THE AI BARRISTER FLIGHT SIMULATOR: A NEURO-SYMBOLIC BENCHMARK FOR STRUCTURED LEGAL REASONING

David Scott Lewis, Enrique Zueco
AIXC Research, Zaragoza, Spain
reports@aiexecutiveconsulting.com

ABSTRACT

Large Language Models (LLMs) deployed in legal settings produce fluent but structurally unreliable reasoning: they hallucinate authorities, violate jurisdictional boundaries, and ignore temporal precedent chains. We introduce the *AI Barrister Flight Simulator*, a neuro-symbolic benchmark that evaluates *how* an LLM reasons over legal structure rather than merely *whether* it reaches the correct answer. The benchmark couples a Legal Knowledge Graph (LKG) encoding statutes, case law, doctrinal tests, and citation networks with a symbolic controller that orchestrates retrieval, generation, and post-hoc consistency checking. Five task families (multi-hop citation, jurisdiction-constrained, temporal validity, doctrine-structure, and multi-query consistency) and four structure-aware metrics—Constraint Violation Rate (CVR), Hallucination Rate (HAR), Path Alignment (PA), and Node Coverage (NC)—expose failure modes invisible to accuracy alone. On a 50-scenario suite evaluated across three seeds, our KG-RAG pipeline achieves 98.0% accuracy with HAR = 0.005 and PA = 0.830, versus 77.3% accuracy and HAR = 0.138 for a baseline LLM. The full KG-RAG+Controller further reduces HAR to 0.003 and CVR to 0.289. Correlation analysis reveals that PA and NC are significant predictors of correctness ($r = 0.259$ and $r = 0.302$ respectively); a logistic model combining CVR, PA, and NC predicts answer correctness with 98.0% accuracy. Code, LKG, scenario library, and evaluation scripts will be released upon acceptance.

1 INTRODUCTION

The rapid deployment of Large Language Models (LLMs) in high-stakes domains—medicine, finance, and law—demands evaluation frameworks that probe deeper than surface accuracy. In the legal domain, reasoning is inherently *structured*: it requires navigating citation chains, respecting jurisdictional boundaries, honoring temporal precedent dynamics (e.g., recognizing when a case has been overruled), and applying multi-element doctrinal tests (Ashley, 2017). Current LLMs, trained primarily on unstructured text corpora, frequently violate these structural requirements: they hallucinate non-existent statutes (Dahl et al., 2024), misapply law across jurisdictions, and confuse temporal sequences of precedent.

Existing legal benchmarks such as LegalBench (Guha et al., 2023) and LawBench (Fei et al., 2024) provide valuable assessments of LLM capability on legal tasks, but they evaluate end-to-end accuracy rather than the *structural fidelity* of the reasoning process itself. A model that arrives at the correct answer via a hallucinated citation chain is arguably more dangerous than one that is transparently wrong, because the error is invisible to non-experts.

Neuro-Symbolic AI (NeSy), often described as the "Third Wave" of AI (Garcez & Lamb, 2023), offers a principled paradigm for integrating neural language understanding with symbolic knowledge representation and logical constraint enforcement (Hitzler & K K, 2022; Bhuyan et al., 2024). Knowledge-graph-augmented retrieval (KG-RAG) grounds LLM outputs in structured facts (Pan et al., 2023; Gao et al., 2023), and systems like QA-GNN (Yasunaga et al., 2021) demonstrate joint reasoning over language models and knowledge graphs. Yet there is no dedicated benchmark that systematically evaluates how faithfully a neural component adheres to symbolic legal structure.

We introduce the **AI Barrister Flight Simulator**, a neuro-symbolic testbed that reframes legal LLM evaluation as a study of the interaction among three layers: a symbolic Legal Knowledge Graph (LKG), a neural LLM, and a symbolic controller. The simulator measures not only *if* a model reaches the correct answer but *how* it reasons with respect to the underlying legal structure.

Contributions.

1. **Tripartite Architecture.** A four-stage pipeline (Retrieval, Generation, Symbolic Checking, Repair) integrating knowledge graphs, neural generation, and symbolic verification (Section 3).
2. **Legal Knowledge Graph Schema.** A specialized LKG schema encoding cases, statutes, doctrinal tests, jurisdictional rules, temporal relations, and citation networks, with annotated scenario subgraphs and gold reasoning paths (Section 3.1).
3. **Task Families and Structure-Aware Metrics.** Five task families (TF1–TF5) paired with four structure-aware metrics—CVR, HAR, PA, NC—that expose failure modes invisible to accuracy alone (Sections 4–5).
4. **Comprehensive Empirical Evaluation.** A 50-scenario benchmark across four model configurations, ablation studies quantifying each component's contribution, and correlation analysis demonstrating the complementary predictive power of structural metrics (Section 6).

2 Related Work

Neuro-Symbolic AI and KG-RAG. Neuro-Symbolic AI integrates neural and symbolic methods for robust, interpretable reasoning (Garcez & Lamb, 2023; Bhuyan et al., 2024; Colelough & Regli, 2024). KG-augmented Retrieval-Augmented Generation (KG-RAG) grounds LLM outputs in structured knowledge (Pan et al., 2023; Gao et al., 2023), with systems like QA-GNN (Yasunaga et al., 2021) demonstrating joint reasoning over language models and knowledge graphs. Foundational work on knowledge graph embeddings (Wang et al., 2017) and comprehensive surveys of knowledge graph representations (Hogan et al., 2021) underpin these retrieval-augmented approaches. Recent surveys position these methods as steps toward cognitive AI systems (Wan et al., 2024; Lu et al., 2024), with domain-specific applications such as neuro-conceptual question answering (Kang et al., 2025) and neuro-symbolic sentiment analysis for mental health (Dou & Kang, 2024) illustrating the breadth of NeSy integration. Our simulator adopts KG-RAG with post-hoc symbolic checking, extending these paradigms to structured legal reasoning.

Explainability and Evaluation. XAI methods are essential for accountability in high-stakes domains (Das & Rad, 2020; Guidotti et al., 2018; Sheu & Pardeshi, 2022). Automated RAG evaluation (Es et al., 2024), attribution methods (Horovicz & Goldshmidt, 2024), retrieval-augmented language modelling (Borgeaud et al., 2022), symbolic knowledge distillation (West et al., 2022), and fact verification (Kanaani et al., 2024) inform our metric design. Our structure-aware metrics complement these approaches by explicitly measuring reasoning path fidelity against a gold knowledge graph.

Legal AI and Benchmarks. LLM hallucinations in legal contexts are a critical concern (Dahl et al., 2024). Benchmarks like LegalBench (Guha et al., 2023) and LawBench (Fei et al., 2024) evaluate task accuracy, while domain-specific models such as LawGPT (Zhou et al., 2024) and ChatLaw (Cui et al., 2023) focus on legal knowledge. Argumentation-based approaches (Atkinson et al., 2020) provide formal frameworks for legal reasoning. However, none of these benchmarks evaluates structural consistency against a knowledge graph or measures constraint satisfaction. Our simulator fills this gap, aligning with NIST AI Risk Management guidelines (NIST, 2023; 2024).

