TECHNICAL PAPER · E · WORKING PAPER

Instrument Validation at Corpus Scale: 393 Million Lines of Change

A measurement instrument is only as good as its proof of correctness. This is the evidence that CodeDelta measures what it claims, unchanged, from a 165-file library to a 43.5-million-line browser.

Working paper · first published 14 July 2026 · every claim carries its check

ABSTRACT

The metrics in Papers C and D are only worth their validation. This paper is that validation. A deterministic churn engine was exercised across a **1,437-window corpus (136 codebases, 393 million logical statements of churn)** and a single **43.5-million-line, 326,948-file Chromium** scan, and its correctness was established four ways. **Value preservation:** the full study-subject set reproduces bit-identically across engine revisions, and the Chromium window reproduces bit-identically across independent runs. **Quantified error:** the corpus campaign doubled as a stress campaign — a small number of inputs triggered latent memory-safety defects, all diagnosed and fixed; the one run known to have traversed a defect differed from the corrected engine by a single statement-level count in 46,146 (0.002%), and the higher-level metrics behind every published figure were identical. **Coverage:** after the fixes, every previously-failing window measures cleanly — the corpus has no holes attributable to instrument failure. **Accuracy at scale:** a synthetic two-million-line single file with planted churn was measured exactly. The instrument does not merely survive scale; it stays correct across it.

IN PLAIN TERMS

Ask CodeDelta

> Anyone can print a number. The hard part is proving the number is right — and staying honest when it is not. While measuring the corpus, a handful of unusual files exposed old bugs deep in the engine. We found them, fixed them, and then proved two things: the fixes changed no correct answer, and where a bug had once nudged a result, it moved it by one count in forty-six thousand. Then we tested the extremes — the whole of Chromium, and a made-up two-million-line file where we knew the right answer in advance. The engine got them right. This paper is the receipts.

01 · Why validate

A measurement tool that is confidently wrong is worse than none: it launders a mistake into a citation. Papers C and D make quantitative claims about millions of statements across hundreds of codebases; those claims are only as trustworthy as the instrument that produced them. This paper sets out the evidence that the instrument is sound — and, where it once was not, exactly how far off it was and how that was established.

02 · Value preservation

The strongest property a measurement instrument can have is that a change to the tool never silently changes a result. Two proofs:

- **Study subjects, bit-identical across revisions.** The full set of Paper C subjects, measured on their pinned commit pairs, reproduces byte-for-byte across successive engine builds — every fix in this paper is value-preserving for every input that already worked.
- **Chromium, bit-identical across runs.** The 43.5-million-line 2024 window, re-scanned independently on the released engine, reproduces its full per-file classification exactly. A result that repeats to the byte on the largest input available is a result that does not depend on timing, memory layout, or luck.

03 · Quantified error — the honest part

The corpus campaign was also a stress test. Running the engine over 136 real codebases surfaced a small number of inputs that triggered latent memory-safety defects in twenty-year-

lineage parsing code — the kind of faults that only pathological real files expose. All were diagnosed to their root cause and fixed. The validation-relevant facts:

- **The measured effect on results was one count in 46,146.** The single run known to have completed while traversing a defect differed from the corrected engine by exactly one statement-level count (0.002%); the logical-statement metrics that underlie every published REWORK figure were identical.
- **Degradation is never silent.** Where an input cannot be fully read — for example, archive collisions on case-insensitive filesystems — the engine proceeds only when the tree is substantially complete, counts and names what it could not read, and stamps the run as degraded. A partial measurement can never impersonate a clean one.
- **Verification catches its own residue.** The final correctness check was itself run twice; the second pass caught a one-count classification artifact in the first and corrected it before publication. The discipline is: measure, fix, re-measure, and let the re-measurement find what the first pass missed.

04 · Coverage

A corpus with silent gaps is a biased corpus. After the fixes, every window that had ever failed was re-measured, and all of them now complete — the corpus of 1,437 measurements has no holes attributable to instrument failure. The contract the engine now meets is simple: no input kills a scan, and no scan reports a clean result it did not earn.

05 · Accuracy at scale

Reproducibility proves consistency; it does not prove correctness. For that, the answer must be known in advance. Two scale credentials:

SPECIMEN	SIZE	TIME	PEAK MEMORY	RESULT
Chromium 2024 (many files)	43.5M LOC · 326,948 files	~1h50m	—	bit-identical across runs
SQLite amalgamation (one file)	255,636 lines	10.6s	87 MB	clean under sanitizer

SPECIMEN	SIZE	TIME	PEAK MEMORY	RESULT
Synthetic monster (one file)	2,000,000 lines	14m42s	1.37 GB	**planted churn measured exactly**

The last row is the decisive one. A single two-million-line file — roughly eight times the largest real source file in common circulation — was built with a known number of edits, deletions and additions planted in it. The engine reported those numbers exactly. Chromium was measured on an ordinary 8 GB laptop reading from an external disk, one command; single-file memory scales at roughly 0.7 GB per million lines, placing the practical ceiling far beyond any real source file.

06 · What this earns

Papers C and D rest on measurements. This paper is the ground they rest on: the numbers reproduce to the byte, the instrument's worst measured error was one count in forty-six thousand, its coverage has no silent gaps, and its accuracy holds against known ground truth at two million lines. The engine is **CodeDelta** (codedelta.app); the corpus it validates is the subject of Paper D.