



Private AI Evaluation Checklist

ONE WORKFLOW. KNOWN EVIDENCE. A CLEAR DECISION.

Use this before comparing models or expanding a pilot. Mark evidence observed on the tested configuration.

ORGANIZATION

WORKFLOW

DECISION DATE

1 DEFINE THE WORKFLOW

- ☐ Name one repeated task and the team that performs it.
- ☐ Name the workflow owner and the person judging output.
- ☐ Record today's baseline: time, quality, cost, or risk.
- ☐ Write the decision this evaluation must support.

2 BUILD A REPRESENTATIVE TEST SET

- ☐ Use real-shaped, approved samples - never convenient demos only.
- ☐ Include known answers and examples requiring multiple sources.
- ☐ Include missing-information and difficult edge cases.
- ☐ For coding, include review criteria and relevant tests.

3 MAP THE COMPLETE DATA PATH

- ☐ Record where documents, code, prompts, and outputs travel.
- ☐ Identify inference, retrieval, client, tool, and hosted services.
- ☐ List every external connection and its purpose.
- ☐ Confirm retention, deletion, backup, and recovery expectations.

4 FREEZE THE TESTED CONFIGURATION

- ☐ Record hardware, model, runtime, client, and exact versions.
- ☐ Confirm that this combination supports the intended capability.
- ☐ Capture configuration, limitations, and reproducible evidence.
- ☐ Re-test after changing any material component.

5 CHECK CONTROL AND OPERATIONS

- ☐ Test identity, roles, source access, and a denied-access case.
- ☐ Verify activity records, health visibility, and failure handling.
- ☐ Name the operator for setup, updates, recovery, and support.
- ☐ Estimate infrastructure cost and operator time separately.

6 SET THE DECISION GATE BEFORE TESTING

- ☐ Define useful output and unacceptable failure in plain language.
- ☐ Set thresholds for source support, review, tests, or accuracy.
- ☐ Measure missing-answer behavior, latency, effort, and reliability.
- ☐ Choose: stop, revise, extend, or proceed to a scoped pilot.

DO NOT ADVANCE WHEN

The workflow has no owner, the test set cannot be judged, the data path is unclear, or the result cannot be reproduced.
A polished response is not evidence by itself. Record the configuration, source, reviewer decision, and observed limitation.

NEXT STEP Discuss one workflow at srasta.ai/demo | Learn more at srasta.ai/learn/private-ai-for-business