Neuro-Symbolic Legal Reasoning with Expert-Verified Customary Law Rules: A Pilot Study on Indonesian Legal Pluralism

Anonymous Authors
Project: Nusantara-Agent

Draft v0.6 – 2026-02-24

Abstract

We present *Nusantara-Agent*, a neuro-symbolic framework for legal reasoning in Indonesian legal pluralism, where national and customary (*adat*) norms can conflict. Our primary contribution is resource-oriented: an expert-verified customary-law rule base encoded in Answer Set Programming (ASP) across three domains (Minangkabau 25 rules, Bali 34 rules, Jawa 36 rules), together with an auditable pilot benchmark protocol. An LLM adjudication layer is used as an interface over symbolic outputs. On 74 dual-expert-labeled scenarios (70 evaluable, 4 disputed), we compare ASP-only (58.6%, Wilson 95% CI [0.469, 0.694]), ASP+Ollama/deepseek-r1 (64.3%, CI [0.526, 0.745]), and ASP+DeepSeek (68.6%, CI [0.570, 0.782]). Observed gains over ASP-only (+5.7 pp and +10.0 pp) are directionally positive but statistically inconclusive at pilot scale: all pairwise McNemar tests are non-significant ($p \geq 0.17$, $n = 70$). Inter-system agreement remains substantial (Fleiss' $\kappa = 0.638$), indicating stable rule-layer behavior across backends. We therefore position this work as a pilot benchmark and methodology paper, not a definitive efficacy claim; reaching statistical power $\geq$0.8 is estimated to require approximately 344 evaluable cases.

1 Introduction

Indonesia is a high-friction test bed for legal AI because legal decisions often involve legal pluralism: codified national law coexists with multiple customary (*adat*) normative systems. In such settings, legal reasoning is not only about retrieving relevant text. It must resolve conflicts across regimes, handle exceptions, and justify why one normative source is prioritized in a given case.

Cross-disciplinary context. This paper is intentionally written for three communities that rarely share technical language: legal scholars, customary-law and culture scholars, and computer scientists. We therefore clarify core terms early. In this paper, *Natural Language Processing (NLP)* means computational methods to process legal text; *Large Language Models (LLMs)* are neural models used here for text-level adjudication; and *Answer Set Programming (ASP)* is a symbolic logic formalism for explicit rule-based reasoning under defaults and exceptions [14, 7]. Our *neuro-symbolic* design combines these roles: neural components map text into usable representations, while symbolic components provide auditable legal constraints [6].

Why this matters for each audience. For legal readers, the core issue is doctrinal defensibility in plural-law conflicts, where opaque end-to-end neural predictions are difficult to audit. For customary-law and cultural readers, the core issue is representation fidelity: *adat* norms cannot

be treated as informal noise if they are normatively operative in practice [10, 3]. For computer-science readers, the core issue is methodological: current legal NLP benchmarks mostly assume one dominant legal regime, which underestimates conflict-handling complexity in plural systems.

Research gap. Prior legal NLP benchmarks focus mostly on monolithic legal systems and text-centric tasks such as judgment prediction, extraction, and summarization. Prior Indonesian legal-pluralism scholarship is rich in doctrinal and anthropological analysis but is typically not encoded as executable reasoning artifacts. The missing bridge is an auditable benchmark line that jointly provides: (1) expert-verified Indonesian *adat* rules, (2) formal executable encoding, and (3) transparent pilot evaluation under explicit governance constraints.

Contributions. This paper addresses that bridge as a resource-first neuro-symbolic study. We use ASP as the rule-centric reasoning core and treat the LLM layer as an adjudication interface. Our concrete contributions are:

- A formally encoded, expert-verified ASP rule base for three Indonesian customary-law domains: Minangkabau (25 rules), Bali (34 rules), and Jawa (36 rules). Of 95 verified rules, 71 are currently active in the reasoning engine.
- A pilot benchmark resource and reporting protocol over 74 dual-expert-labeled scenarios (70 evaluable agreed cases, 4 disputed), including manifest-based provenance and explicit claim gating.
- Cross-backend empirical evidence showing directionally positive but statistically inconclusive gains, plus substantial inter-system agreement (Fleiss' $\kappa = 0.638$), consistent with rule-layer anchoring.
- Transparent negative-result reporting, including the "Rule Over-specification Paradox", backend sensitivity, complete D-label failure (0% recall), and pilot-scale statistical limits.

Scope and limitations. Following the strategic pivot on 2026-02-12, this manuscript focuses on rule formalization, expert verification workflow, and pilot-scale evaluation. The current results are explicitly preliminary. First, statistical power is limited ($n = 70$ evaluable; all pairwise McNemar comparisons are non-significant, $p \geq 0.17$). Second, the same 70 evaluable cases have been used during development and final reporting, so a fully clean held-out test set is not yet available in this pilot cycle. Accordingly, we frame the paper as a benchmark-and-methodology contribution for legal pluralism, while deferring strong efficacy and generalization claims to future unseen-case evaluation.

