# Nusantara-Agent: A Pilot Neuro-Symbolic Multi-Agent Framework for Pluralistic Legal Reasoning in Indonesia

Anonymous Authors
Project: Nusantara-Agent

Draft v0.1 – 2026-02-09

**Abstract**

We present *Nusantara-Agent*, a pilot neuro-symbolic multi-agent system for legal reasoning under plural legal regimes (national and customary law) in Indonesia. The system combines LLM-based role agents, symbolic constraints, and structured expert adjudication. Current evidence should be interpreted as **pilot-scale**: the active evaluation set has 24 cases, and the tie-cases have been closed via independent final arbitration. We release this draft to stabilize methodology and reporting structure, not to claim final generalization performance. **This draft includes architecture diagrams, protocol flow diagrams, formal metric equations, and up-to-date operational tables for transparent reporting.**

## 1 Introduction

Plural legal reasoning in Indonesia often requires reconciling national law and multiple customary law systems. This setting is relevant for NLP and AI because legal outcomes depend on cross-source conflict handling, not only textual retrieval. Our goal is to build an auditable framework that can:

1. separate national and customary perspectives,
2. synthesize competing norms under explicit rules, and
3. document disagreement and uncertainty instead of forcing brittle single-label predictions.

**Current scope.** This manuscript is a **draft preprint candidate** focused on architecture, protocol, and pilot evidence. Final publication-grade claims are deferred until arbitration and dataset promotion are complete.

## 2 Related Work

This work is related to retrieval-augmented generation, multi-agent reasoning, and neuro-symbolic AI. Foundational LLM and prompting work includes Transformers, few-shot scaling, chain-of-thought, and tool-using reasoning [24, 2, 26, 25, 29, 28, 17]. Recent retrieval-centered directions include classical RAG, self-reflective RAG, and hierarchical retrieval designs [10, 1, 4, 16, 6]. For legal-domain NLP, we ground discussion on recent benchmark and workshop studies in legal judgment prediction, contract IE, legal summarization, and legal-domain deployment settings [5, 3, 18, 11, 14, 27, 12, 13, 7, 9, 19, 21, 22, 23].

# 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 judged against expert-voted labels with explicit consensus status (unanimous, majority, tie).

# 4 System Overview

Nusantara-Agent uses a sequential multi-agent graph:

1. National-Law Agent,
2. Customary-Law Agent,
3. Supervisor/Adjudicator Agent.

The pipeline is augmented with symbolic constraints and offline fallbacks for deterministic operation.

## 4.1 Neuro-Symbolic Components

- LLM agents for perspective-specific analysis.
- Symbolic rules for hard legal constraints and contradiction checks.
- Router and safety-net heuristics for offline control.

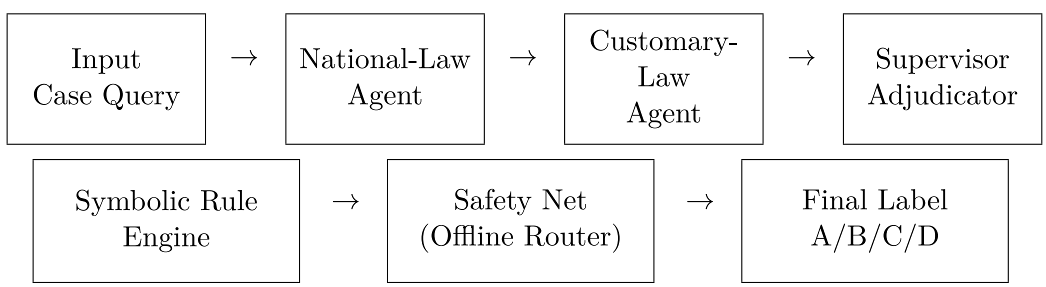


Figure 1: High-level architecture of Nusantara-Agent.

# 5 Dataset and Expert Protocol

## 5.1 Current Snapshot (2026-02-09)

## 5.2 Adjudication policy

For this draft cycle:

- clear majority mismatches were patched,
- tie cases are being arbitrated by an independent final expert,
- all changes are logged with versioned manifests.

| Item | Value |
|---|---|
| Active cases in benchmark file | 24 |
| Cases with expert-4 vote | 16/24 |
| Gold labels with SPLIT value | 0 |
| Remaining tie cases for final arbitration | 0 |
| Reference claim count (document) | 82 |
| Actual cases in active benchmark file | 24 |

Table 1: Operational dataset status used in this draft.

| Gold Label | Count | Percentage |
|---|---|---|
| A | 7 | 29.2% |
| B | 4 | 16.7% |
| C | 13 | 54.2% |
| D | 0 | 0.0% |

Table 2: Gold-label distribution in the current active dataset (N=24).

