# HalluGraph: Auditable Hallucination Detection for Legal RAG Systems via Knowledge Graph Alignment

**Valentin Noël**, Elimane Yassine Seidou, Charly Ken Capo-Chichi, Ghanem Amari

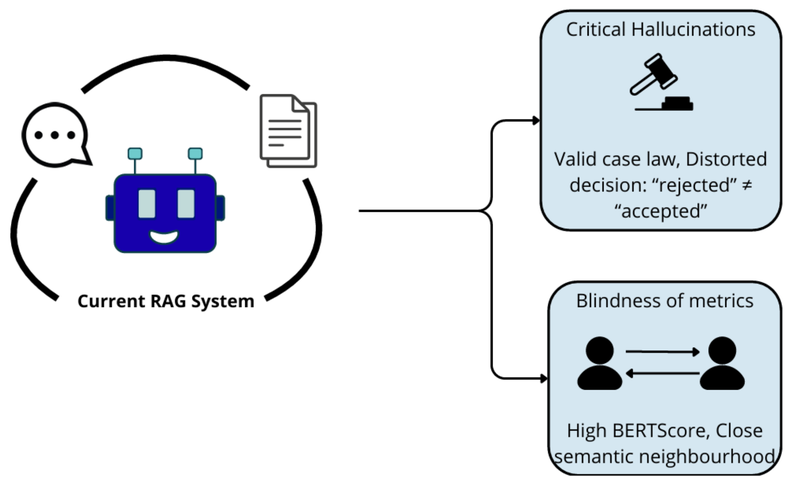
Devoteam

## 1. The Problem: Why Semantic Similarity Fails in Law?

In high-stakes domains like justice, precision is non-negotiable. Current RAG (Retrieval-Augmented Generation) systems suffer from two major flaws:

▶ **Critical Hallucinations:** A model can cite valid case law but misrepresent its holding (”rejected” instead of ”accepted”).

▶ **Blindness of Current Metrics:** Standard metrics (BERTScore [1]) tolerate entity substitutions (swapping ”Plaintiff” and ”Defendant”) because semantic neighborhoods remain close.




**The Challenge:** To ensure Trustworthy AI, we must move from semantic similarity to **structural verification**.

## 3. Auditability Metrics

HalluGraph decomposes fidelity into two interpretable, bounded metrics $[0, 1]$:

▶ **Entity Grounding (EG):** Verifies whether entities mentioned in the response (persons, dates, laws) exist in source documents.

$$\mathrm{EG}(G_a \| G_c, G_q) = \frac{|\{v \in V_a : \exists w \in V_c \cup V_q, \mathrm{match}(v, w)\}|}{|V_a|}$$

*Captures entity substitution hallucinations.*

▶ **Relation Preservation (RP):** Verifies whether asserted relationships (e.g., ”X is liable for Y”) are supported by context.

$$\mathrm{RP} = \frac{1}{|E_a|} \sum_{e \in E_a} 1\left[\exists e' \in E_{\mathrm{ref}} : \mathrm{align}(e, e')\right]$$

*Captures structural and logical hallucinations.*

**Full Audit Trail:** Unlike black-box scores, every flagged error provides concrete evidence:

× **Error Detected**

Entity `'Smith Corp'` missing in Source
Relation `('Defendant', 'liable_for', 'Damages')` unsupported

This ensures **accountability** and **professional responsibility** in legal AI [5].

## 5. Operating Regime & Transparency

### Why Does It Work in Law?

Analysis shows HalluGraph excels on long, dense texts ($> 400$ words, $> 20$ entities), typical of contracts and court opinions. The structural complexity of law becomes an asset for graph-based verification.

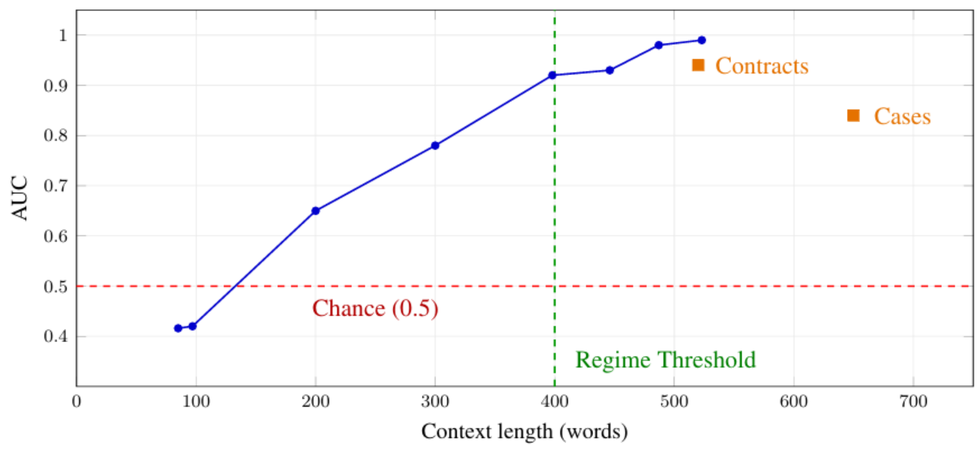



*Figure 1. Performance vs. context length. Legal documents fall in the high-performance regime (AUC $\approx 0.89$).*

## 2. The Solution: HalluGraph

We introduce **HalluGraph**, a graph-theoretic framework that quantifies hallucinations via structural alignment between knowledge graphs (KG) extracted from context, query, and response [2].

