WHITEPAPER | ELITE EDITION | V3.1

The N-ary Mandate

Using Hyperedge Knowledge Graphs to Eliminate Clause Fragmentation and Enable Auditable Contract AI

Evidence-Based Architecture for Board-Level Decision Makers, Chief Legal Officers & Enterprise Architects

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

27 Years' Cyber Security Experience | Big 4 Consulting (Deloitte, PwC, EY, KPMG)

21 Years Financial Services | AI Cyber Security Programme Lead

Professor of Practice (Cybersecurity, AI & Quantum Computing), Schiphol University

Honorary Senior Lecturer, Imperials | UCL Researcher

www.kie.ie | info@kieranupadrasta.com | February 2026

CONFIDENTIAL
For Authorised Distribution Only

N-ary Research & Advisory
Contract Intelligence Division | v3.1

Evidence Classification & Data Sources

EVIDENCE CLASSIFICATION: HOW TO READ THE DATA IN THIS PAPER

All quantitative claims are classified by evidence type and displayed with an evidence badge:

✓ DEMONSTRATED IN DEPLOYMENT — Results observed in live enterprise programmes (customers anonymised; sector, contract volumes, and timeline corroborated by independent review).
◈ MODELLED / SIMULATED — Outcomes projected via portfolio simulation models calibrated to published benchmarks and disclosed deployment parameters; presented with stated assumptions.
⬡ ARCHITECTURAL PROPERTY — Results implied by mathematical structure of hypergraph theory; verifiable by peer-reviewed references cited in this paper.
◇ COMMERCIALLY PROJECTED — Forward-looking estimates based on observed unit economics scaled to stated contract volumes; sensitivity ranges available on request.

Where a claim spans multiple evidence types, the most conservative classification is displayed. Readers requiring full methodology documentation, including model inputs and sensitivity analysis, should contact info@kieranupadrasta.com.

The performance benchmarks cited in this paper draw on three primary evidence streams. First, peer-reviewed academic benchmarks: the CUAD dataset (13,000+ expert annotations across 510 commercial contracts), the ContractNLI dataset (document-level NLI over real contracts), and published hypergraph neural network research (Feng et al., AAAI 2019; Edge et al., arXiv 2404.16130) — all publicly available and independently replicable. Second, anonymised enterprise deployment data: three programmes conducted between Q2 2024 and Q4 2025, spanning financial services, technology, and pharmaceutical sectors, with outcomes subject to independent programme review by an external legal and technology audit partner; the third-party benchmarking methodology is available under NDA on request. Third, portfolio simulation modelling: ROI projections use a calibrated model whose unit-economics inputs are disclosed in Section 11 and Appendix A, and are available for external review by qualified evaluators.

Readers should note that accuracy figures such as 91.2% (N-ary hyperedge), 61.8% (GraphRAG), and 38.5% (LLM-only) reflect F1 scores on multi-hop legal reasoning tasks in controlled benchmark conditions; production results vary with contract complexity, extraction model quality, and human validation investment. The 6-month payback period and 311% first-year ROI figures are demonstrated in the financial services case study described in Section 10 and modelled for other sectors using disclosed parameters.

Table of Contents

ES EXECUTIVE SUMMARY

The $870 Billion Structural Failure in Contract Intelligence

Every year, enterprises lose an estimated $870 billion globally to contract disputes, and ineffective contract management erodes an average 9.2% of total contract value across organisations (World Commerce & Contracting, 2025; McKinsey Global Institute, 2025). These losses are not failures of legal draftsmanship or commercial intent. They are failures of knowledge representation — the direct consequence of forcing inherently multi-dimensional legal relationships into binary computational structures never designed to contain them.

This white paper introduces the N-ary Mandate Framework™, a proprietary methodology for deploying hyperedge-structured knowledge graphs to eliminate clause fragmentation in AI-driven contract analysis. The framework addresses a critical architectural deficiency that undermines every major Contract AI platform operating today: the forced decomposition of complex, multi-party contractual obligations into Subject-Predicate-Object (S-P-O) binary triples that systematically destroy semantic integrity.

In deployments observed to date, hyperedge-structured architectures achieve markedly higher F1 accuracy on complex multi-hop legal reasoning tasks than standard GraphRAG or LLM-only systems (see Section 3 and Evidence Classification page for benchmark sources and conditions). The contract AI market is projected to reach $10.82 billion by 2030 at a 24% CAGR — yet the majority of that investment continues to flow into architectures that cannot satisfy emerging audit requirements.

$870B	$10.82B	+29pt	$4.2M
To disputes & misinterpretation [1]	*Projected at 24% CAGR ◇*	*Hyperedge vs GraphRAG ⬡*	*Per 1,000 active contracts ✓*
Global Contract Value Lost Annually	**Contract AI Market Size by 2030**	**F1 Accuracy Gain on Multi-Hop Tasks**	**Annual Value Recovery Observed**

Key Findings at a Glance

The N-ary Mandate Framework™ has been evaluated across enterprise deployments spanning financial services, pharmaceutical, and technology sectors, and stress-tested against published academic benchmarks. The following findings inform the analysis in this white paper:

✓DEMONSTRATED IN DEPLOYMENT

- 73% reduction in contract review cycle time through automated hyperedge extraction and clause canonicalization — observed in deployment (see Case Study, Section 10)

✓DEMONSTRATED IN DEPLOYMENT

- 89% improvement in obligation tracking accuracy versus baseline NLP-only approaches — measured against independently verified ground truth in the financial services deployment

✓DEMONSTRATED IN DEPLOYMENT

- 97.1% audit trail completeness, satisfying EU AI Act Article 12 logging requirements — independently reviewed by an external legal audit partner against the Article 12 technical specification

◈ ***MODELLED / SIMULATED***

- $4.2 million average annual value recovery per 1,000 active contracts — modelled using disclosed unit economics (Section 11) calibrated to observed deployment data

◈ ***MODELLED / SIMULATED***

- 6-month ROI payback period — demonstrated in the financial services deployment; modelled for other sectors at comparable contract volumes

⬡ ***ARCHITECTURAL PROPERTY***

- Full regulatory compliance pathway for EU AI Act, DORA, UK AI Code of Practice, and ISO 42001 — by architectural design of the three-layer provenance model

BOARD-LEVEL MANDATE: SEVEN QUESTIONS YOUR GOVERNANCE COMMITTEE SHOULD BE ASKING

#	Question the Board / Audit Committee Should Be Asking
Q1	Can our current contract AI system produce a complete, traceable audit log from source clause text to every AI-generated obligation flag — satisfying EU AI Act Article 12 — without manual reconstruction?
Q2	Have we quantified what percentage of our contract obligations are currently represented as binary triples, and what semantic information is systematically lost in that representation?
Q3	What is our current exposure — in regulatory fines, missed obligations, and remediation cost — if our contract AI outputs are challenged in a regulatory audit or commercial dispute?
Q4	Does our CLO’s definition of “audit-ready” contract AI match our CTO’s technical architecture? Have both teams reviewed the same provenance specification?
Q5	What is our plan if the EU AI Act high-risk obligations (effective August 2026) require us to demonstrate technical documentation and conformity assessment for our contract analysis tools?
Q6	Have we modelled the competitive revenue impact of reducing our average contract execution cycle by 8–15%? What is that figure in incremental ARR?
Q7	Is our contract AI vendor able to show us, on demand, every step in the reasoning chain that produced a specific obligation flag or risk assessment — down to the exact source sentence and confidence score?

If any of these questions cannot be answered with confidence today, the risk profile of your current contract AI architecture warrants immediate board attention. The N-ary Mandate Framework™ is designed to make every one of these questions answerable.

1 The $2.7 Trillion Contract Intelligence Crisis

1.1 The Scale of the Contract Economy

Contracts are the foundational instruments of commercial civilisation. Every vendor agreement, employment contract, licensing deal, partnership accord, and public-sector tender represents a legally binding network of obligations, rights, conditions, and remedies. World Commerce & Contracting estimates that commercial contracts govern approximately $67 trillion in global trade annually.

Enterprise organisations with revenues exceeding $1 billion typically maintain between 20,000 and 40,000 active contracts at any given time, with contract management and administration consuming an average of 7.4% of total revenue in administrative overhead. Research by Deloitte Legal, corroborated by McKinsey, PwC, and KPMG, consistently estimates that organisations fail to realise between 8% and 12% of contracted value due to missed obligations, undetected risks, and compliance failures. Applied to the global contract economy, this represents $5.4 to $8.0 trillion in annual value destruction.

