What 510 Real Contracts Taught Us About Measuring Extraction

Why the metric is the hard part of contract extraction

By Joshua Bolick · Awareness Software Group _Published: [set date at publish time]_

"Our model extracts contract data with 86% accuracy" is not a claim. It is the beginning of one. Accuracy on which fields, scored how, against what reference, and what happened to the fields that are not in the number.

We fine-tuned six open-weight models on 510 real, lawyer-annotated commercial contracts. The modelling was the easy part. Deciding what could honestly be measured took longer, changed what we report, and produced a smaller number than the one we could have printed.

This paper is about that decision, because it is the part of an extraction claim a buyer can actually interrogate.

The problem: a contract label is a span, not a value

The corpus is CUAD, 510 commercial agreements annotated by lawyers. For each field, the annotation is a **span of the contract text** that answers it, not a normalised value.

That distinction sounds academic until you look at how long the spans are. Median reference length across 200 contracts:

Field	Median reference	Field	Median reference
Agreement Date	16 characters	Parties	89 characters
Effective Date	18 characters	Governing Law	146 characters
Document Name	28 characters	Renewal Term	252 characters
		Notice Period	**317 characters**

The left column and the right column are not the same task. Pulling a date out of a header is retrieval of a short value. Reproducing a 317-character notice-period clause verbatim is something no model does reliably, and something no user actually wants.

Why we exclude the interesting-sounding fields

The obvious metric for extraction is exact match: did the model produce the reference string. On the left column that is exactly right. A date is either correct or it is not, and after normalising formats there is no partial credit worth arguing about.

On the right column exact match is worse than useless. A model can locate the notice-period clause perfectly, summarise it accurately, and start its answer three words earlier than the annotator did. Exact match scores that **zero**. Score enough clause fields that way and the headline accuracy is dominated by a metric failure rather than a model failure.

So we score the short value-like fields and **exclude the long clause spans**, and we state the exclusion in the same sentence as the number. The supportable claim is:

> *"On 510 public lawyer-annotated contracts, fine-tuning improved short-field extraction from 71.9% to 86.0%. Clause-span location is a different task and is not measured."*

Not *"86% accurate on contract understanding"*.

This is the uncomfortable part of the paper. Excluding fields makes our number smaller and our claim narrower. A vendor who scored everything with exact match would publish a lower headline; a vendor who scored clauses with a generous overlap metric could publish a higher one. Both would be reporting something other than what a buyer will experience.

The rule we hold to: **if a metric cannot fairly score something, exclude it from the score and say so in the same breath.** A buyer who discovers the exclusion themselves will reasonably discount everything else you have told them.

What would be needed to measure the rest

We are not claiming clause location is unmeasurable. We are claiming exact match does not measure it, and we have not yet built what does.

It needs an overlap metric with a defensible threshold, which means deciding how much of a 317-character clause a model must reproduce before the answer is useful. That threshold is a judgement about the workflow, not a property of the model: a lawyer skimming for the right paragraph and a system auto-populating a database want very different answers.

Until that exists, clause location stays out of the number. "Not measured yet" is an honest state for a capability to be in.

Two defects the real corpus found and our tests did not

We built the data pipeline against a hand-written fixture. Every test passed. Two real defects were sitting in it, and both appeared only when it ran against all 510 contracts.

The first: multi-answer fields were being truncated to the first span. 508 of the 510 contracts label *Parties* with several spans, because a contract usually has more than one party. The code took the first

one. The reference for a two-party agreement became `"Distributor"`, and any model that correctly answered with both parties was marked wrong. The fix is to join all spans, and to treat every field as potentially multi-span rather than special-casing the ones we happened to notice.

The second, and worse: about half of all answer spans sat past the first 12,000 characters of their contract. Contracts are long, model context windows are finite, so the pipeline truncates. But it was still scoring the model against answers that lived in the truncated-away part of the document.

That is not a hard metric. It is a measurement of our own truncation, dressed as a measurement of the model. The model was being marked wrong for failing to read text it was never shown.

The fix keeps per-span offsets so a reference can be rebuilt from only the spans the truncated context actually contains. The subtlety worth recording: checking the *first* span's offset is not enough. A multi-span *Parties* field whose first span is on page one would keep its whole joined reference, including parties named on page forty that the model cannot see.

With the context window we use, roughly **84% of scorable references** survive that filter. We report that figure rather than the more flattering one we first published, which was wrong.

The general lesson: fixtures prove shape, not correctness. A test suite built on hand-written examples confirms the code does what its author expected. It cannot tell you what the real corpus contains. Both defects were invisible to a green test run and obvious within minutes of looking at real output.

What this means when evaluating any extraction vendor

None of the above is specific to us or to CUAD. If you are assessing an extraction claim, these questions separate a measured result from a marketed one:

- **Which fields are in the number, and which are not?** Ask for the excluded list. If

there isn't one, ask how long clause fields are being scored.

- **What is the reference?** A normalised value and a span of source text are different

targets, and only one of them can be exact-matched fairly.

- **Was the model shown the text containing the answer?** Truncation is universal for long

documents. Scoring answers that fall outside the window measures the pipeline.

- **Is a multi-part answer scored as one thing?** Parties, jurisdictions and notice periods

are frequently plural.

- **What does the metric do with a nearly-right answer?** Exact match says zero. Whether

that is correct depends entirely on your workflow.

A vendor who has thought about their metric will have answers and will volunteer the limits. A vendor who has not will describe the accuracy number as if the choice of metric were obvious.

The numbers, with their scope attached

Short value-like field extraction on 510 real commercial contracts, 99 held out, three random seeds per model:

Model	Untuned	Fine-tuned
phi-4	71.9%	**86.0%**
qwen3-8b	69.3%	85.9%
granite-3.3-8b	56.9%	81.5%

Scored fields: document name, agreement date, effective date. Clause spans excluded, for the reasons above. Public SEC-filed contracts, not any client's paper: different house style, formats and clause vocabulary, so this measures the method rather than predicting a result on a corpus we have not seen.

Awareness Software Group builds private, fine-tuned models that clients own and run on their own infrastructure.