# 6 Metrics and Equations

To avoid ambiguous reporting, we explicitly define pilot metrics.

## 6.1 Consensus Strength

Let $N$ be the total number of cases, $N_u$ unanimous cases, and $N_m$ majority cases.

$$\text{ConsensusStrength} = \frac{N_u + N_m}{N} \tag{1}$$

For the current snapshot, $(N_u, N_m, N) = (4, 20, 24)$, so:

$$\text{ConsensusStrength} = \frac{4 + 20}{24} = 1.000 \tag{2}$$

## 6.2 Tie Rate

Let $N_t$ be the number of tie cases:

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

With $N_t = 0$ and $N = 24$, TieRate $= 0.000$.

## 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.

| Consensus Status | Count | Percentage |
|---|---|---|
| Unanimous | 4 | 16.7% |
| Majority | 20 | 83.3% |
| Tie (2-2) | 0 | 0.0% |

Table 3: Consensus profile after follow-up ingestion, arbiter vote, and final label patching.

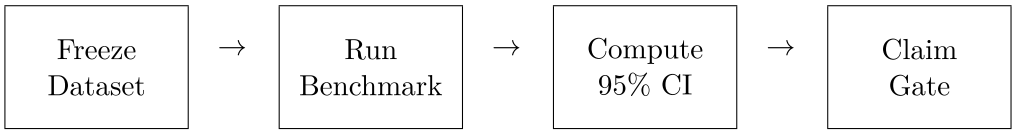


Figure 2: Evaluation protocol and claim gate used for pilot reporting.

# 7 Case-Level Arbitration Resolution

The four previously tied cases have been resolved by a final independent arbiter.

| Case ID | Vote Pattern (A1,A2,A3,A4) | Arbiter Vote | Final Gold |
|---|---|---|---|
| CS-LIN-052 | D, C, D, C | C | C |
| CS-LIN-017 | A, C, C, A | A | A |
| CS-BAL-014 | B, C, B, C | B | B |
| CS-LIN-016 | C, A, C, A | A | A |

Table 4: Resolved tie cases after final arbiter input.

# 8 Experiments and Preliminary Results

## 8.1 What is currently established

- Core deterministic suite passes in current environment (26/26 tests); full symbolic/PDF suite is blocked by missing optional dependencies (`clingo`, `fitz`).
- Historical pilot numbers on earlier dev subset (N=22): 72.73% (LLM mode), 59.09% (offline fallback).
- Post-patch reproducible offline benchmark on final active set (N=24): 41.67% (10/24), Wilson 95% CI $\approx$ [0.245, 0.612].
- These numbers are **not** treated as final or publication-grade generalization metrics.

## 8.2 What is intentionally deferred

- controlled LLM-mode post-arbitration benchmark on the same frozen labels,
- robust held-out evaluation with adequate sample size,
- statistical claims beyond pilot confidence intervals.

# 9 Threats to Validity and Limitations

1. **Sample size**: current N is too small for strong generalization claims.

2. **Label dynamics**: recent expert updates changed gold labels on critical cases.
3. **Dataset promotion gap**: reference claim and active benchmark count are not yet aligned.
4. **Infrastructure variance**: offline and LLM modes yield different behavior profiles.

## 10 Preprint Readiness Statement

At the moment, this work is **ready for an internal/working preprint draft**, but **not yet ready for a strong public NLP preprint claim** centered on performance superiority. It can be shared as a pilot systems paper if the title, abstract, and conclusions explicitly state:

- pilot-scale evidence,
- completed arbitration on the active set with explicit limitations,
- no final generalization claim.

## 11 Data and Code Availability

The dataset and source code are publicly available at:

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

## 12 Conclusion and Next Steps

Nusantara-Agent shows a feasible direction for auditable plural-law reasoning with neuro-symbolic controls. Immediate next steps are:

1. freeze the post-arbitration dataset version as reproducible release candidate,
2. run controlled LLM-mode and ablation benchmarks on the same frozen set,
3. report confidence intervals and claim boundaries consistently.

## 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] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch Roux, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gyorgy Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[16] 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.

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

[18] 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.

[19] 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.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.