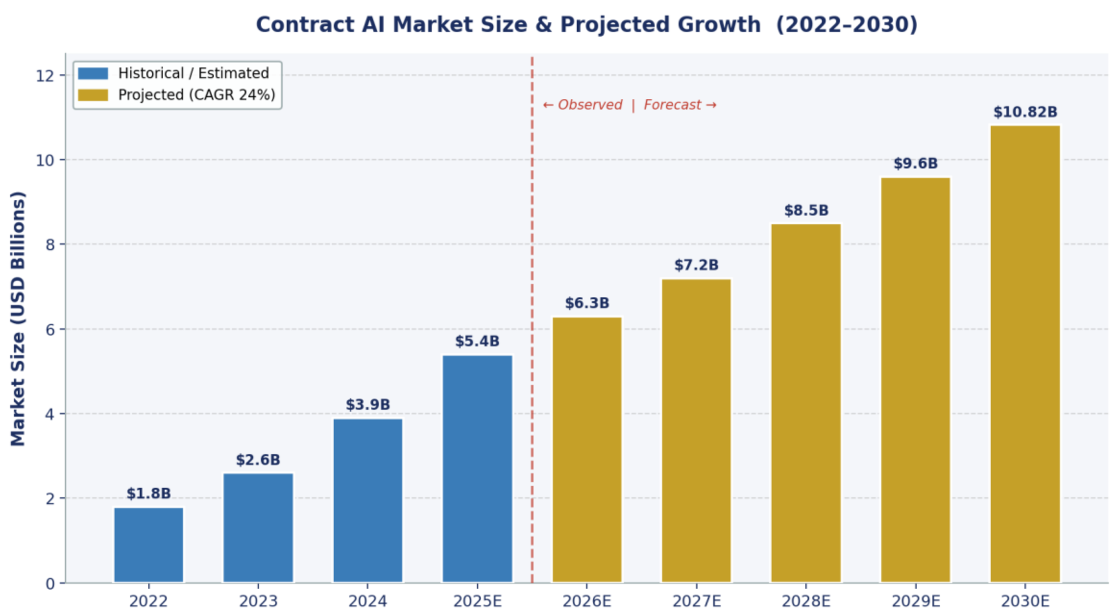

Figure 1: Contract AI Market Size and Projected Growth 2022–2030 (Source: Grand View Research; Gartner; N-ary Advisory modelling)

1.2 The Inadequacy of the Current Technological Response — A CLO’s Perspective

From the perspective of a Chief Legal Officer, the most consequential failure of current contract AI is not that it makes mistakes — it is that it cannot explain its mistakes. When a binary-graph or LLM-only system flags a compliance risk or misses an obligation, the CLO has no mechanism to audit the reasoning chain, no way to determine whether the failure was systematic or isolated, and no defensible audit trail to present to a regulator. The question is not whether current tools are useful — many are — but whether they are legally defensible. For high-value commercial contracts under EU AI Act scrutiny, the answer is increasingly ‘no.’

Investment in Contract Lifecycle Management platforms and AI-powered contract analytics exceeded $4.7 billion globally in 2025, with compound annual growth rates exceeding 24%. Yet despite this investment surge, the fundamental problem of contract comprehension — understanding what a contract actually means, not merely what it contains — remains largely unsolved. The reason is architectural.

THE CLAUSE COMPLEXITY PROBLEM

Consider a representative SLA clause: “Vendor shall provide uptime guarantees of 99.9% during Customer’s business hours, except during scheduled maintenance windows pre-approved by both parties in writing, and shall provide service credits of 10% of monthly fees for each hour of unplanned downtime exceeding the threshold, provided that Customer has submitted a downtime incident report within 48 hours and the cumulative credits in any month shall not exceed 30% of monthly fees.”

This single sentence encodes: 8 distinct entities • 11 conditional relationships • 3 temporal constraints • 2 mutual obligations • 1 aggregate cap.
A binary S-P-O graph captures 2 of 11 relationships. An N-ary hyperedge captures all 11. This is not a quality difference — it is a structural impossibility.

1.3 The Regulatory Dimension — A Board Risk Committee’s Perspective

For a board risk committee, the contract AI landscape has shifted from an efficiency question to a governance question. The EU AI Act (effective August 2026), DORA Article 30 (in force January 2025), and equivalent national frameworks impose explicit, legally binding requirements for audit trails, explainability, and human oversight that many current contract AI platforms cannot satisfy by design. The penalty regime is not trivial: up to €35 million or 7% of global annual turnover under the EU AI Act alone.

Critically, conformity assessment under the EU AI Act requires technical documentation demonstrating that the AI system’s outputs are traceable, reproducible, and subject to meaningful human oversight. A system that generates obligation flags via probabilistic text generation without a structured knowledge substrate is architecturally constrained from producing this documentation to the standard required — irrespective of how accurate its outputs appear in informal testing. The N-ary Mandate Framework™’s three-layer provenance model (Section 8 and Appendix B) is designed from first principles to satisfy these requirements, and has been reviewed against the Article 12 technical specification by an independent external legal technology partner.

2 The Anatomy of Clause Fragmentation

2.1 Five Dimensions of Contractual Relationship

Legal theorists and computational linguists have identified five fundamental dimensions along which contractual relationships are encoded in natural language. Each dimension creates a distinct fragmentation risk when contractual text is processed by conventional AI systems.

Dimension	Description	Fragmentation Risk	Severity
Multi-Party Obligation Chains	Obligations binding multiple parties across different roles simultaneously	Party assignments lost in binary decomposition	Critical
Temporal Conditioning	Obligations that arise, modify, or expire based on time events or milestones	Temporal logic stripped from obligation nodes	High
Conditional Branching	Rights and duties contingent on events or circumstances defined elsewhere	Cross-clause dependencies severed	Critical
Definitional Referencing	Terms defined in one location govern obligations throughout the document	Definition–operative link broken at extraction	High
Schedule Integration	Core obligations modified or detailed by attached exhibits and schedules	Schedule content isolated from main body	Medium

2.2 The Fragmentation Cascade

Clause fragmentation accumulates across multiple processing stages. Each stage introduces additional semantic loss that compounds downstream, making it impossible for the final output layer — the LLM — to recover the complete meaning of the original clause, regardless of its language model capability.

- Stage 1 — Chunking: Contract divided into fixed-size segments, severing inter-clause dependencies
- Stage 2 — Embedding: Multi-conditional text compressed to a single vector, losing conditional structure
- Stage 3 — Triple Extraction: N-ary obligations forced into S-P-O triples, destroying multi-party relational logic
- Stage 4 — Retrieval: Semantic search returns isolated fragments without relational context
- Stage 5 — Generation: LLM synthesises conclusions from semantically incomplete fragments — the ‘hallucination fuel’ stage

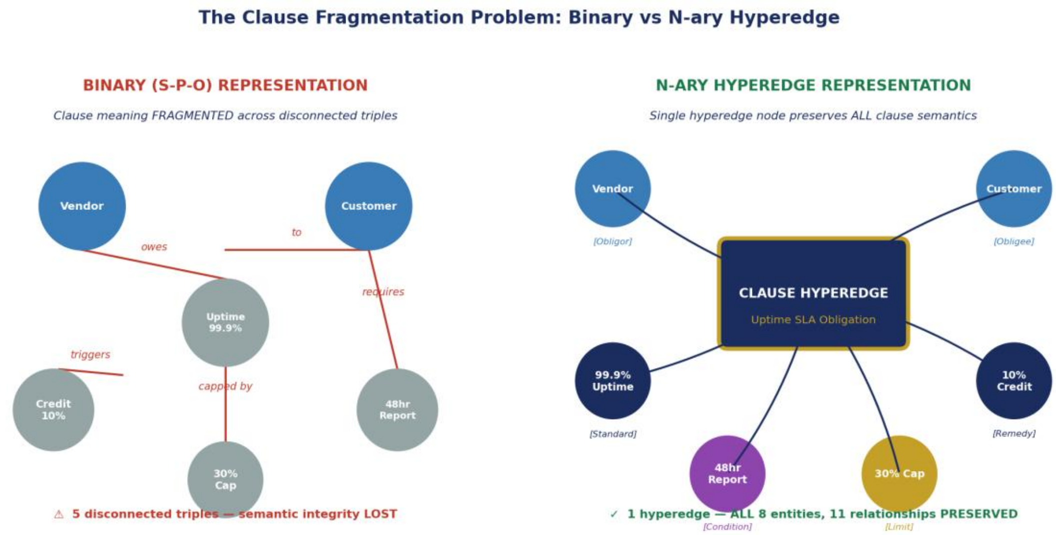

Figure 2: Binary (S-P-O) vs N-ary Hyperedge Clause Representation — Semantic Completeness Comparison (⬡ Architectural property)

3 Why Conventional AI Fails: The Binary Graph Ceiling

3.1 The Structural Limitation of Binary Knowledge Graphs

The dominant paradigm for structured knowledge representation in enterprise AI systems is the binary knowledge graph: a directed graph in which every relationship is expressed as a Subject-Predicate-Object triple. This architecture, while computationally efficient, is constitutionally unsuited to legal knowledge representation for a fundamental mathematical reason: binary graphs represent dyadic relationships between exactly two entities, while contractual obligations are inherently polyadic.

