The Result Model under Inconsistent Knowledge: Theory and Experiments

Yoann MORELLO [a,1] Agata CIABATTONI [a] Morgan GRAY [b]
[a] *TU Wien, Austria*
[b] *University of Pittsburgh, Intelligent Systems Program*

Abstract. The result model provides a foundation for precedent-based reasoning, yet real case bases are often inconsistent, with precedents pointing to opposite outcomes. To address this, we augment the result model with the *Log-Odds Precedent Aggregator (LOPA)*—a Naive-Bayes–style log-odds combiner that treats each applicable precedent as uncertain evidence, learns its reliability from data, and produces both a decision and a confidence score. We evaluate LOPA on the DIAS dataset, comparing it against other extensions of the result model, and a strong machine-learning baseline. Results show that the symbolic and hybrid models perform on par with ML, and in several settings slightly better, while remaining fully interpretable. LOPA is especially robust when many weak precedents compete with a few strong ones, and its calibrated confidence supports coverage–reliability tradeoffs.

Keywords. Result Model, Case Based Reasoning, Inconsistent case base, Bayesian, Naive Bayes, Machine Learning classifiers, Legal AI, LOPA

1. Introduction

In common law, the doctrine of precedent is a form of reasoning where past court decisions guide future ones, while still leaving room to adapt to new circumstances.

The *result model* [1] captures a minimal view of this process: only the outcome and its supporting facts matter, and a later court follows a precedent when the facts before it are at least as strong for the winner as in the original case. More sophisticated models enrich this view by also taking into account, for example, the reasons[2] provided in the case [3], factors valued on ordered sets (aka dimensions) [4, 5], or factor hierarchies [6].

Here we focus on the result model because it aligns with the Drug-Interdiction Auto-Stop (DIAS) dataset [7, 8]: each case consists of one-sided binary factors plus the outcome, the information the result model uses. Even in simple settings (as in DIAS), real data can be inconsistent, with applicable precedents pointing to opposite outcomes.

We study precedent-based prediction under inconsistency in the *result model*. Various approaches have been proposed to handle inconsistencies in case-based reasoning; on DIAS we evaluate three predictors from [9]—*Strict Binary*, *Binary Majority*, and *Bayesian Individual Majority*—and introduce the *Log-Odds Precedent Aggrega-*

[1] Corresponding author: yoann.morello@gmail.com.
[2] This leads to the reason model [2].

tor (LOPA); the first two trace back to [10] (adapted from the reason model to the result model in [9]), while the two Bayesian variants—*Bayesian Individual Majority* and *LOPA*—enrich the result model to operate under inconsistency (cf. [5, 11]) by learning precedent reliability from data and aggregating it transparently. Both *LOPA* and *Bayesian Individual Majority* are hybrid symbolic–subsymbolic methods: they retain the result model's *a fortiori* structure while learning per-precedent reliability from data. *LOPA* replaces the voting heuristic used to aggregate probabilistic evidence in *Bayesian Individual Majority* with a probabilistically grounded *Naive-Bayes–style* [12] combination: it (i) aggregates precedent opinions in *posterior log-odds*, providing a principled alternative to voting, and (ii) prunes redundant chains to retain only maximal (more authoritative) precedents, yielding shorter explanations with calibrated confidence.

Across DIAS, all extensions of the result model perform on par with —sometimes slightly better than —— a strong ML baseline. Among them, *LOPA* offers two practical advantages: (i) a meaningful, calibrated confidence enabling coverage–reliability tradeoffs, and (ii) robustness to "many-weak vs. few-strong" conflicts.

Related work. Both Bayesian models tested here (*Individual Majority* and *LOPA*) extend the notion of authoritativeness from [13, 14] by learning distributions over precedent reliability directly from data and using them in prediction. A complementary line, initiated by [15] and developed, e.g., by [16], studies inconsistencies in probabilistic propositional knowledge bases: they define inconsistency measures and propose optimization-based methods (e.g., distance-to-consistency/Dutch-book coherence) to quantify and, if desired, repair inconsistency in the database. In contrast, rather than eliminating such conflicts, our formalism is designed to make predictions in their presence, providing decisions together with calibrated confidence.

Empirical studies exist in case-based frameworks (e.g., [17]), but evaluations of the result model are relatively few. Aside from [11, 18], which evaluate it on dimensional (ordered) factors rather than binary factors, most prior work is theoretical.

Code and data: `https://github.com/yoannmorello/Bayesian_Result_Model`

2. Preliminaries: Result Model, Beta Distributions and DIAS Dataset

The Result Model [1] is a conservative approach to precedential reasoning. Below, we give the definitions used in this paper; for a detailed model description, see, e.g., [19].