Reader guide and paper roadmap. Readers primarily interested in legal and cultural framing may focus on Sections 2, 3, 5, and 8. Readers focused on computational design and evaluation may focus on Sections 4, 6, and 7. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 defines the multi-label classification task; Section 4 provides a system overview of the *Nusantara-Agent* architecture; Section 5 details the expert annotation protocol and dataset construction; Section 6 defines the evaluation metrics; Section 7 presents experimental results and error analysis; Section 8 discusses limitations of the pilot study; Section 9 provides data and code availability; and Section 10 concludes with future directions.

2 Related Work

This study sits at the intersection of legal pluralism scholarship, legal NLP, and neuro-symbolic AI. The core challenge is that each literature solves only part of the pipeline needed for auditable plural-law reasoning.

Legal pluralism and customary-law scholarship. Indonesian legal anthropology and doctrinal studies have long documented that national law and *adat* norms co-govern many real disputes [10, 3]. This literature explains normative coexistence, local legitimacy, and context-dependent conflict handling. However, it is mostly non-executable: the knowledge is rich for interpretation but difficult to run as machine-checkable reasoning rules.

Legal NLP in predominantly monolithic regimes. Mainstream legal NLP benchmarks have advanced tasks such as legal summarization, contract understanding, and judgment prediction [8, 4, 21, 15, 18, 29, 16, 17, 11, 12, 22, 23, 24, 25]. For computer-science readers, these works establish strong text modeling baselines. For legal readers, however, most of these datasets assume one dominant authority structure and therefore do not stress-test cross-regime legal conflict.

LLM reasoning and retrieval interfaces. The LLM and prompting literature shows strong gains in general reasoning and legal-text interaction [26, 2, 28, 27, 31, 30, 20]. Retrieval-based extensions improve grounding by injecting external context [13, 1, 5, 19, 9]. Yet in plural-law settings, retrieval and fluent generation are insufficient by themselves: the system must also enforce explicit conflict constraints and exception logic.

Neuro-symbolic reasoning for auditable legal decisions. Neuro-symbolic AI combines neural flexibility with symbolic control [6]. ASP, in particular, provides non-monotonic rule semantics that align with legal defaults, exceptions, and contradictions [14, 7]. This is attractive for legal and cultural stakeholders because reasoning traces are inspectable and norm references can be audited, but it still requires robust text-to-rule interfacing to operate on natural language cases.

Positioning and gap closure. The unresolved gap is the lack of an auditable benchmark-and-methodology line for Indonesian plural law that is legible to both legal and computational audiences. Nusantara-Agent addresses this by coupling expert-verified *adat* formalization (symbolic layer) with LLM adjudication interfaces (neural layer), and by reporting pilot evidence with explicit statistical and governance limits.

3 Task Definition

Given a legal scenario x, the system predicts one of four policy labels:

- **A**: mainly national-law resolution,
- **B**: mainly customary-law resolution,
- **C**: synthesis of national + customary law,
- **D**: clarification required.

Predictions are compared against expert labels with explicit agreement status.

4 System Overview

Nusantara-Agent is rule-centric: ASP formalization of expert-verified customary law is the primary modeling contribution. The LLM-based adjudication layer is used as a practical decision interface and is not the central novelty claim.

This design choice is motivated by three requirements specific to legal reasoning in pluralistic jurisdictions: (1) auditability—ASP rules provide explicit, inspectable reasoning traces that satisfy legal domain expectations for transparency; (2) non-monotonic reasoning—ASP's native support for defaults and exceptions accommodates the defeasible nature of customary law norms; and (3) deployment flexibility—the system can operate without API calls in offline mode using deterministic heuristics. Pilot evidence is consistent with this architecture: despite using different LLM backends (local 7B-parameter model versus commercial API), the three system configurations exhibit substantial inter-system agreement (Fleiss' $\kappa = 0.638$), suggesting that the ASP rule layer contributes strongly to shared behavior across backends.

4.1 Core Neuro-Symbolic Components

- **Keyword router and fact extraction** for domain routing and ASP fact grounding. This component maps natural language case descriptions to formal predicates, enabling the symbolic engine to reason over unstructured input without requiring an explicit knowledge graph.
- **ASP rule engine (Clingo)** for hard legal constraints and contradiction signals. Unlike pure retrieval-augmented systems, this engine performs logical inference to detect regime conflicts and apply non-monotonic defaults, distinguishing Nusantara-Agent from standard RAG pipelines.
- **Optional LLM adjudication** for final label synthesis; offline fallback for deterministic operation. The LLM layer serves as a trainable decision interface that can be removed entirely without breaking the pipeline, providing resilience against API unavailability or cost constraints.