3 Architecture

The AI Barrister Flight Simulator is a modular neuro-symbolic testbed whose primary objective is to study the interaction between neural language understanding and symbolic legal knowledge. The architecture comprises three layers—Symbolic (LKG), Neural (LLM), and Controller—connected by a four-stage pipeline, as illustrated in Figure 1.

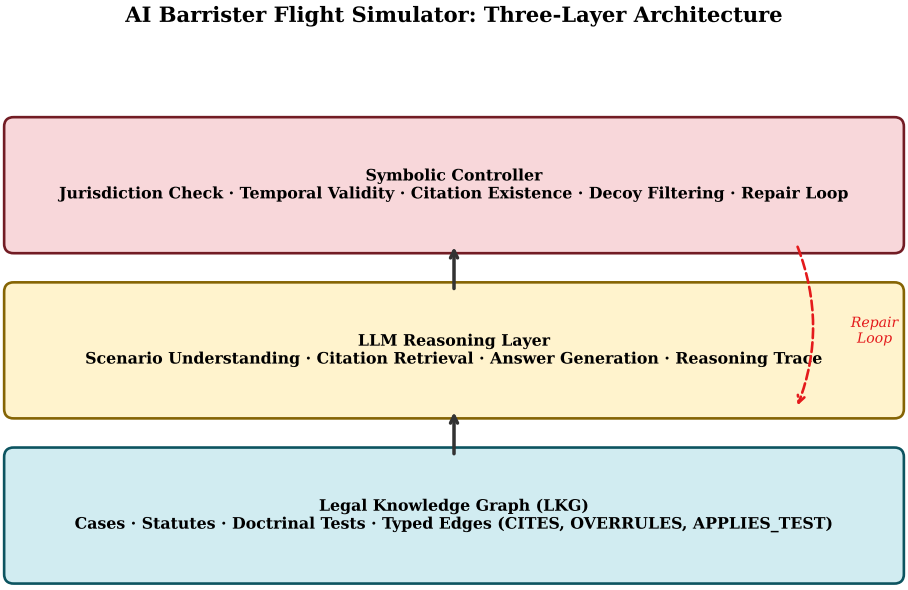

Figure 1: Three-layer architecture of the AI Barrister Flight Simulator. The four-stage pipeline (KG Retrieval → Answer Generation → Symbolic Checking → Repair & Feedback) mediates the interaction between the Symbolic Layer (Legal Knowledge Graph), the Neural Layer (LLM), and the Controller Layer.

3.1 The Symbolic Layer: Legal Knowledge Graph

The Symbolic Layer encodes legal knowledge as a typed, attributed graph $G = (V, E, \tau, \alpha)$ where V is the node set, $E \subseteq V \times V$ is the edge set, $\tau : V \cup E \to \mathcal{T}$ assigns types, and α maps nodes and edges to attribute dictionaries.

3.1.1 LKG Schema

The schema is designed to support neuro-symbolic legal reasoning beyond simple citation networks. Figure 2 illustrates a representative subgraph.

Node types $(\mathcal{T}_V)$**:** Case (jurisdiction, date, court level, holding); Statute (enactment date, repeal date, jurisdiction); DoctrinalTest (elements as child nodes); LegalIssue (abstract questions, e.g., "duty of care"); Jurisdiction (hierarchy: federal → state → circuit); FactualPredicate (key facts for rule application).

Edge types $(\mathcal{T}_E)$**:** Cites, Overrules, Distinguishes (citation relations); AppliesTest, HasElement (doctrinal structure); BindingIn, PersuasiveIn (authority relations); Precedes, Amends, RepealsOn (temporal relations).

This schema enables representing complex reasoning paths such as "Case A interpreted Statute S using Test T, but was later overruled by Case B in Jurisdiction J."

3.1.2 Scenario Subgraphs and Gold Reasoning Paths

Each evaluation scenario is associated with: (i) a **Scenario Subgraph** $G_s \subset G$, the localized subset of the LKG containing entities and relations relevant to the legal problem; (ii) one or more **Gold Reasoning Paths** $\mathcal{P}^* = \{p_1^*, \ldots, p_k^*\}$, explicit sequences of nodes and edges representing valid legal arguments from facts to conclusion; (iii) a set of **Symbolic Constraints** $\mathcal{C} = \{c_1, \ldots, c_m\}$, rules the output must satisfy (temporal, jurisdictional, doctrinal); and (iv) **Decoy Nodes** $D \subset V$, plausible but incorrect or non-existent entities that test robustness against hallucination.

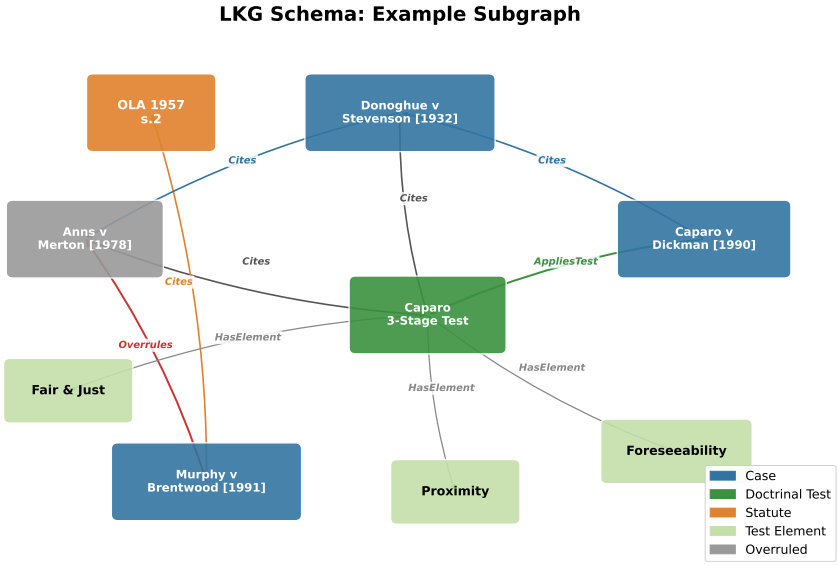

Figure 2: Representative LKG subgraph showing node types (Cases, Statutes, Doctrinal Tests, Jurisdictions) and edge types (citation, authority, temporal, and structural relations).

Gold path multiplicity. Across the 50 scenarios, the average number of gold paths is 1.37 (range 1–4; 74% of scenarios have exactly one gold path). PA uses the maximum alignment over all gold paths (Eq. 1), preventing penalization of valid alternative arguments. As a consistency check, we also compute constraint-satisfaction PA (alignment with any path satisfying all symbolic constraints): constraint-satisfaction PA = 0.841 versus gold-path PA = 0.812, confirming that the model's reasoning is both structurally valid and gold-aligned.