Let $F = F_\pi \cup F_\delta$ be a set of binary factors favoring, respectively, the plaintiff π and the defendant δ. A *case* is a pair $\langle X, s\rangle$ with $X \subseteq F$ and $s \in \{\pi, \delta\}$. Write $X_s = X \cap F_s$ and let $\bar{s}$ denote the opposite side. The *strength* quasi-order for side s is $X \succeq_s Y$ iff $X_s \supseteq Y_s$ and $X_{\bar{s}} \subseteq Y_{\bar{s}}$. A case base is *consistent* if there do not exist two cases $\langle X, \delta\rangle$ and $\langle Y, \pi\rangle$ such that $X \succcurlyeq_\pi Y$. This condition can be equivalently stated in terms of the *a fortiori* constraint: if $\langle Y, s\rangle$ is a precedent and $X \succeq_s Y$, then a constrained court should decide $\langle X, ?\rangle$ for side s. As we will see, in the DIAS dataset all factors favour the plaintiff ($F_\delta = \varnothing$), yet both outcomes occur. Thus, a δ-precedent is simply a past decision for δ, represented by its pro-π factor set; all precedents, whether π or δ, are described using pro-π factors. In this one–sided setting the result–model orders collapse to

$$X \succeq_\pi Y \iff X \supseteq Y, \qquad X \succeq_\delta Y \iff X \subseteq Y.$$

Hence a π-precedent with factors S_i applies to a new case X iff $S_i \subseteq X$; a δ-precedent with factors S_j applies iff $X \subseteq S_j$.

Precedent reliability (Beta *likelihood*s). We associate to each precedent i a Beta distribution [12] Θ_i encoding how reliably its *a fortiori* constraint is respected when it applies to other cases. Whenever i applies to a training case, we count *supports* r_i (the court reaches the same side as i) and *failures* s_i (the opposite side), and set

$$\Theta_i \sim \mathrm{Beta}(r_i, s_i).$$

This is the familiar "successes vs. failures" scheme widely used to model subjective reliability and reputation [20–22]. Its mean $\mu_i = \frac{r_i}{r_i+s_i}$ can be read as the precedent's authoritativeness [13], while the total r_i+s_i reflects how certain that estimate is. How these per–precedent Betas are used in the aggregators is detailed in Section 3; see also [9] for the Bayesian justification.

Dataset. We use a subset of the DIAS dataset (Drug-Interdiction Auto-Stop) identified in [7, 8] sourced from the Harvard Law School Case Law Access Project. It comprises court opinions on whether U.S. police officers have the constitutional grounds to prolong a motorist's detention for suspected drug trafficking or possession. [7] defined the question of reasonable suspicion in terms of factors. Courts assess it under the totality of the circumstances test, which considers "the whole picture." *U.S. v. Powell*, 277 F. App'x 782, 785 (10th Cir. 2008); *United States v. Daniels*, 101 F.4th 770, 776 (10th Cir. 2024) (quoting *U.S. v. Cortez*, 449 U.S. 411, 417–18 (1981)). In the DIAS domain, the factors are *one sided*, requiring the prosecution (aka plaintiff) to show, that all of the observations together support an officer's suspicion.

The full DIAS dataset contains 262 gold-standard annotated judicial opinions; in this work we use a subset of 201 cases. Annotators assessed whether a factor applied based on the court's analysis of reasonable suspicion. They also coded the outcome reached by the court (i.e., whether there was suspicion for detention). The dataset has a class imbalance of roughly 62% of cases were suspicion was found and 38% of cases where suspicion was not found. Although in the DIAS domain "courts at every level of the judicial hierarchy, in every state [in the U.S.], apply the same standard", the dataset contains inconsistencies. For instance some cases with the only factor *"Vehicle contents suggest drugs"* are judged for the plaintiff, while other with the same factor plus additional (plaintiff-favoring) factors are judged for the defendant. Strong indicators (e.g., *"a strong smell of marijuana"*) tend to appear in plaintiff cases, weaker ones (e.g., *"presence of syringes"*) in defendant cases; modeling these factors using a dimensional approach, would likely eliminate some of these inconsistencies. In general, as courts evaluate the totality of circumstances on a case-by-case basis and judges cannot account for every prior decision, the same factors can lead to different outcomes, creating inconsistencies in the dataset.

3. Predictive Models

We describe four approaches to account for inconsistencies in the result model. The first three –the *Strict Binary*, *Binary Majority* and *Bayesian Individual Majority* models– have been introduced in [9]. The first two implement the solutions proposed by [10] for the

reason model, adapted to the result model; the third is their probabilistic variant. The fourth model is new: instead of aggregating *evidence* by stochastic majority voting, it performs a Naive–Bayes–style aggregation of per-precedent reliability estimates (see [12, §3.5]); this can be seen as a weighted-and-biased voting rule.