Domain	Expert-Verified Rules	ASP Rule File	Scope
Minangkabau	25	`minangkabau.lp`	Matrilineal inheritance and pusako norms
Bali	34	`bali.lp`	Purusa/sentana and druwe classifications
Jawa	36	`jawa.lp`	Bilateral inheritance and gono-gini patterns
Total	95	3 domains	Expert-verified customary law base

Table 1: Expert-verified customary-law rules. Of 95 total, 71 are currently encoded in ASP; the remaining 24 were found to cause accuracy regression when added (see Section 7.1).

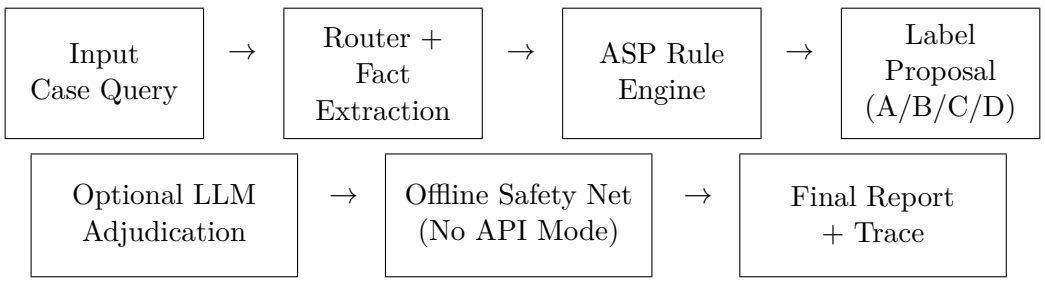

Figure 1: High-level architecture of Nusantara-Agent.

5 Dataset and Expert Protocol

5.1 Current Snapshot (2026-02-19)

The benchmark dataset was constructed in two phases: an initial batch of 24 cases and an expanded batch of 50 new cases. Each case was independently labeled by two qualified legal experts (Ahli-1: Dr. Hendra Kusuma, S.H., M.Hum.; Ahli-2: Dr. Indra Gunawan, S.H., M.H.).

Item	Value
Total cases in benchmark	74
Evaluable agreed cases	70
Disputed cases (excluded from accuracy)	4
Initial batch agreement (24 cases)	58.3% (Cohen's $\kappa = 0.394$)
Expanded batch agreement (50 cases)	94.0% (47/50)
Benchmark manifest source	`data/benchmark_manifest.json`

Table 2: Operational dataset status.

Gold Label	Count	Percentage
A	6	8.6%
B	31	44.3%
C	31	44.3%
D	2	2.9%

Table 3: Label distribution in the evaluable benchmark subset (N=70).

Agreement Metric	Value
Initial batch agreed/total	14/24 (58.3%)
Expanded batch agreed/total	47/50 (94.0%)
Adjudicated (initial disputed)	9/10 resolved
Remaining disputed	4/74 (5.4%)
Accuracy reporting policy	Computed only on agreed subset

Table 4: Agreement profile across annotation phases.

5.2 Rubric refinement

Initial inter-rater agreement in the 24-case batch was moderate ($\kappa = 0.394$; 14/24 agreement), with recurring ambiguity at the A/B/C decision boundary when adat context and national-law triggers co-occurred. The annotation protocol was then tightened in stages: (i) locked A/B/C/D label definitions, (ii) explicit boundary tests for A vs. C, and (iii) a mandatory decision checklist (dominance, dual-requirement, missing-fact check) to justify why synthesis is or is not required. The expanded 50-case packet applied this refined rubric on a disjoint GS case pool and explicitly instructed raters not to inspect prior labels. In the current dataset snapshot, Ahli-1 vs. Ahli-2 agreement on the expanded GS pool is 47/50 (94.0%), with only three remaining

disagreements (GS-0022, GS-0023, GS-0032), all at the B/C boundary and currently marked as disputed. A complete audit trail of this refinement process is maintained in the project methodology log (`docs/methodology/rubric_refinement_log.md`). The same log also records project-level governance sign-off for the rubric protocol.

5.3 Adjudication policy

For the current evaluation:

- only agreed cases are used for quantitative accuracy reporting,
- disputed cases are excluded and queued for third-expert tiebreak,
- benchmark manifest is treated as the reporting authority for snapshot counts.

6 Metrics and Equations

To avoid ambiguous reporting, we explicitly define pilot metrics.

6.1 Agreement Coverage

Let N be total active cases and N_a the agreed/evaluable subset.

$$\text{ConsensusStrength} = \frac{N_a}{N} \tag{1}$$

For the current snapshot, $(N_a, N) = (70, 74)$, so:

$$\text{ConsensusStrength} = \frac{70}{74} = 0.946 \tag{2}$$

6.2 Dispute Rate

Let N_d be the number of disputed cases:

$$\text{TieRate} = \frac{N_d}{N} \tag{3}$$

With $N_d = 4$ and $N = 74$, TieRate $= 0.054$.