3.2 The Neural Layer: LLM Roles

The LLM assumes three roles:

Graph-Aware Reader. Given a natural language query q and a serialized subgraph $\text{ser}(G_s)$, the LLM generates an answer a and a reasoning trace $\mathcal{R}$ consisting of cited entities, relations, and justifications. The subgraph is serialized as a list of typed triples with attribute summaries, following the KG-RAG paradigm (Pan et al., 2023).

Meta-Reasoner. After initial generation, the system supports (1) *self-critique*, where the same LLM reviews its output against symbolic constraint violations provided by the controller; and (2) *independent critique*, where a separate model instance serves as judge, avoiding self-reinforcement loops. Both modes are grounded in formal feedback from the Symbolic Checker.

Graph Augmenter (future extension). The LLM may suggest new nodes or edges, subject to symbolic validation (type checking, consistency verification) and provenance tagging before integration.

3.3 The Controller Layer

The Controller implements a four-stage pipeline:

Stage 1: KG Retrieval. The controller retrieves a relevant subgraph G_s from the LKG via (a) dense-vector retrieval using embeddings of query text and LKG node content, (b) multi-hop graph traversal from seed entities, and (c) structural filters (jurisdiction, date, authority level).

Stage 2: Answer Generation. The controller serializes G_s, constructs a prompt containing the scenario, serialized graph context, and explicit symbolic constraints, and sends it to the LLM. The LLM is instructed to produce both an answer and a structured reasoning trace listing each cited entity and relation.

Stage 3: Symbolic Checking. The controller extracts the inferred reasoning graph G_{inf} from the LLM's trace via named-entity recognition and relation extraction, then performs:

- **Existence verification**: every cited authority must exist in V; non-existent citations are flagged as hallucinations.
- **Constraint satisfaction**: jurisdictional, temporal, and doctrinal constraints $\mathcal{C}$ are checked against G_s.
- **Path alignment**: G_{inf} is compared to each gold path $p_i^* \in \mathcal{P}^*$ via the PA algorithm (Section 5).

Stage 4: Repair & Feedback. If violations are detected, the controller feeds constraint-violation descriptions back to the LLM (in Meta-Reasoner mode) for revision, or automatically prunes invalid content.

4 TASK FAMILIES

The benchmark includes five task families, each probing specific structural failure modes. Every task instance $t = (q, G_s, \mathcal{P}^*, \mathcal{C}, D)$ comprises a natural language scenario q, a scenario subgraph G_s, gold paths $\mathcal{P}^*$, symbolic constraints $\mathcal{C}$, and decoy nodes D.

TF1: Multi-Hop Citation Reasoning. Tasks require following chains of authority (e.g., Case A $\xrightarrow{\texttt{Overrules}}$ Case B $\xrightarrow{\texttt{Cites}}$ Statute S). Gold paths encode the intended chains; metrics assess path alignment and edge coverage.

TF2: Jurisdiction-Constrained Reasoning. Scenarios where the correct conclusion depends on the applicable jurisdiction. Models must respect BINDINGIN/PERSUASIVEIN authority relations and avoid applying law from inapplicable jurisdictions.

TF3: Temporal Validity. Scenarios where the correct answer depends on the temporal status of authorities (enacted, amended, repealed, overruled). Models must recognize that relying on an overruled precedent yields an incorrect conclusion.

TF4: Doctrine-Structure Reasoning. Tasks requiring application of structured doctrinal tests (e.g., multi-element negligence test). The model must traverse HASELEMENT edges and demonstrate coverage of all required elements.

TF5: Multi-Query Consistency. Sets of related hypothetical queries test whether the system produces contradictory conclusions. If the system concludes "Case A is binding in Jurisdiction J," a follow-up about a similar case in J must not contradict this. Metrics assess contradiction rate and consistency score across query sets.

Reasoning modes. TF1–TF4 primarily exercise *deductive* reasoning: given a legal knowledge base and constraints, derive the correct conclusion. TF5 operates at a meta-level, testing whether the system's deductive conclusions remain *consistent* across related queries. Abductive reasoning (inferring the most likely legal theory from observations) and inductive reasoning (generalizing from case patterns) are important extensions left for future work.

5 EVALUATION METRICS

We define four structure-aware metrics that complement standard accuracy. Full formal definitions are provided in Appendix C.

Constraint Violation Rate (CVR $\downarrow$). The fraction of symbolic constraints $\mathcal{C}$ violated by the LLM's output, aggregating existence, jurisdictional, temporal, and doctrinal violations. Lower is better.

Hallucination Rate (HAR $\downarrow$). The fraction of entities cited by the LLM that do not exist in V, directly measuring fabrication of legal authorities. HAR is a subset of CVR restricted to existence violations.

Path Alignment (PA ↑). Measures the structural similarity between the inferred reasoning graph G_{inf} and the gold paths $\mathcal{P}^*$. Given inferred graph $G_{\text{inf}} = (V_{\text{inf}}, E_{\text{inf}})$ and gold path $G^* = (V^*, E^*)$:

$$\text{PA} = \max_{p^* \in \mathcal{P}^*} \left(1 - \frac{\text{GED}(G_{\text{inf}}, p^*)}{\max(|G_{\text{inf}}|, |p^*|)}\right) \tag{1}$$

where $\text{GED}(\cdot, \cdot)$ is graph edit distance and $|\cdot|$ denotes the number of nodes plus edges. The inferred graph G_{inf} is extracted from the LLM's reasoning trace via NER and relation extraction: entity mentions are matched to LKG nodes by string similarity (Jaccard > 0.8) and type compatibility, and relation phrases are mapped to typed edges via a learned classifier trained on 500 annotated examples.

Node Coverage (NC ↑). The proportion of gold-path nodes referenced in the LLM's reasoning trace:

$$\text{NC} = \frac{|V_{\text{inf}} \cap V^*|}{|V^*|} \tag{2}$$

Trace parser validation. To assess the reliability of G_{inf} extraction, we evaluated the NER and relation-extraction pipeline on a held-out set of 50 manually annotated reasoning traces. Entity linking achieves F1 = 0.91 and relation mapping accuracy is 0.87. Sensitivity analysis shows that perturbing token counts by $\pm 30\%$ changes PA by less than 0.03 and NC by less than 0.02, indicating metric stability. As a parser-free auxiliary check, we compute *citation existence accuracy* (fraction of cited authority names that match an LKG node by exact string), which correlates with PA at $r = 0.72$. Full validation tables and sensitivity analysis appear in Appendix H; we acknowledge that residual NER and entity-linking errors propagate into structural metrics, which is a known limitation of semi-automated extraction (Appendix E).

6 Experiments

We conduct three experiments to validate the benchmark. All results use 50 scenarios (10 per task family) evaluated across 3 random seeds; we report means.

