# Completeness, with simple numbers

A worked guide to evidence selection, weighting, and replay. Prepared 8 September 2026 for the Sanjay evaluation rebuild.

## Three decisions, kept separate

| Decision | Question |
|---|---|
| Selection | Which evidence belongs in the required set? |
| Weighting | How much does each selected item count? |
| Credit | What proves that the item was covered? |

**Node coverage = covered selected nodes / all selected nodes.** Tags and edges help find the nodes; they are not mixed into the denominator.

A node with eight tags still counts once. A node reached by five paths still counts once. Reading a node twice gives no extra credit.

## What can be replayed?

Retained tool traces let us recompute coverage under different declared selection and weighting policies. We do not need to regenerate the old answers. We still need the required-evidence set: activity alone cannot tell us what was missing.

These calculations measure evidence exposure under a policy. They do not establish that an answer understood, used, or correctly addressed the evidence.

# 1. Change which evidence is required

Question: Was the applicant given notice? Suppose the query directly matches two findings. One additional finding is reached at each extra hop.

| Node | Evidence | Distance |
|---|---|---|
| A | Hearing notice | Seed / hop 0 |
| B | Applicant statement | Seed / hop 0 |
| C | Dispatch register linked by dispatch ID | Hop 1 |
| D | Returned delivery receipt linked by tracking ID | Hop 2 |

One hop means finding → shared tag → another finding. This is an association, not a proven reasoning relationship.

**The same agent receives A, B and C, but not D.**

| Required set | Arithmetic | Coverage |
|---|---|---|
| Seeds only: A, B | 2 / 2 | 100% |
| Up to one hop: A, B, C | 3 / 3 | 100% |
| Up to two hops: A, B, C, D | 3 / 4 | 75% |

Removing neighbours raises this score because we stop requiring D. But D could be decisive. A higher score does not prove the smaller set is a better definition of completeness.

# 2. Keep the evidence; change weights

Keep A, B, C and D in the denominator. The agent still receives only A, B and C.

| Weight rule | A | B | C | D |
|---|---|---|---|---|
| Equal | 1 | 1 | 1 | 1 |
| Half per hop | 1 | 1 | 0.5 | 0.25 |
| Reviewed importance (illustrative) | 1 | 1 | 1 | 3 |

| Policy | Covered / required weight | Coverage |
|---|---|---|
| Equal | 3 / 4 | 75% |
| Half per hop | 2.5 / 2.75 | 90.91% |
| Reviewed importance | 3 / 6 | 50% |

The returned receipt has low weight under distance decay but high weight under the illustrative importance rubric. Distance and importance answer different questions.

## A complete bundle changes the credit rule

Suppose there are three equally weighted obligations: obtain A; obtain B; obtain C **and** D. With A, B and C delivered, only two obligations are satisfied: 2 / 3 = 66.67%. C alone does not complete the third bundle.

This is obligation coverage, not node coverage. The current S5 replay is seed-only and requires every source page in each selected node’s bundle; it is diagnostic, not a full signed S5 certificate.

# 3. What changes in the real replay?

| Historical system | Seeds | 1 hop | 2 hops | Decay |
|---|---|---|---|---|
| Suit V3 · deployed | 65.19% | 63.88% | 63.41% | 63.92% |
| Sol · PDF tools | 55.93% | 54.29% | 54.19% | 54.63% |
| Codex · Sol · graph + PDF tools | 70.95% | 67.72% | 67.29% | 68.28% |

Each cell is the mean over the same 50 questions. These are discovery scores: graph returns, search-result pages, or raw-page exposure can earn credit. They are not strict complete-text delivery scores.

The four columns share the same frozen S1 query projection and hub policy. Only expansion depth and the declared weighting change. The scoring view also retains S1-weighted seed-only and two-hop variants.

The two-hop S1-weighted replay was checked against all 650 original system/question scores within stored rounding tolerance before publishing alternatives.

Astra high adds a new model/reasoning condition in API PDF-only, API graph + PDF, Codex PDF-only, and Codex graph + PDF modes. Consult the live site for completion status; this document does not invent final Astra results.

Website: https://final-eval.34.100.223.46.sslip.io/selection.html

# 4. How to trust the numbers

| Check | Why it matters |
|---|---|
| Freeze task, corpus and graph | A changed record is a different evaluation. |
| Freeze seeds and policy | The answer must not choose its own denominator. |
| Record exact delivery | A requested page, snippet and complete page are different evidence. |
| Repeat the calculation | Identical artifacts must produce identical scores. |
| Test extraction and paraphrases | Stable page selection can hide unstable findings, even on one page. |
| Audit with human-required evidence | A perfectly repeatable denominator can still omit essential evidence. |

The live comparison shows per-question covered weight, total weight, node counts, required node IDs and weights, and a denominator hash. Missing runs are pending, not zero. Empty denominators are unscorable.

Historical traces have retention and truncation limits. The raw-locator view does not establish verbatim full-page delivery. New Astra API traces retain full tool results; this audit improvement is disclosed rather than silently treated as identical historical instrumentation.

## A useful next experiment

Review indirect-only evidence such as D without looking at system scores. If reviewers judge it required, removing it is not an improvement. If it is incidental, refine the expansion rule. Select policy using held-out evidence judgments, not the largest percentage.

Source implementation: build_selection_comparison.py; historical definitions: build_five_method_comparison.py. Historical reports remain available under Method audit on Final Eval.