6.3 Binomial Confidence Interval (Wilson)

For observed accuracy $\hat{p}$ with sample size n and $z = 1.96$ (95% CI), Wilson interval is:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{4}$$

This is used to report uncertainty on small pilot datasets and avoid over-claiming.

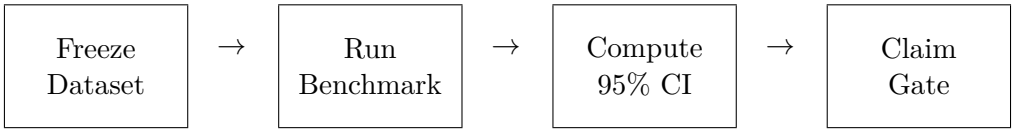

Figure 2: Evaluation protocol and claim gate used for pilot reporting.

7 Experiments and Results

7.1 Ablation: ASP-only vs ASP+LLM (2026-02-19)

We compare two operating modes on the 70 evaluable cases:

- **ASP-only**: keyword router → ASP rule engine → offline heuristic supervisor (no LLM calls).
- **ASP+LLM**: keyword router → ASP rule engine → LLM adjudication via deepseek-r1 (local Ollama).

Mode	Accuracy	Wilson 95% CI	Correct / Total
ASP-only (offline)	58.6%	[0.469, 0.694]	41 / 70
ASP+Ollama (deepseek-r1)	**64.3%**	[0.526, 0.745]	45 / 70
ASP+DeepSeek (API)	68.6%	[0.570, 0.782]	48 / 70

Table 5: Ablation comparison on the 70-case evaluable benchmark.

The LLM layer shows an observed improvement of 5.7 pp (Ollama) and 10.0 pp (DeepSeek) over the symbolic-only baseline. However, as reported in Section 7.4, all pairwise McNemar tests are non-significant ($p \geq 0.17$, $n = 70$); these observed differences should be interpreted as pilot observations pending larger-scale evaluation. Results are reported with `temperature=0` for reproducibility. An earlier non-deterministic run (`temperature=1.0`) achieved 70.0%, illustrating that LLM variance is a significant confound in small-n evaluations.

Rule over-specification. Increasing ASP rule coverage from 71 to 95 rules (100% coverage) paradoxically decreased ASP+LLM accuracy by 7.1 percentage points. Analysis revealed that additional symbolic facts biased the LLM toward adat-dominant (B) classification, causing conflict cases (gold=C) to be misclassified as B. This suggests that neuro-symbolic integration requires careful calibration of symbolic information exposure rather than maximizing rule coverage.

	ASP-only			ASP+LLM		
Label	P	R	Support	P	R	Support
A	0.36	0.67	6	0.36	0.83	6
B	0.66	0.61	31	0.75	0.77	31
C	0.62	0.58	31	0.83	0.65	31
D	0.00	0.00	2	0.00	0.00	2

Table 6: Per-label precision (P) and recall (R) for both modes on the 70-case benchmark.

7.2 Extended Error Analysis: The 12 Hard Cases

Beyond the aggregate error patterns, we conduct a deeper analysis of **12 cases (17.1%)** where all three system variants—ASP-only, ASP+Ollama, and ASP+DeepSeek—fail simultaneously. These "hard cases" reveal systematic limitations in the current rule set and router design.

7.2.1 Gold label distribution in hard cases

Table 7 shows that conflict cases (C) are over-represented in failures, comprising 50.0% of hard cases versus 44.3% overall. This pattern suggests that detecting implicit national–adat conflicts is

likely one of the primary challenges in the current benchmark.

Gold Label	Hard Cases	Overall	Ratio
C (Conflict)	6 (50.0%)	31 (44.3%)	1.13×
B (Adat)	5 (41.7%)	31 (44.3%)	0.94×
A (National)	1 (8.3%)	6 (8.6%)	0.97×
D (Unclear)	0 (0.0%)	2 (2.9%)	0.00×

Table 7: Label distribution in 12 hard cases versus overall benchmark (N=70).

7.2.2 Dominant failure patterns

Table 8 categorizes the failure modes. The dominant pattern is **C→B misclassification** (33.3% of hard cases): conflicts where ethnic keywords dominate the query text, causing the symbolic router to overwhelm conflict signals.

Pattern	Description	Count	Percentage
C→B	Conflict misclassified as Adat	4	33.3%
B→A	Adat misclassified as National	3	25.0%
B→C	Adat misclassified as Conflict	2	16.7%
C→A	Conflict misclassified as National	2	16.7%
A→C	National misclassified as Conflict	1	8.3%

Table 8: Failure patterns in 12 hard cases where all three systems fail.

