ARCHITECTURAL REFERENCE • TECHNICAL SYSTEMS BLUEPRINT

Applied AI Systems & Autonomous Workflow Blueprint

A production-tested blueprint for architecting sub-35ms intent classification, token-constrained data extraction, and event-driven microservices—eliminating operational drag without consuming core developer sprint points.

Author: **Anas Wahab** (Applied AI Systems Operator & Technical Chief of Staff)

Standard: **Zero Hallucination / Production-Grade**

Origin: **Nexlify Systems Laboratory**

PILLAR 01 • INTENT TRIAGE

Hybrid System 1 / System 2 Routing

Routes operational requests through a sub-35ms non-autoregressive encoder before invoking LLMs.

- **System 1 (Laya / ModernBERT):** Fast CPU forward pass classifies intent into typed enums without token generation.
- **System 2 (Generative):** Triggers Gemini or Claude only when high-dimensional synthesis is strictly necessary.
- **Outcome:** Eliminates 70%+ of generative token overhead and eradicates prompt latency on routine triage.

PILLAR 02 • EXTRACTION

Token-Constrained Schema Metaprogramming

Constrains LLM generation logits at the tokenizer level to verified enterprise database ledgers.

- **Runtime Dynamic Enums:** Uses `pydantic.create_model` to generate enums dynamically from live client records (e.g. Chart of Accounts).
- **Zero Hallucination:** Invalid category codes cannot be emitted by the model's sampling head.
- **Idempotency:** Two-phase check-and-claim locks in PostgreSQL prevent duplicate ledger transactions.

PILLAR 03 • INBOUND OPS

Event-Driven Speed-to-Lead Engine

Compresses inbound prospect qualification from hours to sub-25 seconds with streaming resilience.

- **Express 5 SSE Streaming:** Non-blocking real-time token delivery directly to web and messaging interfaces.
- **LoopGuard State Deduplication:** In-memory lock map with disk persistence prevents infinite webhook reply loops.
- **Multi-Model Failover:** Dynamic fallback across model tiers ensures 99.9% uptime during rate limits.

ENGINEERING ARCHITECTURE COMPARISON

VERIFIED BENCHMARKS

Fragile AI Integration (Traditional Anti-Pattern)

Architecture: Generic prompt wrappers sending raw text to multi-billion parameter LLMs for every task.

Latency: 1,500ms – 4,000ms delay per interaction; high cost per operation.

Reliability: Category hallucinations, broken JSON outputs, and webhook retry duplications.

Engineering Tax: Constant tickets assigned to core developers to patch broken prompts and pipelines.

Engineered Autonomous Systems (Anas Wahab Architecture)

Architecture: Dual-speed triage: CPU non-autoregressive encoder (System 1) + Schema-locked LLM (System 2).

Latency: ~33ms deterministic triage; sub-25s end-to-end inbound qualification.

Reliability: Dynamic Pydantic schema validation, LoopGuard state locks, and AES-256-GCM encryption.

Zero Developer Tax: Built, deployed, and maintained independently without consuming sprint points.

Technical Briefings & Strategic Inquiries:
Available for full-time executive deployment as Technical Chief of Staff or Head of AI Operations.

Email: anaswahabofficial@gmail.com
GitHub: github.com/anaswahab231997
LinkedIn: linkedin.com/in/anaswahab97