

A model is a claim about the future. Here is how ours earn the right to make one.

Seven stages, none optional, none waived. This is the same pipeline description we walk through in diligence — nothing here is simplified for marketing, because the pipeline is the product.

SERIES

Technical Note TN-002

VERSION

1.0

UPDATED

9 October 2026

STATUS

Published

AUTHORS

Criteria Intelligence Research

Corresponding author: Jack Cosentino



Real filed records in. Nothing synthetic. Ever.

Every training example begins as a real, filed court record — a docket that exists, in a court that exists, with an outcome that happened. Records are deduplicated on a composite source key and carry their provenance through the entire pipeline. Zero synthetic augmentation in production training: if the data does not exist, we do not manufacture it — we wait, or we do not build the model.

REAL FILED RECORDS · DEDUPLICATED · PROVENANCE-KEYED



Features from the structure of litigation itself.

Jurisdiction, case type, filing posture, party configuration, judicial assignment, procedural history, timing. The feature set is engineered from how litigation actually unfolds — not from text embeddings of what documents happen to say. Structured features are inspectable: you can ask why a prediction moved, and get an answer in litigation terms.

JURISDICTION · CASE TYPE · ASSIGNMENT · POSTURE · TIMING

One model per jurisdiction and case type. Not one model for everything.

A custody outcome in one state and a securities dispute in a federal district are different statistical worlds. We train them separately — thousands of narrow, jurisdiction-specific models rather than one generalist classifier. Narrow models capture the venue-level variance that pooled national models average away, and that variance is exactly where decisions are won.

JURISDICTION × CASE TYPE × TARGET = ONE MODEL

Tested on the future, not on the past.

Every candidate model is evaluated on temporally held-out cases — real outcomes from a later period the model never saw during training. A random shuffle can let a model quietly memorize its era; a temporal split cannot. If the world drifted and the model did not, the holdout catches it before production does.

TRAIN: EARLIER FILINGS → TEST: LATER OUTCOMES, NEVER SEEN

Promotion gates with no exceptions.

A model enters production only by clearing discrimination and calibration review on its holdout — every model is cleared individually against a published standard. Models that fail stay experimental or stub, visibly, in the registry. The gate is why the production count means something: only models that earned the label reach client work, not every model that exists.

TEMPORAL HOLDOUT · CALIBRATION REVIEWED · CLEARED MODEL BY MODEL

Versioned, audited, and honest about failure.

Every model lives in a registry with its training window, evaluation results, version history, and current status — production, experimental, stub, or failed. When later review finds a problem, the model is demoted and the registry says so. A registry that never demotes anything is not governance; it is marketing.

PRODUCTION · EXPERIMENTAL · STUB · FAILED — EVERY STATUS RECORDED

The loop closes on real outcomes.

Courts keep deciding cases, and every new resolved outcome is new evaluation data.

Production models are retrained on cadence against the growing corpus and re-cleared through the same gates. The system is built to be wrong in public and corrected on schedule — which is the only honest way to run prediction infrastructure.

NEW OUTCOMES → RE-EVALUATION → RE-CLEARED THROUGH GATES

WHY TEMPORAL HOLDOUT MATTERS

HOLDOUT — LATER OUTCOMES, NEVER SEEN

A random train/test split lets a model be graded by its contemporaries — cases from the same period, the same judicial climate, sometimes the same litigation wave. It flatters. A temporal split grades the model the way production will: on cases decided after everything it learned from. If performance survives the cut, it

is evidence. If it does not, the model never ships. Every production model in the registry crossed this line before it crossed into production.

THE GATE

REGISTRY ENTRIES — every model ever attempted

ALL

CLEARED PROMOTION GATES — production status

PRODUCTION

DID NOT CLEAR — experimental, stub, or failed, and labeled that way

HELD BACK

Entries that never earned production status stay in the registry and are labeled that way. A gate that never rejects anything is not a gate.

One model means one jurisdiction, one case type, one prediction target — a deliberately narrow definition, because narrow is what makes a probability mean something in a specific courtroom. Production status is not a marketing label for features or endpoints. It means a narrow model cleared a temporal-holdout gate and calibration review, individually.

Calibrated models vs. generative AI

Why prediction infrastructure cannot be an LLM — and where LLMs genuinely help.

READ →

Our training data

Provenance, deduplication, and what we deliberately exclude.

READ →

The registry

Production counts, gates, and the changelog — the honest denominator.

READ →

CITE THIS DOCUMENT

Criteria Intelligence Research (2026). How We Build Models. Technical Note TN-002, version 1.0.

Criteria. <https://critericaintelligence.com/how-we-build-models>

```
@techreport{criterica_tn002_2026,  
  author = {{Criteria Intelligence Research} and Cosentino, Jack},  
  title = {How We Build Models},  
  institution = {Criteria},  
  type = {Technical Note},  
  number = {TN-002},  
  year = {2026},  
  month = oct,  
  note = {Version 1.0},  
  url = {https://critericaintelligence.com/how-we-build-models}  
}
```

DOI: pending