The prevalence of C→B errors suggests that when customary-law keywords (e.g., "Minangkabau," "pusako," "gono-gini") appear prominently, the keyword router defaults to the pure-adat pathway even when national-law dimensions are implicitly present. This is particularly evident in Minangkabau cases involving joint property (*harta bersama*), where ownership disputes span both customary inheritance norms and national civil code provisions.

7.2.3 Domain-specific challenges

Minangkabau cases are over-represented in failures (33.3% of hard cases), particularly around:

- *Harta pusako* (inherited property) classification,
- Inter-ethnic marriage property regimes,
- Joint business property ownership disputes.

This reflects the complexity of matrilineal inheritance rules and their intersection with national civil law, which the current ASP rule base captures incompletely.

7.2.4 D-label complete failure

All three systems fail on both D ("clarification required") cases in the benchmark (0/2 recall). The system consistently assigns substantive labels (A, B, or C) even when the scenario contains insufficient information for legal determination. This indicates that **uncertainty detection** is not supported by the current rule set or prompt design—a critical gap for real-world deployment where refusing to classify is often the correct response.

7.2.5 Root cause summary

The fact that all three systems fail identically on these 12 cases suggests that the dominant error sources are in the **shared ASP rule set** and **router/prompt design**, rather than in backend-specific differences. The primary failure modes are:

1. **Implicit conflict detection**: When conflict keywords are subtle (e.g., "mendahului," "tidak sah") and ethnic keywords dominate, the router misses the conflict signal.
2. **Domain ambiguity**: "General" cases without clear domain markers default to national-law pathways, missing adat dimensions.
3. **No abstention mechanism**: The system cannot output D even when confidence is low.

Two-layer error diagnosis. A finer-grained audit of C→B errors shows two distinct failure layers. Across 29 total C→B events (ASP-only=10, ASP+Ollama=10, ASP+DeepSeek=9), 23/29 (79.3%) occur when no ASP conflict signal is produced, indicating an upstream router/fact-extraction failure rather than an LLM reasoning failure. The remaining 6/29 events (20.7%) occur despite an ASP conflict signal, meaning the adjudicator still collapses to label B; this pattern is concentrated in DeepSeek (5/6 events). Therefore, mitigation should be split into two tracks: improve router keyword-to-fact mapping for conflict recall, and recalibrate adjudicator prompts for conflict-preserving decisions, both without adding ASP rules (to avoid repeating the F-018 regression).

7.3 Per-Domain Performance Analysis

Table 9 breaks down accuracy by customary law domain. Two patterns are noteworthy.

Domain	N	ASP-only	ASP+Ollama	ASP+DeepSeek
Minangkabau	21	71.4%	76.2%	71.4%
Bali	21	71.4%	81.0%	76.2%
Jawa	17	35.3%	29.4%	52.9%
Nasional	7	42.9%	71.4%	71.4%
Lintas	4	50.0%	50.0%	50.0%
Overall	70	58.6%	64.3%	68.6%

Table 9: Per-domain accuracy across three system configurations.

LLM helps Bali and Nasional, hurts Jawa. ASP+Ollama shows larger observed gains over ASP-only for Bali (+9.6pp) and Nasional (+28.5pp) (domain-level McNemar not computed due to small per-domain n), but **decreases** accuracy for Jawa (35.3% → 29.4%, −5.9pp). This is a counter-intuitive negative finding: adding an LLM adjudication layer makes Jawa cases *worse*. Analysis of Jawa errors reveals that bilateral inheritance rules (gono-gini, sigar semangka) are frequently misclassified as national-law (B→A), suggesting the LLM over-weights civil code terminology present in Jawa case descriptions. DeepSeek partially recovers this deficit (52.9%), indicating LLM capability is a factor, but the underlying routing ambiguity for Jawa remains unresolved.

7.4 Cross-Validation Across LLM Backends

To assess whether results are robust to LLM choice, we evaluate three operating configurations on the identical 70-case benchmark:

- **ASP-only**: offline heuristic supervisor (no LLM calls);
- **ASP+Ollama**: deepseek-r1 via local Ollama;
- **ASP+DeepSeek**: DeepSeek-Chat via API.

7.4.1 Accuracy comparison

Table 10 reports accuracy and agreement metrics. ASP+DeepSeek achieves the highest accuracy at 68.6%, followed by ASP+Ollama at 64.3%, with ASP-only at 58.6%. To maintain methodological comparability, Table 10 reports only the canonical three-configuration snapshot (2026-02-20); additional exploratory backends from 2026-02-23 are discussed as sensitivity checks in Section 8.