Strict Binary. This model predicts s iff there exists at least one applicable s-precedent and no applicable $\bar{s}$-precedent; otherwise it abstains, and its coverage can be low in the presence of inconsistencies.

Binary Majority, This model predicts s whenever the number of applicable s-precedents exceeds that of applicable $\bar{s}$-precedents, and abstains on ties; this typically increases coverage but treats all precedents as equally authoritative.

Bayesian Individual Majority. Using the Beta likelihood counts (r_i, s_i) (Sec. 2), and place a uniform prior $\mathrm{Beta}(1,1)$ modeling our lack of knowledge to get $\Theta_i \sim \mathrm{Beta}(r_i+1, s_i+1)$. For each applicable precedent, draw one Bernoulli vote with success $\theta_i \sim \Theta_i$; count the votes for each side and predict the side with more votes (abstain on ties).

Remark 1. *Aggregating many weak precedents can dominate a single very strong one: in a balanced dataset, even an uninformative π-precedent (e.g., the empty π-case) contributes an expected $1/2$ for π, so a handful of such precedents can swamp an excellent δ-precedent. This motivates a different aggregation in our next model.*

3.1. LOPA model

Each applicable precedent is treated as a noisy piece of evidence about the winning side. From training data, every precedent receives a data–driven *reliability*—how often its *a fortiori* application is correct (cf. [9]). For a new case we:

1. identify the applicable precedents;
2. prune redundant ones along inclusion chains to avoid double counting;
3. combine the remaining evidence with a simple Bayesian update.

More reliable precedents exert greater influence; starting from the empirical base-rate priors of the two sides, the aggregation yields posterior probabilities for π and δ. We predict the side with the higher posterior and report that probability as a calibrated confidence. The only simplifying assumption is the Naive Bayes idealization that, *conditional on the true outcome*, the retained (non-redundant) precedents act as independent signals; this makes the combination rule transparent and computationally light [12, §3.5].

Definition 1 (Applicability events E_i)**.** *Fix a new case with factor set X. For each precedent i with decided side $s_i \in \{\pi, \delta\}$ and factor set S_i, define E_i ("i applies to X") by*

$$E_i \text{ holds} \iff \begin{cases} S_i \subseteq X & \text{if } s_i = \pi, \\ X \subseteq S_i & \text{if } s_i = \delta. \end{cases}$$

Let $\mathscr{E}_\pi = \{\, i : \ s_i = \pi \text{ and } E_i \text{ holds} \,\}$ and $\mathscr{E}_\delta = \{\, j : \ s_j = \delta \text{ and } E_j \text{ holds} \,\}$ be the index sets of applicable π and δ precedents (after pruning).

Priors. Side priors are the empirical class proportions in the training set,

$$\pi_0 = \frac{N_\pi}{N_\pi + N_\delta}, \qquad \delta_0 = \frac{N_\delta}{N_\pi + N_\delta}.$$

Step 1 (Bayes on joint evidence).

$$\Pr(Y = \pi \mid \{E_i\}) \propto \Pr(\{E_i\} \mid Y = \pi)\, \pi_0, \qquad \Pr(Y = \delta \mid \{E_j\}) \propto \Pr(\{E_j\} \mid Y = \delta)\, \delta_0.$$

where "$\propto$" means "is proportional to"
Step 2 (Conditional independence & Bayes inside each factor). Assuming that, given Y, the retained applicability events are (approximately) independent,

$$\Pr(\{E_i\} \mid Y = \pi) = \prod_{i \in \mathscr{E}_\pi} \Pr(E_i \mid Y = \pi) \prod_{j \in \mathscr{E}_\delta} \Pr(E_j \mid Y = \pi),$$

and symmetrically for $Y = \delta$. Applying Bayes' rule inside each factor and cancelling common terms yields the posterior–odds identity (with $K = |\mathscr{E}_\pi| + |\mathscr{E}_\delta|$):

$$\frac{\Pr(Y = \pi \mid \{E_i\})}{\Pr(Y = \delta \mid \{E_i\})} = \left(\frac{\pi_0}{\delta_0}\right)^{1-K} \prod_{i \in \mathscr{E}_\pi} \frac{\Pr(Y = \pi \mid E_i)}{1 - \Pr(Y = \pi \mid E_i)} \prod_{j \in \mathscr{E}_\delta} \frac{1 - \Pr(Y = \delta \mid E_j)}{\Pr(Y = \delta \mid E_j)}. \tag{1}$$

