BILYTICA AI OPERATING SYSTEM

AI Delivery Framework

Seven gated stages from framing to operations, each with evidence you cannot skip.

DELIVER · Build

WHAT IT IS FOR

The gap between a demo and a production AI system is evaluation, hardening and operations. DELIVER makes those explicit stages with exit evidence, which is why workloads reach production instead of a second pilot.

WHO RUNS IT

Run by the delivery lead; each gate is signed by the business owner and by risk where the tier requires it.

TYPICAL TIMELINE

First governed release typically in weeks, not quarters, once the data domain is certified.

STAGES AND EXIT GATES

1. Frame — Name the decision, the user, the failure modes and what success would look like.

Exit gate: A written definition of a wrong answer exists.

Artefacts: Workload brief · Failure-mode list

2. Ground — Connect to certified data, define retrieval scope, entitlements and refresh.

Exit gate: Every answer can name its source; entitlements are enforced at query time.

Artefacts: Grounding design · Entitlement rules

3. Prototype — Build the thin slice end to end, including the human hand-off.

Exit gate: The slice runs on real data inside your controls.

Artefacts: Working slice · Prompt and tool spec

4. Evaluate — Score against a golden dataset: groundedness, accuracy, refusal behaviour, Arabic quality.

Exit gate: The workload beats the agreed threshold, and regressions block release.

Artefacts: Golden dataset · Evaluation report

5. Harden — Red-team, bound the tools, add rate and cost limits, plan the rollback.

Exit gate: Known prompt-injection and data-exfiltration paths are closed and tested.

Artefacts: Red-team report · Rollback runbook

6. Launch — Release to a bounded population with monitoring and a human escalation path.

Exit gate: Attestation is current and on-call ownership is named.

Artefacts: Launch checklist · Escalation path

7. Operate — Run evaluations continuously, watch cost and drift, and feed fixes back to the data layer.

Exit gate: Weekly evaluation and cost readings exist for every live workload.

Artefacts: Operations dashboard · Change log

STANDARDS CROSSWALK

STANDARD	WHAT IT SATISFIES
CRISP-DM / TDSP	The classic analytics lifecycle, extended with evaluation and hardening for generative systems.
MLOps / LLMOps	Continuous evaluation, versioning, monitoring and rollback in Operate.
OWASP LLM Top 10	Harden stage test list for injection, exfiltration and insecure tool use.
NIST AI RMF	Measure and Manage functions realised as gates rather than documents.

WHAT YOU GET

- **Golden dataset and evals** — Your questions, your expected answers, re-runnable on every change.
- **Gate evidence pack** — Signed exit evidence for each of the seven stages.
- **Operations runbook** — Monitoring, escalation, rollback and change procedure.

WHAT WE REFUSE TO DO

- **Demo-driven delivery** — A convincing demo with no golden dataset cannot be defended when it is wrong.
- **Skipping the hand-off** — If nobody owns the escalation, the first bad answer stops the programme.
- **Prompt tuning as strategy** — Quality gains that are not measured cannot be kept through the next change.

FREQUENTLY ASKED

What is a golden dataset?

A fixed set of real questions with agreed correct answers and sources, used to score every change before release.

Can stages be run in parallel?

Frame and Ground often overlap. Evaluate and Harden cannot be merged into launch — that is where the risk lives.

How is Arabic quality measured?

Through the Arabic Quality Standard: terminology, register, RTL rendering and answer equivalence against the English evaluation set.

How fast is the first governed release?

Typically in weeks rather than quarters, once the data domain the use case depends on is certified. The gates are what keep that speed safe.

Book the Data Foundation Assessment

Full framework detail: https://bilytica.com/frameworks/ai-delivery-framework/

bilytica.com
info@bilytica.com