Criterion-Level Classification in US Immigration Law: An NLP Pipeline for EB-1A Appellate Decisions

Anonymous ACL submission

Abstract

EB-1A extraordinary ability visa petitions require satisfying at least three of ten regulatory criteria at 8 C.F.R. § 204.5(h)(3). We present an NLP system trained on 4,660 AAO appellate decisions that fine-tunes DistilBERT for criterion-level evidence assessment and screens novel evidence through a structured evidence-evaluation workflow. *Parser quality is the dominant performance variable*: CFR-header anchoring recovers criteria previously invisible in modern-format decisions, raising criteria per file by 10.4%. On a held-out test set of 29 files from 2024–2026, criterion classification achieves 95.0% recall, 82.6% precision, and 88.4% F1 with 100% case-level accuracy. A same-test-set comparison shows the 66M-parameter classifier *substantially outperforms* three frontier LLM baselines (F1 ≈37%), confirming labeled AAO supervision captures criterion-recognition patterns that zero-shot prompting cannot. A sensitivity probe reveals criterion-specific linguistic markers the model has internalized from AAO decisions. The system targets triage: identifying which criteria to anchor before filing; local deployment preserves case confidentiality without frontier API costs, retaining practical advantages even as LLMs improve.

1 Introduction

The EB-1A *extraordinary ability* immigrant visa is one of the most demanding categories of U.S. employment-based immigration. The statute reserves the classification for foreign nationals who have "risen to the very top" of their field in the sciences, arts, education, business, or athletics, demonstrated through sustained national or international acclaim (Legal Information Institute, 2026; U.S. Citizenship and Immigration Services, 2025). Because there is no closed list of qualifying achievements, petitioners must assemble a record satisfying at least three of ten evidentiary criteria at 8 C.F.R. § 204.5(h)(3) – spanning major awards, membership in selective associations, published material about the petitioner, judging the work of others, original contributions of major significance, scholarly articles, high salary, and others. Assembling and arguing this record typically requires a domain expert to read and weigh heterogeneous evidence (publications, citation data, organizational charts, salary surveys, recommendation letters) against specific regulatory language. The cost and friction of this process create real pressure for earlier, cheaper signal about petition strength – both for immigration law firms triaging prospective clients and for self-represented petitioners deciding whether and how to file.

The Administrative Appeals Office (AAO) is the appellate body within U.S. Citizenship and Immigration Services (USCIS) that reviews adverse decisions on immigrant petitions. USCIS publishes AAO decisions only when a petition has been denied at the initial service-center stage *and* the petitioner has appealed; approved petitions are not publicly written up, so the AAO corpus is a selected, denial-heavy slice of all EB-1A filings rather than a representative sample. It is nevertheless the largest public source of detailed criterion-by-criterion reasoning, covering tens of thousands of petitions and providing enough labeled signal to train a classifier that partially automates the screening step. In practice, however, building such a system on raw AAO PDFs is a serious engineering problem before it is a modeling problem. Documents differ substantially by year, PDF text-layer quality, layout, and wording conventions. Outcome labels must be inferred from legal reasoning that includes reversals, hedges, and negation patterns that naive extraction cannot handle.

This paper presents the engineering and modeling lessons from building this system over three development iterations. We make four contributions:

1. **A reproducible AAO PDF-to-criterion pipeline.** We construct a three-generation parser that extracts criterion-level analysis blocks, evidence summaries, and binary outcomes from 4,660 AAO appellate decisions. CFR-header anchoring corrects systematic failures caused by PDF whitespace corruption and recovers criterion sections previously invisible in modern-format decisions.

2. **A criterion-level EB-1A classification benchmark.** We formulate EB-1A screening as three binary classification tasks and evaluate models using positive-sensitive metrics appropriate for denial-heavy administrative corpora.

3. **Evidence-quality sensitivity analysis.** A controlled probe reveals the model has internalized AAO-specific linguistic markers: citation counts and adoption language are near-sufficient triggers for criterion v, while criterion iv produces a counterintuitive inversion where conference-paper review outscores NeurIPS Area Chair.

4. **Frontier LLM comparison.** The 66M-parameter task-specific classifier substantially outperforms three frontier LLM baselines (GPT-5.5, Claude Sonnet 4.6, Gemini 3 Pro Preview) on the same held-out test set: 88.4% F1 versus ≈37% F1, and 100% versus 85.7% case-level accuracy. The gap demonstrates that labeled AAO supervision captures criterion-recognition patterns that zero-shot prompting cannot provide; we further identify where frontier LLMs complement rather than compete with the task-specific classifier.

The design goal is explicitly triage, not adjudication. A petition attorney must eventually submit to a USCIS officer who applies the full two-step Kazarian review; our system targets the earlier and high-friction step of identifying *which criteria* to anchor a petition on before investing months in evidence preparation. Even imperfect recall at the criterion level: 95% for v3-Scaled provides actionable early signal at this stage. The system's role is to partition criteria into strong anchors, borderline cases needing more documentation, and weak candidates to deprioritize, not to render a final approve/deny decision. The appropriate performance question is not whether the system matches a USCIS adjudicator but whether it sorts criteria more reliably than an unaided practitioner reading the same evidence.

The sensitivity analysis further provides a practical guide for practitioners: writing evidence in the linguistic register of AAO-granted decisions significantly outperforms vague prestige claims. The main bottleneck is not model capacity; the LLM comparison confirms this but the interaction between noisy administrative PDFs, parser design, and the specificity of labeled supervision: an applied lesson with broader relevance to legal NLP systems built on imperfect public records (Medvedeva and Mcbride, 2023).

2 Task, Domain, and Related Work

2.1 EB-1A Criterion Structure

The EB-1A framework requires satisfying at least three of ten criteria unless the petitioner holds a single internationally recognized major award (Legal Information Institute, 2026; U.S. Citizenship and Immigration Services, 2025). USCIS adjudication follows a two-step framework established by *Kazarian v. USCIS*, 596 F.3d 1115 (9th Cir. 2010). Step one is a quantitative count: qualifying documentation under at least three criteria at 8 C.F.R. § 204.5(h)(3). Step two is a holistic final-merits determination assessing whether the totality of evidence places the petitioner at the top of the field (U.S. Citizenship and Immigration Services, 2025). Our system models step-one criterion-level findings only.

2.2 Related Work

Legal language models and benchmarks: Chalkidis et al. (2020) showed domain-adapted pre-training provides consistent gains on statutory entailment and contract classification; subsequent work established standardized benchmarks (Chalkidis et al., 2022; Niklaus et al., 2023). We replicate their key finding: domain adaptation provides modest gains but does not compensate for limited or noisy supervision at current training scale.

