Glass-Box Arbitrators: An Explainable Neuro-Symbolic AI Framework for International Commercial Arbitration Proceedings

David Scott Lewis and Haley Yi
AIXC
research@aiexecutiveconsulting.com

Abstract

Large language models (LLMs) are rapidly entering legal workflows, from contract review to dispute resolution, yet recent empirical work shows that state-of-the-art models provide unstable and logically inconsistent answers even on carefully specified legal questions, undermining their suitability for decisions that must be reproducible, appeal-proof, and auditable. In parallel, the London Court of International Arbitration (LCIA) has refined its 2020 Rules into a tightly specified, largely "closed-world" procedural regime for international commercial arbitration. We argue that LCIA arbitration is an ideal testbed for *explainable neuro-symbolic* (E-NS) architectures that combine LLMs with explicit logical formalisms and verifiable reasoning. Rather than treating LLMs as quasi-judges, we treat them as front-end language interfaces and explanation generators while delegating core legal reasoning to symbolic rule engines and satisfiability- or theorem-proving back-ends. We close with a three-year research agenda: LCIA-aligned micro-benchmarks, explanation-centred evaluation metrics, and human-in-the-loop studies of E-NS systems that support—rather than replace—arbitrators in high-stakes, cross-border disputes.

Introduction

LLMs have spurred legal-technology prototypes for drafting, evidence triage, and outcome prediction, but empirical studies show that frontier models remain unstable and logically inconsistent on realistic legal questions. Blair-Stanek and Van Durme find that leading models reverse conclusions on hard appellate questions even under deterministic settings, raising concerns about treating raw outputs as authoritative legal advice or decisions (Blair-Stanek and Durme 2025). Surveys of LLM logical reasoning similarly document unfaithful chains of thought and violations of basic inferential constraints (Cheng et al. 2025; Chen 2025).

Hybrid *neuro-symbolic* approaches aim to combine the pattern-recognition strengths of neural models with the explicitness and verifiability of symbolic reasoning. Logic-augmented frameworks such as SatLM and Logic-LM show that requiring declarative encodings and solver checks can improve logical faithfulness over pure prompting (Ye et al. 2023; Pan et al. 2023). Work at the intersection of theorem proving and natural language explanation further suggests pipelines in which natural-language rationales are checked and refined against machine-verifiable proofs (Quan et al. 2025b,a; Yang et al. 2024).

In this position paper we argue that international commercial arbitration under the London Court of International Arbitration (LCIA) Rules provides a particularly attractive domain for such work. The Rules define a relatively self-contained procedural microcosm governing appointment, jurisdiction, evidence, costs, and awards (London Court of International Arbitration 2020). Because many determinations can be framed as rule-governed inferences given the pleadings, evidence, and prior orders, LCIA arbitration offers a promising "closed-world" setting for logic-heavy LLM reasoning research.

Our central claim is that *explainable neuro-symbolic* (E-NS) architectures—which treat formal reasoning and explanation as first-class citizens—offer a practical route to safer, more reliable LLM use in LCIA arbitration over the next three to five years. By "explainable" we mean more than post-hoc feature attributions: we focus on systems in which legal conclusions are derived via explicit proof obligations and in which user-facing explanations can be traced to symbolic derivations (d'Avila Garcez, Lamb, and Gabbay 2019; d'Avila Garcez and Lamb 2020; Xu and Sun 2021; Quan et al. 2025b; Sadowski and Chudziak 2025). The LLM handles semantic parsing, document understanding, and drafting, but it does not act as the sole locus of legal reasoning.

Contributions

This paper makes four contributions:

1. **Problem framing.** We characterise LCIA arbitration as a closed-world legal reasoning domain with unusually strong procedural constraints, and we specify explainability requirements that go beyond surface-level textual rationales.
2. **E-NS design principles.** We extract design principles for hybrid systems in which symbolic solvers and proof engines provide a correctness backbone, while LLMs handle natural-language understanding and explanation (Ye et al. 2023; Pan et al. 2023; Kirtania, Gupta, and Radhakrishna 2024; Chen 2025).
3. **LCIA-focused pipeline.** We outline an end-to-end E-NS

architecture tailored to LCIA procedure, from rule formalisation and clause parsing through fact extraction and proof search to explanation synthesis and human review.

