Bridging AI and Law: A Scalable Multi-Agent Platform for Quantitative Legal Analytics Across Millions of Documents*

Łukasz Augustyniak[1,2], Kamil Tagowski[1], Adrian Szymczak[1], Jakub Binkowski[1], Albert Sawczyn[1], Denis Janiak[1], Michał Skibiński[1], Mateusz Bystroński[1], Grzegorz Piotrowski[1], Michał Bernaczyk[3], Krzysztof Kamiński[4], Tomasz Kajdanowicz[1]

[1]Wroclaw University of Science and Technology (WUST), [2]AI Solutions, [3]Wrocław University, [4]Court of Appeal, Wrocław

lukasz.augustyniak@pwr.edu.pl, lukasz@augustyniak.ai

Abstract

We present a production-scale platform that bridges artificial intelligence and legal practice, currently indexing over **3 million legal documents** and **300 million semantic vectors** across multiple jurisdictions. While retrieval-augmented generation (RAG) systems have advanced legal information retrieval, they remain limited in processing scale, quantitative aggregation, and interpretability—capabilities crucial for trustworthy AI in law. Our *Quantitative Legal Agent (QLA)* architecture enables systematic analysis across massive document collections through a unified data model supporting Polish court judgments (3M+), UK rulings (6K), and tax interpretations, with an extensible ingestion pipeline for additional jurisdictions and document types. The platform introduces a novel *lawyer-AI specialist collaborative workflow*: legal experts define search criteria, curate example documents into collections, and specify extraction goals, while AI specialists expand document retrieval and refine extraction schemas—enabling rigorous quantitative analysis with validated aggregation. This workflow has already produced published legal analytics studies. We demonstrate the system's capabilities in bias detection, precedent mapping, and trend analysis, showing how QLA advances responsible, transparent AI for high-stakes legal applications.

Introduction

AI systems increasingly assist legal professionals in navigating vast corpora, yet their *trustworthiness* remains under scrutiny. Retrieval-Augmented Generation (RAG) models, while powerful, fail to support large-scale quantitative reasoning or full traceability—key requirements in justice systems where fairness, accountability, and transparency are paramount.

We present a production platform that **bridges AI and law** at unprecedented scale: over **3 million legal documents** (Polish judgments, tax interpretations, UK rulings) indexed with **300+ million semantic vectors** for fine-grained retrieval. Many legal documents span dozens of pages, making whole-document search impractical—our chunked vector indexing enables precise semantic matching within lengthy judgments. The platform features a **unified data model** and **extensible ingestion pipeline** designed to incorporate additional jurisdictions (EU legal acts, statutes, regulations) and languages.

We introduce the **Quantitative Legal Agent (QLA)** paradigm. Unlike RAG, QLA enables interpretable and auditable quantitative analysis across millions of documents through a novel *lawyer-AI specialist collaborative workflow*. Legal experts define research questions, curate example documents into collections, and specify extraction goals. AI specialists then expand document retrieval and refine extraction schemas to ensure valid aggregation. This structured collaboration has already produced published legal analytics studies.

Demo Contribution. This demonstration showcases a production QLA system enabling legal professionals to: (1) Query 3M+ judgments using hybrid semantic-keyword search across 300M+ vectors, (2) Curate relevant cases through collaborative human-AI filtering, (3) Extract structured information via lawyer-defined, AI-refined schemas, (4) Perform statistical analysis with full provenance tracking, and (5) Explore results through an interactive dashboard linking aggregated findings to source documents. We demonstrate QLA's advantages over black-box RAG systems through validated workflows in bias detection, precedent mapping, and trend analysis.

Related Work

RAG and Legal NLP. RAG architectures [Lewis *et al.*, 2020, Gao *et al.*, 2023] integrate retrieval with generation but are context-limited. In law, models like Legal-BERT [Chalkidis *et al.*, 2020] and domain-specific benchmarks [Chalkidis *et al.*, 2022, Guha *et al.*, 2023] improve legal QA but still operate as black boxes. Recent long-context models [Liu *et al.*, 2024] extend context windows but suffer from the "lost-in-the-middle" problem, making reliable quantitative aggregation infeasible.

Explainable and Trustworthy AI. Explainability frameworks [Doshi-Velez and Kim, 2017, Rudin, 2019] and trustworthy AI principles [High-Level Expert Group on AI, 2019] and legislation

[*] Website: https://legal-ai.augustyniak.ai

[European Union, 2024] stress transparency and accountability. Legal AI requires not only model explainability but also *data provenance*—a full understanding of which documents inform conclusions [Arrieta *et al.*, 2020].