Step 3 (Plug–in with uncertainty). For each precedent we learn a Beta posterior for its *a fortiori* reliability from training data, $\Theta_i \sim \mathrm{Beta}(r_i, s_i)$, using $\mathrm{Beta}(r, s)$ counts (no unit pseudo–counts) so that class priors enter only via the prior–odds term above (rather than at the per–precedent level; cf. [9]). We then substitute Monte Carlo draws θ_i, θ_j for $\Pr(Y = \pi \mid E_i)$ and $\Pr(Y = \delta \mid E_j)$ in (1), obtaining a label (the larger posterior) and a calibrated confidence (that posterior probability).

Note on dependence mitigation. Exact independence is violated when precedents nest. We explain below a strategy to mitigate these dependencies, called *max pruning*.

Remark 2. *Following [10], the models in [9] use a (stochastic)* voting *rule. Our LOPA instead forms a linear sum in* log–odds *with a prior–correction term. For each Monte Carlo draw r, sample* $\theta_i^{(r)} \sim \mathrm{Beta}(r_i, s_i)$ *then compute*

$$V^{(r)} = (1-K) \log \frac{\pi_0}{\delta_0} + \sum_{i \in E_\pi} \mathrm{logit}\big(\theta_i^{(r)}\big) - \sum_{j \in E_\delta} \mathrm{logit}\big(\theta_j^{(r)}\big),$$

and the predicted probability $p_\pi = \mathbb{E}_r[\sigma(V^{(r)})]$. *This yields a probabilistic weighted vote, where uncertainty is propagated via the Beta draws.*

Max pruning (independence mitigation). The only structural assumption in the LOPA is conditional independence of the $\{E_i\}$ given Y. In our setting, this assumption is often violated: if $c_1 \succcurlyeq c_2$ (by abuse of notation, $\mathrm{factors}(c_1) \supseteq \mathrm{factors}(c_2)$), then $E_1 \Rightarrow E_2$. Consequently, for any side $s \in \{\pi, \delta\}$,

$$\Pr(E_1, E_2 \mid Y = s) = \Pr(E_1 \mid Y = s),$$

To mitigate these dependencies, we first deduplicate precedents that share the same factor set (keeping one per set), and then retain the *maximal* elements under the side-specific order ($\subseteq$ on π, $\supseteq$ on δ), i.e., those with no strictly superior element. [3] We then apply equations (1) to this pruned set. This does not guarantee full independence but empirical evidence shows that Naive Bayes often performs well even under attribute dependence [23]. *Why it helps:* (i) Log-odds down-weight weak precedents, so many poor π-precedents cannot easily overrule a few decisive δ ones. (ii) Max pruning reduces correlation and double-counting. (iii) MC leverages the full Beta to produce calibrated confidence.

We have introduced four predictors for inconsistent case bases, which we test on DIAS in the next section. The new one, *LOPA* differs from *Bayesian Individual Majority* only in (i) the aggregation method—posterior *log-odds* instead of mean voting—and (ii) pruning applicable precedents to *maximal* ones, yielding more concise explanations.

4. Experimental Protocols

We compare the performance of the LOPA model with the Strict Binary, the Binary Majority, and the Bayesian model as well as with a ML classifier.

4.1. Evaluation Protocol and Majority-Class Imputation Baseline

We test on the DIAS dataset with 30 independent, stratified random splits [24]. In each split, we hold out 20% of the cases for testing and train on the remaining 80%, preserving class proportions. This setup yields robust estimates under varying train/test partitions.

Metrics (reported per split, then averaged). Let N be the number of test cases and let P be the number of non-`NaN` predictions a model makes. We report:

- *Coverage*: $100 \times (P/N)$, the percent of test cases for which the model returns a label (non-abstention).
- *Macro–F1* [4] *(pred. only)*: macro-averaged F1 computed on the subset of test cases with non-`NaN` predictions.
- *Macro–F1 (+ π completion)*: to compare fairly with learners that never abstain, any `NaN` prediction is replaced by a default majority-class label (π); macro–F1 is then computed on all N test cases.

Random-forest baseline.

Following [26], which found Random Forest strongest on an alternative DIAS pruning, we use it as the sole ML baseline.

Cases are 18-bit factor vectors; hyperparameters are tuned by 5-fold CV on the training fold; the model always predicts (coverage 100%).

