

Construction Intelligence Benchmark Protocol

RRI-M02 · Version 0.1

A Test Structure for Comparing Construction Intelligence Workflows Against Defined Baselines

Rhogan Research Institute

Construction Intelligence Research & Standards

Classification: Benchmark Protocol

Status: Developing Research Methodology

Version: 0.1

Abstract

Construction Intelligence workflows are increasingly compared by speed, volume of findings, model capability, or demonstration quality. Those measures can be useful operational signals, but they do not establish whether one workflow produces more reliable construction intelligence than another.

A valid comparison requires a controlled evidence set, a defined decision context, an independently established reference condition, equivalent access to source information, and measures that distinguish correct findings from unsupported output.

The Construction Intelligence Benchmark Protocol (CIBP) is a developing methodology for testing Construction Intelligence workflows against defined baselines using adjudicated construction conditions, evidence-level scoring, controlled workflow execution, and professional validation.

The protocol is designed for commercial construction and may be applied to manual review, search-assisted review, AI-assisted review, domain-specific Construction Intelligence systems, or hybrid human-machine workflows.

The benchmark does not ask which workflow produces the most output. It asks which workflow identifies the right conditions, misses the fewest material conditions, supports its findings with the correct evidence, controls uncertainty, and improves a defined construction decision.

1. Purpose

The purpose of the Construction Intelligence Benchmark Protocol is to establish a repeatable method for answering a question that construction teams, technology developers, contractors, estimators, owners, and researchers increasingly confront:

When two or more workflows review the same construction information for the same decision, which workflow produces the more valid, complete, traceable, and decision-useful result?

The protocol is intended to replace weak comparison statements such as:

- “The AI found more issues.”
- “The review was 80% faster.”
- “The output looked more detailed.”
- “The model understood the drawings.”
- “The system produced fewer pages.”
- “The reviewer liked this one better.”

with measurable and inspectable evidence.

CIBP is not intended to certify an AI model, declare a universal winner, replace professional judgment, establish legal reliance, or prove that a workflow will perform equally on every project.

It measures comparative workflow performance under defined construction conditions.

2. The Controlled-Comparison Principle

Construction Intelligence performance is context dependent. A workflow that performs well on a small, well-coordinated bid set may fail on a revision-heavy hospital renovation, an occupied phased project, or a document package containing multiple design authorities and superseded information.

Accordingly, no benchmark result should be reported without defining the evidence set, decision context, reviewer population, workflow configuration, and baseline.

**Same evidence. Same decision. Same test condition. Independent truth set.
Measurable outcome.**

Without those controls, a comparison is a demonstration rather than a benchmark.

3. Benchmark Scope

CIBP evaluates the Construction Intelligence process itself. It does not score whether the project documents are good, whether the design is correct, or whether the project is low risk.

The benchmark may evaluate:

- manual professional review;
- manual review supported by native PDF search and document-navigation tools;
- generic AI-assisted review where confidentiality and data governance permit;
- specialized construction software;
- domain-specific Construction Intelligence workflows;
- hybrid workflows combining automated detection with qualified professional validation.

The protocol may be applied at the whole-project level, trade level, document-family level, or decision-specific level, provided the benchmark unit and scope are declared before testing begins.

4. Comparison Baseline Architecture

A benchmark requires a defined comparator. No single comparator is appropriate for every research question, so CIBP uses baseline classes.

Class	Baseline	Purpose
BL-1	Qualified Manual Review	Measures performance against a competent professional reviewing the same evidence without AI.
BL-2	Search-Assisted Review	Measures incremental value beyond conventional PDF search, bookmarks, hyperlinks, and document-navigation tools.
BL-3	Generic AI-Assisted Review	Measures domain-specific workflow value beyond a controlled general-purpose AI workflow where use is permitted.
BL-4	Prior Production Workflow	Measures improvement against the organization's existing validated process.
BL-5	Peer Construction Intelligence Workflow	Supports direct system-to-system comparison under identical evidence and test controls.