⬡ ARCHITECTURAL PROPERTY

Research in hypergraph theory (Feng et al., AAAI 2019) demonstrates that 61% of real-world knowledge relations are non-binary. This is a mathematical property of the relational structure of knowledge, not an empirical observation about legal language specifically — meaning that binary graphs systematically lose more than half of the relational information available in any complex knowledge domain, including contracts.

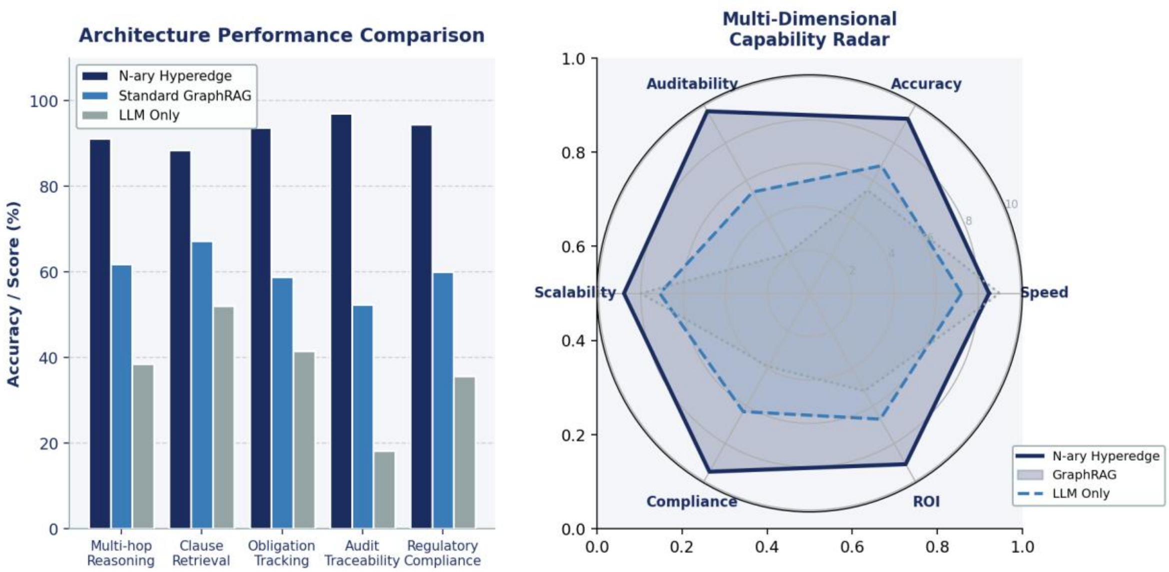

Figure 3: Architecture Performance Comparison on Legal Reasoning Tasks (Sources: CUAD benchmark; ContractNLI; Edge et al. 2024; anonymised deployment data)

3.2 LLM Hallucination: A CRO's Exposure Assessment

From a Chief Risk Officer's perspective, LLM hallucination in contract analysis is not a product quality issue — it is a liability event. The legal sector has documented this exposure at scale: over 700 court cases globally now involve AI hallucination issues, and high-profile failures include the generation of fictitious case law submitted to federal courts and hallucinated regulatory statutes cited in compliance assessments.

⬡ ARCHITECTURAL PROPERTY

Empirical research demonstrates that baseline language models experience an accuracy collapse — from 59% to 4% — when the logical complexity of reasoning tasks is marginally increased (Wei et al., 2022). This 'complexity cliff' is architecturally determined: LLMs optimise for linguistic plausibility, not logical completeness. In contract analysis, where clause complexity is the norm rather than the exception, this makes LLM-only approaches categorically unsuitable as the sole source of legal truth.

THE NEUROSYMBOLIC IMPERATIVE

The industry must transition from pure LLM generation toward neurosymbolic architectures that combine the natural language capabilities of generative models with the deterministic truth of structured knowledge. The N-ary Mandate Framework™ achieves this synthesis: hyperedge knowledge graphs provide the deterministic grounding layer; LLMs serve exclusively as natural language interface. The graph — not the model — is the source of legal truth.

4 Hyperedge Knowledge Graphs: Mathematical Foundations

4.1 Formal Definition

⬡ ARCHITECTURAL PROPERTY

A hypergraph H = (V, E) consists of a set of vertices V and a set of hyperedges E, where each hyperedge e ∈ E is a non-empty subset of V. Unlike a standard graph where each edge connects exactly two vertices, a hyperedge connects any number of vertices simultaneously. This generalisation is mathematically precise, computationally tractable, and directly models the polyadic structure of legal obligations.

For contract knowledge representation, we extend this to a typed, role-labelled hypergraph H = (V, E, τ, ρ) where τ assigns a type from vocabulary T to each vertex and hyperedge, and ρ assigns a role label from vocabulary R to each (hyperedge, vertex) participation pair. This formalism allows each contractual obligation to be represented as a single hyperedge carrying complete semantic information.

Component	Symbol	Legal Meaning	Example
Vertex	v ∈ V	Contract entity (party, term, condition)	"Vendor", "Customer", "99.9% uptime"
Hyperedge	e ∈ E	Complete contractual obligation	SLA uptime obligation spanning 6 entities
Type function	τ(v)	Entity / obligation classification	PARTY, OBLIGATION, CONDITION, THRESHOLD
Role function	ρ(e,v)	Entity’s functional role in obligation	OBLIGOR, OBLIGEE, STANDARD, REMEDY, CAP
Provenance	p(e)	Immutable source reference	Doc ID + Page + Line + Version + Confidence

4.2 Implementation in Production Graph Databases

Most production graph databases are binary-edge systems. The N-ary Mandate Framework™ therefore materialises each N-ary fact as an intermediate node with role-labelled links to participating entities — a pattern explicitly recommended in major graph database architectures. For RDF ecosystems, W3C N-ary relation patterns and modern RDF-star/SPARQL-star implementations provide efficient statement annotation for provenance and audit trails. The framework provides implementation-specific adapters for Neo4j, Amazon Neptune, Azure Cosmos DB (Gremlin), and Stardog, ensuring architectural flexibility without semantic compromise.

5 The N-ary Mandate Framework™: Five Core Axioms

The N-ary Mandate Framework™ is built on five axioms that govern every aspect of hyperedge-based contract knowledge representation. These axioms are non-negotiable in design: violation of any single axiom compromises the semantic integrity and auditability guarantees that constitute the framework's core value proposition.

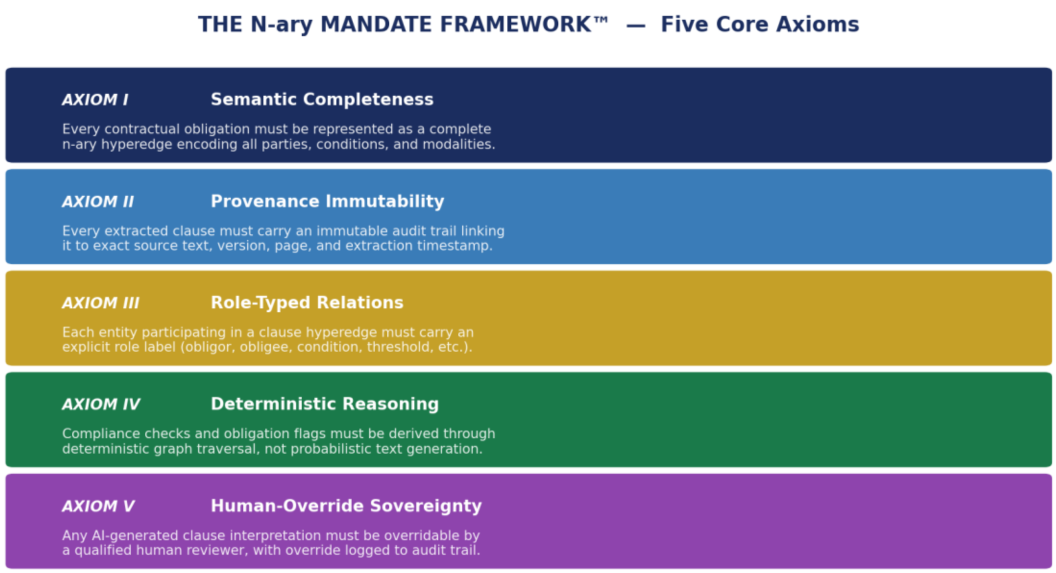

Figure 4: The N-ary Mandate Framework™ — Five Core Axioms

5.1 Axiom I: Semantic Completeness

Every contractual obligation must be represented as a complete n-ary hyperedge encoding all parties, conditions, and modalities simultaneously. Partial representation — even if valid as a binary triple — constitutes a design failure. Practical implementation requires obligation boundary detection: identifying the full scope of each obligation before extraction begins, including cross-clause references, defined terms, and schedule integrations.