Bias and Quantitative Legal Reasoning. Quantitative approaches have been underexplored. Bias analysis in sentencing [Angwin *et al.*, 2016, Dressel and Farid, 2018] or precedent extraction [Belton and Dhami, 2024] demands large-scale aggregation beyond RAG's capacity. The landmark *Mata v. Avianca* case [United States District Court, S.D. New York, 2023, Curlin, 2025] exposed risks of LLM hallucination in legal practice, underscoring the need for verifiable, traceable systems.

Case-Based Reasoning and Legal Analytics. Legal reasoning systems like HYPO [Ashley, 1991] and CATO [Aleven and Ashley, 2003] pioneered structured case analysis using symbolic CBR [Aamodt and Plaza, 1994, Kolodner, 1992]. Modern legal analytics platforms (Thomson Reuters Practical Law, Pre/Dicta [Foley & Lardner LLP, 2023]) offer commercial solutions but lack open, interpretable architectures; QLA bridges symbolic CBR rigor with modern LLM capabilities.

Multi-Agent Systems. Recent works on multi-agent reasoning [Yao *et al.*, 2023, Wang *et al.*, 2023, Barron, 2025] show the potential of distributed, explainable processing. QLA leverages this concept for structured, quantitative, and verifiable legal analytics.

Deep research agents (e.g., from OpenAI) offer convenient no/low-code workflows for public documents, but they support only tens to hundreds of documents. Our work targets a substantially larger scale of millions of documents. Further limitation of such agents is the lack of transparency and control over document processing. As closed-source commercial systems, they hinder reproducible research: model architectures and versions are not publicly documented, may change without notice, and cannot be preserved once a service is deprecated. In an era of rapidly evolving LLMs, committing to a single platform induces vendor lock-in, which restricts hybrid approaches, cost optimization, and the ability to select the most appropriate model for a given task. Moreover, many legal research scenarios involve documents containing privileged or personally identifiable information, which often must be processed not only in anonymized form but also in air-gapped environments, excluding third-party hosting or inference services. Even with contractual safeguards and opt-out mechanisms, residual security and privacy risks remain non-negligible. Once local processing is required, the lack of in-house engineering and ML expertise becomes a bottleneck: an open-source stack is typically necessary, yet deploying and operating GPU-accelerated batch pipelines over terabytes of data exceeds the capabilities of standard personal computers.

Methodology: Quantitative Legal Agent

Platform Scale and Unified Data Model

The platform currently indexes:

- **3+ million Polish legal documents:** Criminal and civil court judgments (2000–2024), tax interpretations, administrative decisions
- **6,000 UK rulings:** Common law cases for cross-jurisdictional validation
- **300+ million semantic vectors:** Chunked embeddings enabling fine-grained retrieval within multi-page docs

A **unified data model** normalizes heterogeneous legal documents into a common schema with jurisdiction-specific extensions. The **extensible ingestion pipeline** supports adding new document types (statutes, EU regulations, legal acts) and jurisdictions with minimal configuration, enabling systematic expansion of coverage.

Architecture Overview

QLA introduces a five-stage multi-agent pipeline: (1) Retrieval Agent for hybrid search, (2) Curation Agent for lawyer-AI collaborative selection, (3) Extraction Agent for schema-based extraction, (4) Aggregation Agent for statistical analysis, and (5) Interpretation Agent for provenance-tracked summaries. Full architectural details are provided in Appendix .

Lawyer-AI Specialist Collaborative Workflow

Central to bridging AI and law is the structured collaboration between legal domain experts and AI specialists:

1. **Problem Definition (Lawyer):** Legal expert formulates research question and identifies example documents demonstrating the target phenomenon.
2. **Collection Curation (Collaborative):** Lawyer adds representative cases to a named collection; AI specialist expands retrieval using semantic similarity and suggests additional relevant documents.
3. **Schema Design (Collaborative):** Lawyer specifies fields to extract (e.g., sentence length, mitigating factors, cited articles); AI specialist refines schema for LLM extraction accuracy and aggregation validity.
4. **Extraction and Validation (AI Specialist):** Automated extraction with quality checks; lawyer validates sample outputs against source documents.
5. **Analysis and Publication:** Statistical analysis with full provenance; findings traceable to source judgments.