At least one qualified professional baseline should be included when the claim concerns construction interpretation, scope, coordination, risk identification, RFI generation, or bid-position control.

The baseline is not the truth set. It is a workflow being compared to the workflow under test.

5. Benchmark Unit and Reference Truth

The primary benchmark unit is the validated construction condition.

A construction condition is an evidence-supported circumstance that is relevant to the defined decision and can be independently evaluated as present, absent, resolved, unresolved, or not determinable from the available information.

Examples include a conflicting finish assignment, missing transition definition, superseded specification requirement, undefined substrate responsibility, incomplete product definition, unpropagated addendum change, ambiguous stair detail, unresolved allowance, or contract requirement that changes trade responsibility.

5.1 Truth-Set Record

Field	Required Benchmark Record
Condition ID	Stable unique identifier.
Decision Context	The decision the condition affects.
Source Evidence	Authoritative drawing, specification, addendum, RFI, contract, schedule, correspondence, or other permitted source.
Location	Sheet, detail, section, page, room, area, clause, or other traceable location.
Condition Statement	Neutral description of the validated condition.
Expected Classification	Condition type and status.
Materiality	Decision-critical, material, secondary, or informational.
Decision Impact	Why the condition matters to the defined construction decision.
Acceptable Evidence	Source combinations that can validly support the finding.
Adjudication Status	Confirmed, disputed, excluded, or pending.

Truth-set construction must be completed without using the tested workflow's output as the sole basis for defining what is true.

A benchmark that defines the answer from the system being tested is circular and cannot establish comparative validity.

6. Test Corpus Design

The benchmark corpus should represent the environments in which the workflow is expected to operate. A convenient or cherry-picked project set can produce impressive results that do not generalize.

6.1 Required Stratification

Dimension	Representative Strata
Project Delivery / Stage	Design development, bid/GMP, award/contract, procurement, construction execution.
Project Type	Commercial office, hospitality, multifamily, education, healthcare, retail, industrial, mixed-use, renovation.
Construction Condition	New construction, renovation, occupied work, phased work, existing-conditions dependency.
Document Condition	Coordinated, incomplete, revision-heavy, addenda-heavy, mixed authority, scan-heavy, correspondence-dependent.
Scope Domain	Trade-specific and cross-trade conditions; Division 09 may be benchmarked as a dedicated subprotocol.
Project Scale	Small, medium, and large document environments defined by corpus distribution rather than arbitrary page-count claims.
Decision Criticality	Informational, material, and decision-critical conditions.
Document Maturity	Multiple maturity levels or decision-readiness states where an independent maturity assessment is available.

6.2 Development, Validation, and Challenge Sets

The corpus should be partitioned before formal testing.

- Development Set — may be used to configure prompts, mappings, rules, extraction logic, or workflow design.
- Validation Set — used to estimate benchmark performance after the workflow configuration is frozen.
- Hidden Challenge Set — withheld from developers and operators until final evaluation to test generalization and reduce benchmark gaming.

No performance claim should be based only on projects used to tune the workflow.

7. Workflow Execution Controls

The tested workflows must operate under equivalent evidence conditions. Differences in source access can dominate the result and falsely appear to be intelligence differences.

7.1 Required Controls

- Identical document package and cutoff timestamp.
- Identical access to drawings, specifications, addenda, RFIs, bid forms, contracts, correspondence, and approved project information included in the test.
- Declared exclusions and unavailable source families.
- Frozen workflow version, model version, configuration, prompt set, retrieval settings, and knowledge-base state.
- Recorded operator identity or role and reviewer qualification.
- Equivalent decision brief and scope instructions.
- No access to truth-set labels, adjudication notes, or competitor outputs during the run.
- Defined treatment of prior project knowledge, templates, historical databases, and organizational memory.
- Recorded start and stop rules for time-based comparisons.