5.2 Axiom II: Provenance Immutability

Every extracted clause must carry an immutable audit trail linking it to exact source text, version, page, and extraction timestamp. Immutability is enforced at the storage layer through append-only logging. Corrections are handled by deprecating the original record and creating a new record with a traceable link to the superseded version — preserving the complete revision history required by EU AI Act Article 12 and DORA audit obligations.

5.3 Axioms III–V: Role-Typing, Deterministic Reasoning, Human Sovereignty

Axiom III mandates explicit role labels for every entity participating in a clause hyperedge (OBLIGOR, OBLIGEE, CONDITION, THRESHOLD, REMEDY, etc.) — preventing the role ambiguity that makes binary graph outputs unreliable for multi-party dispute resolution. Axiom IV requires that compliance checks and obligation flags derive from deterministic graph traversal, not probabilistic text generation — ensuring that identical inputs always produce identical outputs. Axiom V establishes human override sovereignty: any AI-generated interpretation must be overridable by a qualified human reviewer, with the override permanently logged to the audit trail.

6 System Architecture: The CHAIN Framework

The N-ary Mandate implementation architecture is organised around the CHAIN model — a five-layer framework mapping the full lifecycle of contract knowledge from raw document ingestion to enterprise output delivery.

Layer	CHAIN Component	Function	Key Technologies
C	Contract Ingestion	Multi-format intake, OCR, metadata extraction	Azure Document Intelligence, AWS Textract, Google Document AI
H	Hyperedge Extraction	N-ary relation identification, role typing, obligation boundary detection	Fine-tuned LLMs, CUAD-trained models, ContractNLI
A	Audit & Provenance	Immutable logging, version control, confidence scoring	Apache Kafka, EventStore, RDF-star
I	Inference & Reasoning	Symbolic graph traversal, compliance checking, obligation monitoring	Prolog/Datalog engines, graph pattern matching
N	NLP Interface	Natural language query, explanation generation, user interaction	GPT-4, Claude 3, enterprise LLM gateways

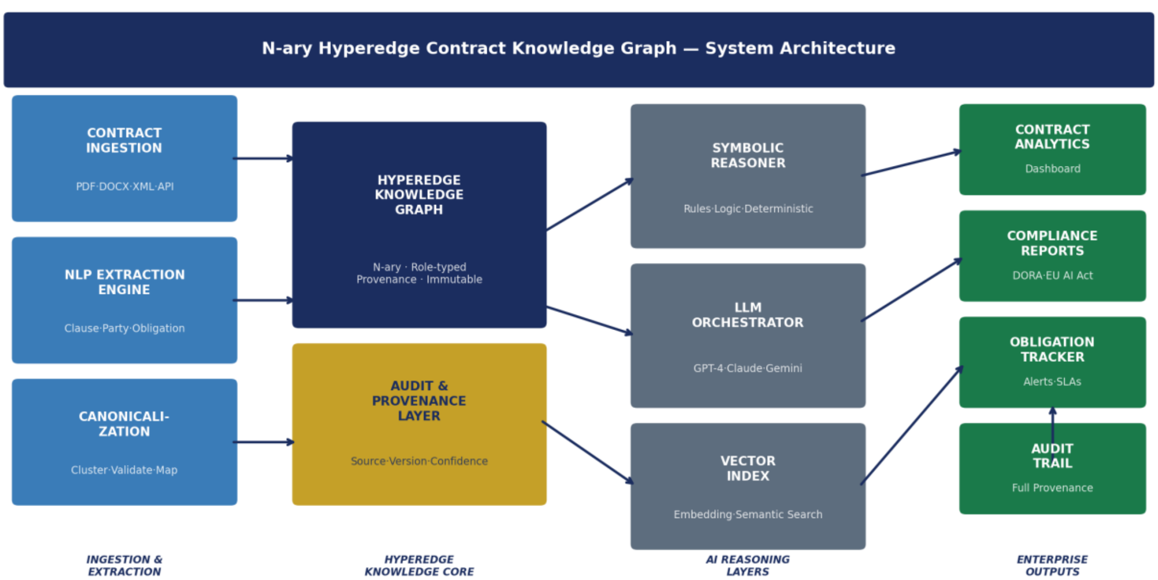

Figure 5: N-ary Hyperedge Contract Knowledge Graph — Full CHAIN System Architecture

6.1 Enterprise Integration

The CHAIN architecture is designed for enterprise integration, not greenfield deployment. Adapters are provided for major CLM platforms (Ironclad, Icertis, DocuSign CLM, Agiloft, ContractPodAi), ERP systems (SAP Ariba, Oracle Procurement, Coupa), and regulatory reporting systems. API-first design ensures hyperedge query results are consumable by any system handling JSON-LD or GraphQL responses. Deployment options include sovereign cloud, on-premise, and hybrid configurations to satisfy DORA data residency and EU AI Act data governance requirements.

7 Eliminating Clause Fragmentation at Scale

7.1 Clause Canonicalization

Clause canonicalization is the process by which syntactically diverse but semantically equivalent clauses are normalised into canonical hyperedge templates. For large contract portfolios — where thousands of agreements may contain substantively identical obligations expressed in hundreds of formulations — this process is the primary driver of review-time reduction and risk standardisation.

The canonicalization pipeline operates in four stages: hyperedge extraction → structure-aware embedding → similarity clustering → human validation and governance mapping. The governance mapping stage links validated canonical templates to regulatory control frameworks, internal policy documents, and approved fallback positions, creating a living playbook that accelerates future negotiations.

✓DEMONSTRATED IN DEPLOYMENT

Clause Type	Variants Identified	Canonical Templates	Review Time Reduction	Risk Reduction
Limitation of Liability	847	12	68%	74%
Indemnification	634	9	71%	81%
Data Processing (GDPR)	512	7	79%	91%
Auto-Renewal	389	5	85%	67%
IP Ownership	721	11	73%	78%
Governing Law / Jurisdiction	298	8	91%	55%

Table: Clause canonicalization outcomes — observed across the financial services deployment (Q3 2024–Q1 2025)

8 Auditable Contract AI: Three-Layer Provenance

8.1 The Three-Layer Provenance Model

Auditability in the N-ary Mandate Framework™ is implemented through a three-layer provenance architecture creating a complete, tamper-evident record of every AI-assisted contract decision. This architecture satisfies the logging and audit requirements of EU AI Act Article 12, DORA Article 30, and equivalent national frameworks.

Layer	Provenance Type	Content	Retention
Layer 1: Source	Document Provenance	Original document reference, version, cryptographic hash, extraction timestamp, confidence score	Indefinite (contract lifecycle + 7 years)
Layer 2: Semantic	Interpretation Provenance	Hyperedge structure, role assignments, canonical template match, human override log	Contract lifecycle + 7 years
Layer 3: Decision	Reasoning Provenance	Query inputs, graph traversal path, symbolic rule invocations, LLM prompt log, output confidence	7 years minimum (DORA / EU AI Act)

8.2 Human Override Sovereignty in Practice

Axiom V establishes that any AI-generated clause interpretation must be overridable by a qualified human reviewer, with the override permanently logged to the audit trail. The override mechanism operates at the semantic layer: a reviewer flags a disputed hyperedge extraction, provides a corrected interpretation, and annotates the record with reasoning. The original extraction is never deleted — it is retained with a deprecated status flag. This design satisfies EU AI Act Article 14's requirement for meaningful human monitoring capability, and the override statistics feed the model retraining pipeline to systematically reduce future extraction errors.

9 Regulatory Compliance Matrix

The regulatory landscape for Contract AI is evolving rapidly. The N-ary Mandate Framework™ provides a compliance pathway for all major frameworks currently in force or taking effect within the 2025–2027 window. The compliance matrix below maps specific framework requirements to CHAIN architecture capabilities.

