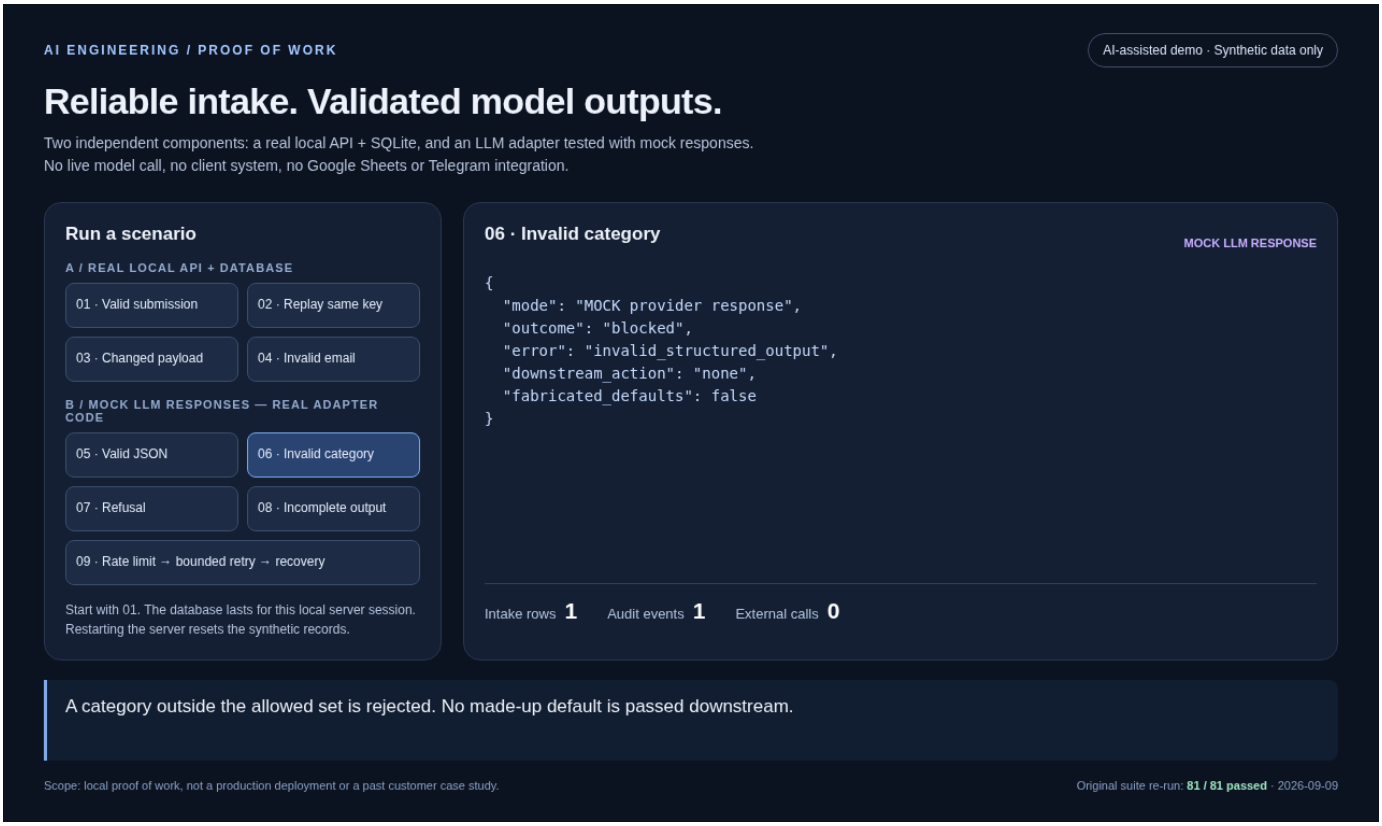
AI ENGINEERING / DEMONSTRATION BRIEF

Reliable intake.
Validated model outputs.

An independent, AI-assisted proof of work. Real local API and SQLite behavior; model responses are controlled fixtures. No customer data, paid model call or production deployment.



Two components, two boundaries

A: FastAPI + Pydantic + SQLite. Validate an inquiry before saving it; use a caller-supplied idempotency key and a transaction for the inquiry and its audit record.
B: Separate LLM adapter. Parse strict structured output, expose refusals/incomplete output, and bound retry attempts. This adapter is not connected to the intake endpoint.

Current local evidence	Observed result
Automated regression suite	81 tests passed; no failed/error cases.
Browser-driven walkthrough	9 scenarios passed against a running loopback service.
External business systems / live model	0 calls. No Sheets or Telegram integration.

EVIDENCE / LIMITS / DELIVERY SCOPE

Demonstrate failure handling, not just a happy path.

Scenario	Observed behavior
Valid request / identical replay	201 then 200; 1 inquiry and 1 audit event remain.
Changed payload / same key	409 conflict; existing records unchanged.
Invalid email	422; no additional records persisted.
Invalid category / model refusal / incomplete	Rejected by the actual parser, using provider-response fixtures.
429 then 200	Real retry code, mock transport: 2 attempts; 1 recorded backoff.

What the sample establishes

The specified application and adapter behaviors pass the enclosed deterministic checks. The tests also cover rollback on audit failure, concurrent duplicate submissions and selected retry/error cases. The demo can be used to discuss how failures should be surfaced and how database invariants should be checked.

What it does not establish

It is not a live AI classification demo, a measured business-accuracy result, a distributed exactly-once guarantee, a security certification or a former client success story. The same content with a different idempotency key is a new request. The two components are intentionally independent. No revenue, time savings or ROI has been measured.

Recording and reproducibility

The 56.64-second video is a silent screen recording with English on-screen explanations; pacing was slowed for readability, so it is not a latency benchmark. The container browser loads the UI in memory and bridges requests to the real localhost service; it does not substitute results. The LLM adapter uses explicit mock fixtures. Source code, the original 81-test suite, runtime versions and all 9 browser results are included in the technical package.

Before commercial use

Agree on the required technical depth, whether narration is required, and the permitted use of this demo-level material. Preserve the demonstration and AI-assistance labels; do not imply a real customer engagement or invent business results. Commercial usage terms must be agreed separately; this brief does not grant exclusive ownership or relicense third-party dependencies.

Technical reference: OpenAI Structured Outputs documentation

https://developers.openai.com/api/docs/guides/structured-outputs

Documentation explains the API contract; it is not evidence that a live integration was tested.