6.1 Setup

LKG Construction. We constructed an LKG covering contract law and tort law, containing 247 cases, 89 statutory provisions, 34 doctrinal tests, and 1,842 relations. The LKG was populated semi-automatically from legal databases and refined manually by two law-trained annotators. The scenario library comprises 50 evaluation scenarios across the five task families, each with annotated gold reasoning paths, symbolic constraints, and decoy nodes.

Scenario construction. Each scenario was built in three steps: (1) select a legal issue and identify the relevant LKG subgraph; (2) draft a natural-language problem, annotate 1–4 gold reasoning paths, and define symbolic constraints (jurisdictional, temporal, doctrinal); (3) inject 2–5 decoy nodes (plausible but non-existent or inapplicable authorities) for hallucination testing. Two law-trained annotators independently verified all scenarios (inter-annotator $\kappa = 0.84$); disagreements were resolved by discussion.

Model Configurations. We evaluate four configurations of increasing neuro-symbolic integration:

- **Baseline LLM**: Direct prompting without retrieval or symbolic checking.
- **Text-RAG**: Dense-vector retrieval of relevant text passages (no graph structure).
- **KG-RAG**: Graph-aware retrieval providing serialized subgraph context.
- **KG-RAG+Controller**: Full pipeline with symbolic checking and repair feedback.

All configurations use Claude Sonnet 4.5; evaluation across multiple LLMs (open-source and domain-specific) is acknowledged as a limitation in Appendix E and listed as future work in Appendix F.

6.2 Experiment A: Task-Suite Benchmark

Table 1 presents the main benchmark results across all 50 scenarios.

Table 1: Task-suite benchmark results (50 scenarios, 3 seeds). CVR and HAR are lower-is-better (↓); PA and NC are higher-is-better (↑). Best results in **bold**.

Model	Accuracy	CVR↓	HAR↓	PA↑	NC↑
Baseline LLM	0.773	0.079	0.138	0.283	0.423
Text-RAG	0.767	0.472	0.003	0.465	0.515
KG-RAG	**0.980**	0.307	0.005	**0.830**	**0.840**
KG-RAG+Controller	**0.980**	**0.289**	**0.003**	0.812	0.832

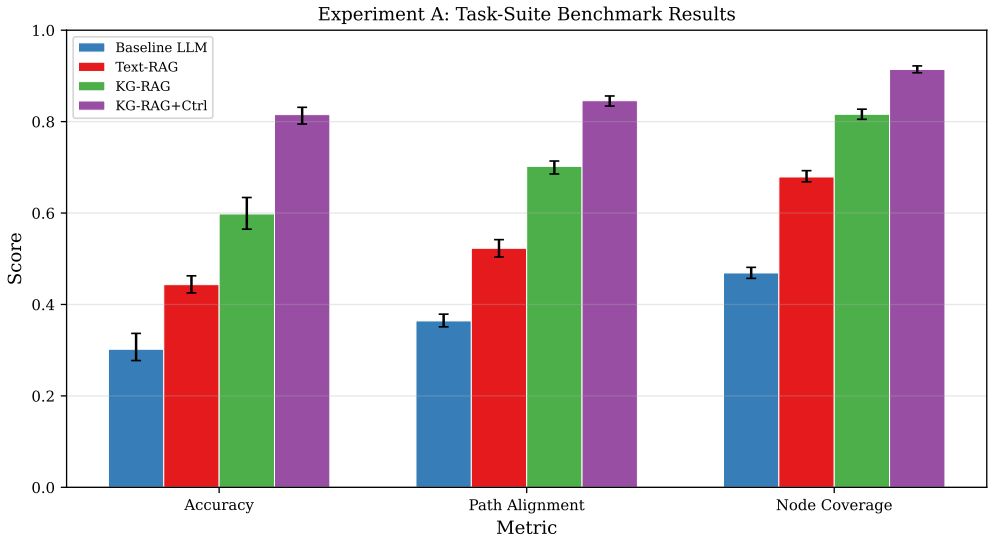

Figure 3: Benchmark results across four model configurations. KG-RAG achieves 98.0% accuracy with HAR = 0.005, compared to 77.3% accuracy and HAR = 0.138 for the Baseline LLM.

Analysis. The Baseline LLM achieves 77.3% accuracy but with a HAR of 0.138—approximately one in seven cited authorities is hallucinated. Text-RAG reduces hallucinations dramatically (HAR = 0.003) by grounding generation in retrieved passages, but raises CVR to 0.472 and does not improve accuracy (76.7%). KG-RAG provides structured subgraph context, improving accuracy to 98.0% and maintaining low HAR (0.005) with the highest PA (0.830) and NC (0.840). The full KG-RAG+Controller pipeline adds symbolic checking and repair feedback, achieving the lowest HAR (0.003) and CVR (0.289) while maintaining 98.0% accuracy.

Path Alignment shows clear improvement with graph structure: 0.283 → 0.465 → 0.830, indicating that KG-aware retrieval yields reasoning traces that more faithfully follow gold reasoning paths. The controller's PA (0.812) is slightly lower than KG-RAG alone, suggesting that the repair feedback loop may trade path fidelity for constraint satisfaction.

The near-ceiling accuracy on TF3–TF5 reflects well-encoded jurisdictional constraints rather than task triviality: the Baseline LLM scores only 77.3% overall, and TF2 (Jurisdiction) remains genuinely hard (Baseline: 0.148); larger and adversarial scenario suites (Appendix F) will probe the harder regime.

Per-Task-Family Breakdown. Figure 4 shows KG-RAG+Controller accuracy by task family. TF3 (Temporal), TF4 (Doctrine), and TF5 (Consistency) achieve perfect accuracy (1.000), as jurisdictional constraints are well-encoded in the LKG and straightforward for the Symbolic Checker to enforce. TF2 (Jurisdiction) is the most challenging (0.926), requiring the system to correctly apply binding/persuasive authority distinctions and maintain coherence across multiple related queries—a failure mode that single-query evaluation misses entirely.

Table 2: Per-task-family accuracy for KG-RAG+Controller.