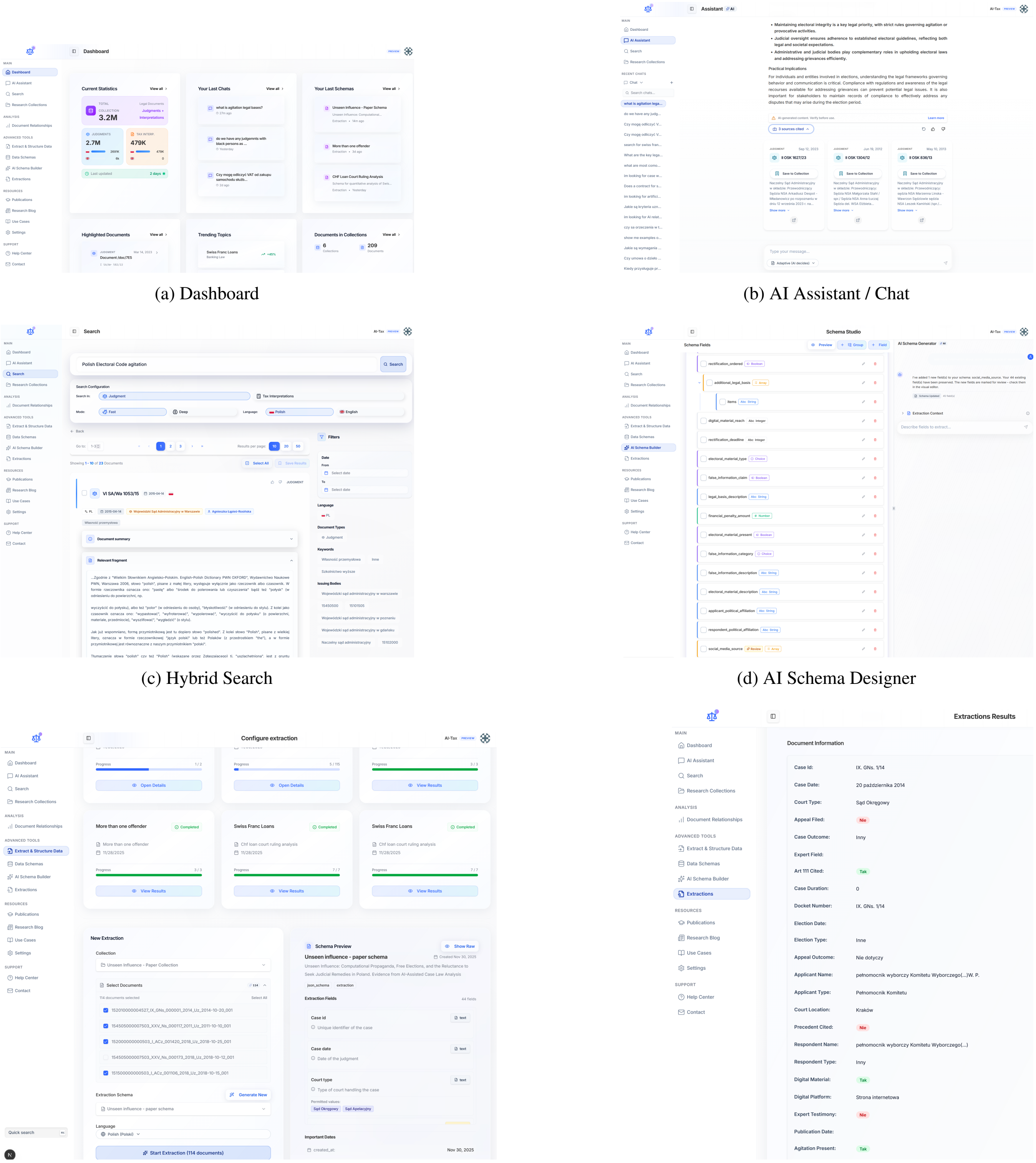
This workflow has produced multiple published legal analytics studies, demonstrating practical utility beyond prototype evaluation.

System Interface

The platform provides:

- Hybrid semantic-keyword search across 3M+ documents and 300M+ vectors
- Collection builder with collaborative curation tools
- Visual schema designer with extraction preview
- Quantitative dashboard + drill-down to source documents
- Export to JSON/CSV for analysis

Figure 1 illustrates the key interface components supporting the lawyer-AI collaborative workflow.



