EVALUATION

AI Evaluation Rubric Builder

A worksheet for evaluating AI outputs, retrieval, tool use, escalation, tone, and reviewer confidence.

WHAT THIS TEMPLATE HELPS YOU DECIDE

Define pass/fail criteria, reviewer instructions, sample cases, escalation thresholds, and improvement loops before AI-supported work reaches production.

BEST FOR

- Agentic workflows
- Document assistants
- Sales, service, and advisory copilots

OUTPUTS

- Evaluation rubric
- Pass/fail thresholds
- Sample test cases

STEP 1

Frame the operating need

Start with the workflow, decision, owner, and business pressure. The template is useful only when it is grounded in a real operating moment.

Operating frame

Output evaluated
Draft, summary, recommendation, classification, tool call, or answer.

Reviewer role
Name who can judge quality and risk.

Failure consequence
Describe what happens if weak output passes review.

Readiness check

- [ ] Evaluation criteria match the workflow risk
- [ ] Reviewers know what evidence to inspect
- [ ] Tool calls and retrieval quality are evaluated separately
- [ ] Thresholds trigger retry, escalation, or rejection
- [ ] Rubric findings update prompts, tools, or source material

STEP 2

Map the architecture questions

Use this page to separate the parts of the system that need design before anyone jumps to tools, prompts, or implementation details.

Design map

Accuracy	What must be correct, sourced, complete, and current?
Grounding	What evidence should the system cite or expose to the reviewer?
Tool behavior	Did the system call the correct deterministic tool with valid inputs?
Escalation	When should the system stop and route to a person?
Tone and fit	Does the output match the workflow, audience, and risk level?

Evaluation turns AI quality from a feeling into an operating contract. The rubric should make weak outputs visible before they scale.

STEP 3

Turn the answers into a brief

A strong brief makes the next decision easier: proceed, defer, redesign, govern, or assess more deeply before implementation.

Decision brief

Pass threshold	What score or defect pattern allows output to proceed?
Reject threshold	Which failures require stopping the workflow?
Regression case	Which cases should be tested every release?
Improvement loop	How do evaluation findings change the system?