What broke, and why it was almost never the model

Building a verification harness for language-model extraction of radiocarbon data from Russian- and Kazakh-language archaeological literature

Conor Reid

Preprint — August 1, 2026

Abstract

Large tracts of archaeological primary data sit in grey literature that aggregating databases do not ingest, largely because it is not published in English. Language models can now read such material, but the resulting pipelines are usually justified by an accuracy figure rather than by an account of how they fail. This paper reports the opposite emphasis. We built an extraction pipeline for radiocarbon determinations in Russian- and Kazakh-language steppe archaeology, wrapped it in a deterministic verification harness, and kept a complete log of every defect encountered. From 46 publications the pipeline produced 1,491 determinations, of which 1,338 (90%) are absent from p3k14c, XRONOS, CalPal, RADON-B and NERD combined. Nineteen defects were logged. *None* was a confirmed transcription error by the model. Seventeen were defects in our own schema, rules or reference data; the remaining two were properties of the sources. We argue that the dominant risk in this class of pipeline is not fabrication but *silent omission*—pages never selected, records dropped without trace, edits that never applied—because a verification harness is built to check what is present, and an absence trips nothing. We report the three omission defects that inflated our own headline result before we caught them, and recommend that harnesses be pointed at absences and at their own assumptions rather than primarily at the model.

1 Introduction

The radiocarbon record of the Eurasian steppe is unevenly visible. Aggregating databases—p3k14c [1], XRONOS, and the sources bundled by `c14bazAAR` [2]—are assembled predominantly from English-language literature. Determinations published in Russian and Kazakh journals are largely absent from them, a gap noted repeatedly in the regional literature but not, to our knowledge, quantified.

Language models capable of reading scanned Cyrillic tables make this material tractable for the first time. The obvious risk is fabrication, and the obvious response is a verification harness. We built one. This paper is mainly about what the harness got wrong.

Our contribution is threefold. First, a dataset: 1,491 determinations with per-record provenance, 90% of which are in no public compilation. Second, a verification design whose central device—requiring the model to emit every value twice, once parsed and once transcribed—detects transcription error without any ground truth. Third, and we think most useful, a complete defect log. Pipelines of this kind are typically reported through a headline accuracy figure. That figure conceals the failure modes that actually threaten a compilation, and in our case it would have concealed all of them.

2 Materials

The corpus comprises 46 open-access publications from Russian- and Kazakh-language journals (*Археология Казахстана*, *Вестник археологии, антропологии и этнографии*, *Поволжская археология*, *Уфимский археологический вестник*, *Теория и практика археологических исследований*) together with a small set of English-language comparators. All carry usable text layers; contrary to expectation, optical character recognition was not required for any document.

Two properties of the corpus shaped the work. Documents obtained from different outlets frequently duplicate one another, so identical articles were detected by comparing the *set of determinations* in each document—two renderings of one article contain the same dates, two different articles essentially never do. And a large fraction of candidate documents contain no determinations at all: of 612 documents retrieved by topical search in one collection pass, 28 were retained, the dominant failure being papers that cite calibrated calendar ranges in prose and never print the underlying measurement.

3 Method

3.1 Pipeline

Pages carrying date tables, or isotope tables, are extracted by a language model (Claude Opus 4.8) into a constrained schema. Records then pass a suite of deterministic rules; nothing is repaired automatically. Figure 1 gives the structure.

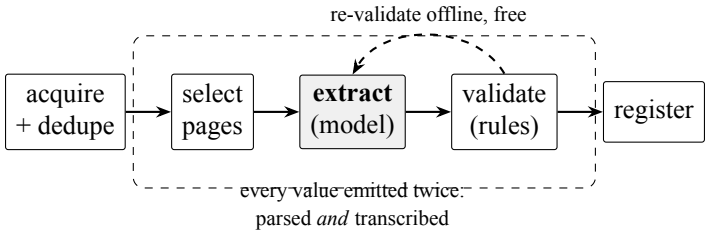

Figure 1: The pipeline. Only one stage involves a model. Extraction is the sole irreversible cost, so raw records are stored before being judged, and the rule suite can be re-run over the whole corpus after any revision at no cost—a property we exercised seven times.

3.2 The central verification device

The schema requires the model to produce every load-bearing value *twice and independently*: once parsed into a typed field, and once transcribed verbatim as printed. A transposed digit makes the two disagree. Detecting that requires no ground truth, no second model and no gold set—only internal consistency—and it scales to the whole corpus for free.

