Step 7 — Evaluation, QA & Regression Control

Purpose

Turn the extraction pipeline into a **reliable, auditable system** by measuring correctness, evidence integrity, and regressions over time.

This step ensures the tool is production-credible, not a demo LLM workflow.

Why This Step Matters

LLM extraction systems tend to fail silently:

- hallucinated clauses,
- paraphrased (non-verbatim) evidence,
- missed conflicts after prompt or model changes.

Step 7 establishes objective trust.

No output is considered valid unless it passes the evaluation harness.

Design principle:

> Accuracy and auditability are measured, not assumed.

Scope (MVP)

- Clause families: veto, liquidity, exit, anti-dilution
- Documents: signed term sheets, SHA, amendments
- Languages: French / English
- Evaluation focuses on **precision and evidence quality**, not recall

---

Gold Set (Ground Truth)

A **small, manually verified reference dataset**.

Composition

- 3–5 companies
- Full authoritative document set per company
- Human-validated extracted rights

For each right

- clause_family, clause_type
- trigger / effect
- present = true / false
- **exact evidence quote (≤25 words)**
- source (file, page, article)

Rule

- Gold set is **never auto-generated**
- Any change requires explicit human validation

Core Evaluation Metrics

1. Extraction Precision (Primary KPI)

> When the system claims a right exists, is it correct?

- False positives are hard failures
- “Not Found” is always preferred to guessing

---

2. Evidence Accuracy (Critical KPI)

An item is invalid if:

- the quote is paraphrased,
- the quote does not support the claim,
- provenance is incorrect or missing.

Target: 100% evidence accuracy

Anything less blocks deployment.

---

3. Conflict Detection Quality

Checks that:

- contradictory clauses across documents are detected,
- no silent override occurs.

Both evidences must be surfaced.

---

4. Ambiguity Discipline

Validate that:

- unclear cases are flagged `ambiguous`,
- the model does not infer unstated rights.

Any hallucinated `present=true` is a hard fail.

Regression Testing

All prompts and schemas are treated as **versioned code**.

For every change (prompt, model, chunking, retrieval):

1. Re-run the pipeline on the gold set
2. Compare outputs vs ground truth
3. Track deltas:
    - added / removed rights
    - changed evidence
    - changed flags

A run fails if:

- precision drops,
- evidence accuracy < 100%,
- new hallucinations appear.

Human-in-the-Loop Review

Human review is required when:

- confidence < threshold,

- flag ∈ {ambiguous, conflicting},
- first run on a new company.

Reviewer actions:

- validate evidence quote and provenance,
- approve / reject the extracted item,
- optionally update the gold set.

The system accelerates review — it never replaces it.

Outputs of Step 7

- Gold set (JSON, versioned)
- Evaluation metrics per run
- Regression diffs (before / after changes)
- Clear audit trail linking outputs to evidence

This makes results:

- defensible internally,
- explainable to legal counsel,
- credible in interviews.

Exit Criteria (MVP)

Step 7 is considered complete when:

- Gold set exists (≥ 3 companies)
- Regression tests run end-to-end
- Evidence accuracy = **100%**
- No hallucinated rights observed
- Prompt and schema versions are tracked

Only after this:

- new clause families can be added,
- portfolio-wide scaling is reasonable,

- Phase-2 features (OCR, benchmarking) are justified.