

MULTI-CHANNEL AI PLATFORM

Advanced AI Automation Platform

Dockerized multi-channel AI engagement for sales, support, qualification, booking, follow-ups, and ops — one shared brain across channels, not a FAQ chatbot or a single linear n8n flow.

M0–M10
~14 WEEKS CLOSED

11/11
UAT DRY-RUN PASS

24/24
PRD DELIVERABLES

API 1.0.0
ARCHITECTURE V1.1

Built by **Abdul Basit** · AI Automation Engineer | n8n, API Integrations & AI Workflows

BUSINESS IMPACT

Faster replies · CRM stays current · clean handoffs

One shared brain across WhatsApp, webchat, email, and voice — grounded answers, automatic CRM sync, structured human takeover with AI pause/resume

OVERVIEW

Every inbound event (WhatsApp, webchat, email, voice) is classified, routed to a specialized agent, grounded in a private KB when needed, persisted to CRM, and resolved by AI or a human with full context — with Redis idempotency, DLQ, booking locks, and Prometheus. FastAPI owns AI/RAG/domain; n8n runs thin side effects only.

TYPE	Multi-channel AI platform (not a single n8n flow)
ROLE	AI Automation Engineer / n8n Architect
DELIVERY	API 1.0.0 · Architecture v1.1 · M0–M10 complete (~14 weeks)
STACK	FastAPI · n8n · React · Postgres · Redis · Qdrant · OpenAI · HubSpot/GHL · Caddy · Prometheus/Grafana

THE PROBLEM

- Fragmented channels** — WhatsApp, web, email, voice bots with no shared brain or CRM sync
- Ungrounded LLM answers** — invent policy/pricing; teams lose trust
- Manual lead loops** — qualify → score → book → follow-up in spreadsheets
- Silent failures** — workflows die without alerts; duplicates reprocess webhooks
- Weak handoff** — raw transcripts; no AI pause/resume or structured package

THE SOLUTION

- Orchestrator + specialists** — classify → route with guardrails
- Qdrant RAG** — citations + confidence gate → escalate, not invent
- CRM factory** — HubSpot primary, GoHighLevel adapter
- Reliability** — Redis idempotency, booking locks, DLQ, Prometheus
- Ops SPA** — JWT RBAC admin / agent / viewer

WHO IT SERVES

- Sales/support teams that need **one brain** across channels
- Ops leads who need KPIs, handoffs, failed jobs, and DLQ in one UI
- Teams replacing four half-built bots with a production-shaped platform
- Dry-run demos without live Meta/HubSpot tokens (mock IDs)

Demo lead: qualify → score → book. **Secondary:** RAG FAQ · handoff pause/resume · viewer routing 403 · webchat · payments.

AGENT ROSTER

AGENT	JOB	AGENT	JOB
Sales	Pitch, objections, lead capture	Support	FAQ via RAG; escalate if ungrounded
Qualification	Score leads; CRM stages	Booking	Slots + Redis lock + Calendar
Follow-Up / RAG	Day 0–14 sequences · grounded citations	CRM / Escalation	Adapter writes · pause AI handoff

Shared brain: one orchestrator across WhatsApp, webchat, email, and voice — RAG-grounded answers, CRM sync, structured handoff, fail-loud ops. Lead score 0–100 → Qualifying / Qualified / book / high-value notify. Stages: New → Qualifying → Qualified → Appointment Booked → Nurture → Won/Lost.

ARCHITECTURE & CAPABILITIES

Solution architecture, guardrails, and handoff ops

Principle: FastAPI owns AI / RAG / domain; n8n owns side effects — Architecture v1.1.