7.2 Stochastic and Adaptive Systems

When a workflow can produce materially different results across runs, repeatability must be measured rather than assumed.

The same frozen evidence set should be evaluated in multiple independent runs. Randomness controls, model temperature, seed controls where available, and any agentic retry or tool-use logic should be recorded.

A workflow that succeeds once and fails repeatedly should not be reported as equivalent to a workflow that consistently reproduces the validated result.

8. Benchmark Test Classes

CIBP separates test classes because Construction Intelligence is not a single task. Retrieval, interpretation, coordination, revision analysis, and commercial control can fail independently.

Test	Construction Intelligence Capability	Primary Question
BT-1	Source Retrieval	Can the workflow locate the evidence required to evaluate the condition?
BT-2	Single-Source Interpretation	Can the workflow correctly interpret an isolated requirement, note, schedule entry, or clause?
BT-3	Cross-Document Coordination	Can independent sources be related and compared correctly?
BT-4	Contradiction Detection	Can the workflow detect when governing sources produce incompatible project interpretations?
BT-5	Scope Identification	Can required work, locations, systems, accessories, and interfaces be found?
BT-6	Responsibility Assignment	Can the workflow establish who is responsible without inventing an assignment?

A workflow should not receive credit for a correct conclusion reached through an unsupported chain if the benchmark claim includes evidence traceability.

Where multiple evidence paths can validly establish the same condition, the truth set should identify all acceptable paths that are known before scoring.

8. Benchmark Test Classes — Continued

Test	Construction Intelligence Capability	Primary Question
BT-7	Revision / Addenda Sensitivity	Can the workflow identify what changed and which prior conclusions are no longer valid?
BT-8	Product / Procurement Definition	Can the workflow determine whether product information is sufficient for pricing, approval, ordering, or release?
BT-9	RFI-Generating Condition	Can unresolved ambiguity be converted into a precise, evidence-grounded clarification request?
BT-10	Qualification / Bid Position	Can uncertainty be translated into a defensible assumption, qualification, exclusion, allowance, or bid position?
BT-11	Uncertainty Control	Can the workflow refrain from asserting a conclusion when evidence is insufficient?
BT-12	Knowledge Continuity	Can validated findings, unresolved conditions, and decision rationale survive workflow handoffs and revisions?
BT-13	Adversarial Document Condition	Can the workflow resist stale, duplicated, contradictory, decoy, or superseded information?

8.1 Negative Controls

The benchmark should contain conditions where the correct answer is that no supported finding exists.

Negative controls are necessary to measure overproduction, hallucinated scope, false contradictions, invented responsibility, and unsupported certainty.

Construction Intelligence must be rewarded for knowing when the documents do not support an answer.

9. Primary Validation Metrics

Primary metrics evaluate whether the workflow correctly identifies validated construction conditions. Raw finding count is not a primary measure.

Metric	Definition
Detection Recall	$TP / (TP + FN)$. Of all validated relevant conditions, the percentage identified by the workflow.
Precision	$TP / (TP + FP)$. Of all scored findings produced by the workflow, the percentage validated as supported.
Missed-Condition Rate	$FN / (TP + FN)$. The percentage of validated relevant conditions the workflow failed to identify.
False Positive Rate	Unsupported or incorrect findings relative to the scored opportunity set defined by the benchmark.
Decision-Critical Recall	Recall calculated only on conditions preclassified as decision-critical.
Evidence Traceability Rate	Accepted findings with a valid trace to supporting construction information divided by all accepted findings.
Evidence Correctness Rate	Cited findings for which the cited source actually supports the stated conclusion divided by all cited findings.

Precision and recall must be reported together. A workflow can increase recall by producing more findings, but excessive unsupported findings increase validation burden and can reduce decision quality.

Decision-critical conditions should be reported separately so strong performance on minor findings cannot mathematically conceal a material miss.

