

AI OPERATIONS PLATFORM

Advanced AI Operations Platform

Dockerized AI workforce layer for sales, finance, support, document intelligence, research, and internal knowledge — agents that act with MCP permissions and HITL, not a FAQ chatbot or disconnected Zapier flows.

M0–M10
MILESTONES CLOSED

PASS
UAT_CLOSEOUT

28/28
PRD DELIVERABLES

M10
API MILESTONE

Built by **Abdul Basit** · AI Automation Engineer

BUSINESS IMPACT

Approvals before money moves · audited tools · one ops console

HITL before ERP/send · MCP allowlists with audit · admin RBAC, eval scorecards, and metrics in one place

OVERVIEW

Business events (email webhooks, CRM hooks, demos, channel ingest) are classified, routed to specialized agents, grounded in RAG (and optional KAG), authorized through an MCP tool gateway, executed against CRM/ERP adapters, and either completed automatically or held for HITL approval — with idempotency, dead letters, audit, eval scorecards, and an admin ops UI. FastAPI owns AI / RAG / MCP / domain; n8n stays thin.

TYPE	AI operations platform (not a single n8n flow)
ROLE	AI Automation Engineer / AI Systems Architect
DELIVERY	API 0.10.0 · Milestone M10 · M0–M10 complete · Closeout 2026-09-11
STACK	FastAPI · LangGraph · MCP · n8n · Postgres · Redis · Qdrant · OpenAI · HubSpot/Odoo · Caddy · Prometheus

THE PROBLEM

- Fragmented ops** — email/finance/CRM/ERP one-off scripts
- Ungrounded LLM answers** — policy/pricing invented
- Uncontrolled tools** — agents call APIs without allowlists
- Unchecked writes** — invoices/quotes without HITL
- Silent failures** — no DLQ, metrics, or ops cockpit

THE SOLUTION

- LangGraph orchestrator** — classify → route with policy
- MCP gateway** — allow/deny + audited tool calls
- Qdrant RAG + optional KAG** — grounded answers; relationships
- HITL admin SPA** — approvals, DLQ, eval, health, JWT RBAC
- Private AI** — sensitive classes never hit public OpenAI

WHO IT SERVES

- Ops/finance/sales teams needing agents that **act safely**
- Operators needing one console for runs, HITL, and eval
- Sandbox-first demos without live vendor credentials
- Lead demo: FR-32 purchase-request → RAG → HITL → ERP/CRM

Demo lead: FR-32 purchase-request → classify → RAG → quotation HITL → ERP + CRM. **Secondary:** invoice IDP → HITL → Odoo · MCP deny · private AI · viewer 403.

AGENT ROSTER + TYPICAL TOOLS

AGENT	TOOLS / FOCUS	AGENT	TOOLS / FOCUS
Support	RAG, notify, memory	Finance / IDP	Extract, Odoo (HITL), private AI
Sales	Lead score, HubSpot, proposals HITL	Research	Browser sandbox, sources, notify
Classifier	Route + confidence ≥85%	Reporting	Metrics / eval scorecards
Document/IDP	Invoice/PO extract ≥90%	Orchestrator	LangGraph classify → route → MCP

Safe action layer: LangGraph routes · MCP allowlists · HITL before money moves · private AI off public LLMs. Confidence bands: High auto · Medium Approve/Reject/Edit · Low stop/escalate. Quote HITL ≥5000 · finance ≥1000 · always HITL for finance mutate/delete.

ARCHITECTURE & CAPABILITIES

Architecture, FR-32, HITL lifecycle, MCP matrix, event envelope

Principle: FastAPI owns AI / RAG / MCP / domain; n8n owns thin side effects.