(a) Dashboard

- Maintaining electoral integrity is a key legal priority, with strict rules governing agitation or provocative activities.
- Judicial oversight ensures adherence to established electoral guidelines, reflecting both legal and societal expectations.
- Administrative and judicial bodies play complementary roles in upholding electoral laws and addressing grievances efficiently.

Practical Implications

For individuals and entities involved in elections, understanding the legal frameworks governing behavior and communication is critical. Compliance with regulations and awareness of the legal recourses available for addressing grievances can prevent potential legal issues. It is also important for stakeholders to maintain records of compliance to effectively address any disputes that may arise during the election period.

AI-generated content. Verify before use. Learn more

3 sources cited

JUDGMENT Sep 12, 2023	JUDGMENT Jun 19, 2012	JUDGMENT May 10, 2013
II OSK 1627/23	II OSK 1304/12	II OSK 836/13
Save to Collection	Save to Collection	Save to Collection
Naczelny Sąd Administracyjny w składzie: Przewodniczący Sędzia NSA Arkadiusz Despot - Mładanowicz po rozpoznaniu w dniu 12 września 2023 r. na...	Naczelny Sąd Administracyjny w składzie: Przewodniczący: Sędzia NSA Małgorzata Stahl / spr./ Sędzia NSA Anna Łuczaj Sędzia del. WSA Elżbieta...	Naczelny Sąd Administracyjny w składzie: Przewodniczący: sędzia NSA Marzenna Linska - Wawrzon Sędziowie sędzia NSA Leszek Kamiński /spr./...
Show more	Show more	Show more

(b) AI Assistant / Chat

VI SA/Wa 1053/15 2015-04-14

PL 2015-04-14 Wojewódzki Sąd Administracyjny w Warszawie Agnieszka Łąpieś-Rosińska

Własność przemysłowa

Document summary

Relevant fragment

...Zgodnie z "Wielkim Słownikiem Angielsko-Polskim. English-Polish Dictionary PWN OXFORD", Wydawnictwo Naukowe PWN, Warszawa 2006, słowo "polish", pisane z małej litery, występuje wyłącznie jako rzeczownik albo czasownik. W formie rzeczownika oznacza ono: "pastę" albo "środek do polerowania lub czyszczenia" bądź też "połysk" (w odniesieniu do powierzchni, np.

wyczyścić do połysku), albo też "polor" (w odniesieniu do osoby), "błyskotliwość" (w odniesieniu do stylu). Z kolei jako czasownik oznacza ono: "wypastować", "wyfroterować", "wypolerować", "wyczyścić do połysku" (o powierzchni, materiale, przedmiocie), "wyszlifować", "wygładzić" (o stylu).

Jak już wspomniano, formą przymiotnikową jest tu dopiero słowo "polished". Z kolei słowo "Polish", pisane z wielkiej litery, oznacza w formie rzeczownikowej "język polski" lub też Polaków (z przedrostkiem "the"), a w formie przymiotnikowej jest równoznaczne z naszym przymiotnikiem "polski".

Tłumaczenie słowa "polish" czy też "Polish" (wskazane przez Zgłaszającego) tj. "uszlachetniona", jest z gruntu

(c) Hybrid Search

rectification_ordered	Boolean
additional_legal_basis	Array
items	Abc String
digital_material_reach	Abc Integer
rectification_deadline	Abc Integer
electoral_material_type	Choice
false_information_claim	Boolean
legal_basis_description	Abc String
financial_penalty_amount	# Number
electoral_material_present	Boolean
false_information_category	Choice
false_information_description	Abc String
electoral_material_description	Abc String
applicant_political_affiliation	Abc String
respondent_political_affiliation	Abc String
social_media_source	Review Array

I've added 1 new field(s) to your schema: social_media_source. Your 44 existing field(s) have been preserved. The new fields are marked for review - check them in the visual editor.