4. **Research agenda.** We propose a three-year programme of LCIA-aligned benchmarks, explanation faithfulness metrics, and human-in-the-loop evaluation, with an emphasis on exportability to other institutional rule sets and to smart-contract arbitration platforms (Han 2025; Zeleznikow 2021).

This is a forward-looking blueprint, not a claim that fully automated awards are desirable or currently feasible. The aim is to make arbitral support tools precise enough to evaluate rigorously and safe enough to deploy under professional oversight.

Background: Explainable Neuro-Symbolic Reasoning in LLMs

Limits of Purely Neural Legal Reasoning

LLMs perform well on many legal NLP tasks, but they struggle with *deterministic* legal reasoning. Blair-Stanek and Van Durme show that frontier models provide unstable answers to difficult appellate questions, with outcomes that vary across repeated or paraphrased prompts (Blair-Stanek and Durme 2025). Logical-consistency studies similarly find violations of basic propositional and first-order constraints, especially under negation, disjunction, and multi-step inference (Xu et al. 2024; Ghosh et al. 2025; Cheng et al. 2025). In international arbitration, where awards must be reproducible from the record and the governing rules, such instability is hard to square with due process. Chain-of-thought rationales are also not guaranteed to track sound inference and can become post-hoc narratives over token-level predictions (Wei et al. 2022; Yang et al. 2024).

Neuro-Symbolic Architectures for Logical Faithfulness

Neuro-symbolic AI integrates statistical learning with symbolic representations and reasoning procedures (d'Avila Garcez, Lamb, and Gabbay 2019). Contemporary hybrid pipelines increasingly use LLMs to translate natural-language problems into symbolic formalisms that are solved by SAT/SMT, answer-set, or theorem-proving engines (Ye et al. 2023; Pan et al. 2023; Kirtania, Gupta, and Radhakrishna 2024; Chen 2025). SatLM and the Logic-LM family require solver-checked symbolic outputs and use solver feedback for iterative repair, improving logical faithfulness over prompting alone (Ye et al. 2023; Pan et al. 2023; Kirtania, Gupta, and Radhakrishna 2024). Related work replaces informal intermediate steps with explicit formulae (symbolic chain-of-thought) and uses theorem provers to verify and refine natural-language explanations, including PEIRCE's LLM-driven refinement loop (Xu et al. 2024; Quan et al. 2025b,a). For legal domains, the design advantage is to delegate correctness-critical reasoning to transparent, deterministic engines while using LLMs for reading, paraphrasing, drafting, and proposing candidate encodings.

Explainability in Neuro-Symbolic Systems

Explainability has long motivated neurosymbolic AI: explicit rules, proofs, and constraints support explanations that can be inspected, challenged, and revised (d'Avila Garcez, Lamb, and Gabbay 2019; d'Avila Garcez and Lamb 2020). Explainable neurosymbolic methods should link symbolic and sub-symbolic levels, not merely highlight input tokens (Xu and Sun 2021). Theorem-prover-assisted pipelines can improve both correctness and perceived quality of natural-language explanations (Quan et al. 2025b), and structured prompting can produce stepwise rationales aligned with rule application (Sadowski and Chudziak 2025). Safety-critical deployments in power systems and medical imaging illustrate how symbolic constraints and domain ontologies can yield audit-ready explanations while retaining expert oversight (Jothimurugan and coauthors 2025; Arrieta and coauthors 2025; Liu and coauthors 2025).

For LCIA arbitration, these developments suggest treating the LCIA Rules and arbitration agreements as a logical substrate, using LLMs to map case materials onto that substrate, and generating explanations grounded in formal derivations yet accessible in natural language.