9. Primary Validation Metrics — Continued

Metric	Definition
Severity / Materiality Accuracy	Agreement between workflow classification and the adjudicated materiality class.
Relationship Accuracy	Correct identification of the specific cross-document relationship producing the finding.
Revision Sensitivity	Affected validated conditions correctly updated after a controlled revision divided by all truth-set conditions affected by that revision.
RFI Conversion Accuracy	RFI candidates that correctly state the condition, cite supporting evidence, and ask a resolvable question.
Qualification Defensibility	Qualifications that accurately control an unresolved condition without silently expanding or narrowing documented scope.
Appropriate Abstention	Insufficient-evidence cases in which the workflow correctly states that the condition cannot be determined.
False-Confidence Rate	Conclusive assertions made where the truth set establishes that available evidence is insufficient or contradictory.
Human Override Rate	AI-supported or automated findings materially changed or rejected during qualified professional validation.
Repeatability	Consistency of findings when the same frozen workflow evaluates the same frozen evidence under repeated controlled runs.

These measures should be reported by test class, project, materiality level, and decision context where sample size permits.

A single blended score should not replace the component metrics during initial validation.

10. Human Validation and Adjudication

Professional validation is part of the benchmark architecture, not a cosmetic review after the result is produced.

10.1 Reviewer Requirements

Reviewers should have demonstrated competence appropriate to the tested domain, decision context, and construction scope. A flooring benchmark should not rely solely on a reviewer whose expertise is unrelated to flooring; a contract-scope benchmark should not be adjudicated only as a drawing-reading exercise.

10.2 Independent Adjudication

Where practical, truth-set conditions should be established by at least two independent reviewers. Disagreements should be resolved through a documented adjudication process by a third qualified reviewer or consensus panel.

The adjudication record should preserve:

- the disputed condition;
- the evidence reviewed;
- each reviewer's initial classification;
- the reason for disagreement;
- the final adjudicated disposition;
- whether the condition remains unsuitable for benchmark scoring.

10.3 Blinding

Adjudicators evaluating correctness should be blinded to workflow identity where feasible. Product names, vendor names, and output formatting should be masked when they could influence judgment.

Unsuccessful findings, false positives, and benchmark disagreements must remain in the research record.

11. Time, Effort, and Workflow Efficiency

Speed is valuable only after result quality is known. A workflow that is fast because it misses difficult conditions is not operationally superior.

11.1 Required Time Measures

- Time to first validated decision-critical finding.
- Total elapsed review time.
- Qualified professional hands-on time.
- Validation and correction time.
- Time spent locating evidence.
- Time spent reconciling contradictions.
- Time required to update the result after a controlled revision.

11.2 Validated Throughput

Validated throughput should be reported as accepted findings or completed benchmark conditions per unit of professional effort, not simply pages processed, tokens consumed, or findings emitted.

Where automated runtime is reported, human validation time should be reported separately. Combining them into one number can obscure a workflow that is computationally fast but expensive to verify.

11.3 Cost Comparison

Cost may be measured when the study objective requires it, but cost comparisons must declare licensing, compute, operator labor, professional validation, setup, integration, and correction costs included in the calculation.

A low-cost workflow that shifts substantial work into hidden review or rework should not be characterized as lower-cost without that burden being measured.

12. Statistical Comparison Structure

Construction conditions are nested within projects. Treating thousands of findings from a small number of projects as thousands of independent project observations creates false precision.

12.1 Primary Sampling Unit

For whole-project benchmark claims, the project should ordinarily be treated as the primary sampling unit. Condition-level statistics may also be reported, but project-level aggregation should be used to test whether performance generalizes across projects.

12.2 Paired Comparison

Whenever practical, competing workflows should evaluate the same projects and truth-set conditions. Paired designs reduce noise caused by project difficulty and document maturity.

12.3 Uncertainty

Comparative results should include confidence intervals or another declared uncertainty measure. Bootstrap intervals resampled at the project level are appropriate for many nonparametric benchmark metrics. Other tests may be used where their assumptions are satisfied.