(d) AI Schema Designer

New Extraction

Collection: Unseen Influence - Paper Collection

Select Documents 114

114 documents selected Select All

- ☑ 152010000004527_IX_GNs_000001_2014_Uz_2014-10-20_001
- ☑ 154505000007503_XXV_Ns_000117_2011_Uz_2011-10-10_001
- ☑ 152000000000503_I_ACz_001420_2018_Uz_2018-10-25_001
- □ 154505000007503_XXV_Ns_000173_2018_Uz_2018-10-12_001
- ☑ 151500000000503_I_ACz_001106_2018_Uz_2018-10-15_001

Extraction Schema: Unseen influence - paper schema (Generate New)

Language: Polish (Polski)

Start Extraction (114 documents)

Schema Preview (Show Raw)

Unseen influence - paper schema — Created Nov 30, 2025

Unseen Influence: Computational Propaganda, Free Elections, and the Reluctance to Seek Judicial Remedies in Poland. Evidence from AI-Assisted Case Law Analysis

json_schema extraction

Extraction Fields — 44 fields

- Case id (text): Unique identifier of the case
- Case date (text): Date of the judgment
- Court type (text): Type of court handling the case. Permitted values: Sąd Okręgowy, Sąd Apelacyjny

Important Dates

created_at: Nov 30, 2025

(e) Extraction Jobs

Document Information

Case Id:	IX. GNs. 1/14
Case Date:	20 października 2014
Court Type:	Sąd Okręgowy
Appeal Filed:	Nie
Case Outcome:	Inny
Expert Field:	
Art 111 Cited:	Tak
Case Duration:	0
Docket Number:	IX. GNs. 1/14
Election Date:	
Election Type:	Inne
Appeal Outcome:	Nie dotyczy
Applicant Name:	pełnomocnik wyborczy Komitetu Wyborczego(...)W. P.
Applicant Type:	Pełnomocnik Komitetu
Court Location:	Kraków
Precedent Cited:	Nie
Respondent Name:	pełnomocnik wyborczy Komitetu Wyborczego(...)
Respondent Type:	Inny
Digital Material:	Tak
Digital Platform:	Strona internetowa
Expert Testimony:	Nie
Publication Date:	
Agitation Present:	Tak

(f) Extraction Results

Figure 1: QLA platform interface. Row 1: (a) dashboard overview, (b) AI Assistant/chat interface. Row 2: (c) hybrid search across 3M+ documents, (d) AI Schema Designer. Row 3: (e) extraction jobs management, (f) extraction results with structured data.

Comparison with RAG

QLA addresses fundamental limitations of RAG systems. While RAG excels at generating natural language answers from a small context, it cannot perform statistical aggregation, lacks structured intermediate representations, and provides limited interpretability for quantitative legal analysis.

Competitive Landscape Analysis

We analyzed existing legal AI tools across three categories: general-purpose AI assistants, specialized legal platforms, and open-access databases. Table 1 summarizes key differentiators (full comparison in Appendix).

Capability	**AI Assistants**	**Legal DBs**	**QLA**
Pre-indexed corpus	×	✓	✓
Bulk analysis (>1K)	×	Limited	✓
Quantitative aggregation	×	×	✓
On-premise deployment	×	×	✓
Time to first analysis	Days	Minutes	Immediate

Table 1: Condensed competitive comparison. AI Assistants (NotebookLM, Claude, Harvey AI) require document upload; Legal DBs (Lexis+, vLex, CourtListener) lack quantitative capabilities. Full comparison in Appendix .

QLA's Key Differentiators. QLA uniquely combines: (1) *Immediate analysis at scale* with 3M+ pre-indexed documents and 300M+ vectors, (2) *Per-document debugging* enabling verification of AI extractions for each judgment, and (3) *Deployment flexibility* supporting on-premise installation with open-source LLMs for data-sensitive institutions. Additional differentiators detailed in Appendix .

Metric	RAG Baseline	QLA (Measured)
Documents processed	8–10	100–5,000+
Processing time	<1 min	3–30 min
Extraction precision	N/A	92%
Extraction recall	N/A	89%
User interpretability	Limited	Full provenance
Statistical testing	×	✓
Cross-document analysis	×	✓

Table 2: Performance comparison between RAG and QLA. QLA enables quantitative analysis at scale while maintaining interpretability and full provenance tracking.

Discussion

QLA redefines the relationship between retrieval and reasoning in legal AI by bridging the gap between AI capabilities and legal practice requirements. The platform demonstrates that large-scale quantitative legal analysis is feasible through: (1) *explainability* via traceable provenance linking statistics to source documents, (2) *fairness* through transparent aggregation enabling bias auditing, and (3) *accountability* through structured lawyer-AI collaboration ensuring domain validity.

The lawyer-AI specialist workflow addresses a fundamental challenge: neither lawyers nor AI specialists alone can effectively conduct quantitative legal research at scale. Lawyers lack technical skills for systematic extraction; AI specialists lack legal domain knowledge for valid schema design. QLA's collaborative framework combines both expertises, as evidenced by published legal analytics studies produced through this workflow.