Interpretation of Table 1. Coverage varies substantially: *Strict Binary* abstains most ($\approx$69% coverage), which inflates its macro–F1 on the easier, predicted subset (0.8411), but lowers macro–F1 once abstentions are completed with π (0.7331). After majority-

[3] Max pruning is *side-symmetric*. After deduplication, we keep *maximal* sets under $\subseteq$ on the π (pro-plaintiff) side and *minimal* sets under $\supseteq$ on the δ (pro-defendant) side (equivalently, maximal for the δ-order). Thus redundancy is removed on *both* sides—not only for π.

[4] Macro–F1 is the unweighted mean of per-class F1 scores (macro-averaging) [25]

Table 1. Averaged over 30 stratified 80/20 splits.

Model	Coverage (%)	Macro–F1 (pred. only)	Macro–F1 (+ π completion)
Bayesian Majority	85.93	0.7795	0.7500
Binary Majority	82.93	0.7862	0.7505
LOPA	85.93	0.7828	0.7535
Strict Binary	68.70	0.8411	0.7331
Random Forest	100.00	0.7400	0.7400

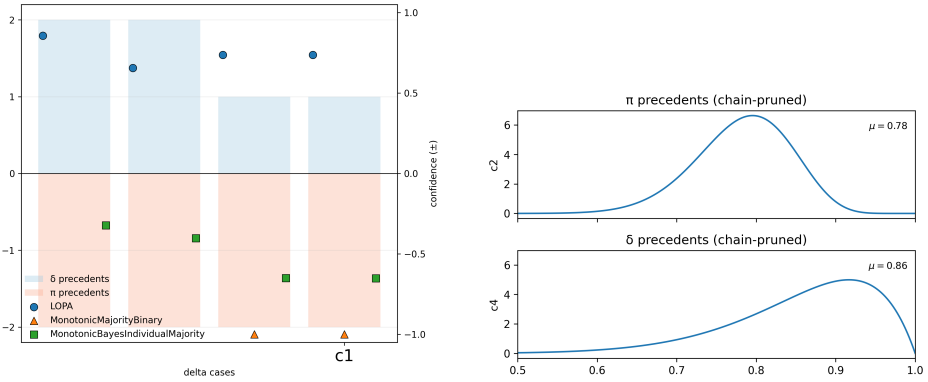

Figure 1. Model disagreements and an illustrative per–precedent explanation. (a) Disagreement set (δ cases). Bars (left y-axis) count applicable precedents—upward support δ, downward π. Markers (right y-axis) show each model's prediction and confidence; above 0 agrees with ground truth and height encodes confidence. (b) Case c_1. Learned Beta posteriors for selected π (left column) and δ (right column) precedents under LOPA.

class imputation, three symbolic models—*Naive-Bayes*, *Binary Majority*, and *Bayesian Individual Majority*—all exceed the Random Forest on macro–F1 (0.7535, 0.7505, and 0.7500 vs. 0.7400), while maintaining competitive coverage (83–86%). This suggests that the *a fortiori* structure captured by our symbolic methods aligns well with the normative regularities in the data, yielding performance on par with or better than a strong ML baseline while remaining transparent and explainable.

4.2. Leave–One–Out (Upper–Bound) Comparison

After establishing parity (even slight superiority) over a strong ML baseline, we compare the symbolic models *to one another* and inspect concrete cases. To avoid dependence on random train/test splits, we use leave–one–out cross-validation (LOO) [12, §1.2.7]: for each case i, we train the model on all other cases $\mathscr{D} \setminus \{i\}$ and evaluate on i. LOO uses $n-1$ training examples per test case, and thus serves as a high-information benchmark for per-case performance (often stronger than results from smaller train/test splits).

LOO Protocol. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the dataset. For each case j: (i) train each model m on $\mathscr{D} \setminus \{j\}$; (ii) compute the model's LOO prediction on x_j (allowing abstention); (iii) for Bayesian models record also a confidence (Monte Carlo estimate).

Where the models differ. Figure 1(a) shows that *Bayesian Individual Majority* [9] aligns in sign with *Binary Majority* on all non-ties, but attains higher coverage by deciding exact

ties. This underscores a limitation of simple counts (Remark 1): many weak precedents can swamp a few strong ones. By contrast, our *LOPA* sometimes prefers fewer, stronger precedents and thus departs from majority when strength and count conflict. Fig. 1(b) illustrates this on the δ case `c1`: despite more π precedents overall, the δ precedent's posterior dominates, and the model correctly predicts δ.

A representative instance is the δ case labeled `c1` in Fig. 1(a), whose per-precedent posteriors are shown in Fig. 1(b).