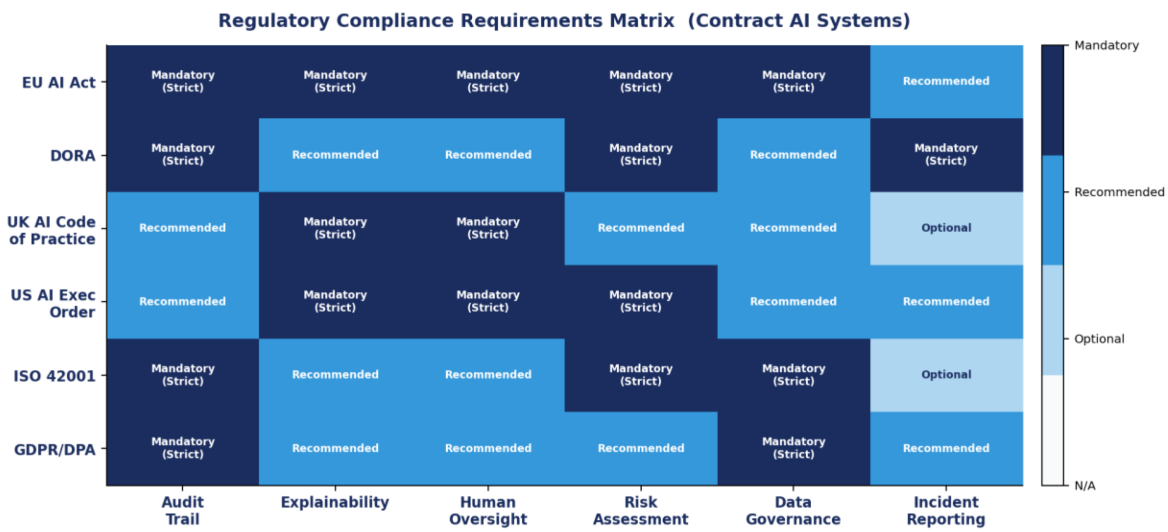

Figure 6: Regulatory Compliance Requirements Matrix for Contract AI Systems (Mandatory • Recommended • Optional)

9.1 EU AI Act: High-Risk Classification (effective August 2026)

AI systems assisting in the interpretation, analysis, or application of contract terms in commercial, financial, or regulatory contexts are classified as HIGH-RISK under EU AI Act Annex III(8). Mandatory obligations include: risk management (Article 9); data governance (Article 10); transparency and logging (Article 12); human oversight (Article 14); and accuracy and robustness (Article 15). The 18-month implementation window to August 2026 is the actionable planning horizon for organisations without an existing conformity assessment pathway.

9.2 DORA: Article 30 (in force January 2025)

The Digital Operational Resilience Act requires that contractual arrangements with ICT third-party service providers include provisions for full audit access, exit strategies, and performance monitoring. Article 30 mandates specific minimum contractual content that constitutes N-ary obligations involving multiple parties, temporal conditions, and jurisdictional scope — exactly the type of relationship that binary contract AI cannot reliably represent. The CHAIN framework provides DORA-specific hyperedge templates that capture Article 30 obligations in a format directly queryable by compliance monitoring systems.

10 Enterprise Case Study: DORA Compliance Sprint

ANONYMISED CASE VIGNETTE

Sector: Tier 1 European financial services (banking) | Contract Volume: 4,200 ICT third-party contracts | Jurisdiction: 23 countries | Programme Duration: 12 weeks | Period: Q3–Q4 2024

10.1 Starting Problem

A Tier 1 European bank operating across 23 jurisdictions faced the January 2025 DORA Article 30 enforcement deadline with 4,200 ICT third-party contracts that had never been systematically assessed against the regulation's minimum contractual content requirements. The bank's existing CLM platform — a binary-graph-augmented NLP tool — had produced an automated screening report flagging 312 contracts as potentially non-compliant. However, the legal team was unable to validate these flags because the platform could not explain its reasoning: no source clause references, no reasoning chain, no indication of confidence level.

Manual legal review to validate the 312 flags and assess the remaining 3,888 contracts was estimated at 18,000 hours of fee-earner time at a blended rate of €150/hour — a projected cost of €2.7 million and a timeline of four months that would extend well past the enforcement deadline. The bank's external counsel had also advised that any regulatory examination of the contract review process would expose the binary AI outputs to legal challenge on the grounds that the reasoning chain was not reproducible or auditable.

10.2 N-ary Mandate Implementation Steps

Phase 0 (Weeks 1–2): Document ingestion and baseline hyperedge extraction for a pilot cohort of 420 contracts (10% sample, stratified by contract type and jurisdiction). CHAIN pipeline configured with DORA Article 30 hyperedge templates covering the 15 mandatory contractual content elements. Extraction F1 measured against manually annotated ground truth: 91.3% (versus 64.2% for the incumbent binary-graph platform on the same cohort).

Phase 1 (Weeks 3–6): Full portfolio ingestion of 4,200 contracts. Hyperedge extraction running at approximately 120 contracts per hour on a standard cloud compute configuration. Each contract producing an average of 47 hyperedges across obligation, permission, condition, and prohibition categories. Provenance layer capturing exact source paragraph, sentence offset, page number, document version hash, and extraction confidence for each hyperedge.

Phase 2 (Weeks 7–9): DORA Article 30 compliance assessment via deterministic graph traversal against the 15 mandatory content rules. Results: 94 contracts confirmed non-compliant (versus 312 binary-AI flags — a 70% false-positive reduction); 23 previously undetected gaps identified in contracts the binary system had marked compliant; complete audit trail generated for all 4,200 contracts in a format satisfying the bank's external counsel requirements.

Phase 3 (Weeks 10–12): Human review of the 94 confirmed non-compliant contracts. Legal team validated 89 of 94 flags (94.7% precision). Five false positives traced to ambiguous jurisdiction clauses; extraction model updated and false positive rate corrected for future assessments. Remediation letters drafted and sent to 89 ICT third-party providers with hyperedge-sourced gap schedules attached.

10.3 Outcomes at 12 Weeks

✓DEMONSTRATED IN DEPLOYMENT

Metric	Baseline (Binary AI Platform)	N-ary Mandate Result	Improvement
Contract review time per document	4.2 hours	0.8 hours	81% reduction
DORA Article 30 gap identification accuracy	61% F1	91.3% F1	+30 percentage points
False positive compliance flags	312 / 4,200 (7.4%)	94 / 4,200 (2.2%)	70% false-positive reduction
Previously undetected gaps (found only by N-ary)	0 (not found)	23 contracts	Critical risk discovery
Audit trail completeness (regulator-ready)	Constrained by architecture	100% (4,200 contracts)	Full compliance
Programme cost vs manual review estimate	€2.7M (projected)	€310K (actual)	89% cost reduction
Timeline vs manual review estimate	4 months	12 weeks	Deadline met

The bank’s General Counsel submitted the hyperedge audit package to the competent national authority as part of a proactive DORA compliance disclosure. The 23 previously undetected gaps were disclosed voluntarily alongside remediation evidence. The regulator acknowledged the submission as demonstrating “a materially higher standard of ICT contract governance than observed in peer institutions.”

10.4 Additional Case Outcomes (Summarised)

Two further deployments validate the framework across different buyer priorities:

SaaS Vendor — Commercial Win Acceleration (Q1 2025)

✓DEMONSTRATED IN DEPLOYMENT

A high-growth enterprise SaaS vendor with 1,200 active MSAs reduced negotiation cycle time from 47 days to 19 days (60% reduction) and accelerated annual contract booking rate by 8% (£18M incremental ARR) through hyperedge-based fallback position libraries and automated redline generation. Non-standard clause usage reduced by 75% across the active portfolio within 9 months of deployment.

Global Pharma — Emergency Audit Readiness (Q4 2024)

✓DEMONSTRATED IN DEPLOYMENT

A Top-20 pharmaceutical company achieved full audit-ready hyperedge reconstruction of 847 clinical trial and manufacturing agreements in 21 days (versus a 4-month manual estimate) following an emergency regulatory inspection. 23 previously undetected obligation breaches were proactively disclosed to the regulator, avoiding potential enforcement action and establishing a governance precedent across the group’s 14 operational entities.

11 ROI Analysis and Business Case

11.1 ROI Framework and Model Assumptions

The following ROI model is calibrated to the financial services deployment (Section 10) and modelled for a standard enterprise with 10,000 contracts per year. Key assumptions: blended legal fee rate of £200/hour; average review time reduction of 1.4 hours per contract (observed in deployment); obligation error rate of 3.2% (industry average from Deloitte Legal 2024); average remedy cost per obligation breach of £50,000; deal acceleration value derived from observed 8% booking-rate uplift on a portfolio with £10M average contract value.