LCIA Arbitration as a Closed-World Reasoning Domain

Structure of the LCIA Rules

The LCIA Arbitration Rules (2020) provide a comprehensive procedural framework covering commencement, tribunal constitution and challenges, hearings, evidence, interim measures, costs, and awards (London Court of International Arbitration 2020). Compared to national-court litigation, proceedings are relatively self-contained: tribunal powers are largely determined by the arbitration agreement and the Rules, with limited reference to the law of the seat or mandatory public policy.

This structure lends itself to formal modelling. Many procedural questions reduce to checking whether the Rules' preconditions are satisfied given the record—for example jurisdiction (Articles 1–7), emergency arbitrator appointment (Article 9B), consolidation or concurrent proceedings (Articles 22.1 and 22.7), or costs apportionment (Articles 24–28). While factual disputes and equitable considerations remain, a substantial class of determinations can be framed as queries over a rule set coupled with a structured representation of the case.

Digital arbitration proposals show how institutional rules can be encoded in smart contracts and paired with AI modules for clause and evidence processing (Han 2025). Work on online courts and intelligent dispute resolution likewise highlights the procedural affordances of digitisation and the long history of rule- and case-based decision support (Susskind 2019; Zeleznikow 2021). Our proposal treats the Rules as the reasoning backbone of a neuro-symbolic architecture rather than a post-hoc checklist.

Explainability Requirements in Arbitration

Arbitral awards must be reasoned, enforceable under the New York Convention, and acceptable to parties and review-

ing courts. For AI-supported arbitration this implies three linked requirements: (i) **rule-level transparency** (which LCIA provisions and clauses were applied), (ii) **inference-level transparency** (inspectable steps connecting premises to conclusions, ideally in a verifiable formalism), and (iii) **narrative-level transparency** (a human-readable account connecting formal reasoning to the factual record). LLM-based legal tools often generate plausible narratives without robust rule- or inference-level guarantees (Blair-Stanek and Durme 2025); E-NS systems are designed to align all three layers.

An Explainable Neuro-Symbolic Architecture for LCIA

We now outline a conceptual architecture for explainable neuro-symbolic LCIA reasoning. The goal is not to automate arbitral decision-making end-to-end, but to provide a research blueprint for hybrid systems in which LLMs and symbolic engines collaborate to support arbitrators.

Layer 1: Knowledge Formalisation

The first layer encodes the LCIA Rules, relevant soft law instruments, and selected doctrinal principles in a formal representation suitable for automated reasoning. Following SatLM and Logic-LM, we assume a family of logic programming or first-order logic fragments that can be fed into external solvers (Ye et al. 2023; Pan et al. 2023; Kirtania, Gupta, and Radhakrishna 2024). Key ingredients include:

- **Rule Schemas.** Each LCIA article is formalised as a rule or rule schema with explicit preconditions and consequences (e.g., "if the arbitration agreement designates the LCIA and the request and registration fee have been received, then proceedings commence on date d").
- **Ontologies and Types.** Parties, claims, tribunals, and procedural acts are typed entities, linked via an ontology that constrains admissible relations and helps detect modelling errors (Xu and Sun 2021; Chen 2025).
- **Soft Constraints.** Some provisions (e.g., fairness, efficiency) are represented as soft constraints or optimisation objectives rather than hard rules, allowing for multi-objective reasoning.

While manual formalisation is labour-intensive, early work on declarative smart contracts and legal rule engines suggests that it is feasible for well-scoped domains (Governatori et al. 2018; Han 2025). Moreover, techniques from symbolic chain-of-thought and quasi-symbolic abstractions may assist in semi-automated rule extraction and validation (Xu et al. 2024; Ranaldi, Valentino, and Freitas 2025).

Layer 2: Case Representation and Fact Extraction

The second layer maps unstructured case materials into structured factual representations compatible with the rule base. Here, LLMs serve as powerful information-extraction and normalisation engines. Building on existing work in legal NLP and arbitration-focused AI, we envisage:

- **Clause Parsing.** LLMs identify and classify arbitration clauses in contracts, including seat, governing law, number of arbitrators, and scope of disputes (Kant et al. 2025).
- **Timeline and Event Graphs.** Pleadings, witness statements, and exhibits are transformed into event graphs capturing who did what, when, and under which contractual obligations.
- **Evidence Typing and Credibility Hints.** Inspired by Han's framework and medical XAI systems, the system tags evidence with provenance, modality, and reliability indicators, which can be surfaced in explanations (Han 2025; Arrieta and coauthors 2025).