The same principle extends to grounding. Each record carries the source row it was read from; the row must appear on the page it claims, and each isotope value must appear in the row it is attributed to. Value-level grounding proved necessary: a model can transcribe a genuine row correctly and still attach isotope values that were never printed, which passes a row-level check and converts an unscreenable determination into a falsely screened one (defect F1).

3.3 Rules and their status

Rules flag; they never repair. A rule that silently corrects data is a rule that silently corrupts it. Flags are divided into *blocking* (an unresolved question about the record) and *informational*. Several informational flags are findings in their own right: that a source published no error term, no laboratory code, or no isotopes.

4 Results

4.1 Extraction accuracy

On the single page measured against an independent hand transcription—a Kazakh date table of 11 determinations, transcribed by hand and never shown to the model—extraction scored 11/11 on laboratory

code, age, error and material: 100% precision and recall. The model additionally noted, unprompted, that one row's laboratory code was a bare prefix with no number and that its age carried an asterisk marking a non-AMS determination.

This is one page. It is not a corpus accuracy figure and we do not claim one. A gold set of 100–150 hand-transcribed determinations with per-field precision and recall remains outstanding, and no record has yet been verified against page images.

4.2 The rule suite was the unreliable component

On that same page the rule suite blocked all 11 correct records. Every flag was a false positive. The dominant cause was a rule that treated the token cal BC as evidence of BP/calendar confusion—but every properly published date table prints the uncalibrated age beside its calibrated interval, so the rule flagged every well-formed table in the corpus.

The rules had been written against synthetic data we invented ourselves. They encoded our assumptions about what date tables look like rather than what they actually look like. Figure 2 shows the pass rate across successive revisions; every gain came from correcting a rule, none from altering data.

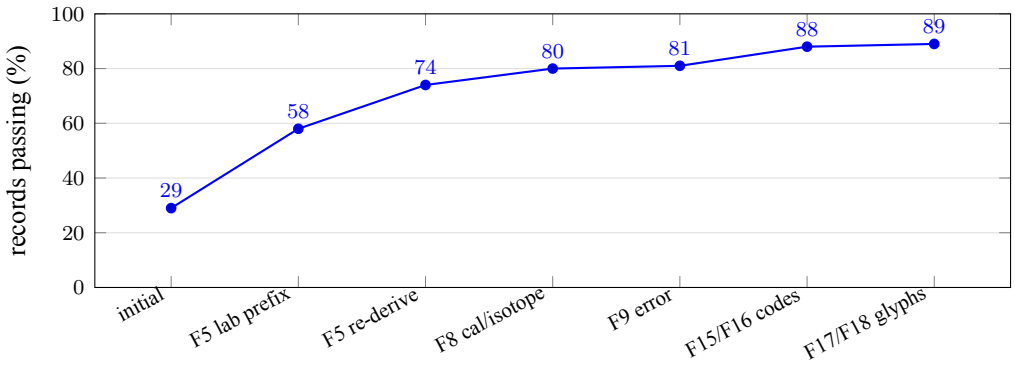

Figure 2: Proportion of records passing the deterministic rule suite across successive revisions of *the rules*. The underlying extractions are unchanged throughout. Each step corrects a false-positive class discovered on real data.

4.3 Coverage relative to public compilations

We matched all 1,491 determinations against p3k14c, XRONOS, CalPal, RADON-B and NERD—217,251 distinct laboratory codes—on a normalised base laboratory code (Table 1).

	determinations	share
Already present upstream	6	0%
Absent from all five compilations	1,338	90%
No usable laboratory code (uncheckable)	147	10%

Table 1: Reconciliation against the standard compilations.

A result this favourable warrants attack before acceptance. Two checks support it. First, the join demonstrably works: it resolves Oxford determinations across databases that write the same laboratory as OxA, OXA and oxa. Second, and more tellingly, the laboratory *numbers* interleave (Figure 3). The same laboratories measured both sets in the same period, the numeric ranges nest inside one another, and the two sets share almost no determination. These are not dates held upstream under a different spelling; they are dates the compilations do not have.