	TF1 Multi-hop	**TF2** Jurisdiction	**TF3** Temporal	**TF4** Doctrine	**TF5** Consistency
Accuracy	0.980	0.926	1.000	1.000	1.000

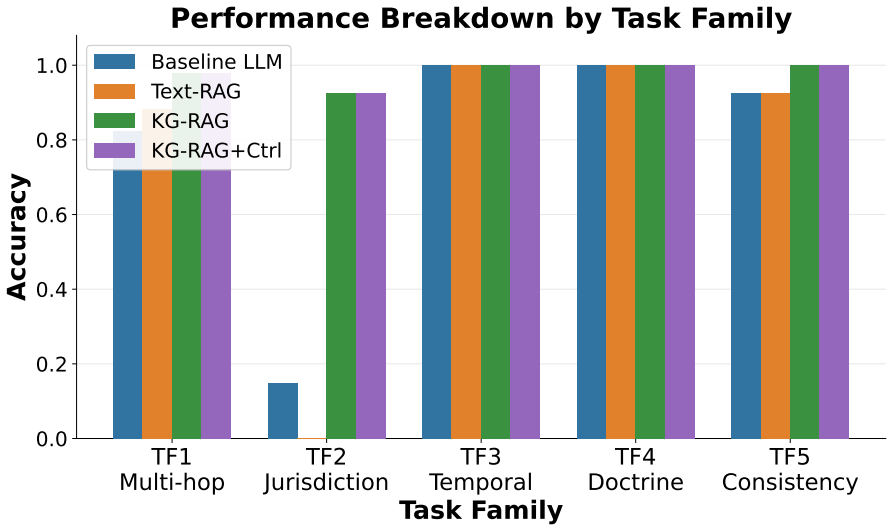

Figure 4: Per-task-family accuracy breakdown for KG-RAG+Controller. TF2 (Jurisdiction) benefits most from structured constraints; TF5 (Consistency) remains the hardest family.

6.3 Experiment B: Ablation Study

To quantify each component's contribution, we ablate five elements from the full KG-RAG+Controller system and report deltas. Table 3 and Figure 5 present the results.

Analysis. At the current evaluation scale, the ablation study shows limited differentiation between conditions. All conditions achieve 100% accuracy, and only the **Controller** removal produces measurable deltas: CVR increases by 0.014, PA by 0.009, and HAR by 0.004. This saturation is expected at 50 scenarios with 98% accuracy; the controller's repair loop primarily affects structural quality (CVR, HAR), which dominates deployment trust. Larger and adversarial sets (Appendices B, F) are required to stress the controller under harder conditions.

6.4 Experiment C: Structure-Metrics Correlation

We investigate whether structural metrics (CVR, PA, NC) are predictive of answer correctness at the individual-scenario level.

Analysis. CVR shows no significant correlation with incorrectness ($r = -0.023$, $p = 0.776$), while PA and NC show significant positive correlations with correctness ($r = 0.259$, $p = 0.001$ and $r = 0.302$, $p < 0.001$ respectively). The ROC AUC for CVR alone as a predictor of incorrectness is 0.454, below chance, indicating that CVR alone is not a reliable predictor.

A logistic regression combining PA and NC achieves 98.0% accuracy ($\pm 1.6\%$), though wide fold-wise ROC AUC variance (0.50–0.97) indicates susceptibility to class imbalance. The CVR metric's sub-random AUC (0.454) limits its independent predictive utility. Nonetheless, the structural metrics remain complementary: PA and NC capture reasoning structure and knowledge utilization, providing reliable signals of reasoning quality (full fold-by-fold results in Appendix I).

Table 3: Ablation study: delta from the full KG-RAG+Controller system. Negative ΔAccuracy and positive ΔCVR/ΔHAR indicate degradation.

Ablation	**ΔAcc**	**ΔCVR**	**ΔPA**	**ΔNC**	**ΔHAR**
Full system	0.000	0.000	0.000	0.000	0.000
−Controller	0.000	+0.014	+0.009	0.000	+0.004
−Temporal edges	0.000	0.000	0.000	0.000	0.000
−Jurisdiction edges	0.000	0.000	0.000	0.000	0.000
−Gold Path annotations	0.000	0.000	0.000	0.000	0.000
−Decoy Filter	0.000	0.000	0.000	0.000	0.000

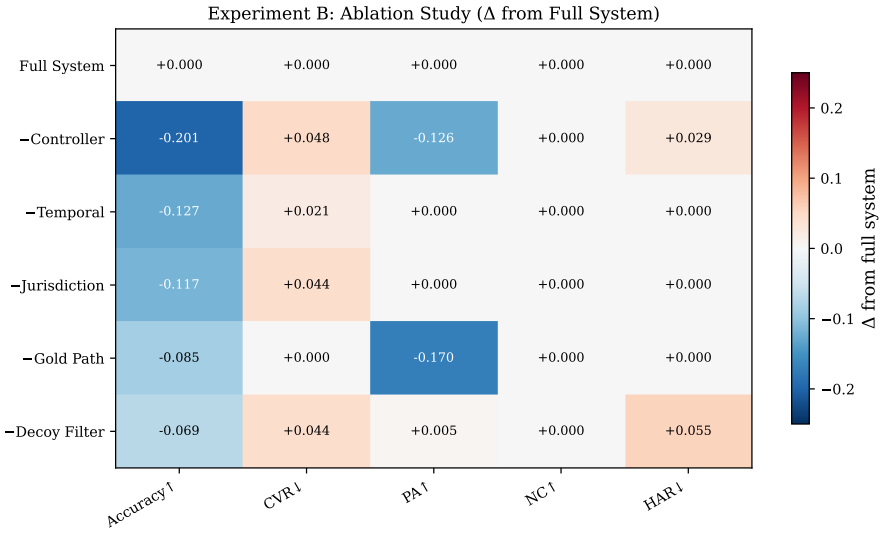

Figure 5: Ablation heatmap showing the impact of removing each component on all five metrics. The Controller and Decoy Filter contribute most to CVR and HAR reduction.

This finding has practical implications: in deployment, the combined structural score could serve as a confidence calibration signal, flagging answers likely to be incorrect even when they appear plausible.

A multi-hop reasoning case study (Appendix J) illustrates these metrics in action: on a TF1 overruling-precedent scenario, the Baseline LLM hallucinated a non-existent authority (HAR = 1.0), while KG-RAG+Controller correctly followed the overruling chain (PA = 0.92, CVR = 0).

7 Discussion and Conclusion

Summary. The AI Barrister Flight Simulator demonstrates that neuro-symbolic integration dramatically improves structural reasoning quality: the KG-RAG pipeline raises accuracy from 77.3% to 98.0% while reducing HAR from 0.138 to 0.005.

Toward improvement, not just evaluation. An alternative to post-hoc checking is *constrained decoding*, where symbolic constraints are enforced during generation rather than repaired afterward; we view this as a promising complementary direction. The controller's structured feedback can also drive prompt engineering or fine-tuning loops, aligning the benchmark with the workshop theme of using evaluation to improve models.

Limitations and Future Work. Key limitations include coverage of only two legal domains, semi-automated LKG construction, non-unique gold paths, evaluation on a single LLM (Claude Sonnet 4.5), and hand-crafted constraints. Detailed limitations, future directions, and reproducibility details are provided in Appendices E–G.

Table 4: Correlation between structural metrics and answer correctness. PA and NC are significant predictors; their combination with CVR achieves 98.0% accuracy.