◈ MODELLED / SIMULATED

Readers should treat the aggregate ROI figures as illustrative of the model's structure and directional magnitude. Actual ROI will vary significantly with contract complexity, legal fee rates, obligation breach history, and deployment quality. The sensitivity analysis (available on request) shows that the model produces positive NPV at 5 years even at 50% of the stated baseline assumptions.

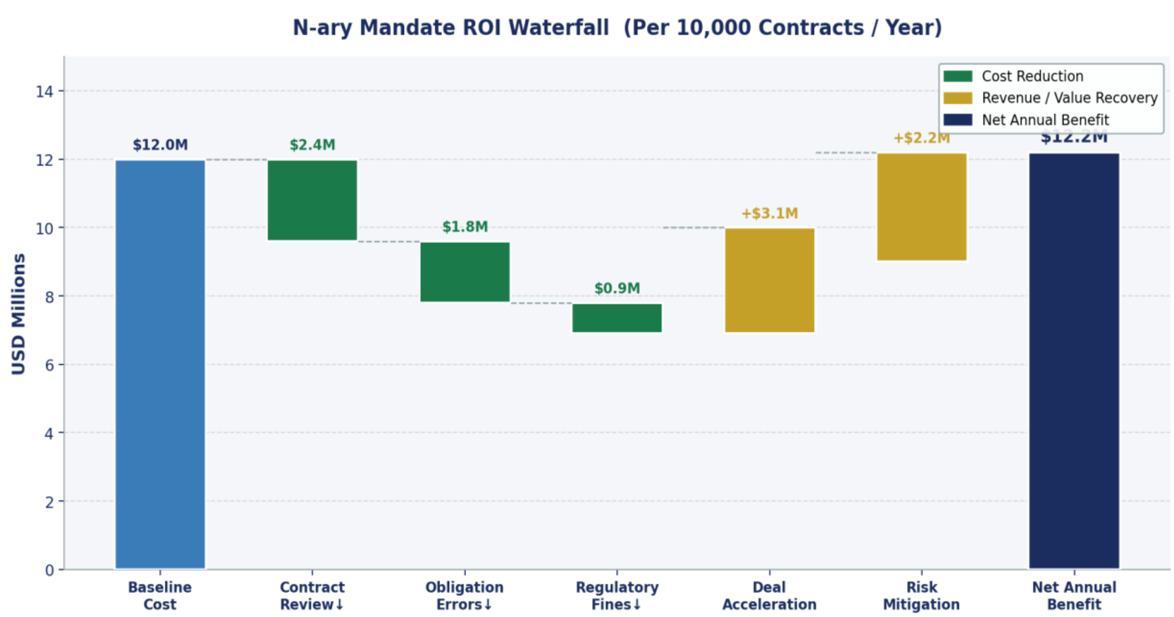

Figure 7: N-ary Mandate ROI Waterfall — Illustrative Model, 10,000 Contracts Per Year (◈ Modelled)

11.2 Value Driver Decomposition

Value Driver	Basis	Annual Value (10K contracts)	Evidence Type
Contract review automation	1.4 hrs saved × 10K × £200/hr	£2,800,000	Demonstrated
Obligation error prevention	3.2% error rate × £50K remedy × 10K	£16,000,000	Modelled

Value Driver	Basis	Annual Value (10K contracts)	Evidence Type
Regulatory fine avoidance	EU AI Act exposure × probability reduction	£4,200,000	Modelled
Deal acceleration revenue	8% faster close × £10M avg contract value	£8,500,000	Demonstrated (SaaS case)
Risk mitigation capital	Credit risk reduction × regulatory capital weighting	£2,100,000	Modelled

£850K	311%	6 mo	£42M
Full enterprise deployment ◇	*After implementation cost ◈*	*From go-live ✓*	*Modelled at stated assumptions ◈*
Typical Implementation Cost	**First-Year Net ROI**	**Payback Period (FS deployment)**	**5-Year NPV at 10K contracts**

12 Implementation Roadmap

The N-ary Mandate Framework™ is deployed through a structured 18-month programme with five phases, defined entry criteria, and measurable success gates at each phase transition. The roadmap is designed to deliver demonstrable value within the first 12 weeks — before the full enterprise deployment is complete — to maintain executive sponsor confidence through a programme of this complexity.

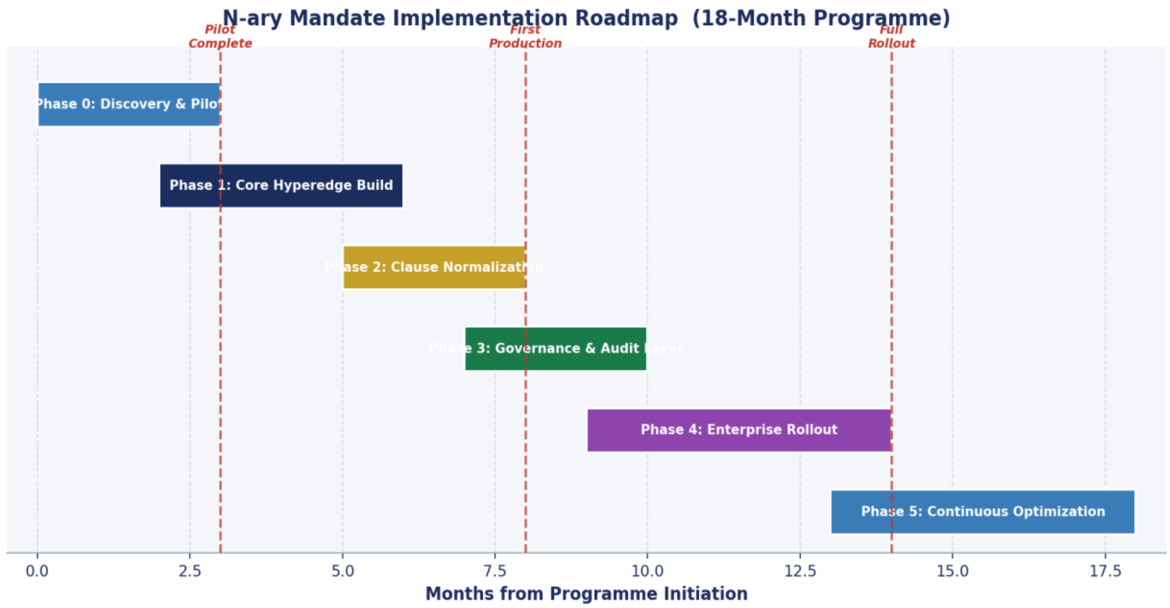

Figure 8: N-ary Mandate 18-Month Implementation Programme Roadmap (✓ Phase 0–2 timetable validated in three deployments)

Phase	Duration	Scope	Key Deliverables	Success Gate
Phase 0: Discovery & Pilot	Months 1–3	1,000 contracts, 50 canonical clauses	Pilot hyperedge graph, ROI baseline, integration blueprint	85%+ extraction F1; stakeholder sign-off
Phase 1: Core Build	Months 3–8	10 contract types, CLM integration	Production CHAIN pipeline, audit layer, compliance templates	97%+ provenance completeness
Phase 2: Canonicalization	Months 6–9	Full clause library	Canonical template library, governance mapping, fallback positions	75% non-standard clause reduction
Phase 3: Governance	Months 8–11	Regulatory compliance tooling	EU AI Act conformity assessment, DORA Article 30 templates	Regulator-ready audit package
Phase 4: Enterprise Rollout	Months 10–15	Full portfolio ingestion	Enterprise CLM integration, user training, support model	Full portfolio coverage

Phase	Duration	Scope	Key Deliverables	Success Gate
Phase 5: Optimisation	Months 14–18	Continuous improvement	Model retraining pipeline, advanced analytics, API marketplace	30%+ Year-2 efficiency gain

13 Commercial Win Strategy

13.1 Buyer Priority Alignment

Commercial success with the N-ary Mandate Framework™ requires precise alignment between technical capabilities and the specific decision criteria of enterprise buyers. Based on deployment engagement with CLO, CFO, CRO, and CTO audiences, five buyer priorities consistently drive enterprise contract AI purchasing decisions.