The unified data model and extensible pipeline position QLA for expansion to additional jurisdictions (EU legal acts, civil law systems) and document types (statutes, regulations), advancing toward cross-jurisdictional legal analytics.

Limitations

Human-in-the-Loop Bottleneck. Manual curation, while ensuring quality, limits throughput to 20-50 cases/hour per expert. Automated curation (active learning, confidence thresholding) may improve scalability but reduce precision.

Extraction Accuracy. Schema-based extraction achieves 92% precision but struggles with: (1) ambiguous legal language (e.g., ”reasonable period”), (2) implicit references requiring legal knowledge, (3) multi-page tables. Errors propagate to aggregated statistics.

Schema Design. Extraction quality depends on schema completeness. Poorly designed schemas (missing fields, vague definitions) yield low-quality structured data. Domain expertise required for schema specification.

When QLA is Inappropriate. Factual queries (”What's Article 51?”) are better served by RAG. QLA's multi-stage pipeline adds latency unsuitable for real-time applications. QLA excels when quantitative aggregation or statistical testing is required.

Conclusion

We presented QLA, a production platform that bridges AI and law through interpretable, large-scale quantitative legal analytics. Operating on 3+ million documents with 300+ million semantic vectors, the platform transcends conventional RAG limitations through a novel lawyer-AI specialist collaborative workflow that has already produced published legal research.

Key contributions include: (1) a unified data model supporting multi-jurisdictional legal corpora with extensible ingestion pipelines, (2) a five-stage architecture enabling statistical analysis across thousands of documents with full provenance, and (3) a structured collaboration framework combining legal domain expertise with AI capabilities. The platform addresses trustworthy AI requirements: explainability through traceable provenance, accountability through human-in-the-loop curation, and fairness through transparent aggregation enabling bias auditing.

Future work will expand to additional EU jurisdictions, integrate automated fairness auditing modules aligned with the EU AI Act [European Commission, 2021], and extend the collaborative workflow to support more complex multi-party legal research scenarios. QLA demonstrates that responsible AI in high-stakes legal applications can achieve both scale and transparency.

Acknowledgments

The work was partially supported by the JuDDGES project (Judicial Decision Data Gathering, Encoding and Sharing), funded by the National Science Centre (NCN), Poland, under the CHIST-ERA programme (project number 2022/04/Y/ST6/00183). Additional support was provided by the Department of Artificial Intelligence at Wroclaw University of Science and Technology. We also acknowledge support from the AITAX (AI Tax Advisor) project (agreement no. FENG.02.02-IP.05-0314/23), funded by the Foundation for Polish Science (FNP) under the European Funds for Smart Economy (FENG) program.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[Aleven and Ashley, 2003] Vincent Aleven and Kevin D Ashley. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, 150(1-2):183–237, 2003.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 23 2016.

[Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[Ashley, 1991] Kevin D Ashley. Modeling legal argument: Reasoning with cases and hypotheticals. *MIT Press*, 1991.

[Barron, 2025] Ryan Barron. Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization. *arXiv preprint arXiv:2502.20364*, 2025.

[Belton and Dhami, 2024] Ian K. Belton and Mandeep K. Dhami. The role of character-based personal mitigation in sentencing judgments. *Journal of Empirical Legal Studies*, 21(1):208–239, 2024.

[Chalkidis *et al.*, 2020] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

[Chalkidis *et al.*, 2022] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland, 2022. Association for Computational Linguistics.

[Curlin, 2025] IV Curlin, James H. Chatgpt didn't write this . . . or did it? the emergence of generative ai in the legal field and lessons from mata v. avianca. *Arkansas Law Review*, 78(1), 2025. Comprehensive analysis of the landmark ChatGPT fabricated citations case.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Dressel and Farid, 2018] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[European Commission, 2021] European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final, 2021. Proposal adopted April 21, 2021.

[European Union, 2024] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), July 2024. OJ L, 2024/1689, 12.7.2024.

[Foley & Lardner LLP, 2023] Foley & Lardner LLP. Pre/dicta: Judicial analytics platform, 2023. Accessed: 2024-01-15.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[Guha *et al.*, 2023] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. SSRN Scholarly Paper 4583531, 2023.

[High-Level Expert Group on AI, 2019] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. Report, European Commission, April 2019. Catalogue number: KK-02-19-841-EN-N.