Channels (WhatsApp / Voice / Email / Webchat)

→ Edge (Caddy TLS, FastAPI webhooks, Redis idempotency, Dashboard SPA)

→ AI Core (Orchestrator → specialized agents → RAG → memory → routing)

→ n8n (DLQ, follow-up tick, CRM/calendar/notify) → External (CRM factory, Calendar, Stripe, Slack/Teams)

→ Data (Postgres, Qdrant, Redis) · Observability (Prometheus → Grafana)

COMPONENT	ROLE
API Gateway / webhooks	Signature verify, normalize, idempotency, enqueue
Orchestrator	Classify → route; guardrails; routing configs; memory scope
Agents	Sales, Support, Qualification, Booking, Follow-Up, RAG, Document, Voice, CRM, Escalation
RAG / CRM / Ops	Qdrant grounded generate · HubSpot GHL factory · Vite SPA JWT RBAC

CHANNELS

- **WhatsApp** — signed webhooks; text/media; template/24h; idempotent by message ID
- **Webchat** — POST /webhooks/webchat; widget /webchat/
- **Email** — classify, lead extract, gated auto-reply; attachments → JSON
- **Voice** — Vapi/Retell; STT → LLM+RAG → TTS; transfer + summary

INTELLIGENCE

- **RAG** — PDF/URL/text; citations; escalate on low confidence
- **CRM** — contacts/deals/notes/tasks; scoring; mock IDs when empty
- **Booking** — Google Calendar + Redis locks
- **Follow-ups** — Day 0/1/3/7/14 with suppressions

OPS & SAFETY

- **Handoff / HITL** — pause AI; full context package
- **Dashboard** — KPIs, handoffs, DLQ, KB, routing
- **Monitoring** — aap_http_* · DLQ · voice · payments
- **RBAC** — admin / agent / viewer

GUARDRAILS MATRIX

Injection	Sanitize user + chunks; escalate on detect
Unsupported claims	Low RAG confidence → escalate (no invent)
High-risk booking	Cancel/reschedule → handoff when flag set
Financial / deletion	Keyword scan → high_risk_approval
CRM schema	Validate before HubSpot writes

HANDOFF REASON CODES + PACKAGE

user_request	Customer asks for human
low_confidence	Ungrounded RAG
injection_suspected	Prompt injection
high_risk_approval	Cancel / financial / deletion
crm_validation_failure	Schema reject

Package: summary · transcript · customer · crm · lead · last_intent · recommended_next_action · blocked_action.

Architecture rule: FastAPI owns AI / RAG / domain · n8n stays thin · classify before CRM mutate · RAG before factual claims · escalation wins. WhatsApp: verify → Redis idempotency → orch → agent(+RAG) → n8n → CRM/reply.

RELIABILITY · DELIVERY · DEPTH

Reliability, milestones, notify, memory, email/docs/voice, n8n

RELIABILITY AND SECURITY

AREA	WHAT SHIPPED
Idempotency	Redis + provider_message_id
Fail loud	Retries, DLQ, ops notify
Booking locks	Redis around slot reservation
Auth / RBAC	JWT admin/agent/viewer
Guardrails	Injection, HITL, confidence, CRM validate
Memory / Audit	PII allowlist; sensitive mutations audited

PHASE GATES (MET)

- Phase 1 (M4) — silent fail/dup/unapproved → 0; handoff 100%; booking ≥95%; RAG ≥80%
- Phase 2 (M7) — follow-up exit 100%; voice CRM; safe email
- Phase 3 / M10 — dashboard RBAC; docs + KT; deliverables 24/24

NOTIFICATION MATRIX

lead.high_value	Slack + Teams
handoff.opened	Slack + Teams
appointment.booked	Slack
automation.failed / DLQ	Slack + Teams (critical)
rag.low_confidence	Slack
followup.sent / exited	Log only

MILESTONES M0–M10 (ALL COMPLETE)

ID	THEME	WHAT UNLOCKED
M0–M1	Kickoff + skeleton	Architecture v1, Compose, schema, Redis, error WF, /health
M2–M3	WA + CRM + booking	Orchestrator, RAG, HubSpot, Calendar locks, lead journey
M4–M5	Handoff + follow-ups	HITL, DLQ, sequences, notify matrix, memory allowlist
M6–M7	Email + voice	Email/docs/Drive, voice provider interface, Phase 2 gate
M8–M10	Dashboard + closeout	RBAC SPA, routing, GHL/Stripe/metrics, UAT 11/11

MEMORY PII ALLOWLIST

Allowed: interest, budget, urgency, location, company, booking_state, language, lead_score, qualification, last_intent, offered_slots, last_appointment_id.