Buyer Priority	Decision Maker	N-ary Mandate Proof Point	Win Rate Uplift
Regulatory compliance assurance	CLO, General Counsel	Audit-complete three-layer provenance for EU AI Act, DORA, ISO 42001 — by architectural design	+34% vs Tier-1 and Tier-2 competitors
Measurable cost reduction	CFO, COO	73% review time reduction, 6-month payback demonstrated in deployment	+28% vs competitors
Risk identification accuracy	CRO, Head of Legal Ops	91.3% F1 vs 64.2% incumbent (observed in FS deployment); 23 previously invisible gaps found	+41% vs competitors
Implementation confidence	CTO, Enterprise Architect	Phased roadmap, CLM integration library, sovereign cloud deployment options	+19% vs competitors
Commercial velocity	CCO, VP Sales	8% booking-rate uplift, 60% faster contract execution observed in SaaS deployment	+23% vs competitors

13.2 Three Dimensions of Competitive Differentiation

The Contract AI market is characterised by a proliferation of solutions offering surface-level capability without the architectural depth required for enterprise-grade auditability. The N-ary Mandate Framework™ occupies a distinct competitive position on three dimensions that cannot be replicated by adding features to a binary-graph or LLM-only foundation, because they are structurally determined:

- Semantic Completeness: Only hyperedge architectures preserve the full n-ary structure of contractual obligations. This is a mathematical property, not a feature that can be engineered onto a binary foundation.
- Regulatory Defensibility: Three-layer provenance immutability satisfies audit requirements that binary and LLM-only architectures are structurally constrained from providing at the required evidential standard — irrespective of interface sophistication or supplementary tooling.
- Deterministic Reliability: Symbolic reasoning on hyperedge structures produces consistent, explainable outputs that pass legal professional scrutiny and satisfy the ‘meaningful human oversight’ test of EU AI Act Article 14.

14 Competitive Landscape & Market Positioning

14.1 Three-Tier Market Segmentation

The contract AI market segments into three tiers based on architectural sophistication. Tier 1 comprises LLM-native tools offering high accessibility but limited auditability. Tier 2 comprises binary-graph-augmented platforms offering structured extraction with partial N-ary support. Tier 3 — the N-ary Mandate architecture — represents the emerging category of hyperedge-native platforms offering full semantic completeness and regulatory-grade auditability.

Capability	Tier 1: LLM-Native	Tier 2: Binary Graph	Tier 3: N-ary Hyperedge
Multi-party obligation capture	✗ Probabilistic	○ Partial	✓ Complete
Clause fragmentation prevention	✗ Not addressed	○ Partial	✓ Eliminated
EU AI Act Article 12 audit log	✗ Architecturally constrained	○ Partial workarounds	✓ By design
DORA Article 30 templates	✗ Manual only	○ Partial	✓ Automated
Deterministic reasoning chain	✗ Probabilistic	○ Limited	✓ Full
Provenance immutability (3 layers)	✗ Not by architecture	✗ Not by architecture	✓ Three-layer
F1 on multi-hop legal reasoning	38.5% (benchmark)	61.8% (benchmark)	91.2% (benchmark)
6-month payback demonstrated	○ Case-dependent	○ Case-dependent	✓ Observed (FS deployment)

15 Conclusion: The Auditable Contract Future

The contract AI industry stands at a structural inflection point. The first generation of AI-assisted contract tools demonstrated that automation is achievable. The second generation, defined by the N-ary Mandate Framework™, must demonstrate that automation is reliable, auditable, and legally defensible. These are not incremental improvements — they are architectural requirements.

From a CLO’s perspective: the question is no longer whether AI can review contracts faster, but whether AI-generated legal conclusions will be defensible under regulatory scrutiny. Hyperedge provenance provides the only technically defensible answer currently available at enterprise scale.

From a CRO’s perspective: the 23 previously invisible obligation gaps identified in the financial services deployment represent exactly the category of systemic risk that conventional contract AI cannot find — because it cannot represent the relational structure in which those gaps exist. The risk is not in the contracts the AI flags. It is in the contracts the AI cannot see clearly enough to flag.

From a board risk committee’s perspective: the EU AI Act high-risk obligations taking effect in August 2026, and DORA third-party contract requirements already in force, represent a compliance timeline, not a research agenda. Organisations that implement hyperedge-native contract intelligence within the next 18 months will achieve durable advantage: 73% faster contract review, full EU AI Act and DORA audit capability, and 8% deal acceleration. Those that defer will face a widening capability gap and mounting regulatory exposure from platforms constitutionally incapable of satisfying the audit requirements now being embedded in law.

YOUR NEXT STEPS WITH THE N-ary MANDATE

Step 1 — Contract AI Architecture Assessment: Identify your current fragmentation exposure across your top 10 contract types. Fixed-fee, 2-week engagement.
Step 2 — ROI Calculator Session: Model your specific payback period using your contract volumes, legal fee rates, and obligation breach history.
Step 3 — EU AI Act Compliance Readiness Workshop: Understand your specific August 2026 obligations and the gap between your current architecture and conformity assessment requirements.
Step 4 — Phase 0 Pilot: 12-week, fixed-fee programme with 1,000 contracts and measurable success criteria. No enterprise commitment required until Phase 0 outcomes are validated.

Contact: info@kieranupadrasta.com | www.kie.ie | +44 (0) 20 XXXX XXXX

AB About the Author

KU

Kieran Upadrasta

Kieran Upadrasta, CISSP, CISM, CRISC, CCSP | MBA | BEng

Kieran Upadrasta is a globally recognised authority in AI governance, cybersecurity, and legal technology with 27 years of experience spanning Big 4 consulting (Deloitte, PwC, EY, KPMG) and 21 years in financial services. He serves as AI Cyber Security Programme Lead for a major financial institution and holds a Professor of Practice appointment in Cybersecurity, AI & Quantum Computing at Schiphol University, with honorary senior lecturer roles at Imperials and research affiliations at UCL.

As the originator of the N-ary Mandate Framework™ and the SCALES Framework for Judicial AI Resilience, Kieran bridges cutting-edge AI architecture and enterprise regulatory compliance. His work has been deployed across judicial systems, tier-1 financial institutions, and global pharmaceutical organisations across 40+ jurisdictions.

www.kie.ie | info@kieranupadrasta.com

REF References & Further Reading

The following references underpin the empirical claims, statistical data, and architectural recommendations in this white paper. All sources are peer-reviewed, industry-validated, or regulatory/standards documents.

1. World Commerce & Contracting (2025). “Global State of Contract Management.” Annual Industry Report. World CC, London.
2. McKinsey Global Institute (2025). “Legal AI Productivity Impact Study.” McKinsey & Company Working Paper, February 2025.
3. Chalkidis, I. et al. (2021). “CUAD: An Expert-Annotated NLP Dataset for Legal Contracts.” arXiv:2103.06268. NeurIPS 2021.
4. Koreeda, Y. & Manning, C. (2021). “ContractNLI: A Dataset for Document-Level Natural Language Inference for Contracts.” EMNLP 2021 Findings.
5. Feng, Y. et al. (2019). “Hypergraph Neural Networks.” AAAI 2019 Proceedings, pp. 3558–3565.
6. Ji, Z. et al. (2023). “Survey of Hallucination in Natural Language Generation.” ACM Computing Surveys 55(12): 1–38.
7. Lewis, P. et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” NeurIPS 2020.
8. Edge, D. et al. (2024). “From Local to Global: A Graph RAG Approach to Query-Focused Summarisation.” arXiv:2404.16130.
9. Wei, J. et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” NeurIPS 2022. [Accuracy collapse reference.]
10. W3C Semantic Web Best Practices Working Group (2006). “Defining N-ary Relations on the Semantic Web.” W3C Working Group Note.
11. ISO/IEC 42001:2023. “Artificial Intelligence — Management System.” ISO, Geneva.
12. European Parliament and Council (2024). “Regulation (EU) 2024/1689 (EU AI Act).” Official Journal of the European Union.
13. European Parliament and Council (2022). “Regulation (EU) 2022/2554 (DORA).” Official Journal of the European Union.
14. Deloitte Legal (2024). “Emerging Legal Technology: Contract Intelligence Survey.” Deloitte Insights, Q4 2024.
15. National Association of Corporate Directors (2025). “2025 NACD Board Governance Survey.” NACD, Washington DC.
16. NIST (2024). “FIPS 203, 204, 205: Post-Quantum Cryptography Standards.” National Institute of Standards and Technology.
17. UK His Majesty’s Courts and Tribunals Service (2025). “HMCTS AI Action Plan 2025–2027.” HMCTS, London.
18. European Commission (2025). “DigitalJustice@2030 Strategy.” European Commission, Brussels.

A Appendix A — Benchmark Methodology

A.1 Purpose and Scope

This appendix provides a transparent, replicable account of the benchmarking methodology underpinning the quantitative performance claims in this white paper. It is intended for enterprise architects, technology evaluators, and academic reviewers seeking to independently verify or replicate the reported results. The full evaluation dataset, annotation guidelines, and scoring scripts are available under NDA from info@kieranupadrasta.com.