12.4 Inter-Rater Agreement

Where benchmark labels depend on professional classification, inter-rater agreement should be measured. Cohen's kappa, weighted kappa, Krippendorff's alpha, or another appropriate statistic may be used depending on the number of reviewers and label structure.

12.5 Sample Size

Sample size should be set by a predeclared precision or statistical-power objective where possible. Pilot data may be used to estimate variance, but the final validation set should be locked before comparative testing.

Large condition counts do not compensate for a narrow project sample.

13. Critical Failure Gates

Some benchmark failures are too important to be hidden inside an average.

CIBP therefore separates metric reporting from critical-condition gates.

13.1 Examples of Gate Conditions

- A decision-critical condition is missed and the workflow nevertheless states that the affected scope is resolved.
- A finding cites evidence that does not support the conclusion.
- A superseded requirement is treated as governing after the controlling revision is available.
- Responsibility is assigned to a trade or party without document support.
- The workflow invents a product, quantity basis, contractual obligation, code requirement, or field condition.
- The workflow suppresses a known contradiction and presents a single unsupported interpretation as resolved.
- A revision changes a validated decision-critical condition and the workflow fails to update the prior position.

The research program should predefine which gates make a run non-passing for the specific use case. Thresholds should be empirically calibrated rather than selected merely to favor a preferred workflow.

13.2 No Mathematical Concealment

A high overall precision score cannot erase a missed decision-critical scope conflict. A fast average review time cannot erase a false contractual assignment. A strong retrieval score cannot erase unsupported interpretation.

Critical failures are reported as failures, not diluted into averages.

14. Revision, Change, and Longitudinal Testing

Construction Intelligence is not static. Projects change through addenda, bulletins, ASIs, RFIs, substitutions, design development, contract changes, and field information.

14.1 Controlled Revision Test

A controlled revision benchmark begins with a frozen project state, records the workflow's validated findings, then introduces one or more known document changes.

The test measures whether the workflow:

- detects the changed source;
- identifies the affected prior conditions;
- retires superseded conclusions;
- creates new conditions where required;
- preserves unaffected validated findings;
- updates qualifications, RFIs, or bid positions affected by the change;
- maintains traceability to the new governing evidence.

14.2 Knowledge Continuity

A longitudinal workflow should also be tested across handoffs: estimating to pre-construction, pre-construction to contract, contract to project management, project management to field execution, and execution to closeout where the use case extends that far.

The benchmark should measure how much validated project intelligence survives the handoff without being recreated, contradicted, or silently lost.

14.3 Memory Leakage Control

For systems with persistent project memory, the benchmark must distinguish intended retained project knowledge from contamination by truth-set labels, prior answer keys, or unauthorized cross-project information.

15. Benchmark Reporting

Every benchmark report should provide enough information for an independent reviewer to understand what was tested, how it was tested, and what the result does and does not establish.

15.1 Minimum Reporting Record

Category	Required Disclosure
Research Question	The specific comparison the benchmark is intended to answer.
Decision Context	The construction decision the workflow is supporting.
Corpus	Project count, strata, document types, maturity characteristics, and inclusion/exclusion rules.
Truth Set	How conditions were established, reviewed, adjudicated, and versioned.
Workflows	Baseline and tested workflow versions, configurations, model/tool versions, prompts/rules, and operator roles.
Controls	Evidence cutoff, access equivalence, time rules, leakage controls, repeat runs, and blinding.
Metrics	Definitions, denominators, materiality classes, and calculation rules.
Results	Project-level and aggregate results with uncertainty.
Failures	Decision-critical misses, unsupported findings, evidence failures, and adjudication disputes.
Limitations	Known threats to generalization, validity, or reproducibility.

Benchmark reports should retain zero and negative results. Selective reporting of only successful project examples invalidates comparative claims.

16. Comparative Claim Rules

Benchmark language should be narrower than marketing language. The claim must match the metric actually tested.