Crucially, the mapping from text to facts is itself uncertain. Following PEIRCE and related work, we envisage an iterative refinement loop in which candidate fact sets are checked for consistency against the rule base and revised when contradictions or gaps are detected (Quan et al. 2025a; Chen 2025).

Layer 3: Symbolic Reasoning and Proof Obligations

Given a formalised rule base and a structured case representation, the core reasoning layer answers queries such as:

- Does the tribunal have jurisdiction over party P?
- Is a request for consolidation admissible?
- Are the conditions for emergency relief satisfied?

We propose to adapt techniques from SatLM, Logic-LM, and symbolic chain-of-thought:

- **Query Decomposition.** High-level questions are decomposed into sub-queries with explicit proof obligations (e.g., jurisdiction requires a valid arbitration agreement, capacity of parties, and absence of conflicting clauses).
- **Solver-Oriented Encodings.** Sub-queries are encoded as SAT, SMT, or answer-set problems; off-the-shelf solvers provide sound, complete answers within their expressive scope (Ye et al. 2023; Pan et al. 2023).
- **Self-Refinement.** When encodings are ill-formed or unsatisfiable, error messages are fed back to the LLM, which attempts to revise the formalisation, as in Logic-LM++ (Kirtania, Gupta, and Radhakrishna 2024).

From an explainability standpoint, this layer yields proof objects—Derivation trees, unsatisfiable cores, or model assignments—that can be inspected and, in many cases, directly mapped to legal reasoning steps.

Layer 4: Explanation Synthesis and Alignment

The final layer converts symbolic proofs and fact graphs into human-readable explanations, tailored to different audiences (tribunal, parties, reviewing courts). Building on theorem-prover–assisted explanation work, we propose a two-stage process (Quan et al. 2025b; Sadowski and Chudziak 2025):

1. **Draft Explanation Generation.** The LLM receives the proof object and a structured summary of the facts, and generates a candidate explanation that walks through the key steps, citing specific rules and evidence.

2. **Explanation Verification and Repair.** The explanation is parsed back into a lightweight logical representation (e.g., an explanation graph); a theorem prover checks whether each step is supported by the underlying rules and facts; detected mismatches trigger revision prompts, iterating until the explanation is both faithful and fluent.

This loop enforces a strong alignment between what the system "says" and what the symbolic layer has actually proved, mitigating the risk of hallucinated legal reasoning. Explanations can expose, rather than hide, the limits of formalisation by explicitly marking points where soft constraints, discretionary standards, or disputed facts are decisive.

Research Agenda: 2026–2029

The blueprint is aspirational and will require sustained empirical and knowledge-engineering work. We highlight four research thrusts.

LCIA Micro-Benchmarks for Logical and Explanatory Quality

Existing logical reasoning benchmarks are valuable but far from the procedural structure of LCIA proceedings (Pan et al. 2023; Ye et al. 2023). We propose LCIA-focused micro-benchmarks targeting:

- **Procedural entailment.** Given a fragment of the Rules and a simple fact pattern, decide whether a procedural action is permissible.
- **Jurisdiction and admissibility.** Evaluate jurisdiction and admissibility questions under varying clause wordings and party constellations.
- **Cost allocation scenarios.** Reason about costs and interest under Article 28 with different party behaviours and outcomes.

Each item would include a gold-standard formalisation, solver output, and a reference explanation authored by arbitration practitioners. Benchmarks should measure not only accuracy but also stability across paraphrases and explanation faithfulness (Blair-Stanek and Durme 2025; Quan et al. 2025b).