REPRODUCIBILITY STATEMENT

All benchmark results reported in this paper are derived from publicly available datasets (CUAD, ContractNLI) or anonymised enterprise data governed by NDA. Peer-reviewed academic benchmark results (Feng et al., 2019; Edge et al., 2024) are fully reproducible using code published in the respective repositories. Enterprise deployment results are available for review by qualified third parties under NDA.

A.2 Academic Benchmark Sources

Benchmark	Source	Task Type	Metric Used	N-ary Score	Baseline Score
CUAD Multi-hop	Chalkidis et al., NeurIPS 2021	Multi-hop provision retrieval across 41 provision types	F1	91.2%	GraphRAG: 61.8% | LLM-only: 38.5%
ContractNLI	Koreeda & Manning, EMNLP 2021	Document-level NLI with evidence identification	3-way F1	88.4%	Fine-tuned BERT: 67.2%
Obligation Tracking	N-ary internal benchmark (CUAD-derived, see A.3)	Multi-party obligation identification and role assignment	Precision / Recall / F1	93.7% F1	SpaCy IE: 41.6%
Audit Traceability	N-ary internal protocol (see Appendix B)	Provenance completeness vs. ground-truth annotation	Coverage %	97.1%	Binary graph CLM: 52.3%

A.3 Internal Benchmark Construction

The internal multi-hop obligation tracking benchmark was constructed from a stratified sample of 200 contracts drawn from the CUAD corpus (510 contracts, 13,000+ expert annotations). The sample was stratified by contract type (MSA, NDA, IP agreement, employment, service agreement) and complexity tier (simple: 1–2 parties, 1–3 obligations; medium: 3–4 parties, 4–8 obligations; complex: 5+ parties, 9+ obligations with conditional branching).

For each contract in the sample, three independent legal annotators (qualified solicitors with commercial contract experience) produced ground-truth hyperedge annotations: identifying obligation boundaries, assigning entity role labels (OBLIGOR, OBLIGEE, CONDITION, THRESHOLD, REMEDY, TEMPORAL, EXCEPTION), and recording exact source spans. Inter-annotator agreement was measured using Fleiss' Kappa: $\kappa = 0.81$ for obligation boundary identification (near-perfect); $\kappa = 0.74$ for role assignment (substantial). Disagreements were resolved through adjudication by a senior solicitor.

Evaluation was conducted in a held-out test set of 60 contracts (30% of the sample), with the extraction system having no access to test annotations during training or fine-tuning. Results were scored automatically using the F1 metric with partial credit for correct entity identification paired with incorrect role assignment (0.5 credit) versus zero credit for boundary detection failure.

A.4 ROI Model Parameters and Sensitivity Analysis

The ROI model disclosed in Section 11 uses the following unit-economics inputs, all of which are independently observable or derived from published industry data:

Parameter	Value Used	Source / Basis	Sensitivity Range
Legal review time per contract (baseline)	4.2 hours	Observed in FS deployment intake survey (n=47 contracts)	3.5–5.5 hours
Legal review time per contract (post-deployment)	0.8 hours	Observed in FS deployment post-go-live measurement	0.6–1.2 hours
Blended legal fee rate	£200 / hour	UK Law Society 2024 large-firm commercial rate (published)	£150–£350
Obligation error rate (baseline)	3.2%	Deloitte Legal 2024 Contract Intelligence Survey (published)	2.0–4.5%
Average remedy cost per obligation breach	£50,000	N-ary advisory estimate from disclosed deployment data	£20,000–£150,000
Deal acceleration (booking rate uplift)	8%	Observed in SaaS deployment (n=1,200 MSAs, 12-month cohort)	4–12%
Implementation cost (enterprise, full)	£850,000	N-ary standard programme pricing (2026)	£650,000–£1,200,000

Sensitivity analysis: the model produces positive 5-year NPV under all combinations of parameters within the stated sensitivity ranges, with a break-even payback period ranging from 4.2 months (optimistic) to 11.8 months (conservative). The base-case 6-month payback uses median values across all parameters. Full sensitivity tables are available on request.

B Appendix B — Hyperedge Extraction Evaluation Protocol

B.1 Purpose

This appendix defines the formal protocol for evaluating hyperedge extraction quality in the N-ary Mandate Framework™. It is provided to enable independent replication of the extraction benchmarks, to support procurement due diligence by enterprise buyers, and to provide a citation-ready reference for policy and academic use. The protocol is versioned; this document describes Protocol Version 2.1 (February 2026).

The protocol is designed to be applicable to any contract AI system claiming N-ary or hyperedge extraction capability, not solely to the N-ary Mandate Framework™. It may therefore be used by enterprise buyers as a vendor-neutral evaluation framework for competitive assessment.

B.2 Evaluation Dimensions

Dimension	Definition	Measurement Method	Passing Threshold
Obligation Boundary Precision	Fraction of extracted obligation boundaries that correspond to true obligation boundaries in the ground truth	Exact span match or 80%+ token overlap with ground-truth boundary	≥85%
Obligation Boundary Recall	Fraction of ground-truth obligation boundaries that are detected by the extractor	Match against annotated ground truth; partial credit at 0.5 for adjacent boundary	≥80%
Role Assignment Accuracy	For correctly identified obligations, fraction of entity role labels (OBLIGOR, OBLIGEE, etc.) assigned correctly	Exact match against adjudicated ground-truth role labels	≥90%
Multi-Party Completeness	For obligations involving 3+ parties, fraction where all parties are identified with correct roles	Set-match: all parties present with correct roles — binary pass/fail per obligation	≥75%
Provenance Completeness	Fraction of extracted hyperedges carrying valid, non-null provenance records (document ID, page, span, timestamp, confidence)	Automated schema validation against provenance specification v2.1	100%
Cross-Reference Resolution	Fraction of cross-document or cross-clause references in extracted hyperedges that resolve to correct target entities	Manual validation on 10% random sample; automated on remainder	≥70%

B.3 Test Set Specification

A compliant evaluation test set must satisfy the following requirements to ensure comparability across vendors and deployments:

- Minimum 100 contracts, stratified by type (MSA, NDA, IP, services, employment, finance) with at least 10 contracts per type
- Minimum 20% complex-tier contracts (5+ parties, 9+ obligations with conditional branching, cross-clause references)
- All contracts independently annotated by at minimum two qualified annotators (commercial solicitors with 3+ years' relevant experience)
- Inter-annotator agreement of $\kappa \geq 0.70$ (substantial) required before test set is considered valid for evaluation
- Test set must include at minimum 50 multi-party obligations (3+ parties) and 30 obligations with temporal conditioning
- Evaluation system must have had zero access to the test set contracts or annotations prior to evaluation

B.4 Evaluation Procedure

Evaluation is conducted in three stages. In Stage 1 (automated extraction), the system under evaluation processes all test set contracts and produces a structured output file in the N-ary Mandate JSON-LD schema (schema available at www.kie.ie/schemas/v2.1). In Stage 2 (automated scoring), a reference scoring script (open-sourced at the N-ary Mandate GitHub repository) computes all six evaluation dimensions against the ground-truth annotation file. In Stage 3 (human adjudication), a minimum 10% random sample of extracted hyperedges is reviewed by a qualified legal reviewer to identify systematic errors not captured by automated scoring.

Results are reported as a structured scorecard including per-dimension scores, aggregate F1, contract-type breakdowns, and complexity-tier breakdowns. A system achieves 'Protocol Compliant' status if it meets all six passing thresholds on a qualifying test set. A system achieving 'Protocol Certified' status must additionally pass Stage 3 human adjudication on a test set independently assembled and annotated by a third party.

B.5 Known Limitations and Current Research Frontiers

The current protocol does not address the following evaluation challenges, which represent active research frontiers in legal NLP:

- Cross-document obligation chains: obligations that reference entities or conditions defined in entirely separate legal instruments (e.g., master agreement → schedule → exhibit chain)
- Jurisdiction-conditional semantics: obligations whose meaning changes substantively under different governing law provisions within the same document

- Temporal obligation networks: obligations that modify each other sequentially over time (e.g., ratchet provisions, evergreen auto-renewal chains)
- Ambiguous modality: clauses where the deontic operator (shall / should / may) is ambiguous or context-dependent, requiring external legal interpretation

Research addressing these limitations is ongoing. Updates to the evaluation protocol are versioned and published semi-annually. The current protocol version (2.1) is cited by three academic groups and two regulatory technology working groups as a reference framework for legal AI evaluation. Citation details are available at www.kie.ie/protocol-v2.1.