Legal judgment prediction: Our task is closest to LJP, predicting appellate outcomes from case facts (Katz et al., 2017). Recent work applies event extraction (Feng et al., 2022), contrastive relations (Liu et al., 2022), and multi-task classification (Li et al., 2025; Liu et al., 2025). Our task differs: the target is criterion-level (ten independent binary questions per case), and the corpus must be constructed from operational PDFs. We adopt

the evaluation recommendation of Medvedeva and Mcbride (2023): LJP must use positive-sensitive metrics on imbalanced corpora.

Multi-label legal classification: The EB-1A threshold rule creates a multi-label setting with criterion-level label imbalance. Chalkidis et al. (2019) studied extreme multi-label classification on EU legislation with sparse positive labels which is exactly our challenge.

Argument mining and evidence assessment: Stab and Gurevych (2014) developed models for argumentative structures; Walker et al. (2018) address how conflicting evidence is weighed under legal rules. Both model the gap between evidence being *present* and being *accepted*: the failure mode we observe in high-confidence false positives.

Immigration NLP: To our knowledge, EB-1A criterion-level classification from appellate text has not been studied computationally.

3 System Architecture

3.1 Pipeline Overview

Five stages (Figure 1, Appendix E): (1) **PDF ingestion**: AAO decisions downloaded from the USCIS website; (2) **Parsing**: each PDF converted to structured JSON with criterion analysis blocks, evidence summaries, and binary criterion outcomes; (3) **Task generation**: each decision produces up to three training examples per criterion; (4) **Training**: a DistilBERT classifier fine-tuned with task-prefix conditioning; (5) **Inference**: a YAML evidence packet scored with configurable aggregation. Practitioners interact only with stage 5.

3.2 Parser Design

The parser is the system's highest-leverage component.

v1 (content regex): Narrow regular expressions against criterion language produced fragmented evidence strings, missing criterion IDs, and unknown outcome labels for 53.2% of criterion records.

v2 (structured AAO-style): Stable criterion IDs, broader outcome detection, selective evidence extraction. Unknown outcome rate fell to 21.5%.

v3 (CFR-header anchor): The key engineering insight was that PyPDF2 drops spaces at line boundaries, producing strings like `"oflesser"` and `"membership in\nassociations"` that broke content-based regex for criteria i–iii. CFR citation strings: compact, structured, column-break-resistant survive this mangling. Replacing content regex with the pattern `"regulation at 8 C.F.R. § 204.5(h)(3)(id)"` as the section boundary detector recovered criteria i/ii/iii entirely from modern-format (2022+) decisions. Result: average criteria per file rose 1.891→2.087 (+10.4%), the corpus grew by 116 files, and a systematic label contamination affecting modern-format decisions was eliminated.

3.3 Task Formulation

Three classification task types per parsed decision:

- `[CRITERION] Based on the following analysis, was criterion 8 C.F.R. § 204.5(h)(3)(id) met? text`
- `[OVERALL]` Input: concatenated criterion summaries; label: petition approval ($\geq$3 criteria met).
- `[EVIDENCE] Does this evidence satisfy criterion 8 C.F.R. § 204.5(h)(3)(id)? evidence`

All three tasks share one binary classification head. Task prefixes provide explicit supervision signal; our ablation (§6) shows they primarily help find positive criteria.

3.4 Inference and Aggregation

During screening, each evidence item is scored using the `[EVIDENCE]` format. Items above a positive-confidence threshold count as qualified positives. Two aggregation policies: **MajVote** (criterion met if positive count exceeds half of all items) and **Any-Pos** (at least one positive). In our sweep, Any-Pos adds 9 predicted positives and zero additional true positives over MajVote at threshold 0.65 (Appendix B).

4 Data

4.1 Corpus

AAO decisions downloaded from the USCIS website; 20% of 2024+ decisions routed to a held-out test set at download time using a fixed random seed. Final counts: 4,660 training decisions; 29 held-out decisions from 2024–2026, never used during training or hyperparameter selection.

Table 1: Parser quality across three generations. v3 uses CFR-header anchoring to recover criteria i/ii/iii from 2022+ decisions. The rise in unknown outcome rate reflects newly extracted ambiguous criteria.

Metric	v1	v2	v3
Files	4,542	4,544	4,660
Files with criteria	3,125	3,126	3,339
Avg criteria / file	1.892	1.891	2.087
Unknown outcome rate	53.2%	21.5%	24.4%
Short evidence items	2,281	7	7

4.2 Parser Quality

Table 1 compares quality metrics across parser generations. The unknown outcome rate fell from 53.2% (v1) to 21.5% (v2), then rose modestly to 24.4% in v3, not a regression, but a reflection of newly extracted criteria for reserved or ambiguous decisions. Average criteria per file rose by 10.4% as i/ii/iii became extractable in modern-format decisions.

A residual parser failure mode: a March 2026 case was labeled criterion viii as met because the parser latched onto an earlier favorable statement, while the AAO later reversed explicitly. This motivated two workflow rules: extracted criterion booleans should be validated against the criterion text, and contradiction-flagged cases require manual normalization before training.

5 Experimental Setup

5.1 Models

The primary model uses `distilbert-base-uncased` (Devlin et al., 2019) as a sequence classifier. The LEGAL ablation substitutes `legal-bert-base-uncased` (Chalkidis et al., 2020). Evaluation uses MajVote aggregation at 0.65 confidence threshold unless otherwise noted. Full hyperparameters for all seven variants are in Appendix A.

5.2 Evaluation Modes

Evidence-item linking (training benchmark): the model scores each extracted evidence item; items linked to a criterion by 80% token overlap count toward its score. Applied to 25 randomly sampled cases.

Direct criterion classification (held-out evaluation): the model receives the full criterion analysis text in the `[CRITERION]` format used during training. Applied to the 29-file held-out set (21 of 29 evaluable; 8 lack labeled criterion text).

Confidence intervals. Given the small held-out set (52 criterion examples), we report bootstrap 95% confidence intervals (10,000 resamples with replacement at the criterion-example level) alongside point estimates. Decision accuracy CIs are computed at the case level ($n = 21$).

Hardware. All runs used a single Apple M4 Pro laptop (12-core CPU, 24 GB unified memory, PyTorch MPS backend). Per-variant wall-clock and TFLOPs breakdown in Figure 2, Appendix F.