Mode	Accuracy	Wilson 95% CI	Cohen's κ	Agreement
ASP-only (offline)	58.6%	[0.469, 0.694]	0.331	Fair
ASP+Ollama (deepseek-r1)	64.3%	[0.526, 0.745]	0.418	Moderate
ASP+DeepSeek (API)	**68.6%**	[0.570, 0.782]	0.483	Moderate

Table 10: Cross-validation across three LLM backends (N=70 evaluable cases). All runs use identical rule state (post-rollback, 71 rules) and `temperature=0`.

7.4.2 Statistical significance

Table 11 presents pairwise McNemar tests. **None of the differences reach statistical significance** at $\alpha = 0.05$: ASP-only vs. Ollama yields $p = 0.344$, ASP-only vs. DeepSeek yields $p = 0.167$, and Ollama vs. DeepSeek yields $p = 0.549$. The largest effect is ASP-only vs. DeepSeek (19 discordant pairs, DeepSeek favored by +7), limited by small sample size ($n = 70$).

Comparison	p-value	Discordant	Favored
ASP-only vs. Ollama	0.344	10	Ollama (+4)
ASP-only vs. DeepSeek	0.167	19	DeepSeek (+7)
Ollama vs. DeepSeek	0.549	11	DeepSeek (+3)

Table 11: McNemar exact binomial test results for pairwise accuracy differences (N=70).

7.4.3 Inter-system agreement

Despite using different LLM backends, the three systems exhibit **substantial agreement** with each other. Fleiss' $\kappa = 0.638$ indicates that the ASP+LLM architecture produces consistent predictions regardless of the specific LLM used. Table 12 breaks down cross-model agreement patterns.

The high unanimous agreement (67.1%) with only 2.9% complete disagreement is consistent with the symbolic rule layer acting as a **strong prior** across backends. In pilot terms, this is a reproducibility-positive signal: system behavior appears more anchored by expert-verified ASP rules than by backend-specific variation.

7.4.4 Majority vote analysis

For cases where two systems agree and one differs, majority vote achieves only **52.4% accuracy** (11/21). This indicates that when systems disagree, the majority is **no more likely to be correct**

Agreement Pattern	Count	Percentage
All 3 agree (correct)	34	48.6%
All 3 agree (wrong)	13	18.6%
All 3 agree (total)	**47**	**67.1%**
2 agree, 1 different	21	30.0%
All 3 different	2	2.9%

Table 12: Cross-model agreement among ASP-only, Ollama, and DeepSeek (N=70).

than random chance. Consequently, simple ensemble strategies offer no improvement over single-system predictions in this setting.

7.4.5 Implications

The cross-validation results provide pilot-scale evidence for the following observations:

1. **Reproducibility feasibility**: In this pilot, a local Ollama backend (deepseek-r1) reaches performance close to a commercial API (DeepSeek), with a 4.3 percentage point gap that remains non-significant. This is feasibility evidence for resource-constrained deployment, not a definitive parity claim.
2. **Rule-centric design hypothesis**: The substantial inter-system agreement (Fleiss' $\kappa = 0.638$) is consistent with the ASP rule layer acting as a dominant anchor across backends. At current scale, this should be interpreted as directional support for a rule-centric design, not as a causal proof.
3. **Exploratory backend sensitivity**: Two additional local Ollama backends run after the canonical freeze show divergent behavior under the same rule state and prompt. ASP+gpt-oss:20b reaches 45/70 (64.29%), matching the canonical local Ollama/deepseek-r1 result (45/70), while ASP+Qwen3-14B drops to 38/70 (54.29%). Against their same-day ASP-only paired baseline (42/70), these correspond to +4.29 pp (McNemar $p = 0.646$) and −5.71 pp ($p = 0.480$), respectively, indicating that neuro-symbolic gains remain backend-dependent.
4. **Open-source B→A bias**: gpt-oss:20b and Qwen3-14B share 6 identical B→A failures (GS-0010, GS-0014, GS-0019, GS-0020, GS-0021, GS-0031), where strong national-law signals in ASP traces are over-interpreted as national-law dominance despite gold label B. This repeated pattern was not observed with the same concentration in the DeepSeek API run, suggesting backend calibration differences in legal conflict handling.

8 Limitations

Despite the observed performance improvements in the neuro-symbolic configurations, several critical limitations must be acknowledged to frame the validity of this pilot study correctly.

Statistical Power and Sample Size. The primary limitation of this study is the small sample size ($n = 70$ evaluable cases). While the ASP+DeepSeek configuration showed a +10.0 percentage point (pp) improvement over the ASP-only baseline, this difference is statistically inconclusive at the current scale. All pairwise McNemar tests yielded non-significant results ($p \geq 0.17$), reflecting a lack of statistical power rather than a confirmed absence of effect. Based on the observed number of discordant pairs, we estimate the current study's statistical power to be approximately 0.3 for a

10pp difference. To achieve a more robust power of ≥ 0.8 at the observed effect sizes, we estimate that approximately 344 evaluable cases would be required. Consequently, the findings presented here should be treated as preliminary pilot observations.