Claim	Required Evidence
"Faster"	Lower validated review time or professional effort under equivalent evidence and output requirements.
"More accurate"	Improvement in predefined correctness measures such as precision, recall, evidence correctness, or classification agreement.
"Finds more"	Higher recall on the locked truth set, not simply a larger output count.
"More complete"	Higher validated coverage, including decision-critical recall, without unacceptable precision loss.
"More traceable"	Higher evidence traceability and evidence correctness rates.
"Requires less review"	Lower measured human correction and validation burden for an equivalent accepted output.
"More reliable"	Repeatable performance across independent runs and multiple projects, with uncertainty reported.
"Better"	A preregistered multi-metric superiority rule with no material regression on critical safety, scope, evidence, or uncertainty-control gates.

16.1 Superiority

A workflow should not be declared superior because it wins one favorable metric while materially regressing on another metric that is critical to the declared use case.

Where tradeoffs exist, the report should show the tradeoff rather than collapsing it into an opaque score.

16.2 Composite Scores

Composite scores may be explored after validation, but initial research should report component metrics independently. Weighting schemes can create the appearance of superiority simply by assigning more value to metrics on which one workflow already performs well.

17. Relationship to Document Maturity and Project Risk

CIBP evaluates workflow performance. The Document Maturity Index evaluates the quality and sufficiency of project information for a defined decision. Project risk evaluates documented exposure. These are separate measurement families.

A workflow can perform well on an immature document package by correctly identifying that the information is unresolved. A workflow can perform poorly on a mature package by missing available evidence.

17.1 Recommended Cross-Analysis

Where DMI or another independently validated information-maturity measure is available, workflow performance may be stratified by document maturity.

This can test questions such as:

- Does recall decline as information becomes more fragmented?
- Does false confidence increase when authority is unresolved?
- Does revision sensitivity change with addenda burden?
- Does human correction burden increase on low-maturity projects?
- Does a workflow preserve performance across different decision contexts?

The maturity score must not be generated by the tested workflow and then used as independent evidence that the same workflow performed well.

17.2 SQUARES RA or Other Risk Models

Risk-model outputs should remain separate from benchmark performance. A workflow does not perform poorly merely because it identifies a high-risk project, and it does not perform well merely because it generates a high risk score.

Project difficulty is a test condition. Workflow performance is the measured response.

18. Validation Program and Anti-Gaming Controls

A credible benchmark must be capable of disproving the system's own performance assumptions.

18.1 Anti-Gaming Controls

- Freeze the primary metrics before the validation run.
- Freeze benchmark inclusion and exclusion criteria before results are known.
- Do not tune the tested workflow on the hidden challenge set.
- Include negative controls and no-answer conditions.
- Include superseded, duplicated, and contradictory source conditions.
- Include difficult but valid evidence paths such as notes, schedules, details, addenda, correspondence, and cross-document dependencies.
- Mask workflow identity during correctness adjudication where feasible.
- Retain all failed runs and reruns in the research record.
- Report model or software updates that occur between benchmark rounds.
- Separate exploratory analysis from confirmatory claims.

18.2 Benchmark Drift

Construction workflows, AI models, software versions, document practices, and project delivery methods change. Benchmark performance should therefore be versioned.

A score from one workflow version should not be silently attributed to a later version. A changed model, retrieval system, prompt architecture, rules engine, validation layer, or document parser may require revalidation.

18.3 External Replication

The strongest evidence occurs when an independent party can reproduce the protocol on a separate project set and reach materially similar conclusions.

Proprietary systems may protect implementation details while still disclosing enough test conditions, metrics, and procedures for the benchmark claim to be independently inspected.