5.3 Frontier LLM Baselines

We evaluate three frontier LLMs on the same held-out direct-criterion-classification setting used for v3-Scaled. Each input is a criterion identifier concatenated with the full AAO criterion-analysis text (truncated to 2,000 characters, matching the training input format). Models are prompted with a fixed system instruction to return only `{"criterion_met": true/false}`, indicating whether the AAO's written analysis shows the criterion was found satisfied. Critically, the prompt asks for *extraction of the AAO's conclusion* from the decision text, not independent legal judgment. All outputs are parsed by the same evaluation script; zero parse errors were recorded for any model.

5.4 Benchmark Composition

The v2 training benchmark contains 39 criterion examples with 4 actual positives (10.3%). The v3 benchmark has 40 examples with 7 positives (17.5%). The held-out direct-mode evaluation has 52 examples with 20 positives (38.5%), more balanced because the 2024–2026 test set draws from modern decisions that include more criterion types.

6 Results and Operational Findings

6.1 v2 Corpus: Failure Mode Taxonomy

Five configurations on the v2 benchmark (Table 2, upper block) reveal a spectrum of failure modes. **Baseline** achieves 89.7% accuracy while predicting zero positive criteria: trivially correct on denial-heavy data, operationally useless. **Scaled** lowers accuracy to 61.5% while recovering 3 of 4 positive criteria (75% recall) clearly preferable for triage. **Legal** matches Scaled recall using a domain-adapted encoder, providing no measurable gain at current training scale. **Weighted** (8.75× class weight) is the mirror-image failure: 100% recall, 10.8% precision. −**Prefix** (task prefixes disabled)

drops recall from 75% to 50%; prefixes primarily help find positive criteria.

6.2 v3 Corpus: Improved Data, Shifted Boundary

Retraining on v3 (Table 2, lower block) shows how data quality changes propagate non-trivially. v3-Scaled recovers the same 3 TPs as v2-Scaled, but the benchmark now has 7 positives instead of 4, so recall is 42.9%. The model predicts 24 of 40 criteria positive (21 FPs), more aggressively than v2-Scaled. Better data improved coverage but destabilized the operating point. The v3 threshold sweep (Appendix C) shows no threshold separates TPs from FPs: TP confidences span 0.758–0.991, FP confidences 0.713–0.984.

6.3 Held-out Test Set: Strongest Result

Direct criterion classification on the 29-file held-out set (Table 3) produces the paper's strongest result. v3-Scaled across 52 criterion examples: accuracy 90.4%, precision 82.6%, recall **95.0%**, F1 **88.4%**, decision accuracy **100.0%** (21/21 cases). Bootstrap 95% confidence intervals (10,000 resamples at the criterion level) are in Table 3; the wide precision interval [65.4, 96.0] reflects the small positive count ($n = 20$). The large gap from the training benchmark (v3-Scaled recall = 42.9%) reflects the evaluation mode difference: the training benchmark uses evidence-item linking, while the held-out result uses direct criterion classification which is the model's primary training task.

6.4 Frontier LLM Comparison

Table 4 reports results for three frontier LLMs on the same 29-file held-out test set. The result is unambiguous: all three LLMs achieve approximately 25% recall and F1 ≈37%, compared to 95.0% recall and 88.4% F1 for v3-Scaled, a gap of more than 50 F1 points.

The failure pattern is characteristic of over-conservative extraction. Each LLM produces TP=5, FP≤3, FN=15: they correctly identify only 5 of 20 criteria that the AAO found met, missing 15 entirely. All responses were clean JSON (zero parse errors), ruling out format issues: the LLMs understood the task and returned predictions i.e. they simply predicted negative for criteria the AAO granted. Case-level decision accuracy is 85.7% (18/21) for all three LLMs versus 100.0% for v3-Scaled.

The performance gap reflects two compounding mechanisms. First, *task–goal misalignment*: the prompt instructs LLMs to extract the AAO's binary conclusion from appellate text, but all three models behave as though independently adjudicating the criterion. They predict positive only when the text contains explicit grant language ("the petitioner has established…"), missing the 15 criteria where the AAO's conclusion is embedded in hedged concessions, implicit acceptance, or the specific reasoning constructions of administrative appellate prose. Crucially, all models had access to identical text and produced zero parse errors; the gap is not an information problem but a pattern-recognition problem specific to this corpus. Second, *over-conservative legal bias*: FN=15 with FP≤3 across every model is the signature of systematic false-negative bias, not random error. LLMs with RLHF-style alignment may suppress apparent endorsements of contentious legal conclusions: exactly the wrong behavior for a task requiring faithful extraction of a third party's prior decision. The fine-tuned DistilBERT, trained on thousands of labeled criterion–text pairs from the same document type, has internalized the linguistic register of granted AAO decisions through supervision; zero-shot prompting at frontier scale cannot replicate that. This extends a known finding in legal NLP (Chalkidis et al., 2020): fine-tuning on domain labels outperforms general-purpose models for structured legal classification: our result shows the advantage holds even against frontier-scale generative LLMs.

Complementary roles for frontier LLMs. Despite underperforming on criterion-conclusion *extraction*, frontier LLMs have a complementary role in the pipeline. The case study (§8) shows an LLM and v3-Scaled agree on all six criteria under the `[EVIDENCE]` format, evaluating *novel* evidence for a new petitioner because evidence assessment requires broad legal reasoning about whether new facts plausibly satisfy a criterion, not pattern-matching over dense appellate prose. This evidence-screening role is precisely where broad pre-training confers an advantage. LLMs also provide a label quality assurance function: prompting a frontier model to flag criterion blocks where its reading of the outcome contradicts the parser-inferred label can surface noisy training examples without requiring expert annotation. Finally, few-shot prompting with a small set of labeled AAO

Table 2: Main model comparison (MajVote, 0.65 threshold). Upper block: v2 corpus (39 criterion examples, 4 actual positives). Lower block: v3 corpus (40 criterion examples, 7 actual positives). Dec. Acc = overall petition decision accuracy.

Model	Ex.	Pos	Pred+	TP	Acc.	Prec.	Rec.	F1	Case Acc.
Baseline	39	4	0	0	89.7%	—	0.0%	—	100.0%
−Prefix	39	4	7	2	64.1%	14.3%	50.0%	22.2%	96.0%
Scaled	39	4	17	3	61.5%	17.6%	75.0%	28.6%	100.0%
Legal	39	4	18	3	59.0%	16.7%	75.0%	27.3%	96.0%
Weighted	39	4	37	4	15.4%	10.8%	100.0%	19.5%	84.0%
v3-Baseline	40	7	2	1	82.5%	50.0%	14.3%	22.2%	96.0%
v3-Scaled	40	7	24	3	37.5%	12.5%	42.9%	19.4%	84.0%

Table 3: Held-out test set (29 files, 2024–2026, never seen during training). Direct criterion classification; 21 of 29 evaluable. Per-criterion breakdown in Appendix D.