Development Contamination and Overfitting. The 70 evaluable cases were utilized for both prompt tuning and final evaluation. During the development phase, iterative refinement of the LLM adjudication layer saw accuracy rise from 54.0% to 68.6%. This introduces a risk of overfitting to the specific scenarios in the pilot benchmark. At the time of writing, a completely held-out test set has not yet been established. Future work must validate these results on a larger, unseen corpus to ensure that the neuro-symbolic integration generalizes beyond the development set.

Inter-rater Agreement and Rubric Refinement. The jump in inter-rater agreement from 58.3% in the first batch (24 cases, Cohen's $\kappa = 0.394$) to 94.0% in the second batch (50 cases) requires methodological transparency. This improvement was achieved through a mid-study rubric refinement process designed to resolve ambiguities at the boundary between customary law (B) and national–customary conflict (C). While this ensured a high-quality gold standard for the pilot, it also indicates that the classification task is highly sensitive to the definitions provided to the experts. The high agreement in the second batch may partially reflect the raters' alignment with a more prescriptive rubric rather than an inherent clarity in the legal scenarios themselves.

Rule Over-specification and Parity Gaps. Our experiments revealed a "Rule Over-specification Paradox" (Failure F-018): increasing the ASP rule coverage from 71 to 95 rules paradoxically decreased system accuracy by 7.1 pp. The additional rules were found to be too general, causing the ASP engine to over-report customary law dominance in conflict cases. This suggests that for neuro-symbolic systems, maximizing rule coverage is secondary to ensuring rule precision and proper calibration between symbolic and neural layers. Furthermore, parity audits (F-015, F-016, F-017) indicate that the current ASP rule base focuses primarily on "hard" inheritance constraints and lacks coverage for procedural and contemporary customary norms, particularly in the Jawa domain.

Backend Dependence in Local Deployment. Additional post-freeze backend runs indicate that local performance is sensitive to model calibration even under the same neuro-symbolic interface. While canonical ASP+Ollama (deepseek-r1) and exploratory ASP+gpt-oss:20b both reached 45/70 (64.29%), ASP+Qwen3-14B dropped to 38/70 (54.29%). All paired McNemar comparisons against their same-day ASP-only baseline remained non-significant (gpt-oss: $p = 0.646$; Qwen3: $p = 0.480$), so these results should be interpreted as directional evidence rather than confirmed effects. The repeated B→A overlap across open-source backends (6 shared cases: GS-0010, GS-0014, GS-0019, GS-0020, GS-0021, GS-0031) suggests a structural prompt-calibration issue when national-law cues co-occur with adat-governed outcomes.

Failure in Uncertainty Detection and Domain-Specific Weaknesses. The system exhibited a complete failure in uncertainty detection (D-label), with 0% recall across all configurations. The extreme class imbalance (only 2 D-label cases) and the lack of an explicit "abstention" mechanism in the prompt design prevented the models from identifying scenarios where clarification was required. Additionally, the Jawa domain remains a significant weakness, with an ASP-only accuracy of only 35.3%. The LLM layer frequently misclassified Jawa customary rules as national law (B→A errors), suggesting that the models over-weight Civil Code terminology present in the

case descriptions. These domain-specific disparities indicate that the framework's effectiveness is currently inconsistent across different Indonesian customary systems.

Scope and Generalizability. All benchmark cases are in Indonesian, and the ASP rule base encodes customary law from three specific regional systems (Minangkabau, Bali, Jawa). Generalization to other legal pluralism contexts—including other Indonesian customary systems, multilingual settings, or non-Indonesian jurisdictions—is not established by this pilot study.

9 Data and Code Availability

The dataset and source code are publicly available at:

`https://github.com/neimasilk/nusantara-agent.git`

10 Conclusion and Next Steps

This paper contributes a pilot benchmark-and-methodology resource for AI reasoning in Indonesian legal pluralism: an expert-verified customary-law rule base encoded in ASP, an auditable neuro-symbolic pipeline, and transparent reporting of both positive and negative findings. On the current 70-case evaluable benchmark, ASP+LLM configurations show observed accuracies of 64.3%–68.6% versus 58.6% for ASP-only (directional gain: +5.7 pp to +10.0 pp). These differences remain statistically inconclusive at pilot scale: all pairwise McNemar tests are non-significant ($n = 70$, $p \geq 0.17$), and power is limited. Substantial inter-system agreement (Fleiss' $\kappa = 0.638$) is consistent with a rule-layer anchoring effect, but should be interpreted as preliminary evidence pending unseen-case validation. Immediate next steps are:

1. expand the benchmark to at least 100 dual-expert-labeled cases to achieve sufficient statistical power for McNemar testing ($\geq$0.8 at the observed effect sizes);
2. establish a clean held-out test set from new Ahli-2 annotations (post-2026-02-23) to validate generalization without development contamination;
3. redesign router and prompt to address implicit conflict detection failures (C$\rightarrow$B errors) and enable D-label abstention;
4. resolve the 4 remaining disputed cases via Delphi Round 2 adjudication and document the rubric refinement process for methodological transparency.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[3] Peter Burns. *The Leiden Legacy: Concepts of Law in Indonesia*. KITLV Press, Leiden, 2004.