Example 1 (Case `c1`)**.** *Let $F(c)$ be the factor set of case c. Case `c1` is a δ decision with*

$$F(\texttt{c1}) = \begin{cases} f_1 = \textit{Nervous Behavior or Appearance} & f_2 = \textit{Refused Consent} \\ f_3 = \textit{Suspicious or Inconsistent Answers} & f_4 = \textit{Other} \end{cases}.$$

There are two applicable π precedents, `c2` with $F(\texttt{c2}) = \{f_1, f_2, f_3\}$ and `c3` with $F(\texttt{c3}) = \{f_1, f_3\}$, and one applicable δ precedent, `c4` with $F(\texttt{c4}) = \{f_1, f_2, f_3, f_4\} = F(\texttt{c1})$. Since $F(\texttt{c3}) \subset F(\texttt{c2})$, max pruning discards `c3` on the π side, keeping only the maximal subset `c2`. Fig. 1(b) shows the learned Beta posteriors for the retained precedents: the δ density (for `c4`) places most of its mass to the right of the π density (for `c2`), reflecting greater confidence that an a fortiori *application of `c4` supports δ. Our LOPA therefore (correctly) predicts δ for `c1`. In contrast,* Binary Majority *counts two π precedents versus one δ and would vote π, and* Bayesian Individual Majority *tends to agree with that majority sign (cf. Fig. 1(a)).*

5. Discussion

Beyond accuracy numbers, what makes our LOPA model distinctive is its ability to explain, justify, and calibrate its decisions. In what follows, we discuss some key aspects.

Explainability. ML captures correlations, not explanations; it estimates feature–label associations, multiply them, and predict. But it cannot say which precedent carried the decision or why it should be trusted. Our extended model addresses this limitation by tracing the precedents that drive the decision. In the presence of inconsistencies, each applicable precedent is treated as an explicit, auditable *a fortiori* explanation, with a learned reliability score (a Beta distribution). Decisions are then based on a small set of such explanations, with calibrated confidence. Instead of "this pattern correlates with π," we can say "this case follows precedents $\{p_1, p_2\}$ *a fortiori*, and we are x% confident."

Compact justifications. We prune redundant chains of precedents (e.g., when one set of facts strictly includes another). Indeed when several applicable precedents stand in an *a* fortiori chain (e.g., $S_i \subset S_j$ on the π side or $S_i \supset S_j$ on the δ side), we keep only the extremal (maximal/minimal) elements. This yields shorter, human-scale explanations: our model cites, on average, **2.47** precedents per decision, compared to **6.14** for the Bayesian Individual Majority. This is consistent with certain judicial guidance [27, p. 159]; see also [28, p. 17], which values the weight of authorities over their number.

Beyond means. In the *Bayesian Individual Majority* [9], the decision depends only on the means of the per–precedent Betas. Indeed, if we write the stochastic vote margin as

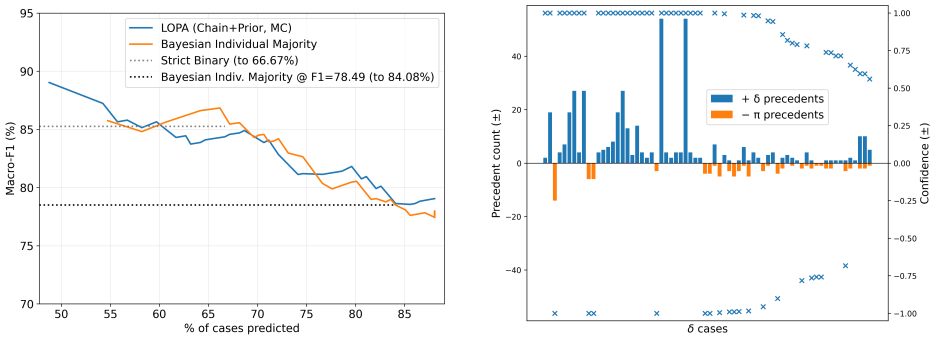

Figure 2. Confidence as reliability. (a) Macro–F1 versus coverage: as the confidence threshold increases, coverage drops and Macro–F1 typically rises (up to noise). (b) δ cases ordered by LOPA prediction confidence: bars show applicable precedents (upward for δ, downward for π); overlaid crosses mark model predictions; Markers above 0 indicate correct predictions; those below 0 indicate errors.