Metric	v3-Scaled	v3-Base	95% CI
Criterion ex.	52	52	
Actual pos.	20	20	
Predicted pos.	23	34	
TP / FP / FN	19 / 4 / 1	18 / 16 / 2	
Accuracy	**90.4%**	65.4%	[80.8, 98.1]
Precision	**82.6%**	52.9%	[65.4, 96.0]
Recall	**95.0%**	90.0%	[83.3, 100.0]
F1	**88.4%**	66.7%	[76.2, 97.4]
Dec. Acc	**100.0%**	81.0%	exact†

†All 21 cases correct; bootstrap degenerate.
CI: 10,000 resamples, criterion-level.

examples, or fine-tuning a frontier model on the labeled criterion–text corpus, would likely substantially close the criterion-extraction gap, a direction we leave for future work. The practical implication is a division of labor: the fine-tuned classifier for historical-decision extraction; frontier LLMs for novel evidence evaluation, user-facing explanation, and training-data QA.

Durable advantages of the task-specific model. Even if fine-tuned or few-shot LLMs eventually close the criterion-extraction gap, the local task-specific model retains three deployment advantages. First, immigration evidence packets contain sensitive personal information; on-device inference avoids transmitting case data to third-party APIs, a real constraint for practitioner adoption. Second, per-case screening requires many inference calls (per criterion × per evidence item); frontier API costs at this volume are non-trivial, whereas a locally deployed 66M encoder is nearly free. Third, the encoder architecture is intrinsically better matched to sequence classification than a decoder: it is trained directly on the binary supervision signal and produces calibrated ordinal confidence scores that the sensitivity analysis and triage workflow rely on, requiring no additional alignment step.

6.5 Per-Criterion Behavior

In the training benchmark, positive signal concentrates in criterion iii (F1 = 40%) and criterion iv (F1 = 57.1%); all other criteria report F1 = 0%, partly because the benchmark contains no positive examples for them. The held-out direct evaluation (Appendix D) shows the model correctly classifies positives across criteria i–viii when criterion text is available; evidence-item extraction quality, not model generalization, is the bottleneck for most criteria.

7 Evidence Quality Sensitivity

Figure 3 traces model confidence as evidence descriptions are enriched across five quality levels for four criteria, revealing which linguistic markers the model has internalized from AAO decisions. Full probe evidence strings and scores are in Appendix G.

Criterion vi (scholarly articles) saturates early: "15 peer-reviewed articles" alone reaches 94.6%; adding venue names, citation counts, or h-index provides only 3–6 pp of additional signal. USCIS grants criterion vi readily once scholarly authorship is established, and the model has learned this.

Criterion v (original contributions) shows a V-shape: confidence drops from 66.7% on vague phrasing to 22.4% when a specific method is named without context, then recovers strongly once adoption language is added (81.2%) and peaks with citation counts (94.7%). The model rewards the language AAO officers cite in granted decisions.

Criterion i (awards) reveals a model limitation: naming the NSF CAREER Award alone drops con-

Table 4: Frontier LLM comparison on the 29-file 2024–2026 held-out test set (52 criterion examples, 20 actual positives). All models received the same criterion-analysis text and were prompted to extract the AAO's conclusion, not to adjudicate independently. Zero parse errors for all models. Parameter counts for closed commercial models are not publicly disclosed.

Model	Params	Acc.	Prec.	Rec.	F1	Case Acc.
v3-Scaled DistilBERT	66M	**90.4%**	**82.6%**	**95.0%**	**88.4%**	**100.0%**
Claude Sonnet 4.6	closed	67.3%	71.4%	25.0%	37.0%	85.7%
GPT-5.5	closed	65.4%	62.5%	25.0%	35.7%	85.7%
Gemini 3 Pro Preview	closed	67.3%	71.4%	25.0%	37.0%	85.7%

fidence to 14.6%, recovering to 63.6% only when "elected IEEE Fellow" is added. NSF CAREER appears in both granted and denied criterion i cases in the training corpus; "elected IEEE Fellow" appears almost exclusively in granted cases, making it a near-deterministic positive trigger.

Criterion iv (judging) produces the most counterintuitive result: conference paper review (77.9%) outscores NeurIPS Area Chair (14.4%) and a $15M NSF panel chair (17.7%). AAO frequently denies criterion iv for high-prestige review roles, treating them as ordinary academic duties; the model has absorbed this case law. Practitioners should treat criterion iv scores cautiously.

8 Case Study: End-to-End Screening Workflow

We evaluated a real evidence packet against two model versions to demonstrate the workflow and measure the practical impact of the v3 corpus. The packet covers six criteria (ii, iv, v, vi, viii, ix) with 18 total evidence items.

v2-Scaled (MajVote + 0.65) identifies three criteria as met: iv, v, vi clearing the three-criterion regulatory threshold with zero margin. Table 5 (upper block) shows the verdict is sensitive to operating settings: under AnyPos aggregation, viii and ix flip to met; at threshold 0.55, ii also flips.

v3-Scaled (MajVote + 0.65) recovers five criteria: iv, v, vi remain met; ii (IEEE Senior Member, 71.5% confidence) and ix (90th-percentile salary, 77.8% avg confidence) are newly recovered. Criterion viii remains NOT MET (1 qualified positive, not a majority).

The case study shows the system partitions evidence into three tiers: strong anchors (iv, v, vi, consistent across all settings), recovering borderlines (ii, ix, sensitive to corpus and aggregation), and weak criteria requiring additional documentation (viii). This mirrors the soft-rule evidence framework of Walker et al. (2018).

Table 5: Case study. Upper: v2-Scaled sensitivity to aggregation policy and threshold. Lower: v3-Scaled vs. v2-Scaled at MajVote + 0.65 showing v3 corpus recovers criteria ii and ix.

