



AI Evaluation Rubric Builder

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

STRUCTURE - CLARITY - BETTER DECISIONS

Practical architecture for accountable AI-enabled work.

What this template is for

Define pass/fail criteria, reviewer instructions, sample cases, escalation thresholds, and improvement loops before AI-supported work reaches production. Use this guide to make the operating choice explicit, test it with evidence, and record who owns the next decision.

BEST USED WHEN

Agentic workflows, advisory tools, document assistants, and sales or service copilots.

WHO SHOULD BE IN THE ROOM

Include the workflow owner, a person who performs the work, a reviewer with authority to approve or stop it, and the relevant data or tool owner.

CANADIAN CONTEXT

Confirm applicable federal, provincial, territorial, sector, professional, contractual, accessibility, and bilingual-service requirements. Use this workbook as operational preparation, not legal advice.

A five-step working rhythm

01

Accuracy

What must be correct, sourced, complete, and current?

02

Grounding

What evidence should the system cite or expose to the reviewer?

03

Tool behavior

Did the system call the correct deterministic tool with valid inputs?

04

Escalation

When should the system stop and route to a person?

05

Tone and fit

Does the output match the workflow, audience, and risk level?

Build the working record

Complete one record for one workflow. Write enough to support a decision, not a generic technology wish list.

PROMPT	YOUR WORKING NOTES
01 Output evaluated	
02 Reviewer role	
03 Failure consequence	
04 Accuracy	
05 Grounding	
06 Tool behavior	
07 Decision, owner, evidence, and review date	

Evaluate intake briefs

STARTING FRICTION

A fluent brief may omit a client constraint that changes the proposed work.

WORKFLOW MAP

CASES

RUBRIC

TEST OUTPUT

EVIDENCE

REVIEWER SCORE

REMEDiation

AI-SUPPORTED ROLE

Score representative outputs against explicit criteria and attach source and tool evidence. Flag uncertainty around accuracy.

HUMAN AUTHORITY

Reviewers define pass thresholds, adjudicate disagreements, and approve remediation. Resolve the accuracy question for this case.

PILOT MEASURE

Track rubric pass rate, severity of defects, reviewer agreement, and escalation accuracy. Record accuracy exceptions separately.

DISCUSSION PROMPT

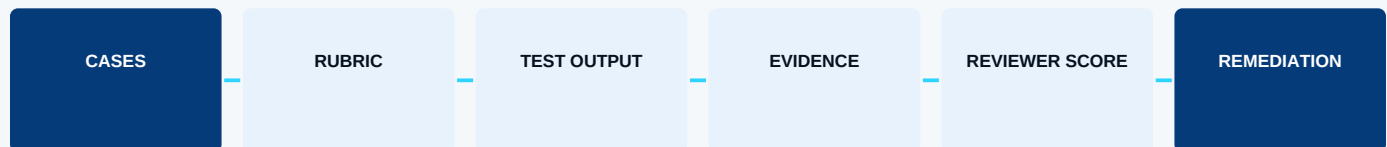
What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Evaluate customer answers

STARTING FRICTION

A reply can sound helpful while inventing a price, policy, or availability claim.

WORKFLOW MAP



AI-SUPPORTED ROLE

Score representative outputs against explicit criteria and attach source and tool evidence. Flag uncertainty around grounding.

HUMAN AUTHORITY

Reviewers define pass thresholds, adjudicate disagreements, and approve remediation. Resolve the grounding question for this case.

PILOT MEASURE

Track rubric pass rate, severity of defects, reviewer agreement, and escalation accuracy. Record grounding exceptions separately.

DISCUSSION PROMPT

What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Evaluate financial summaries

STARTING FRICTION

The firm needs a rubric for factual grounding, materiality, clarity, and abstention.

WORKFLOW MAP

CASES

RUBRIC

TEST OUTPUT

EVIDENCE

REVIEWER SCORE

REMEDiation

AI-SUPPORTED ROLE

Score representative outputs against explicit criteria and attach source and tool evidence. Flag uncertainty around tool behavior.

HUMAN AUTHORITY

Reviewers define pass thresholds, adjudicate disagreements, and approve remediation. Resolve the tool behavior question for this case.

PILOT MEASURE

Track rubric pass rate, severity of defects, reviewer agreement, and escalation accuracy. Record tool behavior exceptions separately.

DISCUSSION PROMPT

What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Before the pilot grows

Use these prompts to test whether the proposed workflow can operate responsibly in its Canadian business context.

01

Accuracy

Evaluation criteria match the workflow risk. Name the owner and the operating decision before selecting a tool.

02

Grounding

Reviewers know what evidence to inspect. Record the approved sources, permissions, and missing information.

03

Tool behavior

Tool calls and retrieval quality are evaluated separately. Test this boundary with representative and adverse cases.

04

Escalation

Thresholds trigger retry, escalation, or rejection. Preserve review, escalation, and correction evidence.

05

Tone and fit

Rubric findings update prompts, tools, or source material. Set a measurable stop condition and the next review date.

06

Readiness to proceed

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

CANADIAN REFERENCE POINTS

Office of the Privacy Commissioner of Canada - Privacy and artificial intelligence
priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/

Government of Canada - Artificial Intelligence guidance, principles, and policy
canada.ca/en/services/science/innovation/artificial-intelligence.html

General operational guidance only. Confirm applicable legal, privacy, cybersecurity, professional, contractual, and regulatory requirements with qualified advisors.