$$S = \sum_{i \in E_\pi} X_i - \sum_{j \in E_\delta} Y_j, \qquad X_i \mid \Theta_i \sim \text{Bernoulli}(\Theta_i),\ Y_j \mid \Theta_j \sim \text{Bernoulli}(\Theta_j),$$

then by the linearity of expectation $\mathbb{E}[S] = \sum_{i \in E_\pi} \mu_i - \sum_{j \in E_\delta} \mu_j$ depends only on the means $\mu_i = \mathbb{E}[\Theta_i]$. *LOPA* is instead *nonlinear* in the per–precedent parameters. Writing

$$\mathscr{O}_\pi(\Theta) = \left(\frac{\pi_0}{\delta_0}\right)^{1-K} \prod_{i \in E_\pi} \frac{\Theta_i}{1-\Theta_i} \prod_{j \in E_\delta} \frac{1-\Theta_j}{\Theta_j},$$

the side and confidence are computed from

$$p_\pi = \mathbb{E}\left[\frac{\mathscr{O}_\pi(\Theta)}{1+\mathscr{O}_\pi(\Theta)}\right],$$

which *cannot* be reduced to means. Thus, while the majority model's expected margin uses means, the LOPA intrinsically leverages the full Beta distributions for both *side* and *confidence*. In either case, the reported confidence is genuinely distributional and would not be available without the Beta posteriors.

Handling inconsistency. A criticism to Bayesian approaches is that they are "just probabilistic." But in inconsistent corpora, all models –including Strict Binary– face uncertainty. Our Bayesian model make this uncertainty explicit, enabling principled tradeoffs between coverage and reliability.

Confidence as a signal. We tested whether model confidence aligns with actual reliability by sweeping a confidence threshold τ and treat predictions with confidence $< \tau$ as abstentions. As τ increases, coverage decreases, but macro–F1 improves smoothly for both Bayesian models, showing confidence is informative. Figure 2(a) shows smooth, improving macro–F1 as coverage is reduced for *both* Bayesian models, indicating that confidence is informative. Using the coverage–F1 curves, we set the smallest confidence threshold τ that matches *Strict Binary*'s $\approx 66.6\%$ coverage; ranking cases by LOPA confidence. Fig. 2(b) then shows that many one–sided cases fall *below* the threshold while

several conflicted cases remain *above*. Thus, the model abstains on potentially misleading one–sided instances yet stays confident on some conflicted ones.

In summary, our LOPA uses per-precedent distributions to determine outcome and confidence. By treating inconsistency as uncertainty, it turns confidence into an informative signal: it not only explains why a prediction is made, but also how sure the system is, providing a principled way to balance coverage against reliability.

6. Conclusion

We introduced a LOPA method that enhances the result model with ML techniques to effectively handle inconsistent case bases. We compared it against three existing extensions and a strong ML baseline (tuned Random Forest). All consideredsymbolic models are on par with –and in two cases slightly better than– ML when evaluated with macro–F1. In particular, our new aggregator achieved a modest performance gain w.r.t. all other models, but its main advantages are explanatory:

1. **Resistance to swamping.** Strong precedents can outweigh many weak ones, avoiding the "majority always wins" problem.
2. **Focused precedent sets.** Redundant chains of precedents are pruned, leaving concise, human-scale justifications closer to judicial practice.
3. **Quantified uncertainty.** It produces calibrated confidence scores, enabling users to abstain from low-confidence cases while maintaining high accuracy elsewhere.

We conclude that *a fortiori* precedent reasoning is competitive with ML while offering superior transparency and actionable uncertainty, properties that are essential for trustworthy legal AI. To close, we outline directions for future research, which include:

Model generality and extensibility. We treat the *a fortiori* constraint as one admissible rule for reasoning with precedents but the framework could also incorporate further rules, e.g. adapting argument moves from [29]. In that setting, strategies such as *downplaying a distinction* or *emphasising a strength* could be encoded as rules analogous to *a fortiori* and learned in the same way. It is also noteworthy that, while our experiments used a simplified (degenerate) instantiation of the result model, the approach is readily adapted to the full result model and, mutatis mutandis, to the reason model.

The conditional–independence assumption underlying Naive Bayes is only partly mitigated by max pruning; richer dependence modeling is a natural next step.

Temporality. The DIAS data set lacks temporal metadata, so our experiments are time-agnostic; incorporating temporality is left for future work.

Dataset Bias. The dataset shows two biases: one from π being the majority class, addressed by our choice of macro–F1; another, more subtle, stems from the fact that many borderline cases are decided for π, rewarding models that lean toward the plaintiff, as gains on borderline cases outweigh the corresponding errors. In practice, the *Bayesian Individual Majority* model inherits this tilt, whereas our *LOPA* stays closer to balance. As a result, the LOPA appears only *slightly* better on DIAS, but we expect a clearer superiority on datasets without this π bias.

Acknowledgments: Work supported by WWTF [Grant ID: 10.47379/ICT23030].

References

[1] Alexander L. Constrained by Precedent'. Southern California Law Review. 1989;63:1.