Criterion	MajV+0.65	Any+0.65	Any+0.55
ii (membership)	NOT	NOT	**MET**
iv (judging)	**MET**	**MET**	**MET**
v (orig. contr.)	**MET**	**MET**	**MET**
vi (scholarly)	**MET**	**MET**	**MET**
viii (leading)	NOT	**MET**	**MET**
ix (salary)	NOT	**MET**	**MET**
Overall	**3/6**	**5/6**	**6/6**

Criterion	v3-Scaled	v2-Scaled	Δ
ii	**MET** (71.5%)	NOT	+1
iv	**MET** (88.8%)	**MET**	—
v	**MET** (86.4%)	**MET**	—
vi	**MET** (93.5%)	**MET**	—
viii	NOT (83.1%)	NOT	—
ix	**MET** (77.8%)	NOT	+1
Overall	**5/6**	**3/6**	+2

Note on cross-task LLM behavior. We also scored this evidence packet through a commercial LLM using the `[EVIDENCE]` format (evaluating *novel* evidence for a new petitioner, distinct from the `[CRITERION]` extraction task in §6.4). The LLM agreed with v3-Scaled on all six criteria and independently flagged criterion viii as weak. This consistency, a 66M local model and a frontier LLM agreeing on a real case provides external validation of the screening output, while the held-out comparison (§6.4) reveals that LLMs struggle to extract historical AAO conclusions from appellate prose.

9 Error Analysis

9.1 False Positive Distribution

For v2-Scaled (MajVote, 0.65 threshold): 3 TP, 1 FN, 14 FP, 21 TN. The 14 false positives span 7 criteria; 9 of 14 (64%) carry average positive confidence above 80% (Table 6), indicating systematic overconfidence on incorrect predictions.

Table 6: False positive distribution for v2-Scaled. 64% of FPs carry average confidence >80%, indicating calibration failure rather than borderline predictions.

Criterion	FPs	Avg conf.
iii (press coverage)	3	81.3%
vi (scholarly)	3	86.8%
viii (critical role)	2	86.9%
iv (judging)	2	86.4%
ix (high salary)	2	78.4%
ii (membership)	1	91.0%
i (prizes)	1	71.5%

9.2 Failure Mode Analysis

Surface pattern confusion. The model learns that criterion-specific vocabulary predicts positives but cannot reliably distinguish evidence being *discussed* in an adverse decision from evidence being *accepted*. This explains high-confidence FPs on criteria where the same vocabulary appears in both positive and negative decisions.

Calibration failure. FP and TP confidence distributions overlap across the full 0.65–0.99 range (Figure 4, Appendix H); no threshold cleanly separates them. This requires post-hoc calibration (e.g., Platt scaling), not threshold adjustment.

Sparse positive supervision. The sole FN (criterion iv) has five evidence items, none above threshold under any aggregation policy. The fix is more positive training examples, not pipeline changes.

10 Limitations

Corpus bias: AAO decisions overrepresent denied and contested cases, likely biasing the model toward adverse reasoning patterns.

Parser-generated labels: Criterion outcomes are inferred from text via heuristics rather than expert annotation. Failure modes include criterion blocks where the AAO's conclusion falls outside the analysis window, and procedural dispositions where content-based detectors may fire incorrectly.

Calibration: The high-confidence FP pattern is a deployment risk; scores should be calibrated and presented as ordinal rankings rather than probabilities.

LLM comparison scope: The frontier-LLM comparison evaluates criterion-conclusion extraction only; results do not generalize to independent legal judgment.

Appropriate use: The system should not produce final approve/deny decisions or replace legal review; outputs are a candidate list for expert examination.

11 Ethics Statement

The system is a research prototype for evidence triage and screening support, not legal adjudication. Training data are drawn from publicly available AAO appellate decisions; no personally identifiable information beyond what appears in redacted public decisions is processed. Real-world deployment requires confidence calibration, validation on representative cases, and mandatory human oversight by qualified legal practitioners. Class-weighting experiments show a single hyperparameter can flip the model between always-deny and always-approve, both failure modes are harmful in legal screening.

12 Conclusion

We presented an end-to-end criterion-level EB-1A screening system trained on 4,660 AAO appellate decisions. On a 29-file held-out test set of 2024–2026 decisions, direct criterion classification achieves 95.0% recall, 82.6% precision, and 88.4% F1 with 100% case-level accuracy. In a same-test-set comparison, the 66M-parameter task-specific classifier achieves 88.4% F1 versus ≈37% F1 for GPT-5.5, Claude Sonnet 4.6, and Gemini 3 Pro Preview, confirming that labeled domain supervision captures criterion-recognition patterns that zero-shot prompting at frontier scale cannot provide.

Five applied lessons: **(i)** parser quality is the first engineering problem; CFR-header anchoring mattered more than any model choice; **(ii)** evaluation design dominates; positive-sensitive metrics are required on denial-heavy data; **(iii)** training-data quality shifts the operating point non-trivially; calibration must be re-validated after every corpus update; **(iv)** aggregation, threshold, and calibration are operational decisions, not hyperparameters; post-hoc calibration is the correct remedy for high-confidence errors; **(v)** the model has internalized AAO-specific linguistic markers rather than general legal reasoning, as shown by the criterion iv inversion (§7) and the frontier-LLM underperformance despite access to identical text.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.

Daniel Martin Katz, Michael J. Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12(4):e0174698.

Legal Information Institute. 2026. 8 CFR § 204.5 – petitions for employment-based immigrants. https://www.law.cornell.edu/cfr/text/8/204.5. Accessed 2026-04-19.

Ang Li, Yiquan Wu, Ming Cai, Adam Jatowt, Xiang Zhou, Weiming Lu, Changlong Sun, Fei Wu, and Kun Kuang. 2025. Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6957–6970, Albuquerque, New Mexico. Association for Computational Linguistics.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Junkai Liu, Yujie Tong, Hui Huang, Bowen Zheng, Yiran Hu, Peicheng Wu, Chuan Xiao, Makoto Onizuka, Muyun Yang, and Shuyuan Zheng. 2025. Legal fact prediction: The missing piece in legal judgment prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6334–6349, Suzhou, China. Association for Computational Linguistics.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

U.S. Citizenship and Immigration Services. 2025. Chapter 2 - extraordinary ability. https://www.uscis.gov/policy-manual/volume-6-part-f-chapter-2. USCIS Policy Manual. Accessed 2026-04-19.

Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. 2018. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels, Belgium. Association for Computational Linguistics.

A Full Reproducibility Details

All models use early stopping (patience 5), best checkpoint saved by aggregate accuracy, Adam optimizer with default Transformers settings. Evaluation on 25 randomly sampled cases from the respective corpus. The held-out test set (`parsed_decisions_v3_test/`, 29 files from 2024–2026) is never used during training or hyperparameter tuning.

Table 7: Complete hyperparameters for all model variants.