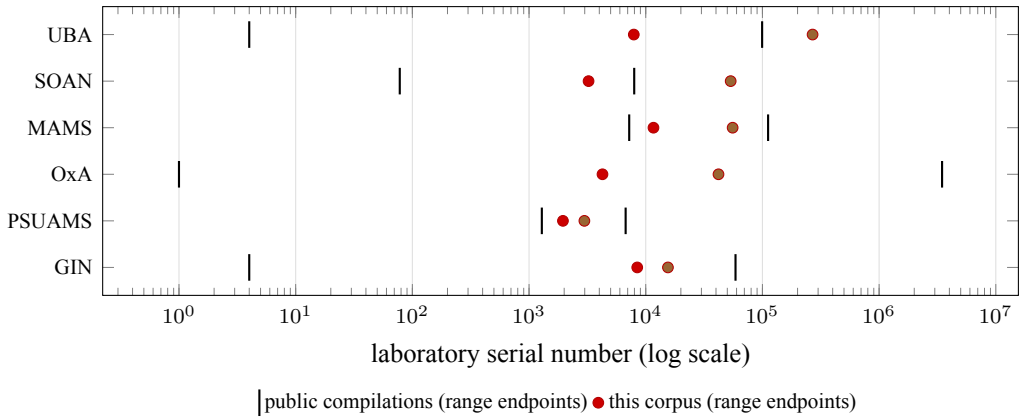

Figure 3: Laboratory serial-number ranges for six laboratories, in the public compilations and in this corpus. The ranges overlap and in several cases nest, yet the sets share almost no individual determination—evidence that the gap is one of coverage rather than of matching.

4.4 Flag distribution

Figure 4 shows the final flag distribution over 1,535 records. The largest category, isotopes absent, is not a defect but the substantive observation that motivated the project: 381 determinations rest on material that can carry a freshwater reservoir offset [3, 4] and were published without the δ^{13}C or δ^{15}N needed to screen them.

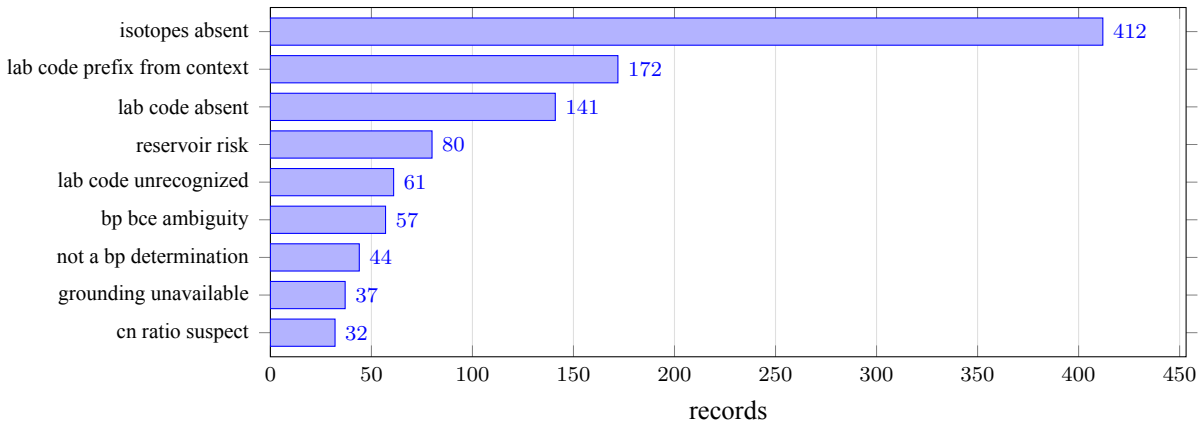

Figure 4: Flag distribution after rule revision. Several categories are properties of the sources rather than defects in extraction.

5 What broke

Nineteen defects were logged. None was a confirmed transcription error by the model. We group them by where the fault lay.

5.1 Omissions: the dominant risk

Three defects share a signature, and we regard this as the paper's central practical finding.

F12 — pages never selected.
Isotope data is frequently published in its own table, on a page containing no determinations at all. The page selector required determinations, so those pages were never extracted, and every date in

the affected documents was counted as unscreenable. The site ranked first in our own register had published its isotopes one page away from its dates.

F13 — records dropped without trace.
Records failing a schema guard were discarded by a bare `continue`. The only symptom was a page reporting zero records where the model had reported seeing twenty-four.

F15 — a partial write that looked clean.
A failed run wrote 1,239 records with an empty report block. Nothing errored; the diagnostic printed "0 blocked", which is indistinguishable from a clean corpus.

A corrupted value trips a rule. A missing page, a dropped record and an empty report block trip nothing, and all three *resemble success*. Verification harnesses are constructed to check what is present; the failures that threaten a compilation are omissions.

The only instrument that detected any of them was a completeness signal we had included almost incidentally: the model's own count of determinations visible on a page, compared against the number it transcribed. That signal is free, needs no ground truth, and is the sole component of the harness capable of noticing an absence. We recommend it as standard.

