data-handling procedure.

06 Review supplier and operating model

Question	Vendor response and verified evidence	Risk owner / disposition
What training and retention rights apply to prompts, outputs and logs?		
Where are requests, logs, backups and support access processed?		
Which subcontractors and concentration dependencies exist?		
What model and safety changes can occur without customer approval?		
What service levels, support and incident notice terms apply?		
How can data and logs be exported/deleted and the model replaced?		
Which controls are available for tools, retrieval and permissions?		

07 Record decision and monitoring

Decision field	Entry
Selected candidate, deployment configuration and permitted scope	
Rejected candidates and reasons	
Residual risks, owner, acceptance authority and expiry	
Pilot stop conditions and rollback owner / tested time	
Monitoring metrics, denominators and cadence	
Review triggers: model, prompt, data, vendor, regulation, incident	
Decision authority / date / system of record	

Decision: ☐ Recommend pilot ☐ Conditional review ☐ Reject / defer. No procurement, pilot or production approval is implied by completion of this document.

Sources and use notes

References support evaluation design and risk identification; none prescribes the weights, cutoffs or legal applicability for a particular firm. Confirm applicable requirements with internal Legal, Compliance, Risk and Privacy owners at the time of use.

NIST, Generative AI Profile AI 600-1 (2024) — Open source

OWASP, 2025 Top 10 risks for LLM applications — Open source

ICO, AI and data protection risk toolkit (under review) — Open source

UK Government, Guidelines for AI procurement (public procurement context) — Open source

Illustrative decision walk through

Fictional internal document summary use case. The figures below illustrate the calculation only and do not benchmark any real provider. Both candidates use the same held-out cases, reviewer protocol and deployment configuration.

Dimension and weight	Model A score	Model B score
Task accuracy 25%	4	5
Safety and misuse 20%	5	2
Tone and instruction fit 10%	4	4
Latency and reliability 10%	3	5
Total operating cost 10%	3	4
Governance suitability 25%	5	3

A weighted arithmetic result would be 85/100 for A and 74/100 for B. If B has a material safety gate failure, B is not recommended regardless of its task accuracy or speed. Even A remains subject to the named authority’s decision, pilot evidence and any applicable firm policy.