Events (email / CRM webhooks / channels)
→ Edge (Caddy, FastAPI, Redis idempotency, Admin SPA)
→ Orchestrator (LangGraph) → specialized agents
→ RAG (Qdrant) + optional KAG → MCP gateway → CRM/ERP/browser/notify
→ HITL approvals → side effects → Data (Postgres, Redis, Qdrant) · /metrics

COMPONENT	ROLE
API / webhooks	Auth, normalize, idempotency claim keys, enqueue
Orchestrator / Agents	LangGraph classify → route; Support, Finance, IDP, Sales, Research, Reporting, Classifier
RAG + KAG / MCP	Grounded generate; relationships; allow/deny + audited tools
CRM / ERP / Admin	HubSpot (+ SF), Odoo (+ QB); Runs, Approvals, DLQ, Eval, Escalations

TOOLS & INTELLIGENCE

- Email classify ≥85%; confidence → HITL
- RAG + optional KAG relationships
- MCP:
CRM/ERP/workspace/memory/notify/browser/leads/proposals
- IDP extract ≥90%; Odoo behind HITL

INTEGRATIONS

HubSpot / Odoo	Sandbox → live when set
SF / QB	M9 sandboxes
Slack / channels	Stub → webhook / sandbox
Ollama / Neo4j	Optional Compose profiles

SAFETY & OPS

- Proposals HITL; browser → DLQ
- Private AI for finance.invoice / PII
- Eval scorecards + Prometheus
- Viewer cannot approve (RBAC 403)

FR-32 PURCHASE-REQUEST (ONCE)

- Inbound purchase-request email (or demo inject)
- Classify (≥85%) → RAG products/policy
- Draft quotation → HITL (viewer 403)
- On approve: ERP create + CRM update
- Idempotency keys block duplicate records on retry

HITL LIFECYCLE (ONCE)

- High-risk action detected (ERP / proposal / outreach)
- Package: what · why · context · payload · risk
- Admin → Approvals → Approve / Reject / Edit
- Edit executes edited payload only
- Downstream failure after approve → DLQ + alerts

MCP ALLOW/DENY BY AGENT

operations	CRM r/w (no delete), memory, workspace.read, notify
support	CRM read only — writes denied
document	workspace.read, memory.retrieve — no CRM write/notify
system	*

Deny → 403 + mcp.tool.denied audit — never silent skip

EVENT ENVELOPE + SANDBOX ADAPTERS

Fields: event_id · type · source · idempotency_key · payload · received_at · correlation_id

HubSpot / Odoo	Sandbox IDs → live when credentials set
Private AI	private_mock default · Ollama optional

Architecture rule: FastAPI owns AI / RAG / MCP / domain · n8n stays thin · money moves only after HITL · tools only through MCP.

Reliability, milestones, IDP/finance, leads/research, DLQ/audit

RELIABILITY AND SECURITY

AREA	WHAT SHIPPED
Idempotency	Webhook/CRM/ERP claim keys
Fail loud	Retries, timeouts, DLQ, alerts
MCP AuthZ	Deny wins; unauthorized audited
HITL	High-risk ERP/email/proposal packages
RBAC	JWT admin/viewer + service API key
Private AI	Sensitive classes never on public LLM

PHASE GATES + CLOSEOUT

Phase 1 (M4) — silent fail/dup/unapproved → 0; FR-32; classify ≥85%; RAG ≥80% Phase 2 (M7) — MCP deny; HubSpot; Odoo HITL; IDP ≥90%; browser DLQ Phase 3 (M8–M10) — RBAC; eval; alerts; docs+KT; PRD 28/28	
uat_closeout.py	PASS
API milestone	M10
PRD §16	28/28
Evidence chain	Phase gates + uat_closeout + m10 docs

MILESTONES M0–M10 (ALL COMPLETE)

ID	THEME	WHAT UNLOCKED
M0–M2	Kickoff → orch + RAG	Architecture, LangGraph, agents, Qdrant seed KB
M3–M4	Email + HITL + reliability	Approvals, audit, DLQ, FR-32 Phase 1 Passed
M5–M6	MCP + HubSpot + IDP/Odoo	Tool gateway allowlists, CRM, Finance HITL
M7	Sales / Research / Phase 2	Leads, proposals, browser DLQ, Phase 2 Passed
M8–M10	Admin → private AI → closeout	Eval/RBAC, private AI, SF/QB stubs, UAT PASS, API M10

IDP 8-STEP + FINANCE HITL

Ingest → checksum dedupe → classify → extract → validate → rules → persist → execute.

