Ragonaut

Email Auto-Labeling at Enterprise Scale

expert@ragonaut.com Prague, CZ

Version 1.4 (June 2025) · *From PoC to Audit-Ready Production – Agile-Enabled Delivery*

! Problem Statement

The Challenge

- Knowledge-workers waste hours manually sorting mail.
- Inconsistent labelling disrupts search, reporting, and compliance e-discovery.

Why It Matters

- Poor triage hides urgent or revenue-related messages.
- Non-standard classification complicates retention and auditing obligations.
- Large Language Models (LLMs) can help – but only if cost, privacy, and security are governed.

Context & Background

A proof-of-concept (PoC) built in *n8n* processed **200** emails and achieved **95%** classification accuracy, validating technical feasibility. Enterprise deployment, however, demands higher scalability, granular logging, FinOps governance, least-privilege access, and disciplined change control beyond what Apps Script or no-code tooling can provide.

Decision Points & Alternatives

Why Not Google Apps Script ?

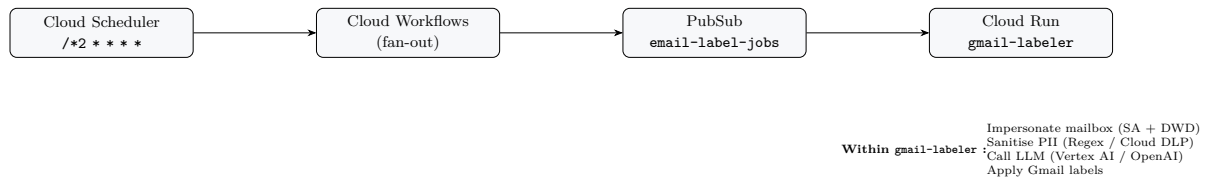
Limitation	Detail	Impact
Execution quotas	6-h total runtime per script per day ; 1 trigger / min.	2-min polling for 100 users exceeds quotas on day 1.
Hard time-outs	Single execution capped at 6 min.	Large backlog processing halts, leaving mail unlabelled.
Background triggers	No PubSub push or Gmail History API.	Forces inefficient polling or complex work-arounds.
Secret handling	Script Properties limited to 500 kB ; readable by any editor.	Cannot adequately protect LLM or DLP keys.
Observability	String logs only ; no structured metrics or error grouping.	Difficult to build FinOps dashboards or SLOs.
Security model	Code runs with each user's identity.	Hard to enforce least-privilege or a central kill-switch during incidents.

Why Not Use n8n for Production ?

Limitation	Detail	Impact
Credential storage	Tokens stored in n8n DB, decrypted at runtime.	Harder to meet ISO 27001 segregation-of-duties.
Change control	Editors can hot-change workflows in UI.	Weak audit trail unless strict governance added.
Performance ceiling	Canvas complexity & throughput degrade beyond 5 k execs / min.	Risk of backlog during mail bursts.
FinOps	No native cost dashboard or quota enforcement.	External scripts required, duplicating effort.
Governance & Privacy	JSON and credentials editable ; PII may log in plaintext.	Difficult to prove immutable change trail ; GDPR concerns.

Outcome – Constraints led to selecting **Google Cloud Run** with IAM integration, keyless workload identity, structured logging, and elastic throughput.

Architectural Diagram



Agile Road-map

Sprint	Increment Goal	Key User Stories	Proof / Metric
0	PoC confirmation	n8n demo accuracy $\geq 95\%$.	Confusion matrix shared.
1	Infrastructure as Code	Terraform baseline ; CI pipeline.	State bucket & plan artefacts.
2	Secure Mailbox Access	SA + DWD ; keyless tokens.	Scope `gmail.modify`.
3	MVP	PII scrub + LLM + label apply.	Regex scrubber ; OpenAI call.
4	Observability + FinOps	Dashboards, alerts.	Log metrics ; cost guard-rails.
5	Pilot	10 users ; rollback drill.	Canary traffic ; feedback loop.
6	Security Hardening	Binary Auth, VPC SC.	Signed image ; perimeter test.
7	Production GA	100 users.	Runbook ; hand-over.

Key Components & Benefits

Component	Benefit	Evidence
Restricted OAuth scope (`gmail.modify`)	Labels only – no send/delete/download.	Access logs show zero sends/deletes in PoC.
Keyless Workload Identity	Removes persistent credentials.	SA keys deleted; audit logs show token issuance only.
PII Sanitisation (Regex / DLP)	GDPR-compliant redaction before LLM.	Test harness shows 100% email masking.
FinOps Budgets & Alerts	Prevents cost overruns.	Mock spike triggered alert in Sprint 4.
Terraform + CI/CD	Repeatable, auditable deployments.	`terraform plan` archived per merge.

LLM-Readable Artefacts Delivery

- **Prompt Template & Taxonomy** – stored in Secret Manager; new versions referenced via alias `prompt_latest`; no redeploy required.
- **Model Version Pinning** – model ID (e.g. `gpt-4o-2025-04-09`) stored in Secret Manager; runtime check ensures match.
- **Evidence Pack Generation** – logs include hashed message ID, prompt version, model ID, redacted PII, and returned label; aggregated daily to Parquet, retained 7 years.

Security & Privacy Controls

1. Restricted OAuth Scope (`gmail.modify`)
2. Keyless Workload Identity
3. Binary Authorization (signed containers)
4. VPC Service Controls
5. PII Sanitisation (Regex / Cloud DLP)
6. Log Redaction
7. Segregation of Duties

$ FinOps Model

Cost Driver	Guard-Rail
LLM tokens	Per-user daily quota via Redis ; monthly alert.
DLP bytes	Budget alert ; auto-throttle.
Cloud Run CPU	Alert at 2 × baseline ; review scaling policy.

Change & Release Management

- **Git Source of Truth** – GitHub Enterprise Cloud with branch protection.
- **CI/CD Pipeline** – GitHub Actions with Cloud Build ; deploys on merge to `main`.
- **Rollback Mechanism** – Cloud Run traffic-split enables 30 s revert (story `REL-3`).
- **Release Cadence** – minor versions every sprint ; major versions quarterly via CAB.
- **Artefact Retention** – `terraform plan`, container digest, and release notes linked to each Git tag.

Maintenance & Handover

Handover Checklist

1. Git repository transfer (history preserved).
2. State-bucket IAM access to client DevOps.
3. CI/CD triggers transferred to client project.
4. Two recorded run-throughs : Ops and FinOps.
5. 30-day warranty sprint ; open defects enter backlog.

Optional Managed Service (Annual Contract)

Service	SLO	Cadence	Deliverable
Health checks & patching	CVE patch $<$ 7 days	Monthly	Updated image + scan report
Prompt & taxonomy tuning	$\geq$ 95% accuracy on sample	Quarterly	Updated secret + test report
FinOps analytics	Budget variance $\leq$ 10%	Monthly	Dashboard + cost insights
24 × 7 incident response	4 h response, 12 h resolution	On-demand	Incident report + post-mortem
Audit assistance	Evidence ready 2 weeks pre-audit	Yearly	Full artefact pack + support

Conclusion

Manual triage is costly ; the PoC proved AI can classify mail accurately. Enterprise deployment demands security, FinOps, and audit controls.

Actionable Recommendations

1. Approve Sprint 0 to refine backlog and success metrics.
2. Allocate service account and Terraform project this quarter.
3. Schedule pilot group onboarding by Sprint 5.

Importance – The proposed pipeline halves email-handling time, ensures regulatory compliance, and delivers predictable costs within three months.

Call to Action

Ready to eliminate manual triage ? Contact us to start implementation.