Autoformalisation and Rule Learning for Institutional Frameworks

Formalising the LCIA Rules and related instruments is a non-trivial knowledge engineering task. Recent work on autoformalisation and quasi-symbolic abstractions suggests several promising directions (Xu et al. 2024; Ranaldi, Valentino, and Freitas 2025; Zhou et al. 2024):

- **LLM-assisted rule drafting.** Use LLMs to propose candidate logical encodings of rule paragraphs, followed by human review and prover- or solver-based consistency checks.
- **Cross-institutional generalisation.** Study how LCIA formalisation patterns transfer to other institutional rules (e.g., ICC, SIAC, UNCITRAL), and identify where they fail.
- **Expressiveness trade-offs.** Quantify how the choice of logical fragment affects both the ease of formalisation and the power of automated reasoning (Donets 2025b; Chen 2025).

Human-in-the-Loop Evaluation and Governance

Explainable neurosymbolic systems must be evaluated not only on benchmark accuracy but also on how they interact with arbitrators and counsel. Inspired by XAI evaluation in safety-critical domains, we envision mixed-method studies in which users review AI-generated reasoning artefacts (Arrieta and coauthors 2025; Jothimurugan and coauthors 2025; Liu and coauthors 2025). Key questions include whether symbolic alignment improves calibrated trust, how users allocate attention between proofs and narratives, and whether interfaces make residual uncertainty legible. Research should also address governance issues such as responsibility for errors, data protection in evidence handling, and environmental costs of large-scale LLM usage (Han 2025; Donets 2025a).

From Support Tools to Co-Reasoning Partners

Finally, we anticipate a shift from static decision-support dashboards towards *co-reasoning* systems in which arbitrators and AI systems iteratively refine representations of rules and facts. Work on multi-step refinement and logical consistency suggests that LLMs can generate and critique alternative reasoning trajectories (Cheng et al. 2025; Ghosh et al. 2025). In an LCIA context, this could involve multi-agent debate between alternative formal models of the case, interactive exploration of "what-if" procedural scenarios grounded in the rule base, and dynamic adjustment of the level of formalisation as the case evolves. Designing such systems will require careful attention to cognitive load, adversarial behaviour, and the risk of over-deference to AI-generated arguments.

Conclusion

LCIA arbitration offers a rare combination of high stakes and relatively closed-world procedure, making it a practical testbed for explainable neuro-symbolic reasoning in language-model–centric systems. By treating the LCIA Rules as a formal substrate and using LLMs for parsing and explanation rather than final judgment, hybrid architectures can produce conclusions that are more stable, transparent, and auditable than end-to-end neural baselines.

Ultimately, the goal is not to automate judgment, but to build tools that make human judgment more informed, more consistent, and more explainable. Explainable neuro-symbolic systems for LCIA arbitration can serve as a testbed for broader efforts to align advanced AI with legal standards of reasoning and justification, and to ensure that the benefits of LLM technology are realised in ways that respect the rule of law.

References

Arrieta, A. R.; and coauthors. 2025. An Explainable AI Framework for Corneal Imaging Interpretation and Refractive Surgery Decision Support. *Diagnostics*.

Blair-Stanek, A.; and Durme, B. V. 2025. LLMs Provide Unstable Answers to Legal Questions. *arXiv preprint*.

Chen, M. K. 2025. A Comparative Study of Neurosymbolic AI Approaches to Interpretable Logical Reasoning. *Proceedings of Machine Learning Research*, 284: 1–16.

Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. *Journal of Artificial Intelligence Research*. Forthcoming.

d'Avila Garcez, A.; and Lamb, L. C. 2020. Neurosymbolic AI: The Third Wave. *Artificial Intelligence Review*, 53(1): 1–22.

d'Avila Garcez, A.; Lamb, L. C.; and Gabbay, D. M. 2019. Neural-Symbolic Computing: An Effective Methodology for Principle Integration of Machine Learning and Reasoning. In *Neural-Symbolic Learning Systems*. Springer.

Donets, A. 2025a. The Environmental Impact of Large Language Model Energy Consumption. *arXiv preprint*.