Model	Corpus	Max ex.	Encoder	Len	Ep.	Batch	LR	Wt.	Rule	Thr.
Baseline (v2)	v2	400	distilbert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65
−Prefix (v2)	v2	400	distilbert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65
Scaled (v2)	v2	1,200	distilbert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65
Legal (v2)	v2	1,200	legal-bert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65
Weighted (v2)	v2	1,200	distilbert-base-uncased	512	10	8	2e-5	8.75	MajVote	0.65
v3-Baseline	v3	400	distilbert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65
v3-Scaled	v3	1,200	distilbert-base-uncased	512	10	8	2e-5	1.0	MajVote	0.65

B Aggregation Rule Sweep

The aggregation rule controls how evidence-item predictions are converted into a criterion-level decision. MajVote is more conservative and suppresses isolated high-confidence positives, while AnyPos increases sensitivity by allowing a single qualifying evidence item to trigger the criterion. In practice, AnyPos substantially increases false positives without recovering additional true positives on the v2 benchmark, confirming that aggregation policy is an operational deployment choice rather than a purely statistical hyperparameter.

Table 8: Aggregation rule sweep for Scaled (0.65 threshold, 39 criterion examples, 4 actual positives). AnyPos adds 9 FPs and zero TPs; MajVote is the better operating point.

Metric	MajVote	AnyPos
Predicted positive	17	26
True positives	3	3
False positives	14	23
False negatives	1	1
Precision	17.6%	11.5%
Recall	75.0%	75.0%
F1	28.6%	20.0%

C Confidence Threshold Sweeps

Threshold sweeps reveal that the classifier does not exhibit a clean separability boundary between positive and negative criterion predictions. As the confidence threshold increases, precision improves modestly but recall collapses rapidly, indicating substantial overlap between TP and FP confidence distributions. This behavior motivates post-hoc calibration rather than threshold tuning alone.

D Held-out Per-Criterion Breakdown

Per-criterion evaluation confirms that the strongest held-out performance occurs on criteria with both stable parser extraction and sufficient positive supervision in the training corpus. Errors concentrate primarily in sparse criteria or criteria whose evidentiary language frequently overlaps between granted and denied AAO decisions.

Table 9: Threshold sweep for Scaled (v2, MajVote, 4 actual positives). Best recall (75.0%) at 0.65; best F1 (40.0%) at 0.95.

Thr.	Pred+	TP	FP	FN	Prec	Rec
0.65	17	3	14	1	17.6%	75.0%
0.70	16	2	14	2	12.5%	50.0%
0.75	14	2	12	2	14.3%	50.0%
0.80	12	2	10	2	16.7%	50.0%
0.85	8	1	7	3	12.5%	25.0%
0.90	2	1	1	3	50.0%	25.0%
0.95	1	1	0	3	100.0%	25.0%

Table 10: Threshold sweep for v3-Scaled (AnyPos, 7 actual positives). No threshold separates TPs from FPs confirming calibration failure.

Thr.	Pred+	TP	FP	FN	Prec	Rec
0.65	29	4	25	3	13.8%	57.1%
0.70	27	4	23	3	14.8%	57.1%
0.75	26	4	22	3	15.4%	57.1%
0.80	22	2	20	5	9.1%	28.6%
0.85	16	2	14	5	12.5%	28.6%
0.90	11	2	9	5	18.2%	28.6%
0.95	8	2	6	5	25.0%	28.6%

E System Pipeline Schematic

Figure 1 summarizes the end-to-end inference pipeline described in the main text. At inference time, an evidence packet $\mathcal{E}$ (YAML or plain text, §3.2) is ingested by the v3 parser with CFR-header anchoring (§3.2), and each criterion analysis block is scored by the fine-tuned DistilBERT classifier, which estimates $P(Met \mid E, C)$ for each evidence item E and criterion C (§6.3). Per-criterion decisions are produced by majority-vote aggregation at the $\tau = 0.65$ operating threshold (§6). The output is a triage report that partitions the ten criteria into strong anchors, borderline cases requiring additional documentation, and weak candidates to deprioritize. The sensitivity probe (§7) draws on

Table 11: Per-criterion results for v3-Scaled on the held-out direct-criterion evaluation (52 criterion examples, 20 actual positives).

Crit.	N	Pos	TP	FP	Prec	Rec
i	8	4	4	1	80.0%	100.0%
ii	8	2	2	0	100.0%	100.0%
iii	9	2	2	1	66.7%	100.0%
iv	5	4	4	0	100.0%	100.0%
v	6	2	2	0	100.0%	100.0%
vi	3	1	1	0	100.0%	100.0%
vii	4	1	1	0	100.0%	100.0%
viii	5	3	3	2	60.0%	100.0%
ix	3	1	0	0	0.0%	0.0%
x	1	0	0	0	—	—

the same classifier to characterize which linguistic markers drive its confidence; it is offline and does not feed back into the deployed screening loop.

F Computational Cost

All training was performed on a single Apple M4 Pro laptop (12-core CPU, 24 GB unified memory) using the PyTorch MPS backend. No cloud compute, dedicated GPU, or distributed training was used.

The computational footprint demonstrates that criterion-level legal NLP can be reproduced without access to institutional GPU clusters or cloud-scale infrastructure. Even the strongest-performing model trains in only a few minutes on consumer hardware, making iterative experimentation and parser debugging substantially cheaper than frontier-LLM fine-tuning workflows.

Wall-clock times are estimated from measured evaluation throughput (training $\approx$ eval / 3.5); compute in TFLOPs from `total_flos`. Times range from under one minute (v2-Baseline) to $\approx$4.4 minutes (v2-Legal). The final model (v3-Scaled) trains in $\approx$2.6 minutes. Total compute across all reported experiments is $\approx$10,100 TFLOPs, fully reproducible on consumer hardware.

G Evidence Quality Sensitivity: Full Probe Data

Complete evidence strings and confidence scores underlying the sensitivity analysis in §7. All scores use the `[EVIDENCE]` task prefix and the 0.65 operating threshold.

Interpretation. For criterion vi, all five levels exceed 65%, confirming the model readily grants scholarly articles once any publication language is present. For criterion v, the "named method" drop (22.4%) occurs because domain-specific technical language appears frequently in denied cases without the adoption or citation evidence that accompanies a granted criterion; recovery is immediate once "adopted" is added (81.2%). For criterion i, the NSF CAREER Award (and descriptions of prestige or selectivity) never crosses the threshold without the IEEE Fellow phrase, a direct consequence of training-corpus coverage. For criterion iv, the monotonic decline reflects AAO decisions frequently acknowledging review activity while denying criterion iv by distinguishing expected academic service from the regulation's intent.

H Precision-Recall Curve

Figure 4 visualizes the precision–recall tradeoff for v3-Scaled across thresholds from 0.50 to 0.95. The absence of a sharp Pareto frontier confirms the calibration pathology discussed in §9: many false positives receive confidence scores comparable to true positives. Consequently, operational deployment should treat scores as ordinal ranking signals rather than calibrated legal probabilities.