Never store: card numbers, SSN, unrestricted free-text dumps.

Preferences: dnc, unsubscribed, timezone, WA opt-in.

EMAIL + DOCUMENT INTELLIGENCE

Pipeline: extract → classify → validate → DB → CRM (if valid) → notify. Formats: PDF/DOCX/XLSX/CSV + vision stub. Schemas: lead_intake, support_ticket, generic_fields. Drive ingest via inactive-by-default n8n workflow.

VOICE PROVIDER PLUG-IN

Contract: start_outbound · end_call · transfer_to_human · synthesize_reply · normalize_webhook. Swap Vapi↔Retell via VOICE_PROVIDER only — no orchestrator rewrite. Live PSTN not required for M7 exit.

FOLLOW-UP SUPPRESSIONS

- DNC / unsubscribed / human-owned
- Appointment booked / customer reply
- Day 0/1/3/7/14 sequence ticks via scheduler

N8N IMPORT + DLQ PROOF

- Activate wf_error_global first
- Leave WhatsApp inbound stub **inactive** (FastAPI owns WA)
- Forced fail: temp throw **or** POST /demo/force-failure → dead_letter list
- Inbound rate-limit → ops notify + throttle ack

Delivery truth: M0–M10 complete · Phase gates met · uat_dry_run 11/11 · test_m10_pack 3/3 · dry-run UAT without inventing client ROI.

OUTCOMES · TECH · DEPTH

Proof, stack, RBAC, routing, payments, UAT map

UAT DRY-RUN

uat_dry_run.py

11/11 Pass

test_m10_pack.py

OK 3/3

PRD §16

24/24 Done (waivers documented)

Waivers (UAT card only): Calendly, Salesforce, M4 prod cutover deferred, live PSTN not required, external IdP out of scope.

RBAC CAPABILITY MATRIX			
CAPABILITY	V	AGT	ADM
List customers/leads/KPIs/handoffs	✓	✓	✓
Resolve handoff / release		✓	✓
Suppressions / follow-up / DLQ ingest		✓	✓
Doc extract / PUT routing / users			✓
JWT HS256 on /admin/* · viewer routing mutate → 403 · denials audited			

TECH STACK + REPRODUCE

FastAPI · n8n · OpenAI · Postgres 16 · Qdrant · Redis 7 · React SPA · Caddy · Prometheus/Grafana · HubSpot+GHL · Calendar · Stripe · Vapi/Retell · JWT · Compose

docker compose up -d --build · uat_dry_run.py · test_m10_pack.py · :8080 /app · /webchat · /health · :3000 · :5678

ROUTING ENGINE + STRIPE/MONITORING/BACKUP

• pipelines[] · score_routes[] · assignee_matrix[] · approval_flags

• Engine: resolve_lead_route · resolve_assignee · approval_required

• Stripe: signed webhook + demo payment_events (no Checkout UI)

• Metrics: aap_http_* · errors · DLQ · voice · payments · Grafana AAP Overview

• Postgres backup/restore drill (staging-safe)

UAT U1-U10 + TESTING PACK LADDER

#	SCENARIO	#	SCENARIO
U1	Demo chat grounded/escalate	U6	Voice qualify + transfer summary
U2	Lead score/stage + mock CRM	U7	Viewer PUT routing → 403
U3	Booking appointment	U8	Webchat channel=web
U4	Handoff open; AI paused	U9	Payment admin row
U5	Email CRM; low-conf no unsafe send	U10	KB re-ingest active

Packs: phase1 → m5–m7 → phase2 → m8/m9 → **test_m10_pack** + uat_dry_run. Cite live /health in demos rather than stale screenshots.

SCOPE HONESTY

• Portfolio case study — no fake logos or fabricated reviews

• Dry-run first without live Meta / HubSpot / Calendar / voice tokens

• Live KT / credential transfer remains environment-specific

DEMO STORYBOARDS A–E

A

Support FAQ via RAG — /demo/chat

B

Lead qualify → score → book

C

High-risk → handoff → release

D

Viewer routing mutate → 403

E

Webchat + payment_events row

SYSTEM VISUALS

Architecture & before / after