19. Limitations

The Construction Intelligence Benchmark Protocol does not establish:

- that the truth set is infallible;
- that benchmark projects represent every construction sector or delivery method;
- that a workflow validated for one trade will generalize to another;
- that a workflow validated for bid review is suitable for contract interpretation, procurement, or field execution;
- that higher benchmark performance eliminates the need for qualified professional review;
- that a system is secure, private, code-compliant, legally sufficient, or contractually authoritative unless those characteristics are separately tested;
- that lower review time produces lower total project cost;
- that identified conditions will cause actual project loss or change;
- that unidentified conditions do not exist.

Truth sets can contain reviewer error. Construction documents can change after the benchmark cutoff. Professional judgment can vary. Proprietary systems can limit reproducibility. Human reviewers can learn between repeated tests. AI systems can change without obvious interface changes.

These limitations should be documented rather than treated as reasons to avoid measurement.

A benchmark is strongest when it makes uncertainty visible.

20. Research Position

Construction Intelligence should not be evaluated by the sophistication of its interface, the size of its context window, the number of findings it produces, or the speed of an unvalidated run.

A Construction Intelligence workflow earns a stronger benchmark position only when it identifies more of the validated conditions that matter, produces fewer unsupported findings, traces conclusions to the correct construction evidence, controls uncertainty, survives revision, requires less corrective professional effort, and does so consistently across representative projects.

The benchmark must be designed so that the preferred workflow can lose.

If the protocol cannot produce credible evidence that a favored system is slower, less accurate, less traceable, less repeatable, or more dependent on human correction than its baseline, the protocol is not a valid benchmark. It is a demonstration framework.

That distinction is central to credible Construction Intelligence research.

21. Research Basis and Standards Alignment

CIBP is a developing research methodology. Its structure is informed by construction information-management practice, cost-estimating validation practice, and current AI test, evaluation, verification, and validation principles. Alignment does not imply adoption or endorsement by any external standards body.

21.1 Research Alignment

- RRI-M01 Document Maturity Index — distinguishes document maturity, project risk, decision readiness, and Construction Intelligence workflow performance; recommends recall, precision, traceability, repeatability, revision sensitivity, and human validation as separate measures.
- NIST AI Risk Management Framework 1.0 — calls for documented test sets, context-relevant metrics, benchmark comparison, measurement uncertainty, independent review, and objective, repeatable TEVV processes.
- NIST AI 200-2 Initial Public Draft, TEVV-Athlon Framework (2026) — proposes customizable AI assessment structures tied to organizational objectives and real-world outcomes. Draft status must be preserved when cited.
- buildingSMART Information Delivery Specification (IDS) 1.0 — demonstrates the value of explicit, computer-interpretable information requirements and automated compliance checking for construction information exchange.
- National BIM Standard–United States Version 4 — emphasizes defined information exchanges, approvals, information requirements, metadata, and common data environment controls.
- U.S. GAO Cost Estimating and Assessment Guide — emphasizes defined purpose, technical baseline, assumptions, independent comparison, sensitivity, documentation, traceability, and reproducibility.
- AACE International Recommended Practice 56R-08 — relates building and general-construction estimate use to the maturity of project definition deliverables and cautions that document maturity alone does not determine estimate accuracy.

21.2 Selected Sources

- Rhogan Research Institute. RRI-M01: Document Maturity Index — Framework. Version 0.1.
- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.
- National Institute of Standards and Technology. The TEVV-Athlon Framework for Evaluating AI Systems, NIST AI 200-2 Initial Public Draft, 2026.
- buildingSMART International. Information Delivery Specification (IDS) 1.0.
- National Institute of Building Sciences. National BIM Standard–United States (NBIMS-US) Version 4.
- U.S. Government Accountability Office. Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Program Costs, GAO-20-195G.
- AACE International. Recommended Practice 56R-08: Cost Estimate Classification System — As Applied in Engineering, Procurement, and Construction for the Building and General Construction Industries.

Rhogan Research Institute

Research • Measurement • Validation • Standards

Methodology RRI-M02 — Construction Intelligence Benchmark Protocol

Version 0.1 — Developing Research Methodology

This protocol is a developing research methodology and is not presented as an adopted industry standard.