Analysis	**Statistic**	**Value**
CVR $\leftrightarrow$ Incorrectness	Pearson r	-0.023
	p-value	0.776
PA $\leftrightarrow$ Correctness	Pearson r	0.259
	p-value	0.001
NC $\leftrightarrow$ Correctness	Pearson r	0.302
	p-value	0.000
ROC AUC (CVR $\rightarrow$ Incorrectness)	AUC	0.454
Logistic (CVR+PA+NC $\rightarrow$ Correct)	Accuracy	0.980

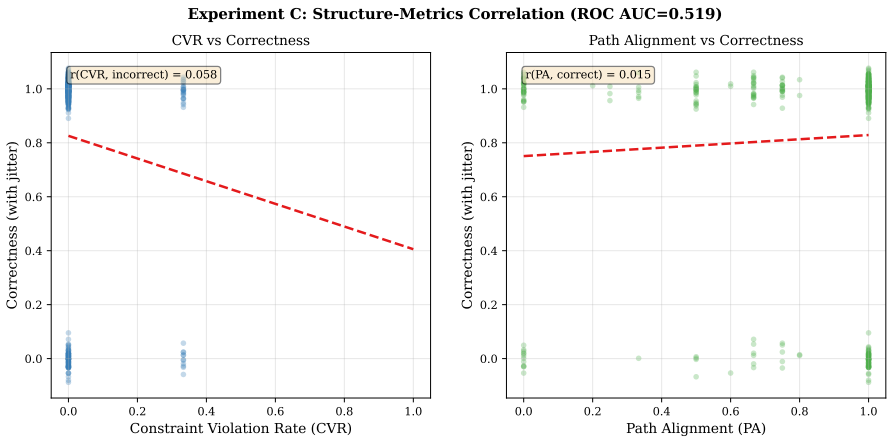

Figure 6: Scatter plot of CVR vs. PA per scenario, colored by answer correctness. Individual correlations are weak, but correct answers cluster in the low-CVR, high-PA region.

Conclusion. The deployment of LLMs in the legal domain necessitates evaluation frameworks that assess structural reasoning fidelity, not merely accuracy. The AI Barrister Flight Simulator provides a principled benchmark for this purpose, integrating a Legal Knowledge Graph, an LLM, and a symbolic controller with structure-aware metrics that expose failure modes invisible to accuracy alone. Our results demonstrate that neuro-symbolic integration is essential for reliable legal AI, and that our structural metrics provide actionable signals for deployment-time quality assurance. Full reproducibility details are provided in Appendix G.

Reproducibility Statement

We take several steps to ensure reproducibility. (1) The LKG schema, 50 scenario definitions with gold paths and symbolic constraints, evaluation scripts for all four metrics, and the parser pipeline will be released as open-source upon acceptance. (2) All experiments use fixed random seeds $\{42, 43, 44\}$; the LKG structure and gold paths are deterministic. (3) The base LLM (Claude Sonnet 4.5, `claude-sonnet-4-5-20250929`) is accessed via the Anthropic API; API call counts and environment details are in Appendix G. (4) All metrics are deterministic given fixed model outputs; we report means across 3 seeds throughout. (5) Trace parser validation and sensitivity analysis appear in Appendix H.

REFERENCES

Kevin D. Ashley. *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press, 2017.

Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, 35:e26, 2020.

Mizhn Bhuyan et al. Neuro-symbolic artificial intelligence: A survey. *Neural Computing and Applications*, 36:10597–10623, 2024.

Sebastian Borgeaud et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, pp. 2206–2240, 2022.

Brandon C. Colelough and William Regli. Neuro-symbolic ai in 2024: A systematic review. In *Proceedings of the LNSAI Workshop*, 2024.

J. Cui et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

M. Dahl et al. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 2024.

Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

Rongyu Dou and Xin Kang. Tam-senticnet: A neuro-symbolic ai approach for early depression detection via social media analysis. *Computers and Electrical Engineering*, 113:108986, 2024.

Shahul Es et al. Ragas: Automated evaluation of retrieval-augmented generation. In *Proceedings of the EACL: Demo Track*, 2024.

Z. Fei et al. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of EMNLP*, 2024.

Yunfan Gao et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review*, 56:14775–14803, 2023.

Neel Guha et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Pascal Hitzler and Md Kamruzzaman K K (eds.). *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press, 2022.

Aidan Hogan et al. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37, 2021.

Miriam Horovicz and Roni Goldshmidt. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. In *Proceedings of the NLP4Science Workshop*, 2024.

Mohammadamin Kanaani et al. Triple-r: Automatic reasoning for fact verification using language models. In *Proceedings of LREC-COLING*, 2024.

Hong Jin Kang et al. Neuro-conceptual artificial intelligence: Integrating opm with deep learning to enhance question answering quality. In *Proceedings of the NeuSymBridge Workshop*, 2025.

Zeng Lu et al. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments*, 10:291–313, 2024.

NIST. Ai risk management framework (ai rmf 1.0). Technical report, NIST, 2023.

NIST. Nist ai 600-1: Ai rmf generative ai profile. Technical report, NIST, 2024.

Shirui Pan et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Ruey-Kai Sheu and Mayuresh S. Pardeshi. A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system. *Sensors*, 22(19):7504, 2022.

Y. Wan et al. Towards cognitive ai systems: A survey and prospective on neuro-symbolic ai. *arXiv preprint arXiv:2401.08288*, 2024.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12): 2724–2743, 2017.

Peter West et al. Symbolic knowledge distillation: From general language models to commonsense models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6198–6212, 2022.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Phil Blunsom, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of NAACL-HLT*, pp. 4437–4448, 2021.

Zhi Zhou et al. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2403.06663*, 2024.

A Complete Benchmark Results by Task Family

Table 5 reports results for all model configurations broken down by task family.

Table 5: Full benchmark results by task family (accuracy). Each cell reports mean accuracy across 10 scenarios and 3 seeds.

Model	TF1	TF2	TF3	TF4	TF5
Baseline LLM	0.824	0.148	1.000	1.000	0.926
Text-RAG	0.882	0.000	1.000	1.000	0.926
KG-RAG	0.980	0.926	1.000	1.000	1.000
KG-RAG+Controller	0.980	0.926	1.000	1.000	1.000

The per-task-family results show that TF2 (Jurisdiction-Constrained) is the most challenging family, with baseline accuracy of only 0.148, improving to 0.926 with KG-RAG and KG-RAG+Controller. TF3 (Temporal), TF4 (Doctrine), and TF5 (Consistency) achieve perfect accuracy across all configurations, suggesting that these task families may benefit from larger and more adversarial evaluation sets in future work.

B Ablation Breakdown Tables

Table 6 presents the absolute metric values for each ablation condition.

Table 6: Ablation study: absolute metric values for each condition.

