@demotoprod · RAG · also LEVERS

Data pipeline checklist

CSV has no formal specification, and an ordinary table in one file-based SQL engine will accept text in an integer column. A shifted column can then reach the prompt as plain rows. Four checks type and validate the data, then let the model query the clean store, each sized for an afternoon.

1 Type every column, and let a failed cast stop the load

In an ordinary table of one file-based SQL engine, 'xyz' inserted into an integer column is stored as text; a strict table raises an error. A type-guessing CSV loader samples 20,480 rows by default and falls back to text when nothing else fits.

- Name and type every column yourself, NOT NULL where a blank is an error, in strict mode.
- Give the loader those types instead of letting it detect them, so a bad value raises an error.

```
CREATE TABLE invoices (
  invoice_id  INTEGER NOT NULL PRIMARY KEY,
  total_cents INTEGER NOT NULL CHECK (total_cents >= 0)
) STRICT;
```

2 Check each file against a written schema before any row loads

RFC 4180 says each line should contain the same number of fields and the header line is optional, its presence signalled outside the file. A tabular schema standard declares required, unique, minimum, maximum, pattern and enum constraints.

- Reject a whole file whose header or field count differs from the schema.
- Keep the rules (type, required, min, max, pattern, enum) in one schema file; reject any row that breaks one.
- Test the gate with a broken export: a shifted column, text in a number field.

3 Quarantine rejects with a reason, and stop the run when they pile up

A loader that skips bad rows hides them. One CSV reader can store each skipped line with its line number, column, error type (such as a failed cast) and message. One published validation system compares each batch with its schema and alerts the on-call to any disagreement.

- Write each rejected row to a quarantine table with file, line, column and reason; drop none silently.
- Stop the run and call no model when rejects pass a share you set from past clean files.

4 Give the model a read-only query tool in place of the rows

In a 2023 study, accuracy could fall significantly when the answer sat in the middle of a long input, even for long-context models; one task was key-value lookup among 75 to 300 pairs. One engine's security page says untrusted SQL needs sandboxing.

- Expose one tool that runs a single SELECT and returns at most 50 rows; the model never sees the table.
- Treat the model's SQL as untrusted: attach read-only, switch off external access, lock settings, log every query.

```
ATTACH 'clean.db' AS clean (READ_ONLY);
SET enable_external_access = false;   -- no ATTACH, COPY or file reads after this
SET lock_configuration = true;        -- model SQL cannot change settings
```


Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Paste into your project

ONE SCHEMA FILE THE GATE CHECKS AGAINST

```
{
  "fields": [
    {"name": "invoice_id", "type": "integer",
     "constraints": {"required": true, "unique": true}},
    {"name": "total_cents", "type": "integer",
     "constraints": {"required": true, "minimum": 0}},
    {"name": "status", "type": "string",
     "constraints": {"enum": ["<your values>"]}}
  ],
  "primaryKey": "invoice_id"
}
```

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. ### Your demo works and production is stuck.

   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. ### You need senior AI judgement without a full-time hire.

   Fractional AI Lead. Monthly: 90-day minimum, then 30 days' notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. ### An AI feature has to work, by a date.

   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. ### A regulator, an auditor or a customer's security team is asking.

   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL

indica-tech.com/#contact

EMAIL

hello@indica-tech.com

WEBSITE

indica-tech.com

ALL CHECKLISTS

indica-tech.com/checklists/

YOUTUBE

youtube.com/@demotoprod

LINKEDIN

linkedin.com/in/startup-founderx

INSTAGRAM

instagram.com/demotoprod

TIKTOK

tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#data-pipeline