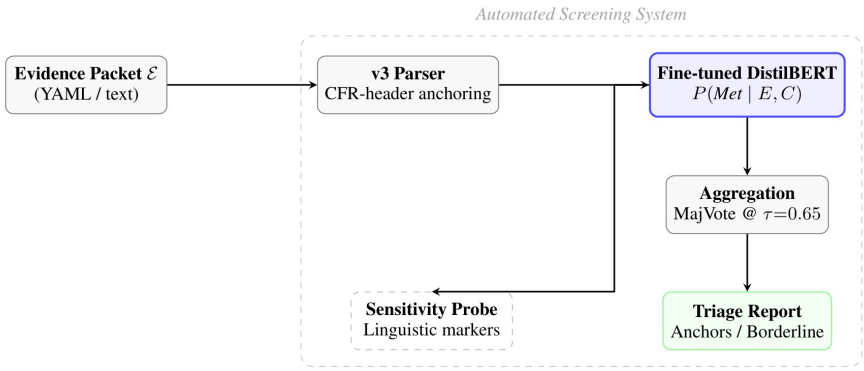

Figure 1: End-to-end inference pipeline. An evidence packet $\mathcal{E}$ (YAML or plain text) is parsed by the v3 CFR-header-anchored parser into criterion analysis blocks, scored by the fine-tuned DistilBERT classifier $P(Met \mid E, C)$, aggregated by majority vote at $\tau = 0.65$, and returned as a triage report partitioning evidence into anchors and borderline cases. The sensitivity probe (dashed) consumes classifier outputs offline to characterize criterion-specific linguistic markers and does not participate in deployed screening.

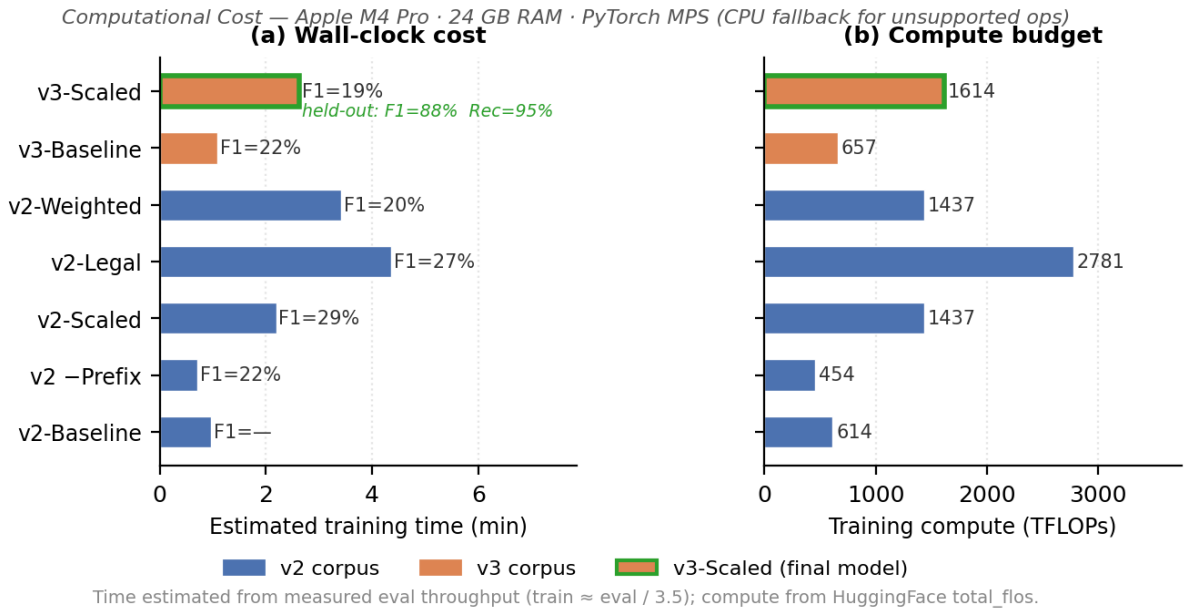

Figure 2: Training cost per variant on an Apple M4 Pro (PyTorch MPS). **Left:** estimated wall-clock time. **Right:** training compute in TFLOPs (`total_flos`). The v3-Scaled bar (green outline) is the final deployed model; its held-out F1 (88.4%) differs from its training-benchmark F1 (19.4%) due to evaluation mode differences (§6.3).