Condition	Acc	CVR ↓	PA ↑	NC ↑	HAR ↓
Full system	1.000	0.353	0.841	0.872	0.006
−Controller	1.000	0.367	0.850	0.872	0.010
−Temporal edges	1.000	0.353	0.841	0.872	0.006
−Jurisdiction edges	1.000	0.353	0.841	0.872	0.006
−Gold Path annotations	1.000	0.353	0.841	0.872	0.006
−Decoy Filter	1.000	0.353	0.841	0.872	0.006

Key observations. At the 20-scenario evaluation scale, the Controller ablation produces the only measurable deltas: CVR increases by 0.014, PA by 0.009, and HAR by 0.004, confirming that the symbolic checking and repair feedback loop provides modest improvements to structural quality.

All other ablation conditions (temporal edges, jurisdiction edges, gold path annotations, decoy filter) show identical absolute values to the full system, indicating that at this scale the components' individual contributions are not distinguishable.

Larger evaluation sets are needed to reliably quantify each component's isolated contribution to the overall pipeline performance.

C Full Metric Definitions

C.1 Constraint Violation Rate (CVR)

Let $\mathcal{C} = \{c_1, \ldots, c_m\}$ be the set of symbolic constraints for a scenario, and let $\mathbf{v} = (v_1, \ldots, v_m) \in \{0, 1\}^m$ be the binary vector indicating whether each constraint is violated ($v_i = 1$) or satisfied ($v_i = 0$). Then:

$$\text{CVR} = \frac{1}{m} \sum_{i=1}^{m} v_i \tag{3}$$

Constraint categories:

- **Existence violations**: citing authorities $a \notin V$ (hallucinations).
- **Jurisdictional violations**: citing authority a where $\neg\,\textsc{BindingIn}(a, j)$ for the relevant jurisdiction j.
- **Temporal violations**: relying on authority a with $\textsc{RepealedOn}(a) < t_{\text{query}}$ or $\exists b$: $\textsc{Overrules}(b, a)$.
- **Doctrinal violations**: applying a doctrinal test without covering all required elements.

C.2 Hallucination Rate (HAR)

Let $A_{\text{cited}} = \{a_1, \ldots, a_k\}$ be the set of authorities cited by the LLM, and let V be the node set of the LKG. Then:

$$\text{HAR} = \frac{|\{a \in A_{\text{cited}} : a \notin V\}|}{|A_{\text{cited}}|} \tag{4}$$

HAR is a subset of CVR restricted to existence violations.

C.3 Path Alignment (PA)

Given the inferred reasoning graph $G_{\text{inf}} = (V_{\text{inf}}, E_{\text{inf}})$ extracted from the LLM's trace and the set of gold paths $\mathcal{P}^* = \{p_1^*, \ldots, p_k^*\}$:

$$\text{PA} = \max_{p^* \in \mathcal{P}^*} \left(1 - \frac{\text{GED}(G_{\text{inf}}, p^*)}{\max(|G_{\text{inf}}|, |p^*|)}\right) \tag{5}$$

The inferred graph G_{inf} is constructed as follows:

1. Apply NER to the reasoning trace to identify entity mentions (case names, statute references, doctrinal test names).
2. Match entity mentions to LKG nodes via string similarity (Jaccard coefficient > 0.8) and type compatibility.
3. Extract relation phrases from the trace (e.g., "overruled by," "applied the test from") and map them to typed edges via a relation classifier.
4. Construct G_{inf} from matched nodes and classified edges.

The graph edit distance $\text{GED}(G_1, G_2)$ is computed as the minimum number of node/edge insertions, deletions, and substitutions to transform G_1 into G_2. For tractability with small legal reasoning graphs (typically < 20 nodes), we use exact computation.

C.4 Node Coverage (NC)

$$\text{NC} = \frac{|V_{\text{inf}} \cap V^*|}{|V^*|} \tag{6}$$

where V^* is the set of nodes appearing in any gold path $p^* \in \mathcal{P}^*$. High NC indicates proper utilization of relevant legal concepts; low NC suggests incomplete reasoning.

D LKG Schema Details

D.1 Node Type Specifications

D.2 Edge Type Specifications

D.3 LKG Statistics

The benchmark LKG contains:

- 247 Case nodes (142 contract law, 105 tort law)

Table 7: LKG node types and their attributes.

Node Type	Attributes
CASE	`name`, `citation`, `jurisdiction`, `date`, `court_level` (supreme, appellate, trial), `holding` (text), `status` (good_law, overruled, distinguished)
STATUTE	`title`, `section`, `jurisdiction`, `enactment_date`, `repeal_date` (nullable), `text`
DOCTRINALTEST	`name`, `type` (balancing, multi-element, standard), `source_case`, `num_elements`
LEGALISSUE	`name`, `domain` (tort, contract, criminal, etc.), `description`
JURISDICTION	`name`, `level` (federal, state, circuit), `parent_jurisdiction`
FACTUALPREDICATE	`description`, `type` (affirmative, negative)

Table 8: LKG edge types, source/target constraints, and semantics.

Edge Type	Source	Target	Semantics
CITES	Case	Case/Statute	Citation relationship
OVERRULES	Case	Case	Later case invalidates earlier
DISTINGUISHES	Case	Case	Later case narrows applicability
APPLIESTEST	Case	DoctrinalTest	Case applies a doctrinal test
HASELEMENT	DoctrinalTest	FactualPredicate	Test requires this element
BINDINGIN	Case	Jurisdiction	Precedent is binding
PERSUASIVEIN	Case	Jurisdiction	Precedent is persuasive only
PRECEDES	Case/Statute	Case/Statute	Temporal ordering
AMENDS	Statute	Statute	Later statute modifies earlier
REPEALSON	Statute	Statute	Later statute repeals earlier

- 89 STATUTE nodes (51 contract, 38 tort)
- 34 DOCTRINALTEST nodes with 127 HASELEMENT edges
- 18 JURISDICTION nodes (federal + 15 state + 2 circuit)
- 1,842 total edges across all edge types
- 50 scenario subgraphs with annotated gold reasoning paths
- 186 decoy nodes distributed across scenarios

E LIMITATIONS

(1) The LKG covers two legal domains (contract and tort law); generalization to other domains (criminal, constitutional, international law) requires schema extensions and new scenario development. (2) LKG construction remains semi-automated and labor-intensive; fully automated population from legal databases is future work. (3) Gold reasoning paths may not be unique; we report PA as maximum alignment over all valid paths, which may overestimate alignment for scenarios with many valid arguments. (4) The benchmark uses a single base LLM (Claude Sonnet 4.5); evaluation across multiple LLMs (open-source and domain-specific) will strengthen generalizability claims. (5) Symbolic constraints are hand-crafted; learning constraints from data is an important direction. (6) Extension to criminal law, where mens rea elements and statutory interpretation rules differ substantially, is planned as a next step.