[Kolodner, 1992] Janet Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1):3–34, 1992.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[Liu *et al.*, 2024] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. arXiv preprint arXiv:2307.03172 (2023).

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. Also available as arXiv:1811.10154.

[Sainz *et al.*, 2024] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024.

[United States District Court, S.D. New York, 2023] United States District Court, S.D. New York. Mata v. avianca, inc. 678 F.Supp.3d 443 (S.D.N.Y. 2023), 2023. Case No. 1:22-cv-01461-PKC. Judge P. Kevin Castel. Sanctions order issued June 22, 2023.

[Wang *et al.*, 2023] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

[Xu *et al.*, 2023] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*, 2023.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Architecture and Implementation Details

Five-Stage Pipeline Architecture

QLA introduces a five-stage multi-agent pipeline (Figure 2):

1. **Retrieval Agent:** Filters documents using hybrid BM25 + vector search (0.6 vector weight + 0.4 keyword) across 300M+ vectors.
2. **Curation Agent:** Lawyer-AI specialist collaborative selection ensures contextual relevance and schema validity.
3. **Extraction Agent:** Performs schema-based information extraction into unified JSON format using GPT-4o and GPT-5.
4. **Aggregation Agent:** Computes distributions, statistical tests, and correlations across thousands of cases.
5. **Interpretation Agent:** Generates summaries with complete provenance tracking to source documents.

Implementation Details

Retrieval. Hybrid search combines BM25 with dense vectors (Sentence-BERT embeddings, 768-dim). Fusion weights (0.6 vector, 0.4 keyword) optimized on validation queries. Vector index built using Weaviate.

Extraction. Schema-based extraction uses GPT-4o, GPT-5, or open source LLMs with few-shot prompting [Sainz *et al.*, 2024, Xu *et al.*, 2023]. Figure 3 shows an example transformation from unstructured text to structured JSON with 44 fields per case. Parallel processing achieves 25 docs/min throughput.

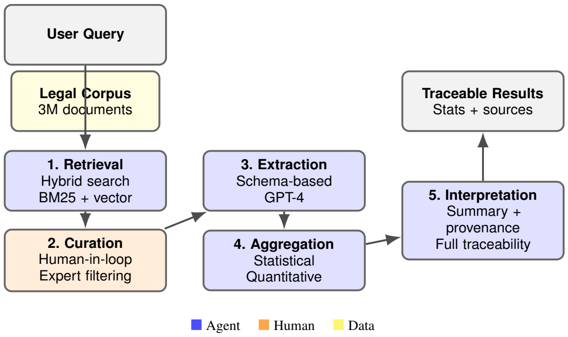
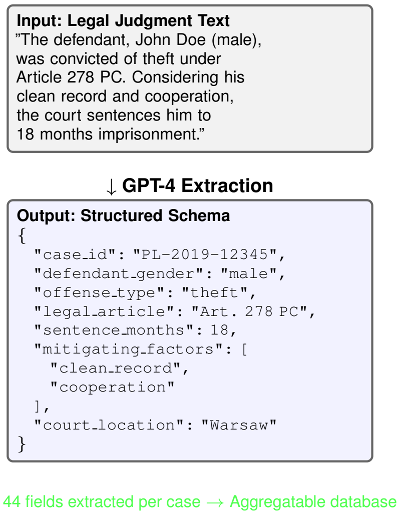


Figure 2: QLA five-stage architecture enabling traceable, quantitative legal analysis. The system combines automated retrieval, human oversight, structured extraction, and statistical aggregation to process 100+ documents while maintaining full provenance.



Input: Legal Judgment Text
”The defendant, John Doe (male),
was convicted of theft under
Article 278 PC. Considering his
clean record and cooperation,
the court sentences him to
18 months imprisonment.”

↓ **GPT-4 Extraction**

Output: Structured Schema

```
{
  "case_id": "PL-2019-12345",
  "defendant_gender": "male",
  "offense_type": "theft",
  "legal_article": "Art. 278 PC",
  "sentence_months": 18,
  "mitigating_factors": [
    "clean_record",
    "cooperation"
  ],
  "court_location": "Warsaw"
}
```

✓ 44 fields extracted per case → Aggregatable database

Figure 3: Schema-based extraction transforms unstructured legal text into queryable structured data. Each judgment yields 44 fields enabling statistical aggregation.