Donets, A. 2025b. On the Expressiveness Limitations of Symbolic Solvers in Safety-Critical Systems. *arXiv preprint*.

Ghosh, B.; Hasan, S.; Arafat, N. A.; and Khan, A. 2025. Logical Consistency of Large Language Models in Fact-Checking. *Proceedings of the International Conference on Learning Representations*.

Governatori, G.; Olivieri, F.; Legge, E. L. G.; and Xu, X. 2018. A Manifesto on Legal-Smart Contracts. *Proceedings of the International Workshop on Juris-informatics*.

Han, P. 2025. AI-powered Digital Arbitration Framework Leveraging Smart Contracts and Electronic Evidence Authentication. *Scientific Reports*, 15(1). Doi:10.1038/s41598-025-21313-x.

Jothimurugan, S.; and coauthors. 2025. Neuro-Symbolic AI for Explainable Decision-Making in Autonomous Grid Operations. *IEEE Transactions on Power Systems*. Forthcoming.

Kant, M.; Nabi, S.; Kant, M.; Scharrer, R.; Ma, M.; and Nabi, M. 2025. Towards Robust Legal Reasoning: Harnessing Logical LLMs in Law. *arXiv preprint*. ArXiv:2502.17638.

Kirtania, S.; Gupta, P.; and Radhakrishna, A. 2024. Logic-LM++: Multi-Step Refinement for Symbolic Formulations. *arXiv preprint*. ArXiv:2403.XXXX.

Liu, Y.; and coauthors. 2025. Neuro-Bridge-X: A Neuro-Symbolic Vision Transformer with Meta-XAI for Interpretable Leukemia Diagnosis. *Cancers*.

London Court of International Arbitration. 2020. LCIA Arbitration Rules. `https://www.lcia.org/Dispute_Resolution_Services/lcia-arbitration-rules-2020.aspx`. Accessed 2025-12-11.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. Y. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In *Findings of the Conference on Empirical Methods in Natural Language Processing*.

Quan, X.; Valentino, M.; Carvalho, D. S.; Dalal, D.; and Freitas, A. 2025a. PEIRCE: Unifying Material and Formal Reasoning via LLM-Driven Neuro-Symbolic Refinement. *arXiv preprint*. ArXiv:2501.XXXX.

Quan, X.; Valentino, M.; Dennis, L. A.; and Freitas, A. 2025b. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Ranaldi, L.; Valentino, M.; and Freitas, A. 2025. Improving Chain-of-Thought Reasoning via Quasi-Symbolic Abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Sadowski, A.; and Chudziak, J. A. 2025. Explainable Rule Application via Structured Prompting: A Neural-Symbolic Approach. *arXiv preprint*.

Susskind, R. 2019. *Online Courts and the Future of Justice*. Oxford University Press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Xu, J.; Fei, H.; Pan, L.; Liu, Q.; Lee, M.; and Hsu, W. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. *Transactions of the Association for Computational Linguistics*, 12: 1–20.

Xu, J.; and Sun, Z. 2021. Explainable Neuro-Symbolic AI: Challenges and Perspectives. *International Journal of Software and Informatics*, 11(3): 1–25.

Yang, K.; Poesia, G.; He, J.; Li, W.; Lauter, K.; Chaudhuri, S.; and Song, D. 2024. Formal Mathematical Reasoning: A New Frontier in AI. *arXiv preprint*. ArXiv:2412.16075.

Ye, X.; Chen, Q.; Dillig, I.; and Durrett, G. 2023. SatLM: Satisfiability-Aided Language Models Using Declarative Prompting. In *Advances in Neural Information Processing Systems*.

Zeleznikow, J. 2021. Using Artificial Intelligence to Provide Intelligent Dispute Resolution Support. *Group Decision and Negotiation*, 30(4): 789–812.

Zhou, J. P.; Staats, C.; Li, W.; Szegedy, C.; Weinberger, K. Q.; and Wu, Y. 2024. Don't Trust, Verify: Grounding LLM Quantitative Reasoning with Autoformalization. In *Proceedings of the International Conference on Learning Representations*.