### Methodology:

**1. Triple Extraction:** Use a Small Language Model (SLM) to extract entities and relations $(s, r, o)$ from source document $(G_c)$ and generated response $(G_a)$ [3].

**2. Graph Alignment:** Compare graphs to detect structural inconsistencies [4].

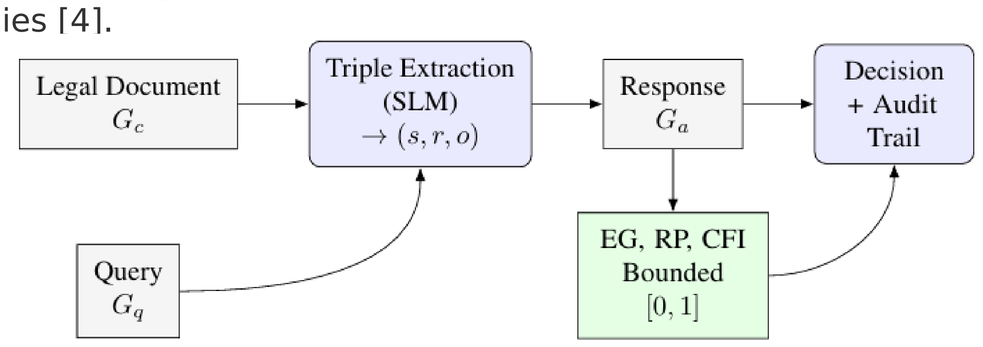



*Figure 2. HalluGraph pipeline: Knowledge graphs extracted from legal documents enable structural verification.*

## 4. Experimental Results: Superior Discrimination

On complex generative legal tasks (Contracts and Case Law), HalluGraph massively outperforms embedding-based methods.

*Table 1. Detection Performance (AUC) on Legal RAG Tasks*

| Dataset | HalluGraph | BERTScore | Gain |
|---|---|---|---|
| Legal Contract QA | **0.94** | 0.60 | **+0.34** |
| Legal Case QA | **0.84** | 0.54 | **+0.30** |
| Coral Biology (Control) | **1.00** | 0.59 | **+0.41** |
| Economics (Control) | **0.99** | 0.55 | **+0.44** |
| **Average (Legal)** | **0.89** | 0.57 | **+0.32** |

**Key Finding:** While BERTScore hovers around chance ($\approx 0.57$) on legal tasks, HalluGraph effectively detects subtle but fatal errors.

## 6. Conclusion: A Bridge to Reliable Legal AI

HalluGraph demonstrates that **structural verification** is critical for deploying AI in legal sectors. By grounding assertions in source text, we provide accountability guarantees for regulatory compliance.

### Key Contributions:

▶ **Bounded metrics** (EG, RP) decomposing fidelity into auditable components.

▶ **Superior performance** on legal RAG tasks (AUC +0.32 vs. baselines).

▶ **Full audit trails** enabling professional accountability.

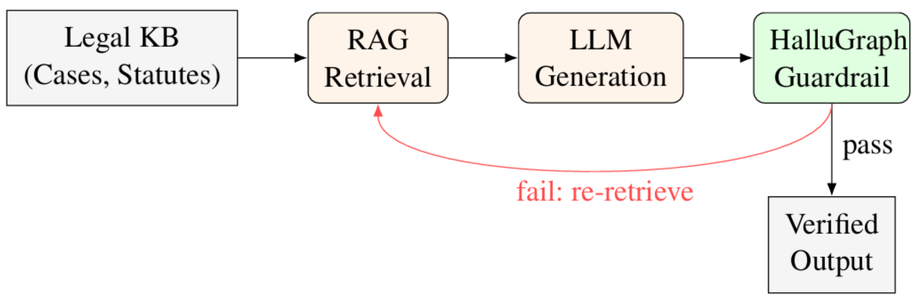



*Figure 3. HalluGraph acts as a post-generation guardrail.*

QR Code Demo

## References

[1] Zhang, T. et al. (2020). BERTScore: Evaluating text generation. *ICLR*.
[2] Lewis, P. et al. (2020). Retrieval-augmented generation. *NeurIPS*.
[3] Huguet Cabot, P.-L. & Navigli, R. (2021). REBEL: Relation extraction. *EMNLP*.
[4] Koutra, D. et al. (2013). Big-Align: Fast bipartite graph alignment. *IEEE ICDM*.
[5] Dahl, M. et al. (2024). LegalBench: Measuring legal reasoning in LLMs. *NeurIPS*.

{valentin.noel, elimane.yassine.seidou, charly.ken.capo-chichi, ghanem.amari}@devoteam.com

AAAI Bridge: Trustworthy AI, Reliability, & Accountability