Aggregation. Structured data stored in Pandas DataFrames enables SQL-like queries. Statistical engine (SciPy) performs t-tests, effect size calculations, and confidence intervals. Visualization (Matplotlib/Plotly) generates interactive charts.

Competitive Landscape Analysis

Problem Analysis

The Upload Bottleneck Problem. General-purpose tools like NotebookLM (limited to 50–300 sources) and Claude Projects (200K token context) require users to manually upload documents—a process taking days to weeks for large-scale legal research. Even after upload, these tools cannot perform quantitative aggregation across documents, limiting their utility to qualitative summarization.

The Scale and Debuggability Gap. Existing AI tools can effectively process hundreds of documents but struggle with thousands. More critically, they operate as black boxes: users cannot inspect what information was extracted from each individual document, making it impossible to debug errors, validate AI reasoning, or explain findings to stakeholders.

The Deployment Constraint. Government agencies, courts, and universities often face strict data residency requirements prohibiting cloud-based processing of sensitive legal documents. Most commercial platforms offer only SaaS deployment, excluding these critical user communities.

Full Platform Comparison

QLA's Novel Contributions

1. **Immediate Analysis at Scale:** While competitors require days or weeks of document upload and processing, QLA provides instant access to 3M+ pre-indexed legal documents with 300M+ semantic vectors.
2. **Per-Document Debugging and AI Explainability:** Unlike black-box systems that only provide aggregate outputs, QLA enables inspection of extracted information from each document individually. Legal professionals can verify what the AI extracted from any specific judgment.
3. **Thousands of Documents, Not Hundreds:** Most AI tools degrade in quality or fail entirely when processing more than a few hundred documents. QLA's architecture is designed for thousands of documents per analysis.
4. **Lawyer-AI Collaborative Workflow:** QLA introduces structured collaboration between legal domain experts and AI specialists. This workflow has produced peer-reviewed legal analytics studies.
5. **Deployment Flexibility:** QLA supports both cloud and on-premise deployment with open-source LLM backends (Ollama, vLLM).
6. **Extensible Multi-Jurisdictional Architecture:** The unified data model and ingestion pipeline support systematic expansion across legal systems.

Validated Use Cases

(1) Sentencing Pattern Analysis: Query: "Analyze sentencing trends for theft offenses in Polish courts (2018–2022)." Legal experts curated 2,400 relevant cases from initial retrieval of 15,000 candidates. Schema extracted 44 fields per case (defendant demographics, offense details, mitigating/aggravating factors, sentence type and length). Statistical analysis revealed significant regional variations in sentencing severity ($p < 0.01$), with full provenance to source judgments.

(2) Citation Network Analysis: Query: "Map precedent citations in employment discrimination cases." System identified 1,800 cases citing relevant EU directives and Polish labor code articles. Extraction captured citation contexts, enabling temporal analysis of doctrinal evolution. Results visualized as interactive precedent networks with drill-down to source text.

(3) Tax Interpretation Trends: Legal researchers explored patterns in tax authority interpretations using collaborative curation. From 50,000 tax rulings, lawyers selected 800 cases on VAT treatment of digital services. Aggregated analysis revealed consistency patterns across regional tax offices.

Capability	**NotebookLM**	**Claude**	**Harvey AI**	**Lexis+**	**vLex**	**CourtListener**	**QLA (Ours)**
Pre-indexed legal corpus	×	×	×	✓	✓	✓	✓
Documents available	0	0	0	Millions	1B+	9M+	**3M+**
Semantic vectors indexed	×	×	×	×	×	2TB	**300M+**
Multi-jurisdiction support	×	×	×	US/UK	100+	US only	**PL+UK+EU**
Bulk analysis ($>$1K docs)	×	×	✓	Limited	Limited	API only	✓
Quantitative aggregation	×	×	×	×	×	×	✓
Per-document debugging	Limited	✓	Limited	Limited	✓	✓	✓
Open-source LLM support	×	×	×	×	×	✓	✓
On-premise deployment	×	×	×	×	×	✓	✓
Lawyer-AI collaboration	×	×	×	×	×	×	✓
Schema-based extraction	×	×	✓	×	×	×	✓
Full provenance tracking	×	✓	Limited	✓	✓	✓	✓
Upload required	✓	✓	✓	×	×	×	×
Time to first analysis	Hours/Days	Hours/Days	Hours/Days	Minutes	Minutes	Minutes	**Immediate**

Table 3: Full competitive comparison of legal AI platforms.