5.2 Schemas that cannot express absence

Two defects (F9, F10) share a cause: a schema that could not represent "not stated" forced the model to fabricate. One source prints `3590±` with the error term genuinely missing; a required integer field left the model no option but `0`. Separately, Russian prose quotes calendar dates—*«1370 ± 30 г. до н. э.»* is 1370 BC, roughly 3320 BP. With only an uncalibrated-BP field available, the calendar year went there. In both cases the model transcribed faithfully and the correct value was unrepresentable.

5.3 Reference data that silently reattributed dates

Laboratory prefixes collide across case: `LU` is St Petersburg State University, `Lu` is Lund; `Ki` is Kyiv, `KI` is Kiel. Our lookup folded case *and* rewrote the prefix to the registry's spelling, so a Lund determination was quietly converted into a St Petersburg one (F7). This was found only by verifying our registry against the laboratory list maintained by *Radiocarbon* rather than trusting our own assembly. A widely circulated third-party laboratory-code list contains the same error.

A separate normalisation defect (F19) mangled codes printed with a spaced separator: `SOAN -4843` was parsed as prefix `SOA` plus remainder `N -4843`, emitting a string that could never match anything. Upstream compilations contain that form in quantity, so this defect was actively suppressing reconciliation matches.

5.4 The model reads the page; the harness read the text layer

One document embeds a subset font with a broken character map: `COAH-6200` extracts as `Ķóξü-6200`, Cyrillic mapped into Greek and Coptic. The rendered page is legible; only the extracted text is corrupt. The model read it correctly, and our grounding check compared that correct transcription against the corrupted text layer and reported 37 mismatches (F17).

We regard this as direct evidence for vision-based extraction: it succeeded precisely where text extraction fails, and the harness penalised it for being right.

6 Discussion

The premise we began with—that a model reading archaeological tables would require policing, and that building the harness to police it was the substantive work—survives only in part. The harness *is* the substantive work. But across roughly 1,500 determinations the digit-transposition check, the strongest

device available because it needs no ground truth, has found no confirmed transcription error. Every hit traced to a schema that offered the model no honest answer.

The engineering problem is therefore not fabrication. It is encoding real-world publishing convention—which no amount of synthetic testing will teach—and noticing what never arrived. We would state the practical guidance as follows.

1. **When a flag fires on real data, assume the rule is wrong.** On our corpus that null hypothesis was correct in every instance.
2. **Calibrate rules on real documents before trusting them.** Precision measured against errors one has invented oneself measures only self-consistency.
3. **Instrument for absence.** Record what the model reports seeing, not only what it returns; log every dropped record with its reason; never let a partial write produce output that resembles a clean run.
4. **Let the schema express uncertainty.** Any field that cannot encode "not stated" will be filled with something.
5. **Verify reference data against an authority.** Our most dangerous defect was in a lookup table we assembled ourselves.

7 Limitations

No record has been verified against page images; a gold set with per-field accuracy remains outstanding, and every figure here should be read as provisional. Reconciliation matches on laboratory code alone, so "absent upstream" is an upper bound on novelty. The register carries no coordinates, which limits its use in spatial queries and is the most valuable outstanding enhancement. Cultural attributions are recorded exactly as printed and are not harmonised across orthographies; consequently they describe publications rather than cultures. Finally, the reservoir-risk thresholds applied where isotopes *are* available are placeholders, not calibrated science: nitrogen enrichment indicating aquatic protein is baseline-dependent, and a single global cut-off is not defensible. The count of unscreenable determinations does not depend on them.

Data and code availability

All code, data and the complete defect log are at https://git.sr.ht/~calgacus/tegiszhol. The register, an export conforming to the c14bazAAR schema, and per-record provenance are published at https://calgacus.srht.site/tegiszhol/. Source PDFs are not redistributed; a manifest records each document's citation, licence and access route.

Every determination is cited to the publication that produced it. This compilation is a pointer to that work, not a replacement for it.

References

[1] Bird, D., Miranda, L., Vander Linden, M., *et al.* (2022). p3k14c, a synthetic global database of archaeological radiocarbon dates. *Scientific Data* 9, 27.

[2] Schmid, C., Seidensticker, D., Hinz, M. (2019). c14bazAAR: An R package for downloading and preparing C14 dates from different source databases. *Journal of Open Source Software* 4(43), 1914.

[3] Svyatko, S. V., *et al.* (2022). Freshwater Reservoir Effects in Archaeological Contexts of Siberia and the Eurasian Steppe. *Radiocarbon* 64(2), 377–388.

[4] Svyatko, S. V., *et al.* (2017). A lack of freshwater reservoir effects in human radiocarbon dates in the Eneolithic to Iron Age in the Minusinsk Basin. *Archaeological and Anthropological Sciences* 9, 1379–1388.