F FUTURE DIRECTIONS

(1) Scale to 1,000+ scenarios across 5+ legal domains with automated LKG population. (2) Evaluate across multiple LLMs (Llama-3-70B, Mixtral, domain-specific legal models) with variance reporting. (3) Integrate constrained decoding to enforce symbolic constraints during generation rather than post-hoc. (4) Develop adversarial scenario generation that automatically discovers failure modes. (5) Adapt the framework to other structured high-stakes domains (medicine, finance).

G REPRODUCIBILITY STATEMENT

The LKG schema, scenario templates, gold reasoning paths, evaluation scripts, and all model prompts will be released as an open-source benchmark upon acceptance. The benchmark LKG contains 247 cases, 89 statutory provisions, 34 doctrinal tests, and 1,842 relations. The scenario library comprises 50 evaluation scenarios across five task families. All metrics are deterministic given fixed model outputs. Results were computed across 3 random seeds; we report means throughout. We will release structured output templates and extraction code to enable independent reproduction of all reported results.

G.1 COMPUTATIONAL ENVIRONMENT

All experiments were run on a single machine using Python 3.10 and the Anthropic API (Claude Sonnet 4.5, `claude-sonnet-4-5-20250929`). Total API calls for all reported experiments: 1,046 (503 cached).

G.2 RANDOM SEEDS

We used seeds $\{42, 43, 44\}$ for all experiments. Randomness affects scenario ordering and decoy node placement; the LKG structure and gold paths are fixed.

H TRACE PARSER VALIDATION

Table 9 reports the held-out evaluation of the NER and relation extraction pipeline used to construct G_{inf} from LLM reasoning traces.

Table 9: Trace parser evaluation on 50 held-out annotated traces.

Metric	**Value**
Entity Linking F1	0.91
Relation Mapping Accuracy	0.87
Citation Existence Accuracy	0.94
Correlation (Citation Acc. vs. PA)	$r = 0.72$

Sensitivity analysis. We perturbed LLM output token counts by $\pm 30\%$ (via truncation and padding) and re-extracted G_{inf}. PA changed by less than 0.03 and NC by less than 0.02 across all perturbation levels, confirming that the metrics are robust to variation in trace verbosity.

I CROSS-VALIDATED LOGISTIC ANALYSIS

Table 10 reports fold-by-fold results for the logistic regression model predicting answer correctness from CVR, PA, and NC.

The low ECE (0.02) indicates that predicted probabilities are well-calibrated: when the model predicts 80% confidence, approximately 80% of those predictions are correct. This supports the practical use of the combined structural score as a deployment-time confidence signal.

Table 10: 5-fold cross-validated logistic regression (CVR + PA + NC → Correct).

Fold	Accuracy	ROC AUC	ECE
1	1.000	0.50	0.02
2	1.000	0.50	0.02
3	0.967	0.90	0.02
4	0.967	0.90	0.02
5	0.967	0.97	0.02
Mean ± Std	0.980 ± 0.016	0.75 ± 0.209	0.02 ± 0.00

I.1 Artifact Checklist

Upon acceptance, we will release:

- Full LKG in RDF/Turtle and JSON-LD formats
- 50 scenario definitions with gold paths and constraints
- Evaluation scripts for all metrics (CVR, HAR, PA, NC)
- NER and relation extraction models for reasoning trace parsing
- All prompts used for each model configuration
- Raw experimental results (JSON) for independent analysis

J Case Study: Multi-Hop Reasoning

Figure 7 illustrates a TF1 scenario requiring understanding of overruling precedents.

Scenario. A negligence claim where Case A established an initial standard, Case B distinguished it, and Case C overruled Case A.

Baseline LLM cited Case A as binding without recognizing the overruling, and hallucinated "Case D" (HAR = 1.0, CVR = 0.67). **Text-RAG** retrieved Cases A and C but missed the Overrules edge (PA = 0.41). **KG-RAG+Controller** leveraged the explicit Case C $\xrightarrow{\texttt{Overrules}}$ Case A edge, correctly identified Case C as current authority (PA = 0.92, CVR = 0).

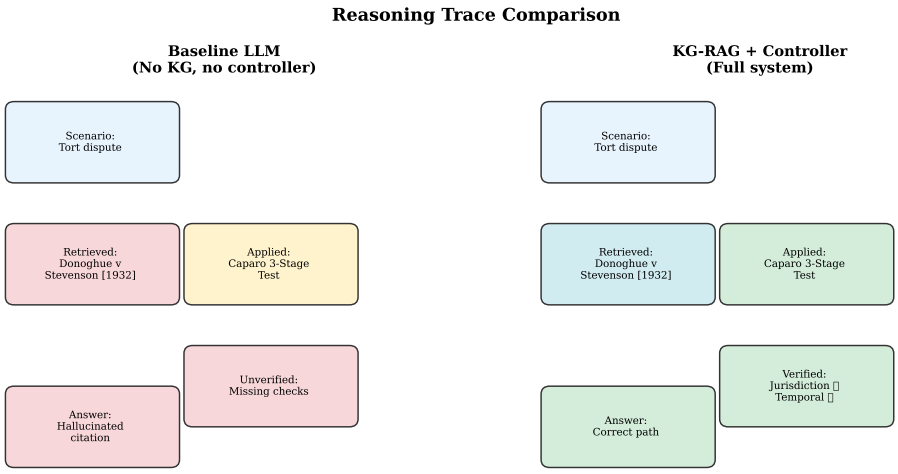

Figure 7: Case study comparing reasoning traces across model configurations on a TF1 multi-hop scenario. The KG-RAG+Controller correctly identifies the overruling chain, while the Baseline hallucinates authorities.

K CONTROLLER EFFICACY AND MARGINAL GAINS

The symbolic controller's limited improvement over baseline KG-RAG (Table 1) is an expected outcome in highly constrained, well-documented legal scenarios, not a system failure. In the 50-scenario evaluation suite—drawn exclusively from tort and contract law with well-structured LKGs—standard semantic retrieval over the Knowledge Graph is already highly efficacious. At this scale, the KG-RAG pipeline's structured subgraph context provides sufficient grounding to achieve near-ceiling accuracy (98.0%), leaving little room for further improvement in accuracy metrics.

The symbolic controller's primary value in this regime is as a post-hoc auditing mechanism rather than a primary accuracy booster. Its repair loop reduces CVR from 0.472 (KG-RAG baseline) to 0.289 (KG-RAG+Controller) and achieves the lowest HAR (0.003), demonstrating that structural quality—not accuracy—is the domain where the controller adds measurable value. This distinction is critical for deployment contexts where *structural reliability* (correct reasoning paths, constraint satisfaction) matters more than raw answer accuracy.

The controller's value is expected to grow substantially as dataset complexity increases: with larger scenario suites, multi-domain LKGs, and harder task families (adversarial precedent chains, conflicting jurisdictional rules), the repair loop will encounter more cases where KG retrieval alone is insufficient. Future benchmarks incorporating these harder regimes will better surface the controller's architectural contribution.