sotto

Local and cloud AI for records work

Choose the configuration that fits your records, people and policies. Deployment and model options vary by the selected Sotto product and approved configuration. Confirm the exact supported platform, model and data handling before using real records.

Compare the complete workflow

This comparison explains general deployment tradeoffs. It is not a list of features available in every Sotto product.

Decision factor	Local model on approved hardware	Cloud model through an approved service
Task quality	Test the exact model, context and runtime. Running locally is not evidence of accuracy.	Test the exact model, endpoint and features. A wider model selection does not establish better results.
Capacity and speed	RAM is shared with the operating system and other apps. Measure memory, document handling and latency on the actual device.	Check model availability, quotas and latency under expected load. Include network delays and retries.
Continuity	Offline inference may be possible when all required components are local. Verify setup, licensing and update dependencies.	Inference depends on connectivity and service availability. Agree outage handling and any permitted fallback.
Cost and upkeep	Count hardware and staff time for installation, model management, updates, backups and support.	Count subscriptions or per-use charges, retries and related services. Set spend limits and assign change ownership.
Data handling	Check app traffic, telemetry, diagnostics, backups and folder sync. Local inference does not prove that no data leaves the device.	Identify each payload and recipient, including logs and support access. Approve permitted data and contract terms first.

Keep human review in every configuration

Use AI outputs as proposals, not legal authority. Review what should remain visible as well as what may need protection. Verify unsupported claims and citations against authoritative sources. The authorized person makes the release decision and checks the actual exported package, not just the preview.

Make the decision on evidence

Local is not automatically compliant; cloud is not automatically unsafe. Compare the same representative examples and assess the complete workflow, including corrections, export and accessibility. Treat a successful small test as evidence for that scope, not proof of general accuracy. [1]

Configuration decision worksheet

Start with fictional or otherwise authorized examples. Record the exact Sotto product, supported platform and available model before testing. Do not switch models or services without the organization's approval.

Organization and workflow _____

Sotto product and version _____ Supported platform _____

Model and local or cloud route _____

Confirm the data and control requirements

[ ] **Payload and routing.** Map documents, prompts, outputs and metadata, including telemetry and diagnostics. Identify destinations, recipients and any fallback behavior.

[ ] **Retention and training.** Check retention, training use, backups, abuse logs and deletion by endpoint, model and feature. No training does not mean no retention. [2]

[ ] **Location and access.** Confirm processing/storage regions, subprocessors, support access, credentials, access logs and encryption responsibilities. Identify required agreements.

[ ] **Operation and exit.** Assign updates, incident handling, support, export, deletion and recovery. Agree what to do after an outage, failed output check or configuration change.

Test the same examples and record the result

Agree reference decisions and thresholds first. Measure missed and unnecessary redactions, unsupported output, corrections, export quality, active time, latency and cost. Check the model is actually available; if not, stop that step. Inspect released files and document searchability and accessibility.

Test packet and pass thresholds _____

Local result _____ Cloud result _____ Evidence _____

[ ] Local selected [ ] Cloud selected [ ] More testing [ ] Neither suitable

Permitted data conditions and retest trigger _____

Record the organization's review

Records lead _____ IT owner _____

Security reviewer _____ Legal or policy reviewer _____

Procurement owner _____ Decision approver and date _____

Public reference notes

[1] NIST Generative AI Profile AI 600-1 Voluntary risk and evaluation guidance, not a product certification.

[2] Anthropic API and data retention Illustrates feature-specific policy differences; not a statement of Sotto provider support or your contract terms.

Reference pages checked November 1, 2026. Confirm current terms for the configuration you select.