[2] Horty JF. Rules and reasons in the theory of precedent. Legal theory. 2011;17(1):1-33.

[3] Horty JF, Bench-Capon TJM. A factor-based definition of precedential constraint. Artif Intell Law. 2012;20(2):181-214. Available from: https://doi.org/10.1007/s10506-012-9125-8.

[4] Horty JF. Modifying the reason model. Artif Intell Law. 2021;29(2):271-85. Available from: https://doi.org/10.1007/s10506-020-09275-z.

[5] Prakken H. A formal analysis of some factor- and precedent-based accounts of precedential constraint. Artif Intell Law. 2021;29(4):559-85. Available from: https://doi.org/10.1007/s10506-021-09284-6.

[6] Bench-Capon TJM. The Role of Intermediate Factors in Explaining Precedential Constraint. In: Grasso F, Green NL, Schneider J, Wells S, editors. Proceedings of (CMNA 2023). vol. 3614 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. p. 21-32.

[7] Gray M, Savelka J, Oliver W, Ashley K. Toward Automatically Identifying Legally Relevant Factors. In: Legal Knowledge and Information Systems. IOS Press; 2022. p. 53-62.

[8] Gray M, Savelka J, Oliver W, Ashley K. Using LLMs to Discover Legal Factors. In: Legal Knowledge and Information Systems. IOS Press; 2024. p. 60-71.

[9] Morello Y, Ciabattoni A. A Bayesian View of the Result Model. Submitted. 2025. Available from: https://www.logic.at/staff/agata/submitted2025.pdf.

[10] Canavotto I. Reasoning with inconsistent precedents. Artificial Intelligence and Law. 2023:1-30.

[11] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: formalisation and experiments. Argument & Computation. 2022;13(2):159-94.

[12] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.

[13] Peters JG, Bex FJ, Prakken H, et al. Justifications derived from inconsistent case bases using authoritativeness. In: Proceedings of ArgXAI 2022. vol. 3209. CEUR WS; 2022. p. 1-13.

[14] Peters JG, Bex FJ, Prakken H. Model-and data-agnostic justifications with a fortiori case-based argumentation. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law; 2023. p. 207-16.

[15] Thimm M. Inconsistency measures for probabilistic logics. Artificial Intelligence. 2013;197:1-24.

[16] De Bona G, Finger M. Measuring inconsistency in probabilistic logic: rationality postulates and Dutch book interpretation. Artificial Intelligence. 2015;227:140-64.

[17] Bruninghaus S, Ashley KD. Predicting outcomes of case based legal arguments. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law. ICAIL '03. Association for Computing Machinery; 2003. p. 233–242.

[18] van Woerkom W, Grossi D, Prakken H, Verheij B. A fortiori case-based reasoning: from theory to data. Journal of Artificial Intelligence Research. 2024;81:401-41.

[19] Horty JF. The result model of precedent. Legal Theory. 2004;10(1):19-31.

[20] Kaplan LM, Şensoy M, Tang Y, Chakraborty S, Bisdikian C, De Mel G. Reasoning under uncertainty: Variations of subjective logic deduction. In: Proceedings of the 16th International Conference on Information Fusion. IEEE; 2013. p. 1910-7.

[21] Kaplan L, Şensoy M, de Mel G. Trust estimation and fusion of uncertain information by exploiting consistency. In: 17th International Conference on Information Fusion (FUSION). IEEE; 2014. p. 1-8.

[22] Josang A, Ismail R. The beta reputation system. In: Proceedings of the 15th bled electronic commerce conference. vol. 5; 2002. p. 2502-11.

[23] Domingos P, Pazzani M. Beyond independence: Conditions for the optimality of the simple bayesian classi er. In: Proc. 13th Intl. Conf. Machine Learning; 1996. p. 105-12.

[24] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31(1):76-7.

[25] Schütze H, Manning CD, Raghavan P. Introduction to information retrieval. vol. 39. Cambridge University Press Cambridge; 2008.

[26] Gray M, Savelka J, Oliver W, Ashley K. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. Royal Society; 2024.

[27] Practitioner's Handbook for Appeals to the United States Court of Appeals for the Seventh Circuit; 2020. Revised through Sept. 28, 2020. Available from: `https://www.ca7.uscourts.gov/rules-procedures/Handbook.pdf`.

[28] Garner BA. Judges on Briefing: A National Survey. The Scribes Journal of Legal Writing. 2001;8. 2001–2002 issue. Available from: `https://scribes.org/wp-content/uploads/2022/12/Scribes_vol8_04_Judges_on_Briefing.pdf`.

[29] Bench-Capon T, Sartor G. A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence. 2003;150(1-2):97-143.