



CATHA PRACTICAL AI FOR CONSTRUCTION LIBRARY

Trust, But Test Construction AI

How to Evaluate AI Before It Runs Construction Workflows

Guide 8

Practical AI for Construction. Responsible. Outcome-driven.

Catha Construction Services | 2026 Edition

How to Use This Guide

A demo shows that AI can work. Evaluation shows whether it works reliably enough for your business.

Construction AI should be tested on representative project work before production. Build evaluation sets from actual drawings, specs, RFIs, submittals, estimates, correspondence and workflow exceptions where permitted.

Best use

Build a test set before production. Measure success, failure and uncertainty against ground truth and business consequences.

Evaluation Stack

Layer	What to test
Model/output quality	Accuracy, completeness, relevance, groundedness
Workflow quality	Correct routing, tool selection, state handling
Safety/control	Permissions, prohibited actions, escalation
Operational reliability	Latency, retries, duplicate prevention, outage behavior
Economic performance	Cost per successful outcome, human review burden
Human usability	Clarity, trust, adoption, correction effort

Build a Representative Test Set

- Normal cases that represent most production volume.
- Hard cases that require judgment or multiple sources.
- Known failure cases.
- Missing/contradictory data.
- Adversarial or policy-sensitive requests.
- Boundary cases just inside and outside allowed authority.

Example Release Thresholds

Metric	Pilot threshold	Production threshold
Critical factual accuracy	Define	Define
Unsupported claims	Define max	Lower max
Correct escalation	Define	Near-100% for high-risk cases
Duplicate external actions	0	0
Unauthorized actions	0	0
Human correction time	Target	Target

Metric	Pilot threshold	Production threshold
Cost/successful case	Target	Target

Regression Testing

Check	Item	Notes
☐	Test suite reruns after prompt changes	_____
☐	Rerun after model changes	_____
☐	Rerun after tool/API changes	_____
☐	Compare new vs. prior version	_____
☐	Block release on critical regressions	_____
☐	Retain failing examples for future tests	_____

Golden rule

If a system can create consequential business effects, evaluation must test the full workflow - not just whether the model writes a convincing answer.

Construction Applications

Measure factual accuracy, completeness, source grounding, format compliance, false confidence, exception handling and whether a qualified reviewer can efficiently verify the result.