- invoice / PO / quotation required fields
- idp_review = no ERP on approve
- Below finance threshold + valid → auto MCP write

LEAD / PROPOSALS / RESEARCH / BROWSER

- ICP score 0–100 + explanation → CRM note
- Proposals: SOW/timeline/pricing; send HITL
- Research sources with URL/timestamp
- Browser sandbox/playwright/firecrawl → DLQ on exhaust

DLQ REPLAY + IDEMPOTENCY SCOPES

- Blind email replay blocked**
- Replay mode `side_effect` only when approval_id present
- Scopes: email · erp:quotation · crm:update/followup · email:send
- Audit by correlation / agent

Concrete reliability: MCP deny → 403 + mcp.tool.denied audit (never silent skip) · same idempotency key → no second ERP/CRM row · blind email replay blocked · private_mock keeps finance/PII off public OpenAI.

Delivery truth: M0–M10 complete · Phase gates met · uat_closeout PASS — agents that act with permissions and human approval.

OUTCOMES · TECH · DEPTH

Stack, UAT U1–U20, RBAC, eval, n8n, open risks

TECH STACK + REPRODUCE

FastAPI · LangGraph · MCP · n8n · OpenAI + private_mock/Ollama · Postgres · Redis · Qdrant · Neo4j opt. · HubSpot/Odoo (+ SF/QB) · Admin SPA · Compose · GitHub Actions

`docker compose up -d --build` · `uat_closeout.py` · `admin/viewer` fixtures · `/admin/ui` · `/health` · `/metrics`

SCOPE HONESTY

- Portfolio case study — no fake logos or fabricated reviews
- Sandbox-first HubSpot/Odoo/SF/QB/WhatsApp without live tokens
- Live KT remains environment-specific

UAT U1–U20 INDEX (UAT_CLOSEOUT)

PHASE	IDS	COVERAGE
1	U1–U4	Email classify · routing · idempotency/DLQ · FR-32
2	U5–U11	MCP · HubSpot · IDP+Odoo HITL · lead · research/proposal · browser DLQ · Phase 2 pack
3	U12–U20	RBAC · eval CI · alerts · private AI · KAG · SF/QB · channel identity · backup · Phase 3 pack

ADMIN RBAC (+ SERVICE)

viewer	Read lists · no approve/DLQ
admin	Approve · DLQ · eval
service	X-API-Key mutate

Modules include Escalations

EVAL SUITES + MONITORING

email_classify	≥85%
idp_extract	≥90%
rag_qa	≥80%
tool_selection	100%
failure_handling	pass

PRIVATE AI + KAG

- PRIVATE_WORKFLOWS never call public OpenAI
- Default private_mock; Ollama profile optional
- KAG: customer↔deal↔document; Postgres edges; Neo4j optional

N8N SUB-WORKFLOWS + EVENT ENVELOPE

wf_email_inbound · wf_document_process · wf_hitl_resume · wf_crm_hubspot · wf_erp_odoo · wf_notify_ops · wf_error_global · wf_rag_ingest · wf_lead_sequence · wf_report_scheduled

Envelope: event_id · type · source · idempotency_key · payload · received_at · correlation_id

OPEN_RISKS R1–R7 + CI/BACKUP

R1 live mailbox · R2 live CRM/ERP · R3 GPU Ollama · R4 Neo4j · R5 IdP · R6 live UAT users · R7 voice/extra vendors — none reopen Phase gates. CI: compose subset + smokes + eval_ci_subset. Backup: backup_restore_dry_run (U19).

SAMPLE JOURNEYS J1–J6

J1 FR-32 purchase-request	J2 Invoice→HITL→ERP	J3 Lead→proposal HITL
J4 Classify + MCP deny	J5 Channel identity	J6 Private AI + RBAC

PORTFOLIO ONE-LINER

Dockerized AI operations platform with LangGraph agents, MCP permissions, Qdrant RAG, HubSpot/Odoo adapters, IDP, HITL, private AI routing, eval/observability, and JWT admin dashboard — **uat_closeout PASS** through **M10**.

SYSTEM VISUALS

Architecture & before / after