[4] Pierre Colombo, Pedro da Costa, Martin Müller, Karolina Stanczak, et al. Saullm-7b: A pioneering language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

[5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kang Jia, Jinliu Pan, Yuxian Bi, Yixiang Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[6] Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

[7] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Clingo = ASP + control: Preliminary report. *Theory and Practice of Logic Programming*, 14(4-5):1–5, 2014.

[8] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*, 2023.

[9] Xiaoxin He et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*, 2024.

[10] Michael Barry Hooker. *Adat Law in Modern Indonesia*. Oxford University Press, Kuala Lumpur, 1978.

[11] Chih-Ping Hou, Siddhant Vashishtha, Chia-Hsuan Chen, Alton Chua, Daman Arora, and Parminder Bhatia. Gaps or hallucinations? NLI metrics for legal LLM summarization. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 54–61, 2024.

[12] Sunhee Kwak, Chris Baraniuk, Kabir Agarwal, Julian Cristia, Dexter Hall, Bowen Pei, Rodrigo Acuna-Agost, and Christos Christodoulopoulos. Classify first then extract: A hybrid approach for improving information extraction from commercial lease agreements. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 62–68, 2024.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[14] Vladimir Lifschitz. *Answer Set Programming*. Springer, Berlin, Heidelberg, 2019.

[15] Sijia Liu, Sanjeev Gupta, Kartik Agarwal, Kunal Kalia, Prakhar Saran, Ramnath Surendran, Saptarshi Karmakar, Suman Das, Pawan Mishra, and Soumya Kundu. Enhancing legal case outcome prediction with LawNeo and mistralAI: A comparative study. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 8–13, 2024.

[16] Pratik Mali, Rodrigo Acuna-Agost, Zhengyi Luo, Donald Braun, Leila Wehbe, Christos Christodoulopoulos, et al. An information extraction system for contracts by leveraging large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 33–45, 2024.

[17] Tejas Narendra, Prashant Tiwari, Vishesh Kedia, Aviral Jindal, Sahil Abrol, and Chandan K. Reddy. Enhancing contract negotiations with fine-tuned legal bert and retrieval-augmented generation. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 46–53, 2024.

[18] Naman Nigam, Ayush M C, Sriniketh K, Rupal K, Vinayak Kakkar, Arshia Kaur, and Danish Pruthi. Rethinking legal judgement prediction in the era of large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 14–24, 2024.

[19] Parth Sarthi, Salman Abdullah, Aman Tuli, Mohit Khurana, Anjan Goyal, et al. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*, 2024.

[20] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

[21] Rianne Sie, Tom Mertens, Henk den Ouden, Erman Erdem, and Frank Schilder. Summarizing long regulatory documents with LLMs. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 2–7, 2024.

[22] Oleg Taranukhin, Prerna Chhikara, Solanki Singh, Nida Shahzad, Mikhail Tokarev, Evgeniya Mineeva, Ahmad Chokri, Dave Kauchak, Smaranda Muresan, and Surya Kallumadi. Empowering air travelers through generative AI: A study of a domain-specific RAG chatbot. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 69–78, 2024.

[23] Andy Tran, Nischal Mainali, Ana Lucic, Bas Terpstra, Eduard Vosselman, Henk Daniels, Jaap van den Herik, Jaap Visser, Geertje Scharnberg, Sina Trapp, Henk Geerlings, Antal van den Bosch, Kazi Sohag Bukhari, and Ricardo van den Ham. When DeBERTa beats behemoths: A comparative study of smaller language models and large language models for legal text classification in low-resource and privacy-focused environments. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 79–86, 2024.

[24] Santosh T.y.s.s, Jacob Knab, Andrew Zachary, Seohyeong Lee, Andrew Hsi, Mihai Surdeanu, Siva Reddy, et al. Lexsumm and lext5: Benchmarking and summarizing legal briefs in bankruptcy cases. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 227–236, 2024.

[25] Santosh T.y.s.s, Siva Reddy, Mihai Surdeanu, Seohyeong Lee, Andrew Hsi, et al. Towards supporting legal argumentation with natural language processing and graph representation learning. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 237–243, 2024.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[27] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[29] Yujie Xie, Jose Camacho-Collados Fernandes, Nikolaos Aletras, Lawrence M. Solan, and Han Xu. CLC-UKET: Explainable charge prediction for UK crown court cases with judgment

documents. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 25–32, 2024.

[30] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[31] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.