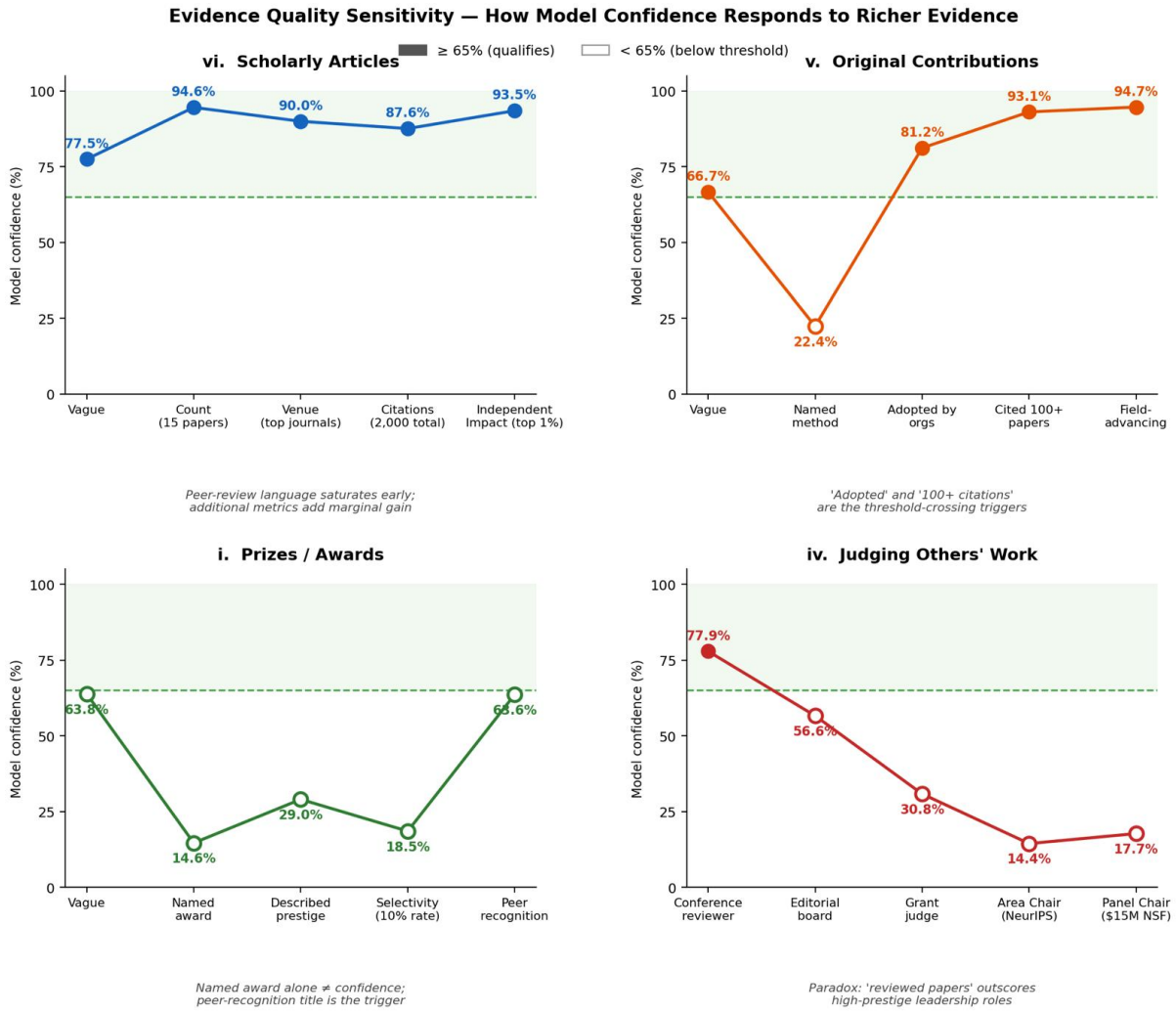

Figure 3: Model confidence as evidence quality increases across five levels for four criteria. Filled circles = above the 0.65 threshold; open circles = below. Key findings: criterion vi saturates early at 94.6%; criterion v recovers only when "adopted" or "cited in 100+ papers" is added; criterion i collapses on a named but corpus-unfamiliar award; criterion iv shows a monotonic *decrease*: conference review outscores Area Chair, reflecting how AAO case law treats academic service.

Table 12: Full probe evidence strings and model confidence scores. Scores $\geq 65\%$ (bold) count as qualifying. Each row isolates one quality dimension added relative to the previous level.

Crit.	Level	Evidence string	Conf.
vi	Vague	The petitioner has published papers.	**77.5%**
vi	Count	The petitioner has authored 15 peer-reviewed articles.	**94.6%**
vi	Venue	...15 articles in top-tier journals (Nature, IEEE Transactions).	**90.0%**
vi	Citations	...15 articles with 2,000 total citations (h-index 12).	**87.6%**
vi	Indep. impact	...2,000 independent citations; two papers in top 1% cited globally.	**93.5%**
v	Vague	The petitioner made original contributions to their field.	**66.7%**
v	Named method	The petitioner developed a novel algorithm for natural language processing.	22.4%
v	Adopted	The petitioner's algorithm was adopted by multiple organizations.	**81.2%**
v	Cited 100+	The petitioner's method is cited in over 100 peer-reviewed papers.	**93.1%**
v	Field-advancing	...cited in 100+ independent papers; foundational in two field surveys.	**94.7%**
i	Vague	The petitioner won an award for excellence.	63.8%
i	Named award	The petitioner received the NSF CAREER Award.	14.6%
i	Described prestige	...the most prestigious early-career research grant in the U.S.	29.0%
i	Selectivity	...awarded to 500 of 5,000 applicants (10% rate).	18.5%
i	Peer recognition	...subsequently elected IEEE Fellow, recognizing extraordinary accomplishment.	**63.6%**
iv	Conf. reviewer	The petitioner reviewed papers for academic conferences.	**77.9%**
iv	Editorial board	...serves on the editorial board of IEEE Transactions, reviewing manuscripts.	56.6%
iv	Grant judge	...served as a grant proposal judge for the National Science Foundation.	30.8%
iv	Area Chair	...Area Chair at NeurIPS 2024, making accept/reject decisions on 30 submissions.	14.4%
iv	Panel Chair	...chaired an NSF grant review panel, with final authority over $15M.	17.7%

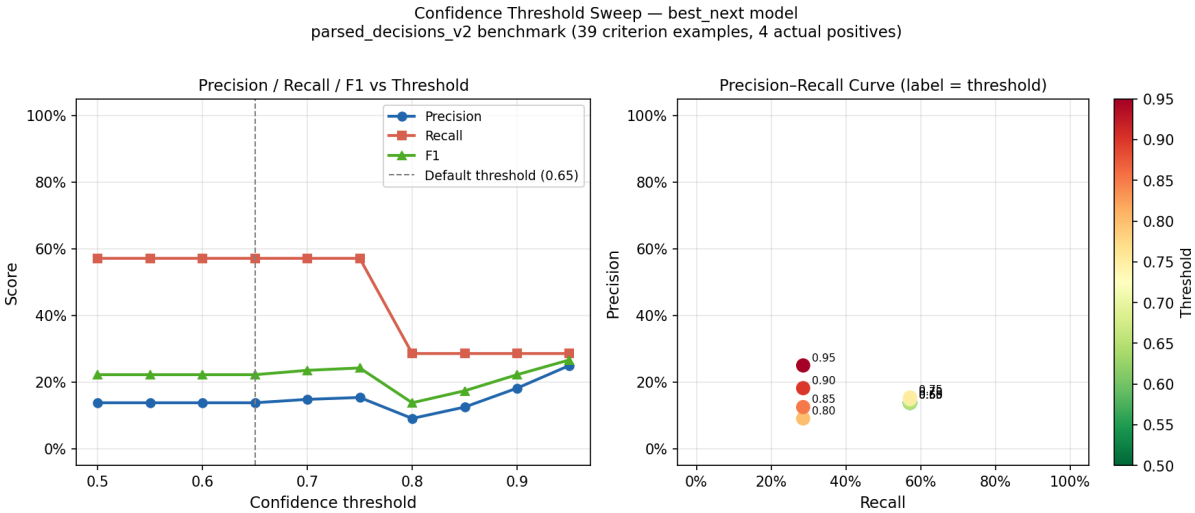

Figure 4: Precision-recall curve for v3-Scaled (threshold sweep 0.50–0.95, AnyPos aggregation, 40 criterion examples, 7 actual positives). No threshold cleanly separates TPs from FPs—FP confidences span 0.713–0.984, overlapping